From Atoms to Devices: Performance Bottlenecks of Nitride Semiconductors Using Multi-Scale Quantum Methods

by

Aagnik Pant

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Applied Physics and Scientific Computing),
in the University of Michigan
2023

Doctoral Committee:

Associate Professor Emmanouil Kioupakis, Chair
Assistant Professor Elaheh Ahmadi
Professor Roy Clarke
Professor Çağliyan Kurdak

Nick Pant
nickpant@umich.edu
ORCID iD: 0000-0002-5450-6045

ACKNOWLEDGEMENTS

Foremost, I would like to thank my advisor, Manos. His constant support and encouragement have been key towards the successful completion of my degree. The Kioupakis Group has been a wonderful place to grow and learn over the last four years, and I am particularly grateful to Manos for building an environment that places greater importance on growth and well-being over productivity.

I would like to thank members of my committee, Cagliyan, Elaheh, and Roy, as well as the Applied Physics family, including Cyndi and Svala.

I am also deeply grateful to my current lab members, including Xiao, Woncheol, Kyle, Emily, Amanda, Mahlet, Yujie, Yuxuan, Kelotchi, and Tiernan, with whom I have developed rewarding friendships. I am also grateful to the Kioupakis group alumni, including Christina, Kelsey, Nocona, Zihao, and Sieun. I will always look fondly back at our conference adventures and our group lunches. Thanks also goes to Srinivas, David, and Kishwar for mentoring me during my first year of grad school, before I joined the Kioupakis group.

My time in graduate school would not have been the same without my friends, many of whom I ended up living with as roommates. You made lockdown not just bearable but enjoyable. In particular, the Beakes and bubble crew: Adarsh, Austin, Amanda, Jana, Jenn, Julia, Julie, Leanne, Parker, Thomas, and Tom. I would also like to thank the many other friends I met during my time at Michigan, who are too enumerable to list here, including my first-year Applied Physics cohort and my first-year AP house roommates. I would also acknowledge my friends from McGill and high school, who I always looked forward to seeing during my breaks and vacations:

Abby, Alex, Alex, Austin, Chelsea, Izzy, Jason, Jenny, Nat, Nina, Uday, Wynn, Richie, and countless more. I cannot wait for all of you to visit me in Austin.

Finally, I would like to thank my family. All of this is possible because of you. Much love to aama, baba, and bahini. I also dedicate this to the memory of my grandmother, who had an insatiable curiosity and would ask me many questions, ranging from "what are atoms?" to "what does quantum mean?" If you had been born a little bit later, or maybe if you had been born in a wealthier country, I would not have been surprised if you had lived another life as a scientist.

Chapter IV was adapted from Appl. Phys. Lett. 117, 242105 (2020), with the permission of AIP Publishing. Chapter V was adapted from Appl. Phys. Lett. 121, 032105 (2022), with the permission of AIP Publishing. Chapter VI was adapted from AIP Advances 12, 125020 (2022), with the permission of AIP Publishing. Chapter V was adapted from a manuscript in progress. I would like to acknowledge the use of the OpenAI ChatGPT language model as a grammar editing tool.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

III-nitride semiconductors have revolutionized modern electronics by enabling high-power radio-frequency and lighting technologies. These materials hold immense potential for new technologies such as miniaturized displays, ultraviolet sterilization, and fast electric-vehicle charging. However, there are still performance bottlenecks that need to be addressed. In this thesis, I investigate the microscopic mechanisms that limit the performance of nitride semiconductors in power-conversion and lighting applications, and propose new solutions using quantum-mechanical methods that connect the microscale physics to macroscale device phenomena.

First, I examine the limitations of III-nitride semiconductors in power-conversion applications. To increase the breakdown voltage of GaN, which is critical for higher power devices, it is promising to alloy it with Al. However, this approach decreases the electron mobility due to alloy scattering. In this thesis, I develop a novel approach to calculate the low-field mobility of semiconductor alloys from first principles. I find that the mobility of AlGaN decreases by a factor of $\sim$7 compared to GaN. Consequently, Al compositions above 75% are required to achieve even a two-fold increase in the Baliga Figure of Merit (BFOM) compared to GaN, at which point impurity doping becomes increasingly difficult. To address this issue, I propose using atomically thin superlattices of AlN and GaN that are free of disorder. My calculations indicate that these nanostructures exhibit a 4.8 eV band gap and a mobility $3-4\times$ higher than random AlGaN. By accounting for the dopant ionization fraction in the BFOM, I show that the superlattices exhibit the highest modified BFOM among prominent competing semiconductors.

Second, I investigate the mechanism that causes InGaN light-emitting diodes (LEDs) to suffer from an emission blueshift and linewidth broadening when operated at high currents, leading to a degradation of their color properties. By systematically considering the effects of polarization-field screening, phase-space filling, and many-body plasma renormalization, I comprehensively explain the current-dependent spectral characteristics of polar III-nitride quantum wells. My analysis overturns the prevailing hypothesis that the emission blueshift is primarily due to the screening of internal polarization fields, as this explanation neglects the contribution of plasma renormalization, which is nearly equal but opposite in magnitude. In contrast, the blueshift is explained only by accounting for a complex interplay of polarization-field screening, plasma renormalization, and phase-space filling. On the other hand, the spectral broadening occurs mainly due to phase-space filling. My analysis suggests that the key to improving the spectral characteristics of InGaN LEDs is to accelerate carrier recombination and transport and reduce the carrier density required to operate them at high power density.

Finally, I investigate the concept of defect tolerance in InGaN emitters. Recent experiments have challenged the widely accepted hypothesis that carrier localization suppresses diffusion and enhances the tolerance of InGaN emitters to defects. By examining the competition between radiative and Shockley-Read-Hall recombination in InGaN alloys using a formalism that I recently developed, I propose that carrier localization and polarization fields enhance the quantum efficiency at low current densities, without invoking the suppression of carrier diffusion. Decreasing the oscillator strength or increasing the quantum-well thickness may improve the quantum efficiency of LEDs for low-power applications but it exacerbates efficiency droop and impair color purity control at high operating powers.

Overall, this thesis paves the way for the continued development of III-nitride technology, and its approach can be generalized to other emerging semiconductors.

CHAPTER I

Introduction

1.1 III-nitride semiconductors

The wide-ranging applications of III-nitrides — including general lighting, optical communication, radio-frequency communication, and power-conversion technologies — depend on the exceptional properties of these compound semiconductors. [1, 2, 3] III-nitrides consist of nitrogen (N) as the anion and group-III metals, such as gallium (Ga), indium (In), and aluminum (Al), and increasingly boron (B) and scandium (Sc), as the cation. Among the III-nitrides, GaN stands out as the most widely used and is ranked as the world's second-most produced semiconductor after silicon (Si). While III-nitrides have revolutionized various fields, there are still applications where their performance falls short of expectations. [4, 5, 6, 7] To uncover the fundamental loss mechanisms in III-nitride alloys and heterostructures, this thesis utilizes computational modeling based on quantum mechanics, with the goal of enabling new designs that significantly enhance their device performance.

The band gap, which is the most fundamental property of a semiconductor, is the energy gap between the highest occupied electronic states (called the *valence band*) and lowest unoccupied electronic states (called the *conduction band*). For electronic applications, the band gap determines the maximum voltage that can be applied to the material before it undergoes dielectric breakdown. For a transistor, the band gap determines the best ratio of on to off current that can be achieved. For optoelectronic

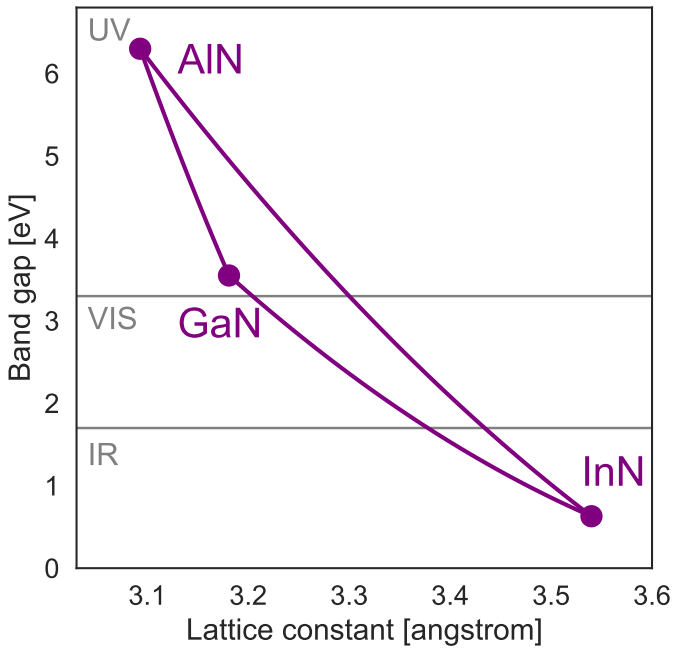

Figure 1.1: Band gap vs lattice constant of the conventional III-nitride semiconductors, as calculated within hybrid-functional density-functional theory. Solid lines are guides to the eye, indicating the possible ternary alloys.

devices, the band gap determines the wavelength of light that can be emitted or absorbed. The widespread use of III-nitride semiconductors is thanks to the fact that they are the only commercial semiconductor system whose band gap can be tuned from the infrared (IR) to the deep ultraviolet (UV) (Figure 1.1). [8, 9] In particular, InN has a band gap of 0.6 eV, GaN has a band gap of 3.5 eV, and AlN has a band gap of 6.3 eV. The *ternary alloys* InGaN and AlGaN have band gaps that span the range from 0.6 eV to 3.5 eV and 3.5 eV to 6.3 eV, respectively.

1.1.1 Alloy disorder

Alloying is a useful way of engineering the band gap, however it has unintended consequences for electronic and optical properties. [10] This is because alloying

breaks a very important symmetry of semiconductors, *translational symmetry*. In the simplest terms, translational symmetry means that a pattern repeats itself over and over again. For example, to create a GaN crystal, one can arrange Ga and N atoms in a repeating pattern to form a bulk solid. Conversely, alloying breaks this symmetry; to create AlGaN alloys, we would randomly replace a fraction of the Ga atoms in the periodic GaN crystal with Al atoms. As a result, there is no specific pattern of Al, Ga, and N atoms that can be repeated to reconstruct the entire AlGaN alloy.

1.1.1.1 Alloy scattering

The breaking of translational symmetry through alloy disorder has significant implications for the electronic properties of semiconductors. [11, 12] In periodic semiconductors, electrons experience a smooth background potential and can propagate through the material without scattering (neglecting lattice vibrations, which can also cause electron scattering). However, in disordered semiconductors, electrons encounter a rough, random background potential that causes them to scatter in all directions (Figure 1.2). This phenomenon is known as the *relaxation of crystal-momentum conservation*, where electrons in semiconductor alloys experience repeated momentum changes due to scattering, even in the absence of external perturbations. This is akin to driving on a newly paved road that is smooth and free of imperfections, allowing the car to maintain a steady momentum, versus driving on an older, pothole-ridden road that causes the car to veer in different directions. The scattering lifetime τ, which is the time between different scattering events, determines the electronic mobility of the material, given by $\mu = e\tau/m^*$, where m^* is the effective mass of the electron in the solid and e is the elementary charge. Mobility reflects how effectively a material conducts electrical energy and is inversely proportional to the resistivity ρ of the material, given by $\rho = 1/en\mu$.

Alloy disorder not only affects the electronic properties but also the optical properties of semiconductors. [13] To generate light, electrons in the conduction band and holes in the valence band of a semiconductor must recombine. When an electron leaves the

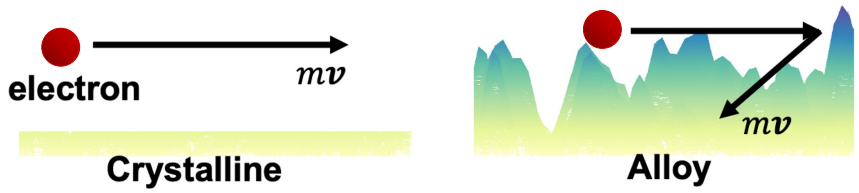

Figure 1.2: In a periodic crystalline semiconductor, electrons propagate with a well defined crystal momentum. In a random-alloy semiconductor, electrons experience repeated crystal-momentum changes due to the disorder perturbation because Bloch states, which have well-defined crystal momenta, are no longer eigenstates of the Hamiltonian. This phenomena is called *relaxation of crystal momentum* and leads to alloy scattering.

valence band, it leaves behind a "hole" in the valence band. Typically, light carries very little momentum, and in most semiconductor applications, it is assumed that the momentum of light is zero. Therefore, for an electron and hole to recombine, they must possess the same momentum. However, in semiconductor alloys, crystal momentum conservation is relaxed. As a result, an electron and hole that would not have recombined in a normal crystalline semiconductor due to differing crystal momenta may have a chance of recombining in an alloy. This is because one or both of the particles may get scattered by the disordered background potential and end up possessing the same momentum. The likelihood of this occurring depends on the material's specific details and the strength of the alloy disorder. Furthermore, relaxation of crystal momentum can affect non-radiative processes such as Auger-Meitner recombination, which is a three-particle scattering process and limits the efficiency of light-emitting diodes at high current densities. [7]

1.1.1.2 Carrier localization

Thus far, our discussion of alloy disorder has neglected the possibility of *carrier localization*. If the disorder is strong enough, it can cause the charge carriers, electrons

or holes, to become localized in place, ultimately inhibiting carrier diffusion. [10] This effect was first hypothesized by Phil Anderson, [14] who won the 1977 Nobel prize for his work on the electronic structure of magnetic and disordered systems. An intuitive analogy to understand localization is that the potholes on the road are so large that the car gets stuck in place and cannot move. While this analogy is helpful at a rudimentary level, it is not entirely correct because localization is ultimately a wave phenomenon. A better analogy would be that of a wave front in the ocean; in the absence of any obstruction, these wave fronts would propagate unperturbed. However, imagine that we stick a very long wooden pole into the bottom of the ocean floor that sticks out on the surface. As the wave front reaches the wooden pole, the pole scatters the wave front and a spherical wave emanates outward from the pole in addition to the original wave front. This is the wave phenomenon of *diffraction*. Now, imagine adding another wooden pole next to our original pole. The wave front would scatter from both poles and create an interference pattern, with some regions showing *constructive interference* and other regions showing *destructive interference*. If we kept adding wooden poles at random locations, what we would eventually find is that the wave front no longer propagates past the wooden poles because destructive interference wins everywhere except at the location where the wave front first came into contact with the wooden poles, where there is massive constructive interference. This phenomenon is called wave localization. In alloyed semiconductors, the same phenomenon occurs except that electron and hole waves are being localized by the alloy disorder.

What determines whether alloy disorder localizes a carrier? We can very roughly estimate the qualitative dependence of the average decay length λ of wave functions on material parameters in a ternary alloy with the WKB approximation,

$$\lambda \sim 1/\sqrt{2m^*(x)\sqrt{x(1-x)}\Delta E},$$

where $m^*(x)$ is the composition-dependent effective mass, x is the composition, and ΔE is the conduction-band offset (for electrons) or valence-band offset (for holes) between the binary compounds. For a given alloy system, there is no localization

for the binary compounds $x = 0$ and $x = 1$ since $\lambda \to \infty$; localization becomes strongest for the intermediate compositions because $x(1-x)$ reaches a maximum for $x = 0.5$. The effective mass is very important; in the III-nitrides, electrons are ten times lighter than holes, $m_e \approx 0.2, m_h \approx 2$, so λ is about $3\times$ larger for holes than electrons. This is why holes in the III-nitrides are typically localized but electrons are not, as seen directly in numerical simulations employing atomistic tight-binding as well as modified $\mathbf{k} \cdot \mathbf{p}$ Hamiltonians. The formula for λ also explains why localization is more important in the III-nitride alloys than in alloys of III-phosphide and III-arsenide semiconductors, despite them sharing similar chemistries. The heavy-hole effective mass in the InGaP system ranges from $0.6m_0$ to $0.8m_0$, and the valence-band offset is approximately 0.3 eV. In the InGaAs system, the heavy hole effective mass ranges from $0.4m_0$ to $0.5m_0$, and the valence-band offset is approximately 0.3 eV. [10] In contrast, in the InGaN system, the heavy-hole mass is about $1.8m_0$ and the valence-band offset is approximately 1 eV. λ is therefore about $3\times$ larger in the III-nitride system than in the III-phosphide system, and $4\times$ larger than in the III-arsenide system. The formula for λ does not account for additional localization effects due to the mismatch in the size of the binary compounds. The mismatch between the binary compounds is around 7% for both the In-containing phosphide and arsenides but about 11% for the In-containing nitrides. This additional component of disorder, which in a tight-binding model appears as disorder in the off-diagonal hopping parameters, is called *off-diagonal disorder*, while the component of disorder due to fluctuations in the energy levels of the binary compounds is called *site-diagonal disorder*.

Overall, carrier localization is an important phenomenon in III-nitrides. However, the physics behind carrier localization remains unclear due to the difficulty in experimentally accessing localized states. On the other hand, while simulations indicate that carriers are localized in III-nitrides, it's challenging to calculate their density of states because of the finite size of simulations. Therefore, it's hard to assess their overall importance to thermally averaged quantities. Additionally, different methods typically disagree on the exact details of localization length in these systems.[15, 16]

The impact of carrier localization on functional properties is also not known. There is no first-principles method to calculate the mobility of localized states, which occurs through variable-range hopping. Moreover, it's not clear how localization affects radiative and non-radiative recombination, with different studies reaching contradictory conclusions.[17, 18, 19, 20] One school of thought suggests that localization is the reason why III-nitride LEDs work at all, despite having relatively high defect densities that would kill arsenide- and phosphide-based LEDs. [21] This hypothesis states that localized carriers are unable to diffuse to defects where non-radiative recombination occurs. However, this idea has been challenged by recent experiments that show long diffusion lengths in III-nitrides, likely because there is always some finite occupation of extended states that do diffuse. [22] Another school of thought suggests that localization reduces the wave-function overlap of electrons and holes, lowering efficiency by preventing them from recombining. [23] However, this result is not widely accepted, and other works suggest that localization actually makes it more likely for carriers to recombine. [18] There is also some evidence that localization exacerbates Auger-Meitner recombination in quantum wells, although it's not clear whether this effect is due to alloy scattering or carrier localization. [24] Surprisingly, no theoretical study to date has investigated the impact of localization on the competition between radiative and Shockley-Read-Hall recombination. Much more work is needed to understand the impact of carrier localization in III-nitride alloys.

1.1.2 Polarization fields

In addition to their alloy disorder, the III-nitrides are characterized by their polarization field. [25, 26] These materials crystallize in the wurtzite phase and belong to the $P6_3mc$ space group, which lacks inversion symmetry. When biaxial strain is applied, dipoles form within the unit cell, resulting in strong piezoelectric fields. The wurtzite structure also exhibits spontaneous polarization fields due to the fact that the dipole of the Ga-N bond does not cancel out along the c-axis, which is typically the direction of growth for commercial devices.

The polarization field has a negligible impact on bulk properties due to the translational symmetry of bulk systems. While bulk properties are not significantly impacted by polarization fields, the lack of translational symmetry in alloys means that polarization fields can have an effect on bulk properties through local fluctuations. However, if we coarse-grain our resolution to a large enough length scale, the average composition of an alloy possesses translational symmetry, and the *macroscopic* polarization field must vanish in bulk alloys. As a result, polarization fields are primarily important only in the presence of compositional gradients, interfaces, and surfaces. Real devices are typically composed of heterostructures of many different materials stacked on top of each other, making polarization fields an important factor in determining their functional properties.

One of the main effects of the polarization fields can be seen in quantum wells, which are used as the active region of light-emitting diodes. A typical quantum well in III-nitride devices consists of an InGaN well region, typically around 3 nm thick, and a GaN barrier region. The well region has a smaller band gap than the barrier region, which allows us to confine electrons and holes in the well region thus promoting the probability of them recombining. The discontinuity of the polarization field between the material of the well and barrier regions gives rise to polarization charges at both interfaces. These polarization charges give rise to gigantic internal electric fields in the quantum well, which are typically of order 1 MV/cm. These electric fields shift the energy levels of the conduction and valence states, and cause the band gap to red shift compared to polarization-free quantum wells. This effect is known as the *quantum-confined Stark effect* (QCSE), in analogy with the Stark effect of the hydrogen atom. It has long been thought that the QCSE is the primary reason for the blue-shift exhibited by InGaN LEDs with increasing current density. As carriers are injected into the quantum well, they screen the polarization field which lessens the QCSE and thus lessens the red-shift, giving rising to an apparent blue-shift. [27] Although this explanation is simple, it neglects many-body plasma renormalization due to the free carriers which is of equal but opposite magnitude as polarization-field screening. Any explanation that neglects this effect is incomplete

and cannot provide a comprehensive and quantitative description of the blue-shift of InGaN LEDs with increasing current density. Some authors have considered the effects of plasma renormalization, but the contribution of the different effects are not clear and the work was performed during a time when even the fundamental gap of InN was not known thus it is not clear which parameters were tuned to match experiment. [28] Moreover, the internal electric fields in quantum wells also separate electrons and holes to opposite sides of the quantum well and severely impact the recombination properties of light-emitting diodes. [29, 5] Indeed, polarization fields are known to exacerbate the efficiency droop phenomenon because they slow the rate of radiative and non-radiative recombination, which forces LEDs to operate at higher carrier densities to maintain the same power density, but the efficiency is lower at high carrier densities.

1.2 Light-emitting diodes

1.2.1 Physics of light emission

LEDs are devices that convert electricity to light. If there exists a non-equilibrium population of electrons and holes in a semiconductor's conduction and valence band, these carriers will recombine to produce light through a process called *spontaneous emission*, thus lowering the total electronic energy of the system. There is a very intuitive but *semi-classical* (therefore, not entirely correct) picture to understand the process of light emission. Classically, a radiating dipole emits electromagnetic waves, with the frequency of the wave being equal to the frequency of the dipole oscillation. In III-nitrides, the valence electrons are in orbitals that closely resemble p-type atomic orbitals. By having even a fraction of the electrons in the s-type orbitals of the conduction band rather than in the valence band, the charge density acquires, at least locally, a net dipole moment. (Conversely, no dipole moment would form and thus no light would be produced if both conduction and valence orbitals were s-type.) This leaves the question: what causes the dipoles to oscillate and radiate charge? In quantum mechanics, the vacuum state of the electromagnetic

field is teeming with fluctuations of virtual particles. These *vacuum fluctuations* occasionally resonate with the dipoles and cause them to oscillate and emit light. The energy of light and the frequency of the dipole oscillation corresponds to the energy difference between the conduction and valence states. It is for this reason that the probability of light emission is proportional to the square of the dipole transition matrix element between conduction and valence states, $e|\langle\psi_c|\,\hat{\mathbf{r}}\,|\psi_v\rangle|^2$. However, since the dipole operator is not uniquely defined under periodic boundary conditions, it is more common to see the rate being written as proportional to the momentum matrix element squared, $|\langle\psi_c|\,\hat{\mathbf{p}}\,|\psi_v\rangle|^2$. The position and momentum matrix elements are related as (in atomic units, $\hbar = e = m_e = 4\pi\epsilon_0 = 1$),

$$\frac{d}{dt}\langle\psi_c|\,\hat{\mathbf{r}}\,|\psi_v\rangle = \langle\psi_c|\,[\hat{H},\hat{\mathbf{r}}]\,|\psi_v\rangle = (\varepsilon_c - \varepsilon_v)\,\langle\psi_c|\,\hat{\mathbf{r}}\,|\psi_v\rangle \tag{1.1}$$

$$\frac{d}{dt}\langle\psi_c|\,\hat{\mathbf{r}}\,|\psi_v\rangle = \langle\psi_c|\,[\hat{H},\hat{\mathbf{r}}]\,|\psi_v\rangle = -\,\langle\psi_c|\,\hat{\mathbf{p}}\,|\psi_v\rangle \tag{1.2}$$

$$\therefore \langle\psi_c|\,\hat{\mathbf{r}}\,|\psi_v\rangle = \frac{\langle\psi_c|\,\hat{\mathbf{p}}\,|\psi_v\rangle}{\varepsilon_v - \varepsilon_c} \tag{1.3}$$

Light emission or *radiative recombination* is always competing with *non-radiative recombination*, which depletes the population of electrons and holes, and dissipates energy through heat. There are two primary mechanisms that lead to non-radiative losses. At low carrier densities, Shockley-Read-Hall (SRH) recombination is the primary loss mechanism. [30] Even in the highest quality LEDs, there are always a finite number of imperfections or *defects* in the crystal. These can be missing or misplaced atoms in the crystal structure or foreign impurities that were incorporated during the growth process. These defects can trap electrons and holes, which lowers the conductivity of the material. This type of carrier capture is typically termed *carrier capture by multi-phonon emission* because each capture event is accompanied by a large distortive relaxation of the defect configuration, which corresponds to the emission of multiple lattice vibration (phonon) modes. Overall, the probability of carrier capture is proportional to the number of free carriers in the vicinity of the defect. A complete SRH cycle consists of a defect sequentially capturing an electron

and a hole. At higher carrier densities, the primary loss mechanism is Auger-Meitner recombination (AMR). [31, 32] This is a three-particle scattering mechanisms whose rate scales with the carrier density to the third power. In the AMR process, the excess energy from an electron-hole recombining transfers to a nearby carrier through the Coulomb interaction. The AMR process is strictly limited by the conservation of crystal momentum between the three scattering carriers, therefore the AMR rate is particularly enhanced by crystal-momentum-conservation breaking by disorder and lattice vibrations. The AMR process is responsible for the efficiency droop phenomenon, which causes the efficiency to decrease with increasing carrier density, limiting the energy efficiency

1.2.1.1 Structure of LEDs

Figure 1.3 depicts a schematic of the different layers that constitute a typical III-nitride LED. The active region, which contains quantum wells, is the most significant component of LEDs as it is where light is generated. The quantum well is composed of a material with a smaller band gap (typically $In_xGa_{1-x}N$, $0 < x < 0.3$) surrounded by a material with a larger band gap (typically GaN), where electrons and holes are confined and recombine to produce light. It is important to grow the quantum well and surrounding barriers with minimal defects to decrease the occurrence of trap-assisted non-radiative recombination. To prevent overcrowding of carrier density that can lead to overheating and degradation of electrical and spectral properties, designs often use multiple quantum wells (MQW).

The MQW active region is located within a *p*-*n* junction, which functions as an on/off switch for the LED. When a forward bias is applied, the LED turns on, and electrons and holes flow from the metal contacts towards the active region, where they combine and produce light. To prevent the overflow of electrons from the quantum well into the hole-injection contact, most LEDs include an AlGaN electron-blocking layer (EBL). This is because electrons are light and have a high kinetic energy, making them difficult to capture in the quantum well. As most metal contacts form a Schottky barrier against the wide bandgap GaN material, the

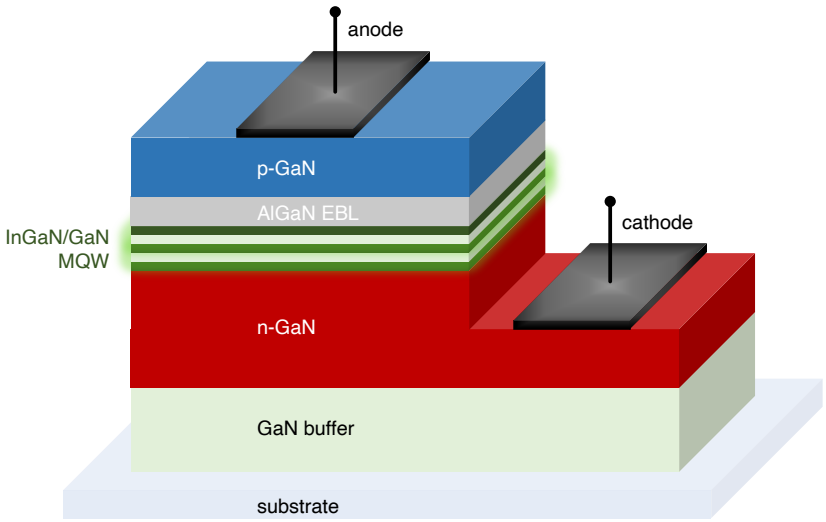

Figure 1.3: Schematic of a III-nitride light-emitting diode.

contact regions of the LED are heavily doped to promote screening and reduce the length of the depletion region of the Schottky barrier. If the depletion region is small enough, carriers can tunnel through the Schottky barrier, leading to linear ohmic characteristics. Additionally, many commercial InGaN LEDs use an In-containing superlattice underlayer below the device (not shown in the figure), which improves the material quality of the active region by trapping or gettering point defects that occur during the growth on the surface of the epitaxial layer. [33, 34] The presence of an In-containing underlayer has been shown to significantly enhance the efficiency of LEDs. Finally, a thick structurally relaxed buffer region separates the device from the substrate (typically sapphire). Large amounts of dislocations are generated at the buffer/substrate interface due to the $\sim 14\%$ lattice mismatch between GaN and sapphire. Great care is taken to minimize the dislocation density at the buffer surface; although III-nitride LEDs are more robust to dislocations than III-arsenide and III-phosphide LEDs, [35] a large dislocation density can still negatively impact device performance. [36]

Once the LED is fabricated, the LED is either directly coupled to an optical fibre for optical communication or packaged in a hemispherical-shaped polymer host, such as poly(methyl methacrylate) (PMMA) or polycarbonate, for general lighting. [37] These polymers are chosen for their transparency in the visible spectrum and their refractive index ($n_r \approx 1.5$), which is intermediate between that of air ($n_r \approx 1$) and GaN ($n_r \approx 2.3$ at ε_G). The importance of the latter criterion can be understood by considering the Fresnel coefficient for reflection at normal incidence, given by

$$R = \frac{n_1 - n_2}{n_1 + n_2}. \tag{1.4}$$

The reflection coefficient $R_{air-GaN}$ between GaN and air is 0.4, meaning 40% of the light that GaN emits is reflected back at the interface. In contrast, the product of reflection coefficients $R_{air-polymer}R_{polymer-GaN}$ between air and polymer and between polymer and GaN, respectively, is much smaller, at 0.04, greatly enhancing the ability to extract light from the LED. This strategy can be thought as a way of minimizing the wave-impedance mismatch at the air-matter interface.

1.2.2 Applications

LEDs are incredibly versatile technologies with a wide range of applications. This section briefly reviews several important applications of LEDs, the current state of the field, and the role III-nitrides play in these applications.

1.2.2.1 General lighting

Owing to the fact that LEDs are 3-4× more efficient than incandescent light bulbs, LEDs are widely used for general lighting, which accounts for approximately 20% of the world's electricity consumption. For general lighting applications, phosphors that emit across a broad spectral range convert light from highly efficient blue LEDs (whose maximum energy conversion efficiency can reach up to 80%) into a spectrum of different colours that is interpreted by the human eye as white light. These LEDs are called *blue-pump LEDs*. The most commonly used phosphor for white LEDs is Ce:YAG or Ce^{3+}-doped $Y_3Al_5O_2$ (yttrium aluminum garnet) because it very efficiently converts blue light to white light. [38] The host garnet material is an ultra-wide-band-gap insulator and the absorption of blue light and luminescence of yellow light arises from transitions between electronic levels of the Ce ion, which substitutes the Y site. Ce:YAG is synthesized using a solid-state reaction at high temperature; it is then milled into a powder and dispersed into a polymer matrix, which is used to coat the LED as a thick hemispherical layer. From an engineering perspective, general lighting is considered to be a mature technology with little room for improvement in terms of energy efficiency. The remaining effort in industry for general lighting is dedicated to engineering the colour of white LEDs so as to emulate natural light. Part of this effort involves the development of efficient green and amber LEDs so as to give more flexibility in tuning the colour spectrum, and another part involves the development of efficient phosphors that produce more natural colours.

1.2.2.2 Consumer electronics and microLEDs

In addition to general lighting, LEDs are commonly used in consumer electronics, including TVs and smartphones. Standard displays use LED-backlit liquid-crystal display (LCD) technology, whose active region consists of long organic molecules embedded in a host polymer matrix that can be polarized by an electric field. LCD displays use two linear polarizers fixed perpendicularly to each other, with a liquid-crystal layer sandwiched between them. The liquid-crystal layer's orientation can be controlled by applying an electric field to control the light that passes through the color filter of each pixel, which is typically an organic dye or quantum-dot material. Despite being widely used, LED-backlit LCD technology cannot be further miniaturized and has limitations in providing ultra-high definition resolution and individual control of pixels.

Micro-LEDs offer a solution to these issues as they are miniaturized LEDs based on III-nitrides, containing InGaN quantum wells with varying indium content for red, green, and blue pixels that can be directly electrically controlled. [39, 40] This allows for much greater control over pixel density and the ability to completely turn off pixels, resulting in deeper blacks. One of the main emerging applications for microLEDs is augmented reality (AR), which is a technology that allows for the overlay of digital information onto the real world, in real time. An early example of AR technology is the heads-up display in cars, which project important information such as speed and safety warnings in the drivers line of sight thus allowing the driver to keep their eye on the road while processing relevant information. AR technology requires miniaturization of LEDs and fast response times.

A significant challenge for microLEDs is achieving high efficiency for all three primary colors (red, green, and blue) at small sizes, particularly for the red wavelength. In addition, microLEDs have a high surface-to-volume ratio, which can lead to increased surface recombination and decreased efficiency. Fabrication of microLED arrays also requires precise control of growth and placement, which can be difficult and expensive. Moreover, microLEDs have to compete with *organic light-emitting diodes* (OLEDs), which also offer control over individual pixels. Compared to mi-

croLEDs, OLEDs can be made into flexible and transparent displays, allowing for a wide range of form factors. However, OLEDs have a shorter lifespan than microLEDs and are typically unstable at high operating powers. [41] Compared to organic LEDs, microLEDs have a significant advantage for response times due to their higher carrier mobility ($\sim 10 - 100$ cm^2/Vs vs < 1cm^2/Vs) and faster recombination lifetimes (ns vs μs-ms). In practice, applications that require flexible substrates will likely benefit from OLEDs while high-performance applications will benefit from microLEDs.

1.2.2.3 Horticulture

As we have discussed for far, LEDs are a highly efficient lighting technology where the cost of producing light is nearly commensurate with the cost of electricity. By integrating high-efficiency LEDs with solar technology, it is possible to directly produce food for human consumption using sunlight. [42] This is precisely what horticulture is accomplishing by growing food from LEDs indoors, a method that is particularly valuable in areas where conventional farming is impractical. Vertical indoor farming could reduce the need for conventional farmland, as agriculture is responsible for 80% of deforestation worldwide. In horticulture, red and blue wavelengths of light are especially important; red light is essential for the flowering and fruiting stages of the plant, while blue light is necessary for vegetative growth and production of green foliage. Although red and blue LEDs are highly efficient, the key lighting challenge is to optimize the growing process by blending other colors of light to mimic the color and brightness of natural sunlight. This can be achieved using InGaN LEDs with varying indium concentrations. Other pressing technical challenges facing horticulture include efficiently managing indoor temperature and humidity and developing methods to cultivate plants effectively without natural nutrient-rich soil.

1.2.2.4 Disinfection

UV LEDs can be used for disinfection by emitting light in the germicidal wavelength range of 200-280 nm, which damages the DNA and RNA of microorganisms, preventing them from reproducing or infecting. [6] One promising material for UV

LEDs is AlGaN, which offers tunability of wavelength due to its adjustable band gap. However, increasing the Al content of the material to achieve higher band gaps also leads to increasing TM polarized light emission (because of p_z-like symmetry of the valence orbital), which makes light extraction difficult. (TM polarized light tends to propagate within the material, leading to reabsorption within the material, instead of propagating perpendicularly, which is necessary for it to escape.) AlGaN alloys with intermediate compositions also have lower mobility due to alloy scattering, which places a fundamental limit on the conductivity for a given carrier concentration. Moreover, doping AlGaN is generally difficult, so impurity scattering further reduces the mobility. Contact resistances are also larger in UV AlGaN devices compared to GaN devices because the band edge is much farther away from the typical work function of contact metals such as Ti or V/Zr contacts. This creates a higher barrier for Ohmic contacts based on field emission, requiring more doping, which is unfortunately challenging in AlGaN. Other materials such as β-Ga_2O_3 and hBN also show promise for UV LEDs, but they have limitations in either dopability or tunability of the specific wavelength range.

1.3 Power electronics

1.3.1 Physics of power conversion

Nearly 20% of the global energy consumed is consumed as electricity. The importance of electricity to modern society cannot be overstated. The vast global electrical infrastructure requires specialized circuits that can convert electrical energy with low losses. [43] Such circuits use *power semiconductor devices*, which are typically *discrete* (as opposed to *integrated*) devices that are designed to efficiently handle the conversion of electrical power. The fundamental power operations are:

- *Rectification*: the process of converting AC signals to DC.
- *Inversion*: the process of converting DC signals to AC.
- *Conversion*: the process of transforming the voltage or frequency of AC or DC

signals.

In addition to these fundamental operations, it may be necessary to *regulate* the power, *filter* the power signal, and *protect* against over-currents and over-voltages. The building blocks required to design circuits that perform these operations are:

- *Rectifiers*: switches that allow current to flow in one direction.
- *Amplifiers*: circuit elements that amplify a small AC signal by transferring power from another DC source.
- *Capacitors*: circuit elements that store and discharge electrostatic energy upon the application and removal of an applied voltage.
- *Inductors*: circuit elements that store and discharge magnetostatic energy upon the application and removal of an electric current.
- *Resistors*: circuit elements that limit the flow of electrical current by dissipating electrical energy as heat.

Among the five fundamental building blocks, rectifiers and amplifiers are typically *active components*, meaning they require external energy to operate, and as a result can add electrical energy to the circuit. On the other hand, capacitors, inductors, and resistors are *passive components*, meaning they can dissipate or store electrical energy. Capacitors are fabricated by sandwiching a high-K dielectric between parallel conducting plates, inductors are fabricated by wrapping a conducting wire about a magnetic core, and resistors are fabricated by moulding resistive material such as carbon powder into a desired shape. On the other hand, rectifiers and amplifiers are fabricated by applying advanced processing and nanofabrication techniques to semiconductors. Typically, rectifiers can be made from *diodes* or *transistors*, while amplifiers are made exclusively from *transistors*. The voltage and frequency requirement for a given application determines the exact type of transistor or diode that is used.

Figure 1.4 shows the characteristics of an ideal rectifier compared to an actual rectifier. An ideal rectifier carries any amount of current with zero voltage drop in the

on state, blocks any value of voltage with zero leakage in the off state, and switches between the on state and off state with zero switching time. In practice, real rectifiers exhibit a finite voltage drop, a finite on-resistance that leads to conduction losses, a maximum voltage they can block due to dielectric breakdown, and a maximum finite frequency at which they can be switched. Figure 1.5 shows the characteristics of an ideal transistor, which may be used as a rectifier (in the saturation region) or as an amplifier (in the action region). An ideal transistor conducts current in the on-state with zero voltage drop, blocks voltage in the off-state with zero leakage, can be operated with a high current and voltage in the active region, has uniform (linear) spacing between characteristics in the active region independent of the forward voltage and current, and has zero time delay in the modulation response, defined as the ratio of the forward (drain to source) AC current and the AC input gate voltage. Of course, real transistors exhibit a finite on-resistance, a maximum voltage they can block in the off state, non-linear characteristics in the active region, and a maximum frequency at which they can be modulated or switched.

Conduction losses in power devices can be minimized by choosing semiconductors with high mobility and high dopability. The device should also be as thin as possible to minimize the overall resistance. Smaller resistances additionally lead to small RC time constants, thus allowing for higher frequency operation. However, thin devices are unable to withstand high voltages, because the voltage ΔV is related to electric field E as, $\Delta V = E/L$, where L is the device thickness. In order to maximize the breakdown voltage, the device should be as thick as possible (or stacked in serial in a module) so that carriers never experience a high electric field that accelerates them to destructive energies. We are clearly at an impasse. It seems that we can *either* have a device with low conduction losses and fast modulation/switching response *or* a device that supports very high voltages before breaking down, but not both.

We can find a way out of this conundrum by elucidating the nature of breakdown in semiconductors. Breakdown occurs when electrons or holes accelerate due to the external electric field and gain sufficient kinetic energy, typically the band gap energy, that they can Coulombically interact with valence electrons and ionize carriers,

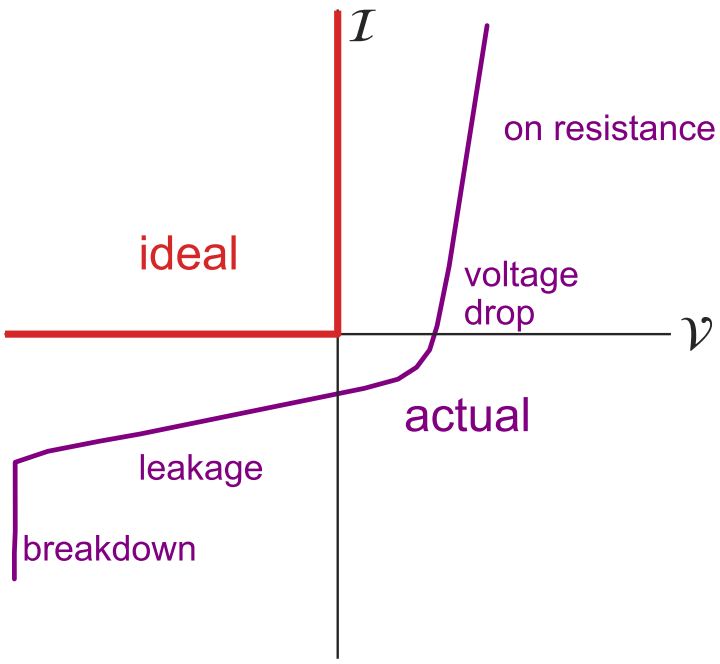

Figure 1.4: Schematic of current-voltage characteristics of ideal rectifiers (red) vs actual rectifiers (purple). Actual rectifiers may be p-n or Schottky power diodes or power transistors that operate in the saturation region. Rectifiers must be able to block high voltages in the off state, have minimum current leakage if off, have no voltage (power) drop if on, and have minimum on-state conduction losses. In many applications, they must be able to operate at high frequencies since voltage and power conversion is typically performed by modulating the duty cycle of the power signal.

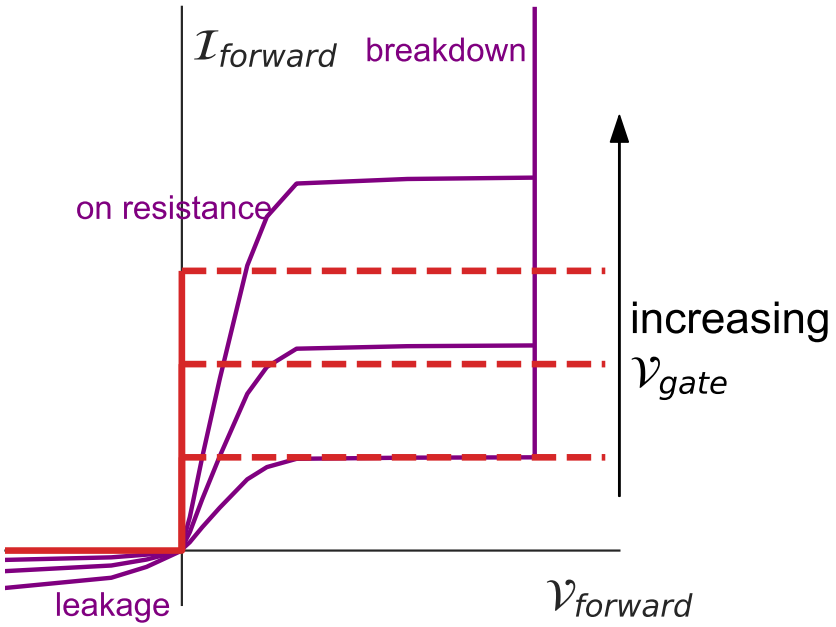

Figure 1.5: Schematic of current-voltage characteristics of ideal amplifying transistors (red) vs actual amplifying transistors (purple). Actual amplifiers exhibit current leakage in the off state, a finite on resistance that leads to conduction losses, a finite forward voltage beyond which breakdown occurs, and non-linear characteristics with respect to the gate voltage. Additionally, their RC time constant must be minimized in order to maximize the range of frequencies in which they can amplify signals.

a process known as *impact ionization*, or produce defects, a process known as *hot-carrier defect generation*. The formation of defects manifests as a slow degradation of the characteristics of the device. Impact ionization leads to *avalanche breakdown*, which is a process mediated by carrier multiplication due to carrier ionization that leads to the generation of large amounts of runaway current and heat in an uncontrolled feedback loop. Semiconductors with wider band gaps can generally sustain stronger electric fields before undergoing breakdown. [44, 45, 46, 47, 48, 49, 50] Empirically, the breakdown field of semiconductors scales quadratically with the band gap (Figure 1.6). Therefore, by using semiconductors with ultra-wide band gaps (loosely defined to be band gaps >3.5 eV), it is possible to aggressively scale the size of semiconductor power devices thus minimizing the resistance while maximizing the breakdown voltage.

The area of the device is another aspect of the design where trade off is required. Larger device areas are necessary to support high currents, since smaller areas overcrowd the current which leads to overheating due to poor thermal dissipation. However, larger device areas increase the junction capacitance, which increases the RC time constant and therefore limits the maximum modulation and switching frequencies. A possible solution is to reduce the current at which the device operates while maintaining the same power output by increasing the voltage. In applications where this is possible, ultra-wide-band-gap semiconductors are promising. Ultra-wide-band-gap semiconductors can additionally operate at much higher temperatures because the thermal population of carriers decreases exponentially with increasing band gap, which reduces leakage current. Another solution is to use materials with high thermal conductivity and custom device designs that enable the efficient extraction of heat from the active region. For applications that necessarily require high current ratings, such as battery charging for electric vehicles, it is often possible to integrate power devices in parallel in a power module so that the overall module can support higher current ratings, without increasing the area of the discrete devices.

Clearly, the development of ultra-wide-band-gap semiconductors with high dopabil-

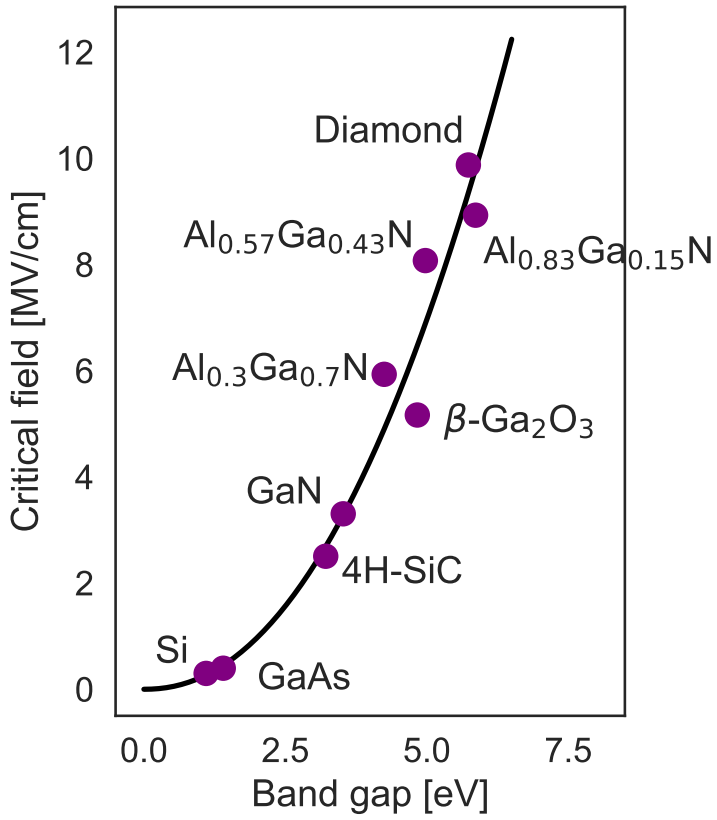

Figure 1.6: Empirical breakdown field as a function of the band gap for a wide range of semiconductors. The solid line is the phenomenological model, $F_{br} = 3.3\text{MV/cm} \times (\varepsilon_{\text{G}}/3.5)^2$.

ity, high mobility, and high thermal conductivity is crucial for the advancement of next-generation power devices. While several ultra-wide-band-gap semiconductors have been discovered, very few exhibit all three desirable properties simultaneously. The search for such materials is ongoing, and the hope is that future research will yield discoveries that will pave the way for efficient and high-performance electronic and power systems.

1.3.2 Applications

1.3.2.1 Power conversion

Power conversion devices are used in applications requiring the conversion of electric power, including consumer power electronics, renewable energy generation, electric cars and rails, HVDC transmission systems, and more. Different types of transistors are typically used for different applications depending on their power and frequency needs. For instance, metal-oxide-semiconductor field-effect transistors (MOSFETs) are used for low power and high-frequency applications such as switch-mode power supplies. Insulated-gate bipolar transistors (IGBTs) are used for medium to high power and high-frequency applications. Thyristors are used for ultra-high voltage and ultra-high current applications but are restricted to low-frequency operation. In all these devices, electrons transport vertically from the source to the drain, running from a metal contact at the bottom of the device to a metal contact at the top of the device. This leads to better distribution of current than lateral designs characteristic of integrated circuits, which means heat is generated uniformly and can be more effectively dissipated. Additionally, vertical devices allow for thicker drift regions, which allows for higher breakdown voltages. Conventionally, silicon (Si) power devices have been used for power conversion applications. However, silicon carbide (commonly 4H-SiC and less commonly 6H-SiC) is increasingly replacing Si technology due to its wider bandgap (3.2 eV vs. 1.1 eV), [43] which makes SiC devices more efficient and capable of operating at higher temperatures.

By moving towards ultra-wide-band-gap semiconductors, the performance of power electronic devices can be greatly improved. Currently, there is no device that per-

forms very well at high voltage and high switching frequencies. This is mostly limited by the performance of the underlying materials. Thus, finding more efficient materials is necessary for enabling even higher efficiencies in emerging technologies such as electric rails, high-voltage DC transmission, and electric vehicles. The Baliga Figure of Merit (BFOM) is a widely used metric for quantifying the performance of semiconductors for power electronics, [43, 51]

$$\text{BFOM} = \frac{V_{BR}^2}{R_{on}} = \frac{1}{4}\epsilon_s \mu E_c^3. \tag{1.5}$$

V_{BR} is the breakdown voltage in the device off state, and R_{on} is the specific resistance of the device in the on state, which quantifies the conduction losses in the on-state. ϵ_s is the dielectric constant, μ is the carrier mobility, and E_c is the critical electric field at which the semiconductor undergoes breakdown. Recently, the importance of including the ionization fraction in the expression for the BFOM has been identified, leading to the modified Baliga figure of merit,

$$\text{Modified BFOM} = \frac{V_{BR}^2}{R_{on}} = \frac{1}{4}\epsilon_s \mu E_c^3 \eta, \tag{1.6}$$

where η is the dopant ionization fraction. [52] Since $E_c \propto \varepsilon_G^2$, we conclude,

$$\text{BFOM} \propto \varepsilon_G^6,$$

which emphasizes the need for ultra-wide band gaps. Although the Baliga Figure of Merit is derived for unipolar rectifiers (specifically, asymmetrically doped abrupt non-punch-through p-n junction diodes), it is now widely used to quantify the performance of potential materials for power electronics applications.

The search for next-generation semiconductors to replace Si and SiC, and drive future power conversion technologies is ongoing. Although several materials have shown promise, including AlGaN, [53] β-Ga_2O_3, [54] GeO_2, [55] cBN, [56] and diamond, [56] the question of which of these materials will ultimately prevail remains unanswered. Each material has its own set of advantages and challenges, with nitrides standing out

for their high thermal conductivity, good dopability, and high mobility, along with an established infrastructure for growth and fabrication. However, challenges such as substrate cost and the need for vertical designs and better contacts remain. A critical comparison of these different semiconductors is presented in Chapter V. Ongoing research is focused on addressing these challenges and improving the performance of III-nitrides and other materials for power conversion applications at both the material and device levels.

1.3.2.2 Radio-frequency amplification

Radio frequencies (RF), which include radio waves and microwaves ($f \sim$ 3KHz to 300GHz, $\lambda \sim 10^{-4} - 10^{8}$ m), have a wide range of applications, including next-generation 5G and 6G mobile networks, security and surveillance radar, automotive radar for car-to-car communication, medical imaging such as MRI, RF frequency sources for plasma, and more. For these applications, high-frequency performance of the transistor is important, which is why HEMTs (high-electron-mobility transistors) are used. HEMTs are unipolar majority carrier devices that use heterostructures to create 2D electron gases at the interface, which creates a high mobility and high density channel. [2] HEMTs can be modulated at high frequencies because they lack a p-n junction capacitance and because unlike bipolar devices they are not limited by the recombination lifetime of minority carriers. The size of HEMTs can be aggressively scaled down without affecting carrier density in the channel region. GaN/AlGaN HEMTs can handle medium power and high frequency, but thermal management is a key issue due to overheating, which will be further exacerbated by scaling down the size. Despite their high performance and widespread use, GaN/AlGaN HEMTs also exhibit breakdown at lower voltages than predicted by their band gap, with the exact mode of breakdown currently unclear. Since GaN/AlGaN HEMTs rarely show the rapid breakdown characteristics typical of avalanche breakdown, it is suspected that their breakdown is due to hot-carrier defect generation.

The need for high-power RF amplifiers for RF communication can be understood by considering how RF signals attenuate in the atmosphere. Attenuation of RF signals

occurs due to the absorption of radiation by gases in the atmosphere. It also occurs due to Rayleigh scattering, *i.e.*, scattering by dielectric spheres whose length scale is much smaller than the wavelength of light. The intensity of Rayleigh scattered light scales with the wavelength as $\sim 1/\lambda^4$, which means higher frequency, shorter wavelength signals are more strongly attenuated. In the presence of rain, Mie scattering, *i.e.*, scattering by dielectric spheres whose length scale corresponds to the wavelength of light, may also contribute to attenuation. The strongest attenuation of RF signals occurs by absorption in the ionosphere, which is a layer in the atmosphere ∼60-600 km above the surface that consists of gases ionized by solar radiation. Although the ionosphere can be avoided for many terrestrial communication applications, it cannot be avoided for space communication.

Overall, RF communication requires high-power amplifiers for communication and sensing, but devices typically either handle very high power or very high frequency, but rarely both. Lateral designs typical of HEMTs, where the channel is parallel to the gate, are better for RF amplifier applications than vertical designs because they minimize the transit length from the drain to the source. However, such designs can lead to current overcrowding, which leads to overheating, and field enhancement effects, which lowers the breakdown voltage. Going to higher band gaps can enable the devices to be scaled further while going to higher breakdown voltages. HEMTs are typically characterized by their *cutoff frequency* (f_T), which is frequency at which the short-circuit current gain h_{21} is unity, where h_{21} is defined as the ratio of the small-signal output current to the small-signal input current of a device whose terminals are shorted. HEMTs are also characterized by their *maximum linear power*, $P_{max,lin}$ which is the highest output that HEMTs can produce while maintaining their linearity in amplification thus faithfully reproducing the input signal.

The Johnson Figure of Merit (JFOM) is widely used to quantify the performance of semiconductors for high-power RF applications. It is defined as the product of the cutoff frequency f_T and the maximum drain-to-source voltage that can be applied before breakdown, [57, 2]

$$\text{JFOM} = f_T V_{DS,max} \tag{1.7}$$

Using small signal circuit models for transconductance amplifiers, it can be shown that the maximum linear power of an amplifier is proportional to the square of the JFOM,

$$P_{max,lin} \propto \text{JFOM}^2.$$

The JFOM can be related to material properties by estimating the cutoff frequency f_T as $f_T = L_{DS}/v_{sat}$, where L_{DS} is the drain-to-source channel length and v_{sat} is the saturation velocity of carriers. This assumes that carriers transit the channel at the saturation velocity, the channel length is shorter than the mean-free path of carriers, and the frequency response of the device is limited by the channel. Moreover, the maximum drain-to-source voltage, which is really the voltage of the external DC source that is transferred to the AC signal, is estimated as $V_{DS,max} = E_c/L_{DS}$, where E_c is the critical field that the semiconductor can sustain before undergoing breakdown. This assumes that electric field is uniform across the channel and the breakdown voltage of the device is limited by the channel. Under these assumptions, the JFOM is,

$$\text{JFOM} = v_{sat}E_c. \tag{1.8}$$

From the relations $E_c \propto \varepsilon_G^2$ and $P_{max,lin} \propto JFOM^2$, we obtain,

$$\text{JFOM} \propto \varepsilon_G^2,$$

$$P_{max,lin} \propto \varepsilon_G^4,$$

thus underscoring the need for ultra-wide-band-gap semiconductors for high-power RF devices.

1.4 Organization of this thesis

This thesis is organized into three parts. The first part introduces the theoretical methods used in this thesis (chapters II and III), which include both first-principles methods for modelling atomic scale phenomena and semi-empirical methods for modelling larger length scales. The second part focuses on the ultra-wide-band-gap

($\varepsilon_G > 3.5$ eV) III-nitride alloys and heterostructures with an emphasis on power-electronics applications (chapters IV and V). Finally, the third part focuses on the wide-band-gap III-nitride alloys and heterostructures ($\varepsilon_G < 3.5$ eV) with an emphasis on light-emitting diode applications (chapters VI and VII).

Chapter II reviews first-principles methods based on density-functional theory for predicting and describing materials phenomena, including the Kohn-Sham equations, exchange and correlation, many-body perturbation theory, electron-phonon interactions, and the method of special quasirandom structures. In chapter III, basic semi-empirical methods for device simulation are discussed based on the envelope-function and effective-mass approximations. Chapter IV presents a first-principles method for calculating the electron mobility of semiconductors, focusing on composition-dependent disorder in AlGaN alloys. Chapter V demonstrates that atomically thin superlattices of AlN and GaN have high mobility due to the absence of alloy disorder, making them particularly promising for power electronics. This chapter also contains a systematic comparison of various semiconductors based on the modified Baliga figure of merit. Chapter VI investigates the injection-dependent emission spectra of III-nitride light-emitting diodes using a Schrödinger-Poisson model that systematically accounts for disorder and many-body effects. Finally, chapter VII investigates the concept of defect tolerance in InGaN emitters, indicating that an apparent defect tolerance emerges in InGaN alloys due to the interplay of carrier localization and polarization fields.

Bibliography

[1] Zetian Mi and Chennupati Jagadish. *III-Nitride Semiconductor Optoelectronics*. Academic Press, 2017.

[2] Patrick Fay, Debdeep Jena, and Paul Maki, editors. *High-Frequency GaN Electronic Devices*. Springer, 2019.

[3] Michael Kneissl and Jürgen Rass, editors. *III-nitride ultraviolet emitters: Technology and applications*, volume 227. Springer, 2015.

[4] Hiroshi Amano, Ramón Collazo, Carlo De Santi, Sven Einfeldt, Mitsuru Funato, Johannes Glaab, Sylvia Hagedorn, et al. The 2020 uv emitter roadmap. *Journal of Physics D: Applied Physics*, 53(50):503001, 2020.

[5] Aurelien David, Nathan G Young, Cory Lund, and Michael D Craven. The physics of recombinations in iii-nitride emitters. *ECS Journal of Solid State Science and Technology*, 9(1):016021, 2019.

[6] Hiroshi Amano, Y. Baines, E. Beam, Matteo Borga, T. Bouchet, Paul R. Chalker, M. Charles, and et al. The 2018 gan power electronics roadmap. *Journal of Physics D: Applied Physics*, 51(16):163001, 2018.

[7] Emmanouil Kioupakis, Sieun Chae, Kyle Bushick, Nick Pant, Xiao Zhang, and Woncheol Lee. Theoretical characterization and computational discovery of ultra-wide-band-gap semiconductors with predictive atomistic calculations. *Journal of Materials Research*, pages 1–22, 2021.

[8] P. G. Moses, M. Miao, Q. Yan, and C. G. Van de Walle. Hybrid functional investigations of band gaps and band alignments for aln, gan, inn, and ingan. *The Journal of Chemical Physics*, 134(8):084703, 2011.

[9] I. Vurgaftman and J.R. Meyer. Journal of applied physics. 94:3675, 2003.

[10] Jasprit Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures*. Cambridge University Press, 2007.

[11] J. W. Harrison and J. R. Hauser. Theoretical calculations of electron mobility in ternary III-V compounds. *Journal of Applied Physics*, 47(1):292–300, 1976.

[12] Michael E. Coltrin and Robert J. Kaplar. Transport and breakdown analysis for improved figure-of-merit for algan power devices. *Journal of Applied Physics*, 121(5):055706, 2017.

[13] Brian K Ridley. *Quantum processes in semiconductors*. Oxford University Press, 2013.

[14] Philip W Anderson. Absence of diffusion in certain random lattices. *Physical review*, 109(5):1492, 1958.

[15] Debapriya Chaudhuri, Michael O'Donovan, T. Streckenbach, O. Marquardt, P. Farrell, Saroj K. Patra, T. Koprucki, and Stefan Schulz. Multiscale simulations of the electronic structure of iii-nitride quantum wells with varied indium content: Connecting atomistic and continuum-based models. *Journal of Applied Physics*, 129(7):073104, 2021.

[16] A. Di Vito, A. Pecchia, A. Di Carlo, and M. Auf der Maur. Simulating random alloy effects in iii-nitride light emitting diodes. *Journal of Applied Physics*, 128(4):041102, 2020.

[17] Daniel SP Tanner, Philip Dawson, Menno J Kappers, Rachel A Oliver, and Stefan Schulz. Polar (in, ga)n/gan quantum wells: Revisiting the impact of carrier localization on the "green gap" problem. *Physical Review Applied*, 13(4):044068, 2020.

[18] Christina M Jones, Chu-Hsiang Teng, Qimin Yan, Pei-Cheng Ku, and Emmanouil Kioupakis. Impact of carrier localization on recombination in ingan quantum wells and the efficiency of nitride light-emitting diodes: Insights from theory and numerical simulations. *Applied Physics Letters*, 111(11):113501, 2017.

[19] Claude Weisbuch, Shuji Nakamura, Yuh-Renn Wu, and James S Speck. Disorder effects in nitride semiconductors: impact on fundamental and device properties. *Nanophotonics*, 10(1):3–21, 2020.

[20] Huan-Ting Shen, Claude Weisbuch, James S Speck, and Yuh-Renn Wu. Three-dimensional modeling of minority-carrier lateral diffusion length including random alloy fluctuations in (in, ga) n and (al, ga) n single quantum wells. *Physical Review Applied*, 16(2):024054, 2021.

[21] Shigefusa F Chichibu, Akira Uedono, Takeyoshi Onuma, Benjamin A Haskell, Arpan Chakraborty, Takahiro Koyama, Paul T Fini, Steven P DenBaars,

and Shuji Nakamura. Origin of defect-insensitive emission probability in incontaining (al, in, ga) n alloy semiconductors. *Nature materials*, 5(10):810–816, 2006.

[22] Aurelien David. Long-range carrier diffusion in (in, ga) n quantum wells and implications from fundamentals to devices. *Physical Review Applied*, 15(5):054015, 2021.

[23] Matthias Auf Der Maur, Alessandro Pecchia, Gabriele Penazzi, Walter Rodrigues, and Aldo Di Carlo. Efficiency drop in green ingan/gan light emitting diodes: The role of random alloy fluctuations. *Physical Review Letters*, 116(2):027401, 2016.

[24] Mehran Shahmohammadi, Wei Liu, Georg Rossbach, Lise Lahourcade, Am'elie Dussaigne, Catherine Bougerol, Rapha"el Butt'e, Nicolas Grandjean, Benoit Deveaud, and Gw'enol'e Jacopin. Enhancement of auger recombination induced by carrier localization in ingan/gan quantum wells. *Physical Review B*, 95(12):125314, 2017.

[25] Fabio Bernardini, Vincenzo Fiorentini, and David Vanderbilt. Spontaneous polarization and piezoelectric constants of iii-v nitrides. *Physical Review B*, 56(16):R10024, 1997.

[26] Cyrus E Dreyer, Anderson Janotti, Chris G Van de Walle, and David Vanderbilt. Correct implementation of polarization constants in wurtzite materials and impact on iii-nitrides. *Physical Review X*, 6(2):021038, 2016.

[27] E Kuokstis, JW Yang, Grigory Simin, M Asif Khan, R Gaska, and MS Shur. Two mechanisms of blueshift of edge emission in ingan-based epilayers and multiple quantum wells. *Applied Physics Letters*, 80(6):977–979, 2002.

[28] LH Peng, CW Chuang, and LH Lou. Piezoelectric effects in the optical properties of strained ingan quantum wells. *Applied physics letters*, 74(6):795–797, 1999.

[29] Emmanouil Kioupakis, Qimin Yan, and Chris G Van de Walle. Interplay of

polarization fields and auger recombination in the efficiency droop of nitride light-emitting diodes. *Applied Physics Letters*, 101(23):231107, 2012.

[30] Audrius Alkauskas, Qimin Yan, and Chris G Van de Walle. First-principles theory of nonradiative carrier capture via multiphonon emission. *Physical Review B*, 90(7):075202, 2014.

[31] Emmanouil Kioupakis, Daniel Steiauf, Patrick Rinke, Kris T Delaney, and Chris G Van de Walle. First-principles calculations of indirect auger recombination in nitride semiconductors. *Physical Review B*, 92(3):035207, 2015.

[32] Demetrios Matsakis, Anthea Coster, Brenda Laster, and Ruth Sime. A renaming proposal:"the auger-meitner effect". *Physics Today*, 72(9):10–11, 2019.

[33] A. M. Armstrong, B. N. Bryant, M. H. Crawford, D. D. Koleske, S. R. Lee, and J. J. Wierer. Defect-reduction mechanism for improving radiative efficiency in ingan/gan light-emitting diodes using ingan underlayers. *Journal of Applied Physics*, 117(13):134501, 2015.

[34] C. Haller, J.-F. Carlin, G. Jacopin, D. Martin, R. Butté, and N. Grandjean. Burying non-radiative defects in InGaN underlayer to increase InGaN/GaN quantum well efficiency. *Appl. Phys. Lett.*, 111:262101, 2017.

[35] A Hangleiter, F Hitzel, C Netzel, D Fuhrmann, U Rossow, G Ade, and P Hinze. Suppression of nonradiative recombination by v-shaped pits in gainn/gan quantum wells produces a large increase in the light emission efficiency. *Physical review letters*, 95(12):127402, 2005.

[36] Andrew Armstrong, Tanya A Henry, Daniel D Koleske, Mary H Crawford, Kurt R Westlake, and Stephen R Lee. Dependence of radiative efficiency and deep level defect incorporation on threading dislocation density for ingan/gan light emitting diodes. *Applied Physics Letters*, 101(16):162102, 2012.

[37] Pallab Bhattacharya. *Semiconductor optoelectronic devices*. Prentice-Hall, Inc., 1997.

[38] S Nishiura, S Tanabe, K Fujioka, and Y Fujimoto. Properties of transparent ce: Yag ceramic phosphors for white led. *Optical Materials*, 33(5):688–691, 2011.

[39] Hongxing Jiang and Jingyu Lin. Development of nitride microleds and displays. *Semiconductors and Semimetals*, 106:1–56, 2021.

[40] Jacob Day, Jing Li, DYC Lie, Charles Bradford, JY Lin, and HX Jiang. Iii-nitride full-scale high-resolution microdisplays. *Applied Physics Letters*, 99(3):031116, 2011.

[41] Yuge Huang, En-Lin Hsiang, Ming-Yang Deng, and Shin-Tson Wu. Mini-led, micro-led and oled displays: present status and future perspectives. *Light: Science & Applications*, 9(1):105, 2020.

[42] Paul Morgan Pattison, Monica Hansen, and Jeffrey Y Tsao. Led lighting efficacy: Status and directions. *Comptes Rendus Physique*, 19(3):134–145, 2018.

[43] B. Jayant Baliga. *Fundamentals of Power Semiconductor Devices*. Springer Science & Business Media, 2010.

[44] M. Higashiwaki, K. Sasaki, A. Kuramata, T. Masui, and S. Yamakoshi. Gallium oxide (ga2o3) metal-semiconductor field-effect transistors on single-crystal -ga2o 3 (010) substrates. *Applied Physics Letters*, 100:013504, 2012.

[45] H. Niwa, J. Suda, and T. Kimoto. 21.7 kv 4h-sic pin diode with a space-modulated junction termination extension. *Applied Physics Express*, 5:064001, 2012.

[46] D. Khachariya, S. Mita, P. Reddy, S. Dangi, P. Bagheri, M. Hayden Breckenridge, R. Sengupta, E. Kohn, Z. Sitar, R. Collazo, and S. Pavlidis. Al0.85ga0.15n/al0.6ga0.4n high electron mobility transistors on native aln substrates with ¿9 mv/cm mesa breakdown fields. In *Device Research Conference (DRC)*, pages 1–2, June 2021.

[47] A. A. Allerman, A. M. Armstrong, A. J. Fischer, J. R. Dickerson, M. H. Crawford, M. P. King, M. W. Moseley, J. J. Wierer, and R. J. Kaplar. Al0.3ga0.7n

pn diode with breakdown voltage ¿1600 v. *Electron. Lett.*, 52(16):1319–1321, 2016.

[48] A. Nishikawa, K. Kumakura, and T. Makimoto. High critical electric field exceeding 8 mv/cm measured using an algan p-i-n vertical conducting diode on n-sic substrate. *Japanese J. Appl. Physics, Part 1 Regul. Pap. Short Notes Rev. Pap.*, 46:2316–2319, 2007.

[49] R. J. Kaplar, O. Slobodyan, J. D. Flicker, and M. A. Hollis. (invited) a new analysis of the dependence of critical electric field on semiconductor bandgap. In *ECS Meet. Abstr. MA2019-02*, pages 1334–1334, 2019.

[50] X. Yan, I. S. Esqueda, J. Ma, J. Tice, and H. Wang. High breakdown electric field in β-ga2o3/graphene vertical barristor heterostructure. *Appl. Phys. Lett.*, 112(3):032101, 2018.

[51] B J Baliga. Semiconductors for high-voltage, vertical channel field-effect transistors. *Journal of Applied Physics*, 53(3):1759–1764, 1982.

[52] Y Zhang and JS Speck. Importance of shallow hydrogenic dopants and material purity of ultra-wide bandgap semiconductors for vertical power electron devices. *Semicond. Sci. Technol.*, 35:125018, 2020.

[53] Robert J Kaplar, Andrew A Allerman, Andrew M Armstrong, Mary H Crawford, Jeramy R Dickerson, Arthur J Fischer, Albert G Baca, and Erica A Douglas. Ultra-wide-bandgap algan power electronic devices. *ECS Journal of Solid State Science and Technology*, 6(2):Q3061, 2016.

[54] Andrew J Green, James Speck, Grace Xing, Peter Moens, Fredrik Allerstam, Krister Gumaelius, Thomas Neyer, Andrea Arias-Purdue, Vivek Mehrotra, Akito Kuramata, et al. β-gallium oxide power electronics. *APL Materials*, 10(2):029201, 2022.

[55] Sieun Chae, Kelsey Mengle, Kyle Bushick, Jihang Lee, Nocona Sanders, Zihao Deng, Zetian Mi, Pierre F Poudeu, Hanjong Paik, John T Heron, and Emmanouil Kioupakis. Toward the predictive discovery of ambipolarly dopable

ultra-wide-band-gap semiconductors: The case of rutile geo2. *Applied Physics Letters*, 118(26):260501, 2021.

[56] Nocona Sanders and Emmanouil Kioupakis. Phonon-and defect-limited electron and hole mobility of diamond and cubic boron nitride: A critical comparison. *Applied Physics Letters*, 119(6):062101, 2021.

[57] E Johnson. Physical limitations on frequency and power parameters of transistors. In *1958 IRE International Convention Record*, volume 13, pages 27–34. IEEE, 1966.

CHAPTER II

First-Principles Methods for Atomic-Scale Modelling

2.1 Motivation

First principles modeling of materials at the atomic scale is a powerful theoretical characterization method that seeks to describe materials solely using the fundamental constants of the universe. [1, 2] By avoiding tunable empirical parameters, this approach can provide predictive insights into the maximum performance limits of materials, which can help screen different materials for a given application. Theoretical insights can also uncover mechanisms that limit the performance of existing materials, particularly those related to non-radiative recombination or phonon collisions, which are difficult to study experimentally because they do not produce light. First principles modeling is particularly advantageous in these cases, as it provides direct insight into these mechanisms. Because these models are meant to be predictive, they are rigorously tested against experimental data whenever possible to ensure their accuracy in describing a wide range of materials. While this field is still developing, tremendous progress has been made in accurately calculating functional properties for materials, as well as predicting new materials. These advances have great potential in accelerating materials discovery, leading to the development of new technologies.

2.2 Many-body Schrödinger equation

In the modelling of solids, we solve the time-independent many-body Schrödinger equation for electrons, assuming that the nuclei are classical and fixed at their equilibrium positions,

$$\left[-\sum_i \frac{1}{2}\nabla_i^2 + \frac{1}{2}\sum_{i\neq j}\frac{4\pi}{|\mathbf{r}_i - \mathbf{r}_j|} - \sum_{i,I} Z_I \frac{4\pi}{|\mathbf{r}_i - \mathbf{R}_I|}\right]\Psi_{\{\mathbf{R}\}} = E(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)\Psi_{\{\mathbf{R}\}}. \tag{2.1}$$

In this chapter, we will use Hartree atomic units ($e = m_e = \hbar = 4\pi\epsilon_0 = 1$) unless specified otherwise. In the square brackets, the first term represents the kinetic energy of the interacting electrons, the second term represents the many-body electron-electron repulsion, and the third term represents the electron-nuclear attraction. $\Psi_{\{\mathbf{R}\}}$ is the many-body electronic wave function at fixed nuclear coordinates, whose modulus squared corresponds to the ground-state charge density. The eigenvalue E on the right-hand side represents the total electronic energy for fixed nuclear coordinates. Equation (2.1) is derived in Appendix A by applying a series of approximations to the time-dependent many-body Schrödinger equation that treats both electrons and nuclei on equal quantum-mechanical footing.

2.2.1 The exponential wall

Unfortunately, equation (2.1) is too complicated to solve directly. To get a sense for why solving the many-body equation is intractable, consider the case of GaN, which has 16 valence electrons in the primitive cell. The volume of the primitive cell of GaN is approximately 19.72 Å^3. Assuming a reasonable discretization length of 0.1 Å, a grid of the primitive cell would have $N_p \sim 20,000$ points. Since the wave function indexes all 16 electronic coordinates, a total of $N_p^{16} \sim 10^{69}$ complex numbers would be required to represent just the ground state of the electronic many-body wave function. Using double precision complex numbers (16 bytes), one would require 10^{55} petabytes of storage. For comparison, the largest modern supercomputers support ~ 10 petabytes of memory storage. Even if somehow we were able

to represent the wave function in memory, performing calculations would still be infeasible. The Hamiltonian corresponding to a wave function with 10^{69} elements would have dimensions $\sim 10^{69} \times 10^{69}$. Assuming matrix-matrix multiply scales as $O(N^3)$, a single matrix-matrix multiply of the Hamiltonian would require $\sim 10^{207}$ floating point operations. The fastest exascale supercomputers today are able to perform 10^{18} floating point operations per second. This means that performing a single matrix-matrix multiply would take $\sim 10^{189}$ seconds. For context, the universe is only $\sim 10^{12}$ seconds old. This problem is referred to as the exponential wall of the many-body Schrödinger equation. Clearly, a radically different approach is required if we have any hope of solving the many-body Schrödinger equation for real systems.

2.3 Density-functional theory

A major milestone in computational materials science came with the introduction of the Kohn-Sham equation, and thus density-functional theory, by Walter Kohn and Lu Jeu Sham in 1965 [3]. Kohn and Sham mapped the interacting many-body Schrödinger equation to a Schrödinger-like equation for fictitious non-interacting particles, with the constraint that the density of the fictitious particles yield exactly the ground-state charge density. The advantage of working with a system of non-interacting particles is that the wave functions depend parametrically on three position coordinates corresponding to the x, y, and z directions, rather than on $3N$ coordinates, corresponding to the x, y, and z directions for each of the N electrons in the system. As such, density-functional theory is a ground-state theory of materials that has made solving the Schrödigner equation of real systems computational tractable.

2.3.1 Kohn-Sham equations

The Kohn-Sham equations are given by,

$$\left[-\frac{1}{2}\nabla^2 - \sum_I Z_I \frac{4\pi}{|\mathbf{r}-\mathbf{R}_I|} + V_{KS}[n]\right]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}), \tag{2.2}$$

$$\langle\psi_i|\psi_j\rangle = \delta_{ij} \tag{2.3}$$

where $\psi(\mathbf{r})$ is the Kohn-Sham wave function (or *orbital*) and ε is the Kohn-Sham energy. $V_{KS}[n] \equiv \delta E_{KS}[n]/\delta n$ is defined as a functional derivative of the Kohn-Sham energy functional and represents the effective potential that maps the fictitious non-interacting system of Kohn-Sham particles onto the real interacting system of electrons. The Kohn-Sham equations are non-linear in the charge density since the charge density depends on the wave functions and the wave functions depend on the charge density through the energy functional. As such, they must be solved self-consistently until the charge density converges to within a desired tolerance.

The Kohn-Sham equations represent a formal exactification of the mean-field theory of interacting electrons. Within mean-field theory, a drastic approximation is made that electrons do not interact with each other through the Coulomb interaction. Instead, each non-interacting electron experiences an effective or *mean field* that approximates the many-body interaction. Of course, *real* electrons are interacting so the non-interacting particles are simply mathematical tools that are invented to make the problem tractable. Mean-field theories, such as Hartree-Fock theory, had existed for a long time prior to the work of Kohn and Sham. What makes the work of Kohn and Sham special is that the Kohn-Sham potential depends only on the ground-state charge density rather than having a non-local dependence on each of the wave functions in the system, as in Hartree-Fock theory. The former is tractable even for large systems, while the latter quickly becomes intractable as the system size increases.

The Kohn-Sham equations build on on previous work by Pierre Hohenberg and Kohn that proved that the *total electronic energy is a functional of the ground-state charge*

density. [4] The **Hohenberg-Kohn theorems** state:

1. The ground state electron density $n(\mathbf{r})$ uniquely determines the nuclear coordinates and thus the attractive nuclear potential that electrons experience. (This is not intuitive and is typically proven using proof by contradiction; the proof requires that the charge density correspond to a non-degenerate ground state.)

2. The nuclear potential V_n uniquely determines the ground-state many-body electronic wave function $\Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e})$. (This is intuitive, given that $\Psi_{\{\mathbf{R}\}}$ is an eigenstate of $H = T_e + V_e + H_{e-n}$, where H_{e-n} contains the nuclear potential.)

3. The total electronic energy is a functional of the ground-state many-body electronic wave function. (This is obvious from the definition of energy, $E \equiv \langle \Psi | \hat{H} | \Psi \rangle$.)

It is important to note that the Hohenberg-Kohn theorems do not apply when the ground-state wave function is degenerate, in which case one needs to invoke the constrained-search formulation of Levy and Lieb. Notwithstanding this problem, the Hohenberg-Kohn theorems provide the link needed to connect the total energy to the ground-state charge density. This is best summarized visually as $n(\mathbf{r}) \rightarrow V_n \rightarrow \Psi \rightarrow E$.

Having established the connection between the total electronic energy and the ground-state charge density, we can systematically search for the ground-state charge density by variationally minimizing the total electronic energy functional $E[n]$. Of course, we have no hope of doing this if we use the definition of the electronic energy in terms of the many-body wave function. As such, Kohn and Sham made the Ansatz that the exact ground-state charge density can be represented by the ground state density of non-interacting particles. This Ansatz goes by the name "non-interacting V-representability." Using this Ansatz, the Kohn-Sham equations are derived in Appendix B.

At this point, we answer a simple question: if energy is a functional of the charge density then why do we need to introduce intermediary Kohn-Sham wave functions?

Why not perform all calculations with the charge density? The reason is simple: we do not know how to reconstruct the total interacting kinetic energy from the charge density alone, forcing us to resort to an orbital representation where the kinetic energy operator is well defined. There are ongoing efforts to construct an orbital-free density-functional theory, whose advantage is that there is no need to diagonalize the Hamiltonian to obtain wave functions. However, these methods tend to rely on approximations of the kinetic energy functional, *e.g.*, the Thomas-Fermi kinetic-energy functional for the homogeneous electron gas, and are less accurate than orbital density-functional theory.

2.3.2 Exchange and correlation

There exists a universal energy functional that exactly maps the Kohn-Sham equations onto the many-body Schrödinger equation for the ground state. Unfortunately, we do not know how to calculate it, thus forcing us to resort to approximations in order to solve the Kohn-Sham equations. The practice of approximating the energy functional is an entire field of study in its own right, and we will only cover the fundamentals.

2.3.2.1 Hartree potential

For conceptual simplicity, the Kohn-Sham energy functional is split into a classical electrostatic component, called the *Hartree* term, and a quantum-mechanical component, called the *exchange-correlation* term, $E_{KS}[n] = E_H[n] + E_{XC}[n]$. The Hartree term represents the classical *average* Coulomb repulsion felt by electrons due to the presence of all other electrons. Consider the charge density $n(\mathbf{r}) = \sum_i |\psi_i(\mathbf{r})|^2$ generated by the occupied Kohn-Sham states. The electrostatic potential generated by this charge density can be obtained by solving Poisson's equation, $-\nabla^2 V_H(\mathbf{r}) = 4\pi n(\mathbf{r})$. The Hartree energy is thus obtained by integrating this classical potential over the charge density,

$$E_H[n] = \int d\mathbf{r} \int d\mathbf{r}' n(\mathbf{r}) n(\mathbf{r}') \frac{4\pi}{|\mathbf{r} - \mathbf{r}'|}. \tag{2.4}$$

2.3.2.2 Self-interaction error

Although the Hartree approximation is a good starting point, we can immediately see a problem with this formulation, namely that of the *self-interaction error*. Consider a single electron localized at position $\mathbf{r} = \mathbf{0}$ that generates the charge density $n(\mathbf{r}) = \delta(\mathbf{r})$. Substituting the delta function into equation (2.4), we see that the corresponding Hartree energy clearly tends to infinity for a localized charge. This shows quite lucidly that the Hartree approximation causes the Kohn-Sham particles to be *unphysically* repelled by themselves. Of course, no real charge density is a delta function, so we can ask what happens in the opposite limit of a completely extended wave function $\psi = 1/\sqrt{V}$. The corresponding charge density is $n(\mathbf{r}) = 1/V$. In the limit of an infinitely large volume, the charge density of a single electron vanishes and the Hartree self-interaction energy also goes to zero. Therefore, the interaction error becomes progressively worse as the charge becomes more inhomogeneous. One can also see this mathematically by directly solving for the Hartree energy, in which case the following identity is useful,

$$\begin{aligned}
\int d\mathbf{r}' \frac{4\pi}{|\mathbf{r} - \mathbf{r}'|} &= 8\pi^2 \int_0^L dr' (r')^2 \int_0^\pi d\theta \frac{\sin\theta}{\sqrt{r^2 - 2rr'\cos\theta + (r')^2}} \\
&= 8\pi^2 \int_0^L dr' \frac{r'}{r} (r + r' - |r - r'|) \\
&= -\frac{2}{3} r^2 + 16\pi^2 r + \frac{16\pi^2}{3r} L^3.
\end{aligned}$$

In general, performing self-consistency with the self-interaction error tends to delocalize the charge density; in other words, the self-interaction error tends to favour delocalization.

2.3.2.3 Neglect of exchange and correlation

Another problem with the Hartree approximation is that it neglects the fact that electrons are fermions, meaning that more than one electron with the same spin can-

not occupy the same state. This is often referred to as *Pauli repulsion* or *exchange*. The effect of *exchange* is that if an electron is known to be found at position $\mathbf{r}$, it is very unlikely for another electron with the same spin to be found in its immediate vicinity, *i.e.*, exchange introduces correlations in the charge density. The resulting cloud of net positive charge that surrounds each electron is termed the *Fermi hole*. In addition to the exchange interaction, the Coulomb interaction itself can introduce correlations in the charge density. These additional correlations exist between electrons of like and unlike spins, unlike the correlations due to the exchange interaction which only occurs between electrons of like spin. These additional correlations have a non-trivial spatial dependence arising from the non-trivial spatial and temporal dependence of screening. In principle, all many-body effects that are not encompassed by the electron exchange is grouped into the correlation term. The positive cloud that surrounds electrons because of both exchange and correlation is called the exchange-correlation hole. Overall, the exchange-correlation hole leads to fluctuations in the pair-wise distribution function of the charge density, which ultimately lowers the total energy of the system. There is a simple picture that gives an intuitive understanding for this. The exchange-correlation hole, by virtue of being positive, creates an effective attractive potential for electrons that lowers their energy. If the attractive potential is strong enough, then electrons can even localize in place. Including these exchange-correlation effects is necessary in order to map the Kohn-Sham system onto the real system of interacting electrons.

Within the Hartree-Fock formalism, the exchange interaction is exactly accounted for by using an anti-symmetric Slater determinant Ansatz for the wave function, and variationally minimizing the total energy under the constraint of orthonormal wave functions. This procedure gives rise to two operators in the resulting eigenvalue problem: a *local* classical operator corresponding to the Hartree potential and a *non-local* operator with no classical analogue called the exchange operator. Within the Hartree-Fock method, the exchange operator exactly cancels out the self-interaction of the Hartree potential. Therefore, it is reasonable to assume that including exchange-correlations effects within density-functional theory would not

only introduce necessary physics relating to anti-symmetry and screening but also help cancel the spurious self-interaction of the Hartree potential; this is further discussed in the next section.

2.3.2.4 Local-density approximation

The simplest approximation to the E_{XC} functional is the so-called "Local-Density Approximation" (LDA) [5]. In this approximation, the exchange-correlation functional at $\mathbf{r}_0$ is approximated by E_{XC} of the homogeneous electron gas (HEG) having density $n(\mathbf{r}_0)$ equal to the local Kohn-Sham electron density $n_{\mathrm{KS}}(\mathbf{r}_0)$. E_{XC} for the homogeneous electron gas can be separated into an exchange energy E_X due to the exchange interaction and a correlation energy E_C due to the Coulomb repulsion (minus the Hartree potential). The exchange energy can be solved exactly by substituting plane waves into the non-local exchange operator (the same opeartor from Hartree-Fock theory),

$$E_X^{\mathrm{LDA}}[n] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int d^3\mathbf{r} n(\mathbf{r})^{4/3}. \tag{2.5}$$

There is no exact form of the correlation functional, but approximate forms are computed using quantum Monte-Carlo simulations. [6] The advantage of this formulation of the exchange-correlation functional is that it only depends on the *local* charge density. However, there is no formal justification for the LDA, since many-body interactions in a single-particle picture are necessarily *non-local* and *time- (or energy-) dependent*. Because of the local XC functional, the self-interaction is also only *partially* cancelled out leading to systematic errors in the LDA. Nevertheless, the LDA turns out to be a good starting point for the calculation of the electronic structure of solids in most cases. The reason for this is a fortuitous cancellation of errors. LDA overestimates exchange-correlation effects because the homogeneous electron gas over-screens compared to electrons in real solids, which compensates for the fact that the LDA has a self-interaction error, leading to overall reasonable total energies. Nevertheless, rigorous interpretation of LDA eigenvalues and wave-

functions is difficult, although *a posteriori* analyses show that there is qualitative agreement with experimental data and many-body perturbation theory calculations of weakly correlated solids. Systematically, LDA tends to underestimate band gaps and overestimate binding energies, and it can be shown mathematically that these are related to the incorrect derivatives of the LDA exchange-correlation at fractional occupations.

2.3.2.5 Generalized-gradient approximation and beyond

A first-order improvement to the LDA would involve the inclusion of gradients in the exchange-correlation functional. This gives the functional in the Generalized-Gradient Approximation (GGA), [7]

$$E_{XC}^{\mathrm{GGA}}[n] = \int d^3\mathbf{r} f(n, \nabla n). \tag{2.6}$$

There is no unique way of constructing a GGA functional, and standard constructions invoke scaling arguments and exact constraints. The Perdew-Burke-Ernzerhof (PBE) functional is the most common choice for the GGA functional.[**?**] The GGA improves many of the short-comings of the LDA, particularly with respect to the overbinding of atoms and molecules. This results in the GGA being a good starting point for first-principles chemistry calculations. However, the GGA still severely underestimates the band gap in gapped systems and ambiguities remain in the interpretation of the Kohn-Sham eigenvalues, resulting in a poor description of the electronic structure of solids. A further improvement to the GGA functional involves incorporating second-order gradient corrections to the exchange-correlation functional. These functionals, known as meta-GGA functionals, also suffer from ambiguities in interpreting the electron spectrum, warranting the need for a different approach. Often, GGA functionals are improved by mixing a fraction (typically 25%) of the exact Fock exchange from Hartree-Fock theory to correct for the self-interaction error. These functionals are called *hybrid functionals*, and typically provide a good description of the band structure, including band gaps and effective masses, in common semiconductors. Since the exact Fock exchange is non-local, the

exchange-correlation functional is also non-local in this scheme, and for this reason hybrid functional calculations are typically much more costly than calculations with local or semi-local functionals. Nevertheless, hybrid functionals offer a way of self-consistently calculating the band structure as well as total energy of semiconductors with a high degree of accuracy, and for this reason they are routinely used for thermodynamics calculations. [8, 9]

2.4 Many-body perturbation theory

In order to calculate functional properties of materials, we need accurate band structures, for which we need a very accurate description of excited states. [10] Standard constructions of the exchange-correlation functional within DFT tend to describe excited states poorly. Many-body perturbation theory based on the Green's function method is a systematic way of improving the description of excited states, by diagrammatically applying many-body corrections to the Kohn-Sham energies and wave functions. [11] As suggested by its name, this correction scheme is perturbative, and only works if the Kohn-Sham description of the system is qualitatively correct. Appendix C provides a brief overview of the Green's function method necessary for understanding this section.

2.4.1 Quasiparticle formulation

In many-body theory, Green's functions are defined as expectation values of field operators. However, the Kohn-Sham equations give non-interacting wave functions. In order to connect Green's functions to Kohn-Sham wave functions, we use the Lehmann representation which expresses Green's functions in terms of Dyson orbitals or quasi-particle wave functions $f_s(\mathbf{r})$. In a non-interacting theory, Dyson orbitals reduce to non-interacting wave functions $\psi_s(\mathbf{r})$ (see Appendix C for proof). The Green's function in the Lehmann representation is,

$$G(\mathbf{r}, \mathbf{r}'; \omega) = \sum_s \frac{f_s(\mathbf{r}) f_s^*(\mathbf{r}')}{\omega - \varepsilon_s + i\eta}, \tag{2.7}$$

where s is a generalized index that tracks both the band index n and the crystal wave vector k. The Dyson equation is,

$$\left(i\frac{\partial}{\partial t_1} - h(\mathbf{r}_1)\right) G(1,1') - \int d2\Sigma(1,2)G(2,1') = \delta(1,1'), \tag{2.8}$$

where h is the non-interacting part of the Hamiltonian (which includes kinetic, electron-nuclear and Hartree terms) and Σ is the non-local and dynamical self energy. One can substitute the Lehmann representation of the Green's function into the left-hand-side of the Dyson equation and the completeness relation into the Dirac delta function on the right-hand-side to obtain the so-called *quasiparticle equation*,

$$\hat{h}_0(\mathbf{r})f_s(\mathbf{r}) + \int d^3\mathbf{r}'\Sigma(\mathbf{r},\mathbf{r}';\varepsilon_s)f_s(\mathbf{r}') = \varepsilon_s f_s(\mathbf{r}), \tag{2.9}$$

where $f_s(\mathbf{r})$ are the quasiparticle wavefunctions, which should be solved self-consistently.

2.4.2 *GW* approximation

The fundamental question for solving equation (2.9) is how to calculate Σ. It turns out that Σ itself depends on $f_s(\mathbf{r})$ and the problem is highly non-linear. In the linear-response regime (linear in the electron-electron interaction), a systematic way of exactly calculating Σ and $f_s(\mathbf{r})$ is provided by Hedin's equations, which are a set of coupled non-linear integro-differential equations that must be solved self-consistently. [12] In practice, Hedin's equations are impossible to solve exactly, and tractable approximations are needed. The most common approximation, which involves neglecting so-called *vertex corrections*, gives the *GW* approximation, which is the same approximation that we obtain by performing a diagrammatic expansion of the Green's function in terms of the screened Coulomb interaction W, and only retaining irreducible diagrams that are linear in W. In practice, the self-energy is written most efficiently is reciprocal space as,

$$\Sigma(\mathbf{k},\varepsilon_s) = i\sum_{\mathbf{q}}\int \frac{d\omega}{2\pi}G^0(\mathbf{k}+\mathbf{q},\varepsilon_s+\omega)W(\mathbf{q},\omega), \tag{2.10}$$

where G^0 is the non-interacting Green's function, which is practically chosen to be the Kohn-Sham Green's function. Conventionally, the GW approximation of Σ is used to solve equation (2.9). Moreover, only a single iteration of the self-consistency is performed if the Kohn-Sham eigenvalues and eigenfunctions are close to the final value; diagrammatically, this corresponds to neglecting repeated diagrams of $\Sigma = iGW$. Typically, self-consistency improves total energies but leads to worse agreement in spectral properties compared to experiment; this has to do with the neglect of vertex corrections, and there is ongoing development work to include vertex corrections in self-consistent GW calculations as prescribed by Hedin's equations.

2.5 Density-functional perturbation theory

Since most devices operate at room temperature, it is essential to include finite temperature effects in calculations of the functional properties of materials. These finite temperature effects arise from the coupling of electrons with lattice vibrations. The quanta of lattice vibrations are *phonons*, which are bosonic particles whose interactions with electrons renormalize their energies as well as lead to finite lifetimes. The former gives rise to effects such as the Varshni effect where the band gap shrinks with increasing temperature. The latter gives rise to electron-phonon scattering, which limits the mobility and typically increases rates of radiative and non-radiative recombination. Phonons also have a dispersion, which can be calculated using density-functional perturbation theory in the adiabatic approximation. [13]

Before proceeding with this chapter, we briefly review the form of the Hamiltonian for nuclei in the adiabatic approximation. Let $\hat{H}$ be the Hamiltonian for nuclear wave functions in the adiabatic approximation (equation (A.15)). $\hat{H} = \hat{T}_n + \hat{U}(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)$, where $\hat{T}$ is the nuclear kinetic energy operator and $U(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)$ is the potential energy landscape. $\hat{U} \equiv \hat{V}_{n-n} + \hat{E}(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)$, where $\hat{V}_{n-n}$ is the operator for electrostatic repulsion between nuclei and $\hat{E}(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)$ is the total *electronic* energy operator that depends parametrically on the nuclear coordinates.

2.5.1 Theory of harmonic crystals

This section briefly reviews the theory of harmonic crystals. Crystals can be thought of as balls (atoms) connected to each other by springs (chemical bonds). This is a valid representation as long as the total energy is close to the minimum, such that we can perform a Taylor expansion of the total potential energy felt by each atom (at equilibrium position r_0) in terms of displacements and retain only the zeroth and second order terms, i.e.,

$$U = U^0 + \frac{1}{2} \sum_{ls\alpha, l's'\beta} \frac{\partial^2 U}{\partial(R_{l\alpha} + \tau_{s\alpha})\partial(R_{l'\beta} + \tau_{s'\beta})} u_{ls\alpha} u_{l's'\beta} \tag{2.11}$$

Here, $u_I(t)$ is the displacement of atom I from its equilibrium position $R_I + \tau_I$, where R is the position of the unit cell and τ is the position of the atom within the unit cell. Let us define the spring constant as,

$$K_{ls\alpha, l's'\beta} \equiv \frac{\partial^2 U}{\partial(R_{l\alpha} + \tau_{s\alpha})\partial(R_{l'\beta} + \tau_{s'\beta})}, \tag{2.12}$$

which we evaluate at the relaxed configuration. Recall that U is the potential energy landscape of the *atoms*, which in the adiabatic approximation (equation (A.15)) is giving by the sum of the nuclear-nuclear interaction and the total electronic energy for a given set of atomic coordinates.

In Einsteins' repeated summation notation, the equation of motion for atom s is,

$$M_s \frac{d^2}{dt^2} u_{ls\alpha} = -K_{ls\alpha, l's'\beta} u_{l's'\beta} \tag{2.13}$$

Since this is a wave equation, we can look for solutions of the form ,

$$u_{ls\alpha} = u^0_{s\alpha} e^{i(\mathbf{q}\cdot(\mathbf{R}_l + \tau_s) - \omega t)},$$

where $u^0_{s\alpha}$ is a constant that is periodic in the unit cell. Notice that we have used tensor notation, so u is actually a vector. Substituting the Ansatz for u into the

equation of motion gives an eigenvalue equation,

$$D_{s\alpha,s'\beta}(\mathbf{q})v^0_{s'\beta} = \omega^2 v^0_{s\alpha} \tag{2.14}$$

$$v^0_{s\alpha} = M_s^{1/2} u^0_{s\alpha} \tag{2.15}$$

$$D_{s\alpha,s'\beta}(\mathbf{q}) = \frac{1}{M_s M_{s'}} \sum_{\mathbf{R}_l} e^{i\mathbf{q}\cdot(\tau_{s'}-\tau_s)} K_{0s\alpha,ls'\beta} e^{i\mathbf{q}\cdot\mathbf{R}_l}. \tag{2.16}$$

Note that we have made use of the fact that $K_{ls\alpha,l's'\beta}$ is invariant to rigid translations by a lattice vector, and written it as $K_{0s\alpha,l's'\beta}$. $D(\mathbf{q})$ is the dynamical matrix; it is the Fourier transform of the spring constants with respect to the lattice. The dynamical matrix plays the central role in the calculation of phonon dispersion. The eigenvalues of the dynamical matrix correspond to $\omega^2_{\nu\mathbf{q}}$, the squared phonon frequencies, and the corresponding eigenvectors $e_{s\alpha}(\mathbf{q})$ (related to but *not equal to* $v^0_{s\alpha}$) represent (mass-reduced) non-interacting vibrational models that each obey the equation of motion of a harmonic oscillator. These collective displacement modes are called *phonon modes* or *normal modes*.

Therefore, the central challenge for density-functional theory is to construct the $3M \times 3M$ dynamical matrix. Once the dynamical matrix is obtained, the phonon frequencies and modes can be easily calculated by matrix diagonalization. Since the dynamical matrix is the Fourier transform of the second-order partial derivative of the total potential energy landscape of atoms with respect to atomic displacements, the challenge is to efficiently calculate these partial derivatives.

2.5.2 Partial derivatives of the potential energy

The first partial derivative of the potential energy landscape with respect to atomic displacements is the *force*. The force on atom s in the unit cell is given by the *Hellmann-Feynman theorem*,

$$F_{s\alpha} = -\frac{\partial}{\partial\lambda_\alpha} \langle\Psi| \hat{U} |\Psi\rangle = -\langle\Psi| \frac{\partial\hat{U}}{\partial\lambda_\alpha} |\Psi\rangle, \tag{2.17}$$

where $\lambda_\alpha \equiv R_{ls\alpha} + \tau_{s\alpha}$, and Ψ is the many-body wave function that includes both nuclear and electronic degrees of freedom. Without using the Hellmann-Feynman theorem, we would have needed to evaluate the potential energy by solving the self-consistent Kohn-Sham equations at *multiple* atomic coordinates and used finite differences to evaluate the force. Thanks to the Hellman-Feynman theorem, we can simply evaluate the expectation value of the derivative of the potential energy operator at a *single* atomic coordinate.

The force on atom s is evaluated as,

$$\begin{aligned} F_{s,\alpha} &= \langle\Psi|\, \partial_{\lambda_\alpha} \hat{V}_{n-n} + \partial_{\lambda_\alpha} \hat{V}_{e-n} + \partial_{\lambda_\alpha} \hat{T}_e + \partial_{\lambda_\alpha} \hat{V}_{e-e}\, |\Psi\rangle \\ &= \langle\Psi|\, \partial_{\lambda_\alpha} \hat{V}_{n-n} + \partial_{\lambda_\alpha} \hat{V}_{e-n}\, |\Psi\rangle \\ &= \langle\chi|\, \partial_{\lambda_\alpha} \hat{V}_{n-n}\, |\chi\rangle + \langle\Psi_e|\, \partial_{\lambda_\alpha} \hat{V}_{e-n}\, |\Psi_e\rangle \end{aligned} \tag{2.18}$$

As is clear from the equation above, the only contributions to the force on the atom is the electrostatic repulsion between the nuclei (first term) and the electrostatic attraction between electrons and the nuclei (second term). In other words, the forces that hold solids together are classical electrostatic forces and the complicated many-body quantum mechanical interactions simply drop out. In the clamped nuclei approximation, the first term has a simple algebraic (inverse polynomial) form, and the term is easy to evaluate. The second term involves an integral and takes the form,

$$\langle\Psi_e|\, \partial_{\lambda_\alpha} \hat{V}_{e-n}\, |\Psi_e\rangle = \int d^3\mathbf{r} n(\mathbf{r}) \partial_{\lambda_\alpha} V_\lambda(r),$$

where $V_\lambda(r)$ corresponds to the Coulomb force between electrons and nuclei, which has a simple algebraic form.

Similarly, the spring constant can be written as,

$$\begin{aligned} K_{ls\alpha,l's'\beta} &= \frac{\partial}{\partial\lambda_\beta} \langle\psi_c|\, \frac{\partial \hat{U}}{\partial\lambda_\alpha}\, |\psi_v\rangle \\ &= \langle\chi|\, \partial^2_{\lambda_\beta,\lambda_\alpha} \hat{V}_{n-n}\, |\chi\rangle + \frac{\partial}{\partial\lambda_\beta} \langle\Psi_e|\, \partial_{\lambda_\alpha} \hat{V}_{e-n}\, |\Psi_e\rangle\,. \end{aligned} \tag{2.19}$$

Once again, the first term is easy to evaluate and is a simple inverse polynomial. We focus on the second term, which is more involved,

$$\begin{aligned}\partial_{\lambda_\beta} \langle \Psi_e | \, \partial_{\lambda_\alpha} \hat{V}_{e-n} \, | \Psi_e \rangle &= \partial_{\lambda_\beta} \int d^3 n(\mathbf{r}) \partial_{\lambda_\alpha} V_\lambda(r) \\ &= \int d^3\mathbf{r} \big(\partial_{\lambda_\beta} n(\mathbf{r}) \partial_{\lambda_\alpha} V_\lambda(\mathbf{r}) + n(\mathbf{r}) \partial^2_{\lambda_\beta, \lambda_\alpha} V_\lambda(\mathbf{r}) \big). \end{aligned} \tag{2.20}$$

In contrast to atomic forces, which only depends on the charge density, the spring constant depends on the partial derivative of the charge density to atomic displacement, as seen by the appearance of $\partial_{\lambda_\beta} n(\mathbf{r})$ in the first term above. It is this term that makes the construction of the dynamical matrix difficult. One option to evaluate it is with finite differences, where one constructs supercells and displaces the atoms by a small amount. Not only is this method computationally expensive but the vibrational modes that can be calculated with this method are restricted to those which can be represented by finite supercells. An alternative approach that gives access to eigenvectors throughout the Brillouin zone with primitive-cell calculations is to use *linear-response theory.*

2.5.3 Linear-response theory

In the spirit of linear-response theory, we ask how how a small atomic displacement changes the charge density. To answer this, we linearize the charge density in terms of the atomic displacement of a single atom s, $\Delta\lambda_\alpha \equiv R_{l's'\alpha} - R_{ls\alpha} + \tau_{s'\alpha} - \tau_{s\alpha}$,

$$n(\mathbf{r}) = n_0(\mathbf{r}) + \chi_\alpha \Delta\lambda_\alpha + O(d\lambda^2), \tag{2.21}$$

where χ_α is the linear response of the charge density to a small atomic displacement, $\chi_\alpha \equiv \partial n/\partial\lambda_\alpha$. The first-order correction to the charge density is then,

$$\begin{aligned}\Delta n &= \frac{\partial n}{\partial \lambda_\alpha}\Delta\lambda_\alpha \\ &= \sum_i \big(\frac{\delta n}{\delta\psi_i}\frac{\partial\psi_i}{\partial\lambda_\alpha} + \frac{\delta n}{\delta\psi_i^*}\frac{\partial\psi_i^*}{\partial\lambda_\alpha}\big)\Delta\lambda_\alpha \\ &= 2\sum_i \psi_i^*\Delta\psi_i + \psi_i\Delta\psi_i^* \\ &= 4\sum_i \mathrm{Re}\left(\psi_i^*\Delta\psi_i\right)\end{aligned} \tag{2.22}$$

where we have defined $\Delta\psi_i \equiv (\partial\psi_i/\partial\lambda_\alpha)\Delta\lambda_\alpha$, which turns out to be the central quantity we need to evaluate. If the nuclear displacements are small, the idea is that we can approximate the derivative of ψ as $\partial\psi_i/\partial\lambda_\beta \approx \Delta\psi_i/\Delta\lambda_i$, where $\Delta\psi_i(\mathbf{r})$ is the first-order correction to the wave function from standard perturbation theory due to the response of the background SCF potential $V_{SCF}(\mathbf{r})$ to a small atomic displacement $\Delta\lambda_i$, which we denote $\Delta V_{SCF}(\mathbf{r})$. There are two challenges with this approach. One challenge is the evaluation of $\Delta\psi_i$, which requires a sum over many empty states,

$$\Delta\psi_i(r) = \sum_{i\neq j} \frac{\langle\psi_j|\,\Delta V_{SCF}\,|\psi_i\rangle}{\varepsilon_i - \varepsilon_j}\psi_j(\mathbf{r}). \tag{2.23}$$

The second challenge is the dependence of $\Delta V_{SCF}(\mathbf{r})$ on the first order correction to the charge density $\Delta n(\mathbf{r})$ and thus the first order correction to the wave function $\Delta\psi_i(\mathbf{r})$, which leads to a non-linear self-consistent equation. Although we simply have to live with the self-consistency problem, we can avoid having to sum over empty states by linearizing the Kohn-Sham equation.

To obtain $\Delta\psi_i$, we linearize the Kohn-Sham equation in terms of the atomic dis-

placement of a single atom ($H_{SCF} \equiv -\frac{1}{2}\nabla^2 + V_{SCF}(r)$),

$$\begin{aligned}
\frac{\partial}{\partial\lambda_\alpha}\big(H_{SCF}\psi_i - \varepsilon_i\psi_i\big) &= 0 \\
\frac{\partial H_{SCF}}{\partial\lambda_\alpha}\psi_i + H_{SCF}\frac{\partial\psi_i}{\partial\lambda_\alpha} - \frac{\partial\varepsilon_i}{\partial\lambda_\alpha}\psi_i - \varepsilon_i\frac{\partial\psi_i}{\partial\lambda_\alpha} &= 0 \\
\big(H_{SCF} - \varepsilon_i\big)\frac{\partial\psi_i}{\partial\lambda_\alpha}\Delta\lambda_\alpha &= -\big(\frac{\partial H_{SCF}}{\partial\lambda_\alpha} - \frac{\partial\varepsilon_i}{\partial\lambda_\alpha}\big)\psi_i\Delta\lambda_\alpha \\
\big(H_{SCF} - \varepsilon_i\big)\Delta\psi_i &= -\big(\Delta V_{SCF} - \Delta\varepsilon_i\big)\psi_i. \qquad (2.24)
\end{aligned}$$

To obtain $\Delta\psi_i$, we simply need to solve the linear equation (2.24), which is also known as the *Sternheimer equation*. Clearly, in this approach, we circumvent the need to sum over empty states to calcualte $\Delta\psi_i$. ΔV_{SCF} and $\Delta\varepsilon_i$ are defined as,

$$\Delta V_{SCF} = \Delta V_{e-n}(r) + \int d^3\mathbf{r}'\frac{\Delta n(r')}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta v_{xc}}{\delta n}\Delta n(\mathbf{r}), \qquad (2.25)$$

$$\Delta\varepsilon_i = \langle\psi_i|\,\Delta V_{SCF}\,|\psi_i\rangle\,. \qquad (2.26)$$

Equations 2.22, 2.24, 2.25, and 2.26 constitute the main equations of density-functional perturbation theory that must be solved self-consistently. Finally, once $\Delta\psi_i$ is known, the dynamical matrix can be constructed by evaluating the inter-atomic spring constants using the Hellman-Feynman theorem, which is then diagonalized to yield the phonon eigenvectors and eigenfrequencies.

2.6 Modelling electronic transport

Modelling electron transport enables us to identify the mechanisms that limit the ability of a material to transport electrical energy. From the classical Drude theory, it is well known that energy losses happen from electron collisions. Newton's equation for an electron in an electric field $\mathbf{E}$, undergoing phenomenological dampening characterized by the lifetime τ, is,

$$\frac{d\mathbf{p}}{dt} + \frac{\mathbf{p}}{\tau} = -e\mathbf{E}(t) \qquad (2.27)$$

By taking the Fourier transform of this equation, we obtain,

$$i\omega\mathbf{p}(\omega) + \frac{\mathbf{p}(\omega)}{\tau} = -e\mathbf{E}(\omega)$$
$$\mathbf{p}(\omega) = -\frac{e\mathbf{E}(\omega)}{(i\omega + \frac{1}{\tau})} \tag{2.28}$$

Equation 2.28 gives us the net momentum 'response' of a system to which a harmonic electric field is applied. For a static field, this reduces to:

$$\mathbf{p} = -e\mathbf{E}\tau \tag{2.29}$$

In the absence of collisions, an electric field would increase the momenta of electrons forever according to $\mathbf{F} = m\mathbf{a}$. Equation 2.29 tells us that collisions will cause the momenta of electrons to 'randomize' or 'relax' every τ seconds, leading to a finite momentum. This naturally leads us to a definition of *mobility*. We define mobility as the linear response of the net velocity of electrons to an applied electric field, and substitute the result of equation 2.29,

$$\mu \equiv -\frac{\mathbf{v}}{\mathbf{E}} = -\frac{\mathbf{p}}{m\mathbf{E}} = \frac{e\tau}{m^*} \tag{2.30}$$

Clearly, the mobility depends on the scattering lifetime τ, so the challenge of first-principles calculations is to estimate τ in materials accurately. Among many things, scattering can arise from interactions of electrons with the lattice, electron-electron interactions, interactions of electrons with a disordered potential landscape, as well interactions of electrons with impurities. The rates of scattering for all these interactions can be calculated from first principles.

2.6.1 Electron-phonon interactions

The fundamental limit to the mobility in ordered materials is set by phonons. In order to calculate phonon-limited mobility, we need to know the rate at which electrons scatter with lattice vibrations. [14] The matrix element needed for calculating this

rate is called the *electron-phonon matrix element* or *electron-phonon vertex*, and defined as,

$$g_{nm\nu}(\mathbf{k},\mathbf{q}) \equiv \langle \psi_{m\mathbf{k}+\mathbf{q}} | \, \Delta V_{SCF}^{\nu\mathbf{q}} \, | \psi_{n\mathbf{k}} \rangle \,, \tag{2.31}$$

where $|\nu\mathbf{q}\rangle$ indexes the phonon mode, the *u.c.* subscript denotes the integration takes place over the unit cell, and $\Delta V_{SCF}^{\nu\mathbf{q}}$ is the first-order correction to the SCF potential due to a collective displacement of atoms by phonon mode $|\nu\mathbf{q}\rangle$,

$$\Delta V_{SCF}^{\nu\mathbf{q}} = \frac{e^{i\mathbf{q}\cdot\mathbf{r}}}{\sqrt{2M_0\omega_{\mathbf{q}\nu}}} \sum_{s,\alpha} \sqrt{\frac{M_0}{M_s}} \mathbf{e}_{s\alpha\nu}(\mathbf{q}) \sum_{\mathbf{R}} e^{-i\mathbf{q}\cdot\mathbf{R}} \frac{\partial V_{KS}}{\partial \tau_{s\alpha}}\Big|_{\mathbf{r}-\mathbf{R}} \tag{2.32}$$

$$\equiv e^{i\mathbf{q}\cdot\mathbf{r}} \Delta v_{SCF}^{\nu\mathbf{q}} \tag{2.33}$$

where $\partial V_{KS}/\partial \tau_{\kappa\alpha}$ is obtained by calculating ΔV_{KS} for a small atomic displacement $\Delta\lambda_\alpha \equiv \Delta\tau_{s\alpha}$ using the Sternheimer equation, M_0 is a reference mass that ultimately cancels out (included for numerical stability), M_s is the mass of atom s, $\mathbf{e}_{s\alpha\nu}$ in cell-periodic part of the phonon eigenvector. It is common to write the electron-phonon matrix element to explicitly show that only integration over the unit cell (as opposed to the entire crystal volume) is needed,

$$g_{mn\nu}(\mathbf{k},\mathbf{q}) \equiv \frac{1}{\Omega_{u.c.}} \langle u_{m\mathbf{k}+\mathbf{q}} | \, \Delta v_{SCF}^{\nu\mathbf{q}} \, | u_{n\mathbf{k}} \rangle_{u.c.} \,, \tag{2.34}$$

where $\Omega_{u.c.}$ is the volume of the unit cell (u.c.).

2.6.1.1 Wannier interpolation

For practical calculations of electron mobility, the electron-phonon matrix elements are needed in very fine k-space grids with $O(10^6)$ grid points. However, density-functional perturbation theory can only reasonably calculate matrix elements for $O(10^3)$ grids. This challenge is overcome by *Fourier interpolating* the coarse DFPT electron-phonon matrix elements to very fine grids using the method of *maximally localized Wannier functions*, which makes use of the fact that both electronic orbitals and atomic displacements are localized in real space and therefore require smaller

basis sets to converge, in order to perform efficient interpolations in k-space. [15, 14] In practice, certain phonon modes near the Brillouin zone center ($\mathbf{q} \to 0$) have divergences or discontinuities due to dipole or quadrupole moments. Because of this, the real space representation of these modes do not have a localized basis; however, their k-space behaviour is exactly described by analytical expressions. In such cases, the long-range parts of the matrix element are treated analytically and the short-range components are Fourier interpolated.

2.6.2 Fermi's golden rule

The connection between carrier mobility and scattering relies on the fact that the mobility is proportional to the scattering lifetime, $\mu \propto \langle \tau \rangle$. In the interaction picture, we are interested in the transition probability a state $|i\rangle$ scattering to some state $|f\rangle$,

$$P_{i \to f}(t) \equiv |\langle f| U(t, t_0) |i\rangle|^2, \tag{2.35}$$

where $U(t, t_0)$ is the time evolution operator in the *interaction picture* sandwiched between states $|i\rangle$ and $|f\rangle$. In the interaction picture, the Schrödinger equation takes the form,

$$i\frac{\partial}{\partial t} U(t, t_0)\psi(t_0) = V_I(t) U(t, t_0)\psi(t_0), \tag{2.36}$$

where $V_I(t) \equiv e^{iH_0 t} V(t) e^{-iH_0 t}$ is the perturbation in the interaction picture. Integrating this equation leads to the following self-consistent equation,

$$U(t, t_0) = U(t_0, t_0) - i \int_{t_0}^{t} dt_1 V_I(t_1) U(t_1, t_0). \tag{2.37}$$

We will use the normalization condition $U(t_0, t_0) = 1$. Similar to the Born series, we can expand this equation into the so-called *Dyson series*,

$$\begin{aligned} U(t, t_0) = U(t_0, t_0) \\ &- i \int_{t_0}^{t} dt_1 V_I(t_1) U(t_1, t_0) \\ &- \int_{t_0}^{t} dt_1 V_I(t_1) \int_{t_0}^{t_1} dt_2 V_I(t_2) U(t_2, t_0) \\ &+ O(V_I^3) \end{aligned} \tag{2.38}$$

To *first order* in the perturbation, the Dyson series is truncated as,

$$U(t, t_0) \approx \mathbf{1} - i \int_{t_0}^{t} dt_1 V_I(t_1) U(t_1, t_0). \tag{2.39}$$

Now, consider a *time-independent* perturbation. To ensure this perturbation is smoothly turned on, we can write it as $V(t_1) = Ve^{\eta t}$. To ensure this perturbation is time-independent for all finite times, we will take the limit $\eta \to 0+$. However, for $t_0 \to -\infty$, $V(t_0) \to 0$ thus the perturbation is zero in the distant past. (In the distant future, $V(t) \to \infty$, however we will not integrate that far into the future.) Taking $t_0 \to -\infty$, the transition probability *amplitude* T_{if} is,

$$T_{if} = \langle f|i\rangle - i \int_{-\infty}^{t} dt_1 \left\langle f\right| e^{iH_0 t_1} V e^{\eta t_1} e^{-iH_0 t_1} \left|i\right\rangle, \tag{2.40}$$

where we have explicitly written out the time evolution operator due to the unperturbed Hamiltonian. Acting on $\langle f|$ and $|i\rangle$ with the time evolution operators, we can

rewrite the transition amplitude as,

$$\begin{aligned} T_{if} &= \langle f|i\rangle - i \langle f| V |i\rangle \int_{-\infty}^{t} dt_1 e^{i(\varepsilon_f - \varepsilon_i)t_1} e^{\eta t_1} \\ &= \langle f|i\rangle - i \langle f| V |i\rangle \frac{e^{i(\varepsilon_f - \varepsilon_i)t} e^{\eta t}}{i(\varepsilon_f - \varepsilon_i) + \eta} \end{aligned} \tag{2.41}$$

If $\langle i|$ and $|f\rangle$ are eigenstates of the unperturbed Hamiltonian, they are orthogonal so $\langle i|f\rangle = 0$. Squaring the transition amplitude to get the transition probability $P_{if} \equiv |T_{if}|^2$, and taking the limit $\eta \to 0$ gives a delta function from the Lorentzian,

$$P_{if} = 2\pi |\langle f| V |i\rangle|^2 \delta(\varepsilon_f - \varepsilon_i). \tag{2.42}$$

This expression states that the transition probability from an initial state to a final state is proportional to the squared matrix element that couples them and a delta function that enforces energy conservation between the initial and final state. For the situation where the perturbation is a harmonic potential, e.g., $V(t) = Ve^{\pm i\omega t}$, the derivation is more complicated but the result is simple,

$$P_{if} = 2\pi |\langle f| V |i\rangle|^2 \delta(\varepsilon_f - \varepsilon_i \pm \omega), \tag{2.43}$$

meaning the perturbation can now couple states with energies differing by the frequency of the harmonic potential. The overall transition rate from an initial state i to any final state f is given by summing over all final states,

$$R_{i\to f} = 2\pi \sum_f |\langle f| V |i\rangle|^2 \delta(\varepsilon_f - \varepsilon_i \pm \omega). \tag{2.44}$$

This is the standard Fermi's golden rule of time-dependent perturbation theory (for the case $\hbar = 1$). We use this expression as the starting point to calculate the rates of scattering processes of not just electron-phonon scattering, but also electron-photon scattering (optical absorption and emission) and electron-electron scattering (e.g., Auger-Meitner recombination). If the occupation of fermions (electrons) or bosons

(phonon, photon) is important to the scattering problem, Fermi's golden rule is typically modified so as to prevent scattering to states already occupied by a fermion as well as to account for the fact that scattering by boson absorption depends on the occupation factor of the boson. While this is typically done phenomenologically, such expressions can be systematically derived starting from a many-body Green's function description of the problem.

2.6.3 Quantum theory of scattering

This section is dedicated to the quantum theory of scattering, which we do not explicitly use to model scattering processes but the insights of which will be useful in understanding the approximations that we have implicitly made upon using Fermi's golden rule to calculate the scattering integral of the Boltzmann transport equation. Consider a Hamiltonian $H = H_0 + V$, where H_0 is the unperturbed Hamiltonian (typically, at the DFT or G_0W_0 level), and V is a perturbation. The Schrödinger equation can be written as,

$$(H_0 - \varepsilon)\phi(r) = V(\mathbf{r})\phi(\mathbf{r}). \tag{2.45}$$

If we view the term $V(\mathbf{r})\phi(\mathbf{r})$ as a source term and replace it with a Dirac δ function, we obtain a partial differential equation for the Green's function (impulse response) of the system,

$$(H_0 - \varepsilon)G(\mathbf{r}, \mathbf{r}') = \delta^{(3)}(\mathbf{r} - \mathbf{r}'). \tag{2.46}$$

We can invert this equation as,

$$G(\mathbf{r}, \mathbf{r}') = (H_0 - \varepsilon - i\eta)^{-1}\delta^{(3)}(\mathbf{r} - \mathbf{r}'). \tag{2.47}$$

Here, the inverse is the *resolvent* of the operator: the inverse is only taken for energies for which the operator is not singular. The poles of the Green's function correspond to the natural frequencies or excitation energies of the system. We have also added a small term $i\eta$, $\eta > 0$, keeping in mind that we will later take the limit $\eta \to 0$. (We have done this to ensure stability of the solutions; recall, the time-dependent

wave function has the term $\exp(-iEt)$, so adding an imaginary term has the effect of dampening it as $\exp(iEt - \epsilon t)$. Once the Green's function is known, the wave function can be reconstructed for any arbitrary potential via a convolution,

$$\phi(\mathbf{r}) = \int d^3\mathbf{r}' G(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \phi(\mathbf{r}'). \tag{2.48}$$

This is simply a restatement of the Schrödinger equation as an integral equation, and is known as the *Lippmann-Schwinger* equation. Since $\phi(\mathbf{r})$ is on both sides, this equation must be solved self-consistently.

2.6.3.1 Born approximation

If the perturbation $V(\mathbf{r})$ is weak compared to the unperturbed Hamiltonian, then it is possible to write this equation as a convergent series that can be truncated up to a desired order in the perturbation. This expansion is known as the *Born series*,

$$\begin{aligned}\phi(\mathbf{r}) = \phi_0(\mathbf{r}) &\\ &+ \int d^3\mathbf{r}' G(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \phi_0(\mathbf{r}') \\ &+ \int d^3\mathbf{r}' G(\mathbf{r}, \mathbf{r}') V(\mathbf{r}') \int d^3\mathbf{r}'' G(\mathbf{r}', \mathbf{r}'') V(\mathbf{r}'') \phi_0(\mathbf{r}'') \\ &+ O(V^3),\end{aligned} \tag{2.49}$$

where ϕ_0 is the incoming wave, and we identify all higher order terms as the scattered waves. We obtain the *first-order Born approximation* if we truncate all terms greater than $O(V)$. The first Born approximation corresponds to considering only incoherent single scattering events. For the purposes of this thesis, it is sufficient to know that one implicitly makes the first Born approximation in calculating scattering rates within Fermi's golden rule. Although the first Born approximation works well for calculating carrier mobility, capturing transport effects at low temperature, such as weak localization, requires going beyond the first Born approximation, which typically requires many-body methods based on non-equilibrium Green's functions.

2.6.4 Boltzmann transport equation

The Boltzmann equation describes how a distribution of particles $f(\mathbf{r}, \mathbf{p})$ traverses through phase space ($\mathbf{r}$, $\mathbf{p}$) under Newton's Laws of Motion. Boltzmann equation is often presented as a law, like Newton's Law, but it can be formally derived for electrons in solids from the many-body Kadanoff-Baym equations. We skip the derivation and write phenomenologically,

$$\frac{df}{dt} = \frac{\partial f}{\partial t}\Big|_{scatt.} + \frac{\partial f}{\partial t}\Big|_{drift}, \tag{2.50}$$

where we have ignored the diffusion term, assuming the material is spatially homogeneous. This equation simply states that the change in the distribution function is due to particles scattering in or out from collision events or because the electric field accelerates carriers and thus changes its occupation in phase space. In equilibrium, $df/dt = 0$, and the Boltzmann Transport Equation takes the form,

$$S[f] + \mathbf{E} \cdot \nabla_{\mathbf{k}} f = 0, \tag{2.51}$$

where $S[f]$ is a scattering integral, which is generally a non-linear functional of the distribution function. The steady-state Boltzmann transport, which is a non-linear integral equation, is rarely solved directly. Instead, it is linearized,

$$\frac{\partial}{\partial E_\beta} S[f] + \frac{\partial f}{\partial k_\beta} \frac{\partial f}{\partial E_\beta} = 0. \tag{2.52}$$

The exact form of this equation obviously depends on the scattering integral. Within the first Born approximation, the scattering integral is generally a variation of Fermi's golden rule. In the linearized Boltzmann tranport equation, one generally solves for $\partial_{E_\beta} f$, from which the mobility can be calculated as,

$$\mu_{\alpha\beta} \equiv \frac{1}{n_c} \frac{\partial j_\alpha}{\partial E_\beta} = \frac{1}{n_c \Omega_{u.c.}} \frac{\partial}{\partial E_\beta} \sum_{n\mathbf{k}} v^\alpha_{n\mathbf{k}} f_{n\mathbf{k}} = \frac{1}{n_c \Omega_{u.c.}} \sum_{n\mathbf{k}} v^\alpha_{n\mathbf{k}} \frac{\partial f_{n\mathbf{k}}}{\partial E_\beta}, \tag{2.53}$$

where $v_{n\mathbf{k}}$ is the band velocity, n_c is the carrier concentration in the unit cell, and $\Omega_{u.c.}$ is the volume of the unit cell. [16]

2.7 Modelling alloy disorder

Modelling alloys from first principles is highly complicated because of their low symmetry. The brute force approach to modelling alloys is to model a very large simulation cell with N atoms, such that $N \to \infty$. In an alloy with N sites, there are 2^N possible permutations. In order to calculate the partition function, in principle one needs to calculate the energy of all 2^N structures, which is quite obviously impractical. Therefore, approximations are needed to calculate observables for alloys.

For the case of random alloys, where the probability of a site hosting atom A or atom B is independent of its position, the simplest approximation is the *virtual-crystal approximation*, where the alloy is modelled as a homogeneous solid whose constituent atoms are taken to be an average of the atoms of the parent compounds. In such a scheme, any observable O of a *random* alloy composed of compounds A and B can often be written as,

$$O_{AB}(x) = xO_A + (1-x)O_B - bx(1-x), \tag{2.54}$$

where x is the alloy composition, and b is the *bowing parameter*, which indicates deviation from a linear interpolation. Typically, b is small for structural parameters but can be significant for electronic and optical parameters, such as the band-gap energy. However, the virtual-crystal approximation cannot capture disorder effects, which are typically quite strong in many random alloy systems. It also cannot capture non-random effects such as short or long-range order.

2.7.1 Cluster expansion

One approach to modelling alloys is to perform a *cluster expansion*. [8] Consider the problem of estimating the total energy of an alloy with N atoms in configuration S, composed of atoms A and B. In the most naive approximation, we could approximate

the total energy of the alloy by simply adding the energy contribution from the individual atoms,

$$E_{AB}(x) \approx N_A E_A + N_B E_B,$$

where E_A is the energy of atom A and N_A is the number of A atoms. This is similar in spirit to the virtual-crystal approximation, but it is obviously not a good approximation because an alloy is not a linear superposition of non-interacting atoms. What we are missing in the energy expression is the terms that represent *interactions between different atoms*. Each of these interactions can be thought of as contributing a correction to the energy. Phrased in a different way, every time there exist certain *structural correlation* between atoms, the energy corrections corresponding to these correlations must be included in the energy expression. The most trivial configuration of atoms possible is a single atom. The next simplest configuration is a *pair* of like atoms; these are not limited to nearest neighbours, but we can easily imagine (due to *locality*) that the energy correction due to pairs of atoms vanishes with increasing separation distance. Higher order configurations include *triplets*, *quadruplets*, and so on. Typically, the word *configuration* is reserved to describe the configuration of atoms in the the entire crystal structure. The configuration of atoms that represent structural correlations are called *figures* or *clusters*, and the act of expanding any property, such as the total energy, in terms of these figures is called a *cluster expansion*.

The full cluster expansion of the total energy of a given configuration σ can be written as,

$$E_\sigma = \sum_f m_f X_{f\sigma} \varepsilon_f, \tag{2.55}$$

where f is an index that sums over all clusters (figures) in the configuration, m_f is the number of times a cluster f appears in a configuration according to the symmetry of the lattice, $X_{f\sigma}$ is called the correlation function and represents the frequency with which a cluster appears in the configuration, and ε_f is the energy contribution of cluster f. This cluster expansion is exact, as long as the sum over f is not truncated (the reason is that $X_{f\sigma}$ forms a complete orthonormal basis, e.g., see Wei *et al*). $X_{f\sigma}$

is defined as,

$$X_{f\sigma} \equiv \frac{1}{m_f} \sum_{\beta \in f} \prod_{i \in \beta} S_i, \tag{2.56}$$

where β sums over all figures f in the configuration, i sums over all lattice points in the figure β, and $S_i = +1$ if site i is occupied by atom A and 0 (or -1) if site i is occupied by atom B. Of course, what we really want is the energy averaged over the ensemble of all allowed configurations σ,

$$\langle E_\sigma \rangle = \sum_f m_f \langle X_{f\sigma} \rangle \varepsilon_f, \tag{2.57}$$

where we have used the fact that the expectation operator is linear, $\langle X + Y \rangle = \langle X \rangle + \langle Y \rangle$. Thus, we have rewritten the problem of calculating the total energy as calculating the energy contribution of individual figures and calculating the expectation value of the correlation functions over all configurations. By truncating the sum over f, we can now systematically estimate the total energy of the alloy without explicitly evaluating the total energy.

2.7.2 Special quasirandom structures

Looking at equation (2.57), we ask the question if we can estimate the ensemble average of the total energy with the energy of a single representative configuration, which we will term the *special structure* (SS). This should be possible if all the correlation functions $X_{f,SS}$ of our special configuration exactly match the ensemble averages $\langle X_{f\sigma} \rangle$. If and only if this condition is met then the energy of the SS,

$$E_{SS} \equiv \sum_f m_f X_{f,SS} \varepsilon_f, \tag{2.58}$$

exactly equals $\langle E_\sigma \rangle$. Thus, the problem is transformed to one of designing the SS such that its structural correlation functions match the desired target correlation function.

In general, we do not know *a priori* the values for $\langle X_{f\sigma} \rangle$. However, for the special case

of *random* alloys, $\langle X_{f\sigma} \rangle_R$ have a simple analytical form. What allows us to derive an analytical expression for $\langle X_{f\sigma} \rangle_R$ is the fact that the probability of a lattice hosting atom A or B is completely uncorrelated or *statistically independent* from another site having atom A or B, which means, $\left\langle \prod_{i \in f} S_i \right\rangle_R = \prod_{i \in f} \langle S_i \rangle_R = \prod_{i \in f} x = x^{k_f}$, where k_f is the number of vertices (lattice points) in the figure f. We have assumed S_i can take values of 1 or 0; if S_i takes values of 1 or -1 then $\left\langle \prod_{i \in f} S_f \right\rangle = \prod_{i \in f} (x \times (+1) + (1-x) \times (-1)) = (2x-1)^{k_f}$. Choosing the convention where S_i is 0 or 1,

$$\langle X_{f,\sigma} \rangle_R = x^{k_f}. \tag{2.59}$$

Therefore, we can employ stochastic optimization methods, such as simulated annealing, to design *special quasirandom structure SQS* whose correlations function match the ensemble average of random alloys, $X_{f,SQS} = x^{k_f}$.

It turns out that even relatively small SQS supercells estimate the properties of random alloys (total energy, band gap, structural properties, etc.) remarkably well. This is because observable properties only depend on *local* structural information. This idea is related to Walter Kohn's concept of the *nearsightedness of electrons*. In practice, when generating SQS's, deviations in pairwise correlations up to the second nearest neighbours and triplet correlations up to the next nearest neighbours are minimized. Clearly, the main advantage of SQS's is that they allow us to capture randomness remarkably well, using relatively small supercells with only several hundred to thousands of atoms, without performing configurational averaging. However, care must be taken since small SQS's will show effects of artificial periodicity for small electron wavevectors.

2.7.3 Spectral function from band unfolding

In plane-wave implementations of density-functional theory, periodic boundary conditions are automatically applied when modelling solids. This means that the real space crystal structure, which has a periodicity corresponding to the unit cell, is represented in reciprocal space using a reciprocal lattice that also has periodicity

corresponding to the reciprocal lattice unit cell. However, not all unit cells of the reciprocal lattice are equivalent. The *Wigner-Seitz cell* of the reciprocal lattice is the smallest possible unit cell that exhibits all symmetries of the crystal structure, and is called the *first Brillouin zone*. When representing the electronic structure in reciprocal space, we typically use the first Brillouin zone but this is not the only choice. If we define the unit cell to be larger than the primitive cell by defining larger lattice vectors $\mathbf{R}$, then the Brillouin zone shrinks in volume because the reciprocal lattice vectors $\mathbf{G}$ become smaller. This means that when modelling alloys using modestly large periodic supercells, which have small Brillouin zones, it becomes difficult to interpret the band structure since all the bands have folded in. To address this issue, we can unfold the band structure from the supercell basis to the equivalent primitive cell basis. For alloys, the primitive cell basis that is chosen corresponds to the primitive cell of an equivalent virtual-crystal alloy. By doing this, we can obtain a more accurate representation of the electronic structure and make it easier to interpret the band structure. [17, 18, 19, 20, 21]

The spectral function A_{SC} in the Brillouin zone of the supercell can be represented as a density of states,

$$A_{SC}(\mathbf{K}, \varepsilon) = \sum_{n} \delta(\varepsilon - \varepsilon_{n\mathbf{K}}), \tag{2.60}$$

where capital $\mathbf{K}$ is a wave vector in the supercell Brillouin zone, and n indexes the band energy. The spectral function can be defined for any $\mathbf{q}$ in the entire reciprocal cell (not restricted to the Brillouin zone) as,

$$A(\mathbf{q}, \varepsilon) \equiv \sum_{n\mathbf{K}} |\psi_{n\mathbf{K}}(\mathbf{q})|^2 \delta(\varepsilon - \varepsilon_{n\mathbf{K}}), \tag{2.61}$$

where $|\psi_{n\mathbf{K}}(\mathbf{q})|^2$ is simply the probability of finding particle $|n\mathbf{K}\rangle$ with momentum $\mathbf{q}$. $\psi_{n\mathbf{K}}(\mathbf{q})$ can be straightforwardly calculated by Fourier transforming the Bloch function, $\psi_{n\mathbf{K}}(\mathbf{r}) = \frac{1}{\sqrt{V}} u_{n\mathbf{K}}(\mathbf{r}) e^{i\mathbf{K}\cdot\mathbf{r}} = \frac{1}{\sqrt{V}} \sum_{\mathbf{G}} c_{n\mathbf{K}+\mathbf{G}} e^{i(\mathbf{K}+\mathbf{G})\cdot\mathbf{r}}$. The Fourier transform

is,

$$\begin{aligned}
\psi_{n\mathbf{K}}(\mathbf{q}) &= \int_V d^3\mathbf{r} e^{-i\mathbf{q}\cdot\mathbf{r}} \psi_{n\mathbf{k}}(\mathbf{r}) \\
&= \frac{1}{\sqrt{V}} \sum_{\mathbf{R}} e^{i(\mathbf{K}-\mathbf{q})\cdot\mathbf{R}} \int_\Omega d^3\mathbf{r} u_{n\mathbf{K}}(\mathbf{r}) e^{i(\mathbf{K}-\mathbf{q})\cdot\mathbf{r}} \\
&= \frac{N}{\sqrt{V}} \sum_{\mathbf{G}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} \int_\Omega d^3\mathbf{r} u_{n\mathbf{K}}(\mathbf{r}) e^{i(\mathbf{K}-\mathbf{q})\cdot\mathbf{r}} \\
&= \frac{N}{\sqrt{V}} \sum_{\mathbf{G}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} \int_\Omega d^3\mathbf{r} \sum_{\mathbf{G}'} c_{n\mathbf{G}'+\mathbf{K}} e^{i\mathbf{G}'\cdot\mathbf{r}} e^{i(\mathbf{K}-\mathbf{q})\cdot\mathbf{r}} \\
&= \frac{N}{\sqrt{V}} \sum_{\mathbf{G}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} \sum_{\mathbf{G}'} c_{n\mathbf{G}'+\mathbf{K}} \int_\Omega d^3\mathbf{r} e^{i(\mathbf{K}+\mathbf{G}'-\mathbf{q})\cdot\mathbf{r}} \\
&= \frac{N\sqrt{\Omega}}{\sqrt{V}} \sum_{\mathbf{G}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} \sum_{\mathbf{G}'} c_{n\mathbf{G}'+\mathbf{K}} \delta_{\mathbf{K}+\mathbf{G}'-\mathbf{q},0} \\
&= \sqrt{N} \sum_{\mathbf{G}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} c_{n\mathbf{q}}. \qquad (2.62)
\end{aligned}$$

Plugging this result into the expression for the spectral function gives,

$$\begin{aligned}
A(\mathbf{q},\varepsilon) &= N \sum_{n\mathbf{K}} \sum_{\mathbf{G}\mathbf{G}'} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}} c^*_{n\mathbf{q}} \delta_{\mathbf{K}-\mathbf{q},\mathbf{G}'} c_{n\mathbf{q}} \delta(\varepsilon - \varepsilon_{n\mathbf{K}}) \\
&= N \sum_n |c_{n\mathbf{q}}|^2 \sum_{\mathbf{G}\mathbf{G}'} \delta_{\mathbf{G},\mathbf{G}'} \delta(\varepsilon - \varepsilon_{n\mathbf{q}+\mathbf{G}'}) \\
&= N \sum_n |c_{n\mathbf{q}}|^2 \sum_{\mathbf{G}} \delta(\varepsilon - \varepsilon_{n\mathbf{q}+\mathbf{G}}). \qquad (2.63)
\end{aligned}$$

Recall that $\varepsilon_{n\mathbf{K}}$ is only defined in the first Brillouin zone of the supercell. Therefore, the sum over $\mathbf{G}$ in the equation above is restricted to those vectors that fold $\mathbf{q}$ back into the supercell Brillouin zone. For every wave vector $\mathbf{q}$ defined in the entire reciprocal crystal volume, there exists only a single reciprocal lattice $\mathbf{G}$ vector that folds it into the first Brillouin zone. (The converse relation does not hold.) Therefore, we can remove the sum over $\mathbf{G}$, and specifically denote $\mathbf{G} \to \mathbf{G}_\mathbf{q}$ as the vector that

folds $\mathbf{q}$ into the first Brillouin zone,

$$A(\mathbf{q}, \varepsilon) = N \sum_n |c_{n\mathbf{q}}|^2 \delta(\varepsilon - \varepsilon_{n\mathbf{q}+\mathbf{G}_\mathbf{q}}). \tag{2.64}$$

This expression shows that the unfolding procedure only requires knowledge of the *supercell* Bloch wave functions. We could stop at this point, but it is often necessary to go one step further and *refold* the spectral function onto a different basis.

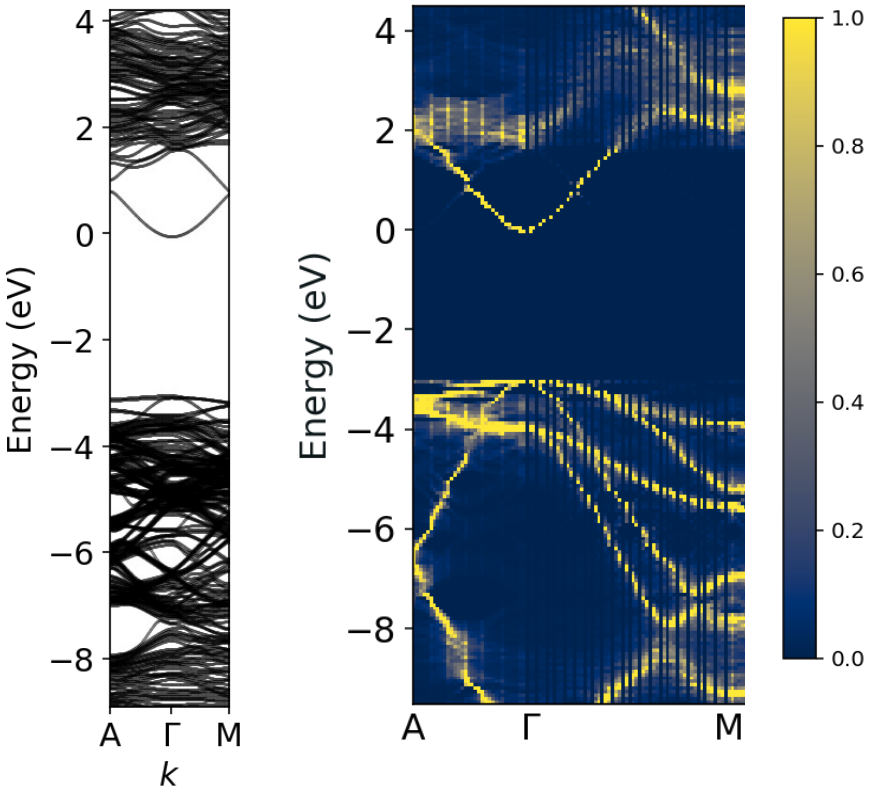

Figure 2.1: Band structure of 128-atom AlGaN supercell from LDA-DFT (left) unfolded onto the primitive cell basis (right).

Alloy disorder can be thought of as a perturbation of a high-symmetry reference structure. For III-nitride alloys, the reference is the virtual-crystal wurtzite structure. Alloy scattering can thus be thought of as a perturbation that causes transitions between eigenstates of the reference high-symmetry structure. To this end, it is necessary for us to refold the fully unfolded spectral function onto the primitive cell basis of the virtual-crystal wurtzite structure. We remark the following folding procedure if considering a *specific* $\mathbf{q}$,

$$\mathbf{k} \leftarrow \mathbf{q} + \mathbf{g}_{\mathbf{q}} \tag{2.65}$$

$$\mathbf{k} + \mathbf{g}_i \rightarrow \mathbf{q}_i; i = 1, 2, ..., N \tag{2.66}$$

Here, $\mathbf{k}$ is a wave vector in the first Brillouin zone of the reference structure and $\mathbf{g}_{\mathbf{q}}$ is a reciprocal lattice vector of the reference structure that folds $\mathbf{q}$ onto the reference Brillouin zone, and i is an index that runs over all N reciprocal lattice vectors g_i. When refolding the spectral function $A(\mathbf{q}, \varepsilon)$, we must consider that *many* $\mathbf{q}$ vectors will fold onto the same $\mathbf{k}$, and we must sum the contribution of all of these $\mathbf{q}$ vectors. This gives us the folding procedure for constructing the spectral function in the primitive cell basis, $A_{PC}(\mathbf{k}, \varepsilon)$,

$$\begin{aligned} A_{PC}(\mathbf{k}, \varepsilon) &= \sum_{\mathbf{g}} A(\mathbf{k} + \mathbf{g}, \varepsilon) \\ &= N \sum_{n\mathbf{g}} |c_{n\mathbf{k}+\mathbf{g}}|^2 \delta(\varepsilon - \varepsilon_{n\mathbf{k}+\mathbf{g}+\mathbf{G}_{\mathbf{k}+\mathbf{g}}}), \end{aligned} \tag{2.67}$$

where the sum over $\mathbf{g}$ is a sum over all the reciprocal lattice vectors of the primitive cell. This is the main equation that is implemented in unfolding the band structure from the supercell basis onto the primitive cell basis. Figure 2.1 shows an example of supercell band structure of a random AlGaN alloy in a 128-atom supercell calculated from LDA-DFT, and its equivalent unfolded band structure in the wurtzite primitive cell basis.

Bibliography

[1] Feliciano Giustino. *Materials modelling using density functional theory: properties and predictions*. Oxford University Press, 2014.

[2] Richard M Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2020.

[3] W Kohn and LJ Sham. Self-consistent equations including exchange and correlation effects. *Physical review*, 140(4A):A1133, 1965.

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

[5] W Kohn. Density-functional theory for excited states in a quasi-local-density approximation. *Physical Review A*, 34(2):737, 1986.

[6] David M Ceperley and Berni J Alder. Ground state of the electron gas by a stochastic method. *Physical review letters*, 45(7):566, 1980.

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

[8] Axel Van De Walle and Gerbrand Ceder. The effect of lattice vibrations on substitutional alloy thermodynamics. *Reviews of Modern Physics*, 74(1):11, 2002.

[9] 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.

[10] Giovanni Onida, Lucia Reining, and Angel Rubio. Electronic excitations: density-functional versus many-body green's-function approaches. *Reviews of modern physics*, 74(2):601, 2002.

[11] Mark S Hybertsen and Steven G Louie. Electron correlation in semiconduc-

tors and insulators: Band gaps and quasiparticle energies. *Physical Review B*, 34(8):5390, 1986.

[12] Lars Hedin and Stig Lundqvist. Effects of electron-electron and electron-phonon interactions on the one-electron states of solids. In *Solid state physics*, volume 23, pages 1–181. Academic Press, 1970.

[13] Stefano Baroni, Stefano De Gironcoli, Andrea Dal Corso, and Paolo Giannozzi. Phonons and related crystal properties from density-functional perturbation theory. *Reviews of modern Physics*, 73(2):515, 2001.

[14] Feliciano Giustino. Electron-phonon interactions from first principles. *Reviews of Modern Physics*, 89(1):015003, 2017.

[15] Nicola Marzari, Arash A. Mostofi, Jonathan R. Yates, Ivo Souza, and David Vanderbilt. Maximally localized wannier functions: Theory and applications. *Reviews of Modern Physics*, 84(4):1419, 2012.

[16] Samuel Poncé, Elena R. Margine, and Feliciano Giustino. Towards predictive many-body calculations of phonon-limited carrier mobilities in semiconductors. *Physical Review B*, 97(12):121201, 2018.

[17] L.W. Wang, L. Bellaiche, S.H. Wei, and A. Zunger. "majority representation" of alloy electronic states. *Physical Review Letters*, 80(21):4725, 1998.

[18] Timothy B. Boykin and Gerhard Klimeck. Practical application of zone-folding concepts in tight-binding calculations. *Physical Review B*, 71(11):115215, 2005.

[19] Wei Ku, Tom Berlijn, and Chi-Cheng Lee. Unfolding first-principles band structures. *Physical Review Letters*, 104(21):216401, 2010.

[20] Voicu Popescu and Alex Zunger. Extracting e versus k effective band structure from supercell calculations on alloys and impurities. *Physical Review B*, 85(8):085201, 2012.

[21] SG Mayo, F Yndurain, and JM Soler. Band unfolding made simple. *Journal of Physics: Condensed Matter*, 32(20):205902, 2020.

CHAPTER III

Semi-Empirical Methods for Modelling Larger Length Scales

3.1 Motivation

Although first-principles atomistic calculations are successful in predicting complex phenomena in materials, they are limited to systems with a few thousand atoms. Semi-empirical theories are necessary to model the interactions between various components of semiconductor devices made of different materials across length scales ranging from nanometers to micrometers. These semi-empirical theories use Hamiltonians that are parameterized to match first-principles calculations of bulk systems. Although less predictive than fully first-principles calculations, semi-empirical theories when parameterized on first-principles data can accurately handle larger systems, which is crucial since qualitatively different phenomena may emerge at larger length scales.

In this chapter, we develop a formalism based on $\mathbf{k} \cdot \mathbf{p}$ perturbation theory to model III-nitride quantum-well heterostructures and alloys used in light-emitting diodes. These systems exhibit phenomena whose length scales greatly exceed the length scale of the unit cell, and thus cannot be captured by standard plane-wave-based first-principles codes. For example, quantum wells of III-nitrides have a polarization-charge discontinuity, which results in strong polarization fields. This effect is known

as the quantum-confined Stark effect and qualitatively changes the physics of recombination compared to the bulk. In chapter V, we will use the techniques developed in this chapter to model InGaN LEDs. We show that the QCSE dominates the optical properties of LEDs and that accurately capturing the physics of large length scales of InGaN LEDs is crucial to modeling them. We also use the techniques developed in this chapter to explain the origin of the green gap in LEDs fabricated with state-of-the-art epitaxy in collaboration with experimentalists. In chapter VII, we model carrier localization within the $\mathbf{k} \cdot \mathbf{p}$ formalism, and investigate the impact of carrier localization on the tolerance of InGaN LEDs to defects. Finally, we combine the $\mathbf{k} \cdot \mathbf{p}$ formalism with supervised machine learning to explore the large configurational landscape of LED designs. This leads us to discover new LED designs that outperform the current state of the art in terms of exhibiting improve efficiency and better spectral characteristics.

3.2 $\mathbf{k} \cdot \mathbf{p}$ perturbation theory

$\mathbf{k} \cdot \mathbf{p}$ perturbation theory is a very powerful technique that allows us to expand wave functions and energies to arbitrary k points from knowledge of wave functions and energies at high symmetry points. [1] The single particle Schrödinger equation is diagonalized at high symmetry point(s) in the Brillouin zone. Using a clever trick, we perturbatively expand the wave functions and energies to k points close to the high symmetry point (which for simplicity we take to be the Γ point) using standard perturbation theory.

We start with the single-particle Schrödinger equation in a periodic solid,

$$\left(\frac{-1}{2}\nabla^2 + V(\mathbf{r})\right)\psi_{n\mathbf{k}}(\mathbf{r}) = \varepsilon_{n\mathbf{k}}\psi_{n\mathbf{k}}(\mathbf{r}), \tag{3.1}$$

where n and $\mathbf{k}$ are band and k-point indices. Noting that the solutions are Bloch functions, $\psi_{n\mathbf{k}}(\mathbf{r}) = u_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}$, and applying the product rule of the Laplacian op-

erator yields,

$$\left(-\frac{1}{2}\nabla^2 - \frac{i}{m}\mathbf{k}\cdot\nabla + \frac{k^2}{2} + V(\mathbf{r})\right) u_{n\mathbf{k}}(\mathbf{r}) = \varepsilon_{n\mathbf{k}} u_{n\mathbf{k}}(\mathbf{r}) \tag{3.2}$$

Recognizing $-\frac{1}{2}\nabla^2 + V(\mathbf{r})$ as the unperturbed Hamiltonian H_0, and performing the canonical substitution $\mathbf{p} \leftarrow -i\nabla$, we can rewrite the equation above as,

$$\left(H_0 + \frac{\mathbf{k}\cdot\mathbf{p}}{m} + \frac{k^2}{2}\right) u_{n\mathbf{k}}(r) = \varepsilon_{n\mathbf{k}} u_{n\mathbf{k}}(\mathbf{r}) \tag{3.3}$$

The $\mathbf{k}\cdot\mathbf{p}$ term, which is small for k close to the high-symmetry point, is cleverly viewed as a perturbation term. The term $k^2/2$ is simply a number and shifts the energy by a constant albeit k-dependent value. We can use the set $\{u_{n\mathbf{k}}(r)\}$, which is complete and orthonormal, as the basis for perturbation theory. For this example, we take $\mathbf{k} = 0$ as the high symmetry point, which is appropriate for direct gap semiconductors. To lowest order, for non-degenerate bands,

$$\left|n\mathbf{k}^{(1)}\right\rangle = \left|n^{(0)}\right\rangle + \sum_{m\neq n} \frac{\left|\mathbf{k}\cdot\left\langle m^{(0)}\right|\mathbf{p}\left|n^{(0)}\right\rangle\right|^2}{\varepsilon_{n0} - \varepsilon_{m0}} \left|m^{(0)}\right\rangle, \tag{3.4}$$

where we have used bra-ket notation to represent the cell-periodic part of the Bloch states, and the superscript denotes the order of perturbation theory. The first-order order energy correction is zero for band extrema (the intraband momentum matrix element is zero at band extrema), and the lowest-order energy correction is the second order correction. For non-degenerate bands,

$$\varepsilon^{(2)}_{n\mathbf{k}} = \varepsilon^{(0)}_n + \frac{k^2}{2} + \sum_{n\neq m} \frac{\left|\mathbf{k}\cdot\left\langle m^{(0)}\right|\mathbf{p}\left|n^{(0)}\right\rangle\right|^2}{\varepsilon_{n0} - \varepsilon_{m0}} \tag{3.5}$$

For non-degenerate bands, the perturbation needs to be diagonalized in the degenerate subspace after which non-degenerate perturbation theory can be applied to the

diagonalized states.

3.2.1 Effective mass from perturbation theory

We can rewrite expression (3.5) in terms of a single *effective* mass,

$$\varepsilon_{k,\alpha,\beta} = \frac{k_\alpha k_\beta}{2m^*_{\alpha,\beta}} \tag{3.6}$$

by recognizing,

$$\frac{1}{m^*_{\alpha,\beta}} \equiv \delta_{\alpha,\beta} + \sum_{n \neq m} \frac{\left\langle n^{(0)} \right| p_\alpha \left| m^{(0)} \right\rangle \left\langle m^{(0)} \right| p_\beta \left| n^{(0)} \right\rangle}{\varepsilon_{n0} - \varepsilon_{m0}} \tag{3.7}$$

where α and β index different directions. Expression (3.7) can be viewed as a definition of the effective mass for non-degenerate bands. Similar expressions exist for degenerate bands, which additionally require diagonalizing the second-order perturbation in the degenerate subspace (e.g., see Ref. [2]). In practice, the summation is rarely performed as it converges very slowly with the number of empty states, and the expression for the mass is obtained by either fitting expression (3.6) to the band structure near the band edge or by applying finite differences to calculate the second derivative. We will use these effective-mass models to construct semi-empirical Hamiltonians for systems where material variation occurs over large length scales. Other models, such as the 6-band and 8-band Kohn-Luttinger models, exist that parameterize multiple bands at a time. [3, 4] The semi-empirical Hamiltonians constructed with these more advanced models are able to account for band non-parabolicity and explicitly treat interactions (mixing) between the valence bands. However, for the properties considered in this thesis, these effects are not important and the effective-mass approximation is sufficient.

3.3 Schrödinger equation in the effective-mass approximation

For situations where the dispersion is accurately described by the effective-mass model, the single-particle Schrödinger equation takes the approximate form,

$$\left(-\frac{1}{2m^*}\nabla^2 + \varepsilon^{(0)}_{n,\mathbf{0}}(\mathbf{r}) + V(\mathbf{r})\right)\varphi_n(\mathbf{r}) = \varepsilon_n\varphi_n(\mathbf{r}), \tag{3.8}$$

where n is a generalized band index, $\varepsilon^{(0)}_{n\mathbf{0}}(\mathbf{r})$ is the band energy at the Γ point for the material at position $\mathbf{r}$, and $V(\mathbf{r})$ is a potential *different from the periodic potential of the lattice*. For example, the potential could represent the electrochemical potential that electrons experience due to dopants or impurities. It could also represent the effective many-body potential, or the *mean-field* potential, that electrons experience due to the electrostatic repulsion from all other electrons in the system. Here, $\varphi(\mathbf{r})$ is the *envelope* wave function rather than the full wave function, in analogy with the envelope of wave packets formed by electromagnetic waves. The envelope wave function is related to the full wave function as $\psi(\mathbf{r}) \approx \frac{1}{\sqrt{V}}\varphi(\mathbf{r})u(\mathbf{r})$, where $u(\mathbf{r})$ is the cell-periodic part of the Bloch function at the high symmetry point where the energy dispersion is perturbatively expanded to second order, and V is the volume of the crystal.

3.3.1 Derivation from the envelope-function approximation

To derive equation (3.8), we start by expressing a general solution to the single-particle Schrodinger equation as a linear combination of Bloch states in the first Brillouin zone,

$$\psi(\mathbf{r}) = \frac{1}{\sqrt{V}}\sum_{n\mathbf{k}} c_{n\mathbf{k}}u_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{3.9}$$

If Bloch states from different bands do not mix, the solutions take the form,

$$\psi_n(\mathbf{r}) = \frac{1}{\sqrt{V}}\sum_{\mathbf{k}} c_{n\mathbf{k}}u_{n\mathbf{k}}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}}. \tag{3.10}$$

This is a valid approximation if the perturbing potential $V(\mathbf{r})$ that breaks the lattice periodicity is slowly varying, and the band is separated energetically from other bands. Plugging this expression into the full single-particle Schrödinger equation and using the fact that $H_0\psi_{n\mathbf{k}}(\mathbf{r}) = \varepsilon^{(0)}_{n\mathbf{k}}(\mathbf{r})\psi_{n\mathbf{k}}(\mathbf{r})$, we obtain,

$$H\psi_n(\mathbf{r}) = (H_0 + V(\mathbf{r}))\psi_n(\mathbf{r}) = \frac{1}{\sqrt{V}}\left(H_0 + V(\mathbf{r})\right)\sum_{\mathbf{k}} c_{n\mathbf{k}} u_{n\mathbf{k}}(\mathbf{r}) e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.11}$$

$$= \frac{1}{\sqrt{V}}\sum_{n\mathbf{k}} c_{n\mathbf{k}}(\varepsilon_{n\mathbf{k}}(\mathbf{r}) + V(\mathbf{r}))u_{nk}(\mathbf{r})e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.12}$$

For III-nitrides (and most conventional semiconductors), the cell-periodic part of the Bloch function varies slowly with the $\mathbf{k}$ vector close to the Γ point. Consequently, we can approximate $u_{n\mathbf{k}}(\mathbf{r})$ with its value at the Γ point, which we denote as $u_n(\mathbf{r})$, allowing us to bring it out of the sum. At this point, we also perform a perturbative expansion near the Γ point, $\varepsilon_{n\mathbf{k}}(\mathbf{r}) \approx \varepsilon_{n\mathbf{0}}(\mathbf{r}) + k^2/2m^*(\mathbf{r})$,

$$H\psi_n(\mathbf{r}) = \frac{1}{\sqrt{V}}u_n(\mathbf{r})\sum_{n\mathbf{k}} c_{n\mathbf{k}}(\varepsilon^{(0)}_{n\mathbf{0}}(\mathbf{r}) + \frac{k^2}{2m^*(\mathbf{r})} + V(\mathbf{r}))e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.13}$$

Recognizing k^2 as the Fourier transform of $-\nabla^2$, we can rewrite the equation as,

$$H\psi_n(\mathbf{r}) = u_n(\mathbf{r})\left(-\frac{1}{2m^*(\mathbf{r})}\nabla^2 + \varepsilon^{(0)}_{n\mathbf{0}}(\mathbf{r}) + V(\mathbf{r})\right)\frac{1}{\sqrt{V}}\sum_{\mathbf{k}} c_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}} \tag{3.14}$$

We now identify $\sum_{\mathbf{k}} c_{n\mathbf{k}} e^{i\mathbf{k}\cdot\mathbf{r}}$ as the Fourier transform of an envelope function, $\varphi_n(\mathbf{r})$. It is by definition slowly varying with respect to the lattice constant because it only contains $\mathbf{k}$ states in the first Brillouin zone. Letting ε_n be the energy eigenvalue corresponding to the eigenstate, $\psi_n(\mathbf{r}) \approx u_n(\mathbf{r})\varphi_n(\mathbf{r})$, we obtain,

$$\varepsilon_n u_n(\mathbf{r})\varphi_n(\mathbf{r}) = u_n(\mathbf{r})\left(-\frac{1}{2m^*(\mathbf{r})}\nabla^2 + \varepsilon^{(0)}_{n\mathbf{0}}(\mathbf{r}) + V(\mathbf{r})\right)\varphi_n(\mathbf{r}) \tag{3.15}$$

Multiplying by $u_n^*(r)$ from the left and integrating over the unit cell gives,

$$\varepsilon_n \varphi_n(\mathbf{r}) = \left(-\frac{1}{2m^*(\mathbf{r})} \nabla^2 + \varepsilon_{n\mathbf{0}}^{(0)}(\mathbf{r}) + V(\mathbf{r}) \right) \varphi_n(\mathbf{r}). \tag{3.16}$$

It is worth noting that corrections to $\varepsilon_{n\mathbf{0}}^{(0)}(\mathbf{r})$ are necessary to account for additional strain energy in heterostructures, which amounts to making the replacement $\varepsilon_{n\mathbf{0}}^{(0)}(\mathbf{r}) \rightarrow \varepsilon_{n\mathbf{0}}^{(0)}(\mathbf{r}) + D_{\alpha,\beta}\epsilon_{\alpha,\beta}$, where D is the deformation potential and ϵ is the strain.

3.3.2 Potential from the Poisson equation

As mentioned in the previous section, the potential $V(\mathbf{r})$ encompasses any potential that carriers feel that is different from the periodic potential of the lattice. In practice, one typically includes within $V(\mathbf{r})$ the electrostatic term due to point charges as well as the mean-field potential that carriers experience due to their many-body interaction with other carriers. These effects are accounted for by solving the Poisson equation,

$$\nabla^2 V(\mathbf{r}) = -\frac{\rho(\mathbf{r})}{\epsilon_r}, \tag{3.17}$$

$$\rho(\mathbf{r}) = \rho_f(\mathbf{r}) + n(\mathbf{r}) - p(\mathbf{r}) \tag{3.18}$$

where $\rho(\mathbf{r})$ is the net charge density, which includes contributions from both fixed charges $\rho_f(\mathbf{r})$, *e.g.*, impurities, polarization charges, and dopants, as well as free electrons $n(\mathbf{r})$ and free holes $p(\mathbf{r})$ due to, e.g., doping, electrical injection, or optical injection. An important physical constraint on $\rho(\mathbf{r})$ is that it must obey charge neutrality within the simulation region. The electron and hole densities are obtained by

summing over the envelope functions calculated from the Schrödinger equation,

$$n(\mathbf{r}) = \sum_{m \in \mathrm{CB}} |\varphi_m(\mathbf{r})|^2 f_m \tag{3.19}$$

$$p(\mathbf{r}) = \sum_{m \in \mathrm{VB}} |\varphi_m(\mathbf{r})|^2 (1 - f_m), \tag{3.20}$$

where CB and VB stand for conduction and valence band, respectively, and $f_m = (1+\exp\left[(\varepsilon_m - \varepsilon_f)/k_B T\right])^{-1}$ is the *quasi-equilibrium* Fermi function that uses the *quasi* Fermi level for electrons or holes. These equations are only valid if the 3D simulation (supercell) volume is sufficiently large that the Brillouin zole has folded over into a very small reciprocal cell volume. For 1D calculations that assume translational symmetry along two directions, we need to additionally sum over explicit k states or, equivalently, integrate the density of states over the carrier's energy. For example, for 1D calculations of a quantum well that is translationally invariant within the plane, the expression for electrons becomes,

$$n(x) = \sum_{m \in \mathrm{CB}} |\varphi_m(x)|^2 \int d\varepsilon f_m(\varepsilon) d_m(\varepsilon) \tag{3.21}$$

where $d_m(\varepsilon)$ is the analytical expression for the 2D density of states in a quantum well.

3.3.3 Self-consistency of the Schrödinger & Poisson equations

The potential $V(\mathbf{r})$ is dependent on the envelope functions through the Poisson equation, and the envelope functions, in turn, depend on $V(\mathbf{r})$ through the Schrödinger equation. As a result, these differential equations are non-linearly coupled and require self-consistent solutions. In practice, this involves assuming an initial form for the free-carrier density, from which an initial guess for $V(\mathbf{r})$ is derived. $V(\mathbf{r})$ is then used as input for the Schrödinger equation to obtain new wave functions. The new wave functions are subsequently used as input to the Poisson equation, which provides a new guess for $V(\mathbf{r})$. This iterative process continues until the charge density

and potential converge to a desired tolerance. Techniques such as under-relaxation, where the charge density is fractionally updated each iteration, can accelerate the convergence. In cases where the free-carrier density is low, self-consistency may not be necessary as the contribution from free carriers to $V(\mathbf{r})$ is minimal. However, in situations where the free-carrier density is high, self-consistency becomes crucial in capturing relevant physical effects, such as electrostatic screening of internal fields in polar heterostructures and p-n diodes.

3.4 Many-body effects in the free-carrier plasma

To capture the many-body electrostatic interaction of the carriers, we have made the mean-field Hartree approximation. However, this approximation includes spurious self-interaction of charge carriers with themselves, leading to systematically overestimated energy renormalization and artificially delocalized wave functions. To address this issue, we need to account for exchange-correlation effects that encompass all many-body effects beyond the Hartree approximation, allowing us to more accurately map the effective single-particle Schrödinger equation onto the many-body Schrödinger equation. [5]

Since free carriers are relatively well extended, the local density approximation (LDA) exchange-correlation functional can be used to account for many-body effects, which involves using the exchange and correlation potential of a homogeneous electron gas having equivalent charge density at every position $\mathbf{r}$. The exchange term arises from the anti-symmetry of the many-body wave function and enforces Pauli's exclusion. The exchange of a homogeneous electron gas is exactly solvable, and the potential has an analytical form (see II). All many-body effects that are not captured by the exchange potential are grouped into the correlation potential, which must be solved numerically. For work presented in this thesis, we use the parameterization of the electron correlation by Perdew and Zunger of the Monte-Carlo calculation of Ceperley and Alder. [6, 7]

Including the LDA exchange-correlation potential in the self-consistent Schrödinger

and Poisson equation has the benefit of providing the correct plasma renormalization of the free-carrier energies, which can be useful in predicting properties such as the plasma renormalization of the band gap due to free carriers. However, for systems with very inhomogeneous charge densities, e.g., disordered alloys, the local-density exchange-correlation becomes less reliable. Therefore, in such systems, more advanced many-body corrections at the level of the GW approximation or explicit calculation of the electron-plasmon self-energy are likely required to obtain accurate energies and wave functions. [8, 9]

3.5 Computational implementation

We have implemented a 1D self-consistent Schrödinger-Poisson solver that takes into account broken translational symmetry along the direction of growth. The code assumes periodic boundary conditions, and we solve the time-independent Schrödinger equation using finite differences and the Poisson equation using the fast Fourier transform. To accelerate the convergence of self-consistency, we have employed under-relaxation, where we set a mixing fraction value between 0 and 1 that determines how much the charge density is updated in every iteration of self consistency. We have also incorporated the renormalization of energies and envelope functions by the free-carrier plasma in the local-density approximation. Additionally, we are able to calculate the ground-state excitonic properties using the variational approach. For calculations that require consideration of broken translational symmetry within the plane of growth, we utilize the software nextnano++, which also uses finite differences. Once the energies and wave functions are obtained, we calculate functional properties such as optical absorption or luminescence spectra, as well as corrections to recombination rates, using a parallelized in-house post-processing code. Detailed descriptions of these calculations are provided in the appropriate chapters of this thesis. For 3D calculations, we use the commercial software nextnano++. [10]

3.5.1 Calculation of material parameters from first principles

The Schrödinger-Poisson method is a semi-empirical approach that relies on accurate knowledge of bulk material parameters. To obtain these parameters, we rely on first-principles calculations. While it is common practice with device simulations to empirically tune input parameters to match experimental results, we strictly advise against this as it can result in a complete loss of predictive accuracy and compromise the reliability of the simulation. In our work, we use material parameters that are calculated using hybrid-functional density functional theory. Hybrid-functional DFT mixes a fraction of the exact Fock exchange to local or semi-local functionals, which helps overcome many of the systematic errors of DFT in conventional semiconductors. We ensure that the excited state properties calculated with hybrid-functional DFT are consistent with calculations from many-body perturbation theory. The material parameters used as inputs to the Schrödinger-Poisson method include the lattice constant, elastic constant, deformation potential, polarization constant, band gaps, band offsets, and effective masses.

3.5.1.1 Lattice constants

The lattice constants are calculated by performing structural minimization using conventional techniques, such as conjugate-gradient minimization. [11] For a given configuration, forces can be calculated using the Hellman-Feynman theorem, provided the charge density is known. These forces, which are simply the gradients of the potential energy, are used by the conjugate gradient algorithm to update the ionic configurations and reach a local minimum in the potential-energy landscape. This approach allows us to determine the optimal lattice constants for the material under consideration.

3.5.1.2 Elastic constants

To calculate the elastic constants, the following relation is used, [11]

$$\frac{\Delta U}{V} = \frac{1}{2}\sum_{ij} C_{ij} u_i u_j, \tag{3.22}$$

where ΔU is the change to the total energy of a structure due to small deformations u_i and u_j, and C_{ij} is the elastic constant (related to the force constant). We use Voigt notation, where $i = 1, 2, 3$ corresponds to the directions xx, yy, zz, and $i = 4, 5, 6$ corresponds to xy, xz, yz. Additionally, u_i and u_j are *engineering* strains, which are related to the normal definition of strain by, $u_i = \epsilon_i$ (for $i = 1, 2, 3$) and $u_i = 2\epsilon_i$ (for $i = 4, 5, 6$). This choice in notation is made due to the degeneracy of the off-diagonal components, e.g., $\epsilon_{xy} = \epsilon_{yx}$. The derivation of the elastic constants involves performing a set of small isotropic, tetragonal, and trigonal deformations to the crystal structure. These deformations result in sets of expressions for the total energy in terms of C_{ij}, u_i, and u_j, which can then be algebraically manipulated to solve for C_{ij}.

3.5.1.3 Band gaps and effective masses

Band gaps and effective masses are sensitive to the choice of functional used in electronic structure calculations. Hybrid functionals, which incorporate a fraction of exact exchange mixed with a standard density functional, are used to calculate band gaps. The theoretical exact exchange mixing fraction for our preferred HSE06 hybrid functional is 0.25, but it can also be adjusted to match the experimental band gap. We have ensured that any band gap that we use from hybrid DFT is validated against against calculations from many-body perturbation theory, specifically the G_0W_0 approximation, which is known to provide accurate band gap values for a wide range of materials. Moreover, finite difference methods or fitting techniques are typically used to calculate the inverse of the second derivative of energy with respect to wave vector, which provides information about the curvature of the energy bands near the band extrema and thus the effective masses. [11, 5]

3.5.1.4 Band offsets and deformation potential

Band offsets and deformation potentials are calculated using the model-solid approach, where the choice of the functional also affects the results. [12] We use values computed with hybrid functionals; it is often not possible to compare these values against many-body perturbation theory calculations since calculating the total energy in the GW approximation is computationally expensive, particularly for large supercells.

The procedure for calculating band offsets at interfaces between material A and B is as follows: First, the bulk band structures of both material A and material B are calculated. However, since the absolute energy scale is not well-defined when pseudopotentials are used, the energy bands need to be aligned to an absolute scale. To achieve this, slab calculations are performed for both materials, where a vacuum region is included in the simulation cell. The thickness of the material slab is chosen such that the electrostatic potential can be accurately averaged within the slab. The potential shift required to reference the band structure energy of the respective bulk materials to an absolute scale is obtained from the difference between the average potential in the slab and the vacuum potential. Once the energies are obtained on an absolute scale, the conduction-band offset is calculated as the difference between the conduction band minimum energies of material A and material B, while the valence-band offset is calculated as the difference between the valence band maximum energies of material A and material B. A similar approach is used to calculate deformation potentials, where material A is the unstrained semiconductor and material B is the strained semiconductor. In the limit of small strain, the deformation potential is calculated as $D_{ij} = \Delta\varepsilon/\Delta\epsilon_{ij}$, where ε is the energy eigenvalue and ϵ is the strain.

3.5.1.5 Polarization constants

The most common approach to calculating polarization constants is using the modern theory of polarization based on the Berry phase approach. [13] The spontaneous polarization field of a desired structure can be calculated by computing the different

in formal polarization between the desired structure and centro-symmetric reference structure. It is important this reference structure be adiabatically connected to the desired structure, which in this context means that the deformation preserves the band gap. For wurtzite III-N crystals, the natural reference structure is the hexagonal phase, the latter of which is accessed from the former by an adiabatic deformation of the internal u parameter. The formal polarization for a given structure λ is defined (in S.I. units) as,

$$\mathbf{P}_f = \mathbf{P}_{ion} + \mathbf{P}_{el} = \frac{e}{\Omega} Z_s^{ion} \mathbf{R}_s^{\lambda} + \frac{ief}{8\pi^3} \sum_j^{occ} \int_{BZ} d^3\mathbf{k} \left\langle u_{j,\mathbf{k}}^{\lambda} \right| \nabla_{\mathbf{k}} \left| u_{j,\mathbf{k}}^{\lambda} \right\rangle ,$$

where Ω is the unit cell volume, Z_s is the ion charge, and f is the spin degeneracy. The piezoelectric constants can be also calculated using the modern theory of polarization by calculating the change in the formal polarization upon the application of strain. There are two choices for the piezoelectric constants: proper vs improper constants. The improper constants should be chosen for device simulations as these account for changes to the spontaneous polarization field due to the application of strain. If strain is applied, the surface area of the unit cell changes which changes the surface density of the polarization charge, and this must be accounted for when calculating the polarization field. [14]

3.6 Limitations

The formalism that I have outlined in this chapter has several limitations. Firstly, it is only applicable when the perturbing potential $V(\mathbf{r})$ varies slowly compared to the lattice constant. However, this limit is not well defined, and there is no clear length scale at which the formalism breaks down. Notably, the assumption of a slowly varying potential is not valid when modeling alloy disorder in a random alloy. Despite this, when compared to atomistic tight-binding calculations, the results exhibit reasonable qualitative agreement of energies and wave functions.

In practice, the presence of a non-periodic perturbation leads to the mixing of dif-

ferent Bloch states, which in turn causes the single-band approximation, employed in deriving the effective-mass Schrödinger equation, to break down. This is particularly true when the energy bands in a given manifold are not separated, as is the case for the valence band of III-nitride materials. In certain cases, e.g., if one is trying to predict the polarization of light that a particular alloy composition emits, it may be necessary to explicitly account for the band degeneracy through multi-band $\mathbf{k} \cdot \mathbf{p}$ models, such as the 6-band or 8-band models. These models, which are generalization of the effective-mass approximation, capture non-parabolicities and valence-band mixing, which is critical for determining optical polarization properties, given the distinct angular momentum symmetries of different valence states. Nevertheless, if the focus is solely on average effects that depend on the density of states rather than the orbital character, the effective mass model, as employed in this formalism, may yield satisfactory agreement with experimental observations and atomistic calculations.

Furthermore, this formalism encounters challenges when dealing with different materials that exhibit dissimilarity in their cell-periodic Bloch functions, *e.g.*, at interfaces. The derivation of the effective-mass Schrödinger equation assumes that the entire device can be described by a single cell-periodic Bloch function, which is generally not valid in such cases. However, for materials with similar chemistries, such as the members of the III-nitrides (e.g., AlN, GaN, and InN), the Bloch functions exhibit sufficient similarity, justifying the approximate validity of the single cell-periodic Bloch function assumption. Nonetheless, it should be noted that there are alternative, albeit more computationally expensive, approaches that account for variations in Bloch functions, and in situations where such variations become significant, a higher level of theory may be warranted. [15]

Finally, it is important to acknowledge that this formalism lacks information about the underlying atomistic structure, which may be crucial in certain cases. In such scenarios, higher levels of theory, such as atomistic tight binding or density functional theory (DFT), should be employed. However, for large-scale devices, where the effects of atomistic structure tend to average out, Schrödinger-Poisson simulations are often

sufficient for capturing the essential physics.

Bibliography

[1] Jasprit Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures*. Cambridge University Press, 2007.

[2] Oleg Rubel, Fabien Tran, Xavier Rocquefelte, and Peter Blaha. Perturbation approach to ab initio effective mass calculations. *Computer Physics Communications*, 261:107648, 2021.

[3] J M Luttinger and W Kohn. Motion of electrons and holes in perturbed periodic fields. *Physical Review*, 97(4):869, 1955.

[4] SL Chuang and CSK Chang. k p method for strained wurtzite semiconductors. *Physical Review B*, 54(4):2491, 1996.

[5] Richard M Martin. *Electronic Structure: Basic Theory and Practical Methods*. Cambridge University Press, 2020.

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

[7] David M Ceperley and Berni J Alder. Ground state of the electron gas by a stochastic method. *Physical review letters*, 45(7):566, 1980.

[8] Mark S Hybertsen and Steven G Louie. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Physical Review B*, 34(8):5390, 1986.

[9] Fabio Caruso and Feliciano Giustino. Theory of electron-plasmon coupling in semiconductors. *Physical Review B*, 94(11):115208, 2016.

[10] Stefan Birner, Tobias Zibold, Till Andlauer, Tillmann Kubis, Matthias Sabathil, Alex Trellakis, and Peter Vogl. Nextnano: general purpose 3-d simulations. *IEEE Transactions on Electron Devices*, 54(9):2137–2142, 2007.

[11] Feliciano Giustino. *Materials modelling using density functional theory: properties and predictions*. Oxford University Press, 2014.

[12] Chris G Van de Walle. Band lineups and deformation potentials in the model-solid theory. *Physical review B*, 39(3):1871, 1989.

[13] Raffaele Resta and David Vanderbilt. Theory of polarization: a modern approach. In *Physics of Ferroelectrics: A Modern Perspective*, pages 31–68. Springer, 2007.

[14] Cyrus E Dreyer, Anderson Janotti, Chris G Van de Walle, and David Vanderbilt. Correct implementation of polarization constants in wurtzite materials and impact on iii-nitrides. *Physical Review X*, 6(2):021038, 2016.

[15] Bradley A Foreman. Exact effective-mass theory for heterostructures. *Physical Review B*, 52(16):12241, 1995.

CHAPTER IV

Electron Mobility of Random AlGaN Alloys Evaluated by Unfolding the DFT Band Structure

We calculate the alloy-disorder-limited electron mobility of $Al_xGa_{1-x}N$ from first principles. $Al_xGa_{1-x}N$ is a technologically important ultra-wide-band-gap alloy with promise in light emitting diodes and high-power transistors. Alloying introduces statistical disorder, which causes electrons to scatter between different crystal-momentum states, leading to a reduction in mobility for intermediate alloy compositions. The corresponding lifetime, which appears as an energy broadening in the band structure, can be evaluated by unfolding the band structure from the supercell basis to the primitive-cell basis. We fit the first-principles band broadening with a model scattering potential, and evaluate the low-field electron mobility using the semiclassical Boltzmann transport equation in the relaxation-time approximation. Our calculated mobility is in agreement with experimental values. We also find the *lowest* alloy-scattering electron mobility (total electron mobility) across the entire composition range to be 186 $cm^2/V \cdot s$ (136 $cm^2/V \cdot s$), which is comparable to the *highest* electron mobility predicted in the competitor system, β-$(Al_xGa_{1-x})_2O_3$. Our results elucidate the intrinsic limits imposed by alloy disorder on electron transport in $Al_xGa_{1-x}N$.
This chapter was reprinted (adapted) with permission from Appl. Phys. Lett. **117**, 242105 (2020).

4.1 Introduction

Thanks to the success of nitride materials for visible light-emitting diodes (LEDs) and high-electron-mobility transistors (HEMTs), there is enormous interest in understanding the properties of ultra-wide-band-gap nitride materials for power electronics and optoelectronics.[1, 2, 3] The alloys of GaN and AlN span band-gap energies from approximately 3.4 eV to 6 eV.[4] This makes $Al_xGa_{1-x}N$ an attractive platform for high-power HEMTs and field-effect transistors (FETs). Moreover, the ability to tune the band gap of AlGaN in the UV range makes it a natural candidate for ultraviolet LEDs (UV-LEDs).[5] The need for efficient, portable UV-LEDs for sterilization is acutely highlighted by the COVID-19 pandemic.

4.1.1 Why AlGaN?

Despite their potential, AlGaN-based devices face several challenges relating to their energy efficiency. Thus, there is a pressing need to understand the microscopic quantum processes that affect the efficiency of energy transport and conversion in these materials.[5] Breakthroughs in the growth of high-quality n-type films through polarization doping, delta-doping, and high-temperature AlN interlayers have enabled direct studies of the electronic transport properties of AlGaN.[6, 7, 8, 9, 10, 11, 5] These experiments unequivocally identify alloy disorder as a primary scattering mechanism that limits the electrical efficiency of AlGaN-based devices. However, it is unclear how the electron mobility in AlGaN compares to the electron mobility in the closest competitor, β-$(Al_xGa_{1-x})_2O_3$.[12, 13, 14, 15, 16]

4.1.2 Previous work on alloy scattering from first principles

Theoretical investigations of the effects of alloy disorder on carrier transport remain sparse. The virtual-crystal approximation (VCA) is a common starting point for understanding the electronic properties of alloys, such as their band gap and carrier effective masses. However, it cannot capture the effects of alloy scattering since it does not take into account the statistical disorder associated with the breaking of transla-

tional symmetry.[17] Recently, Coltrin and Kaplar conducted a comprehensive study of electrical transport and breakdown in AlGaN, based on analytical models.[18] Their work suggests that alloy scattering is the dominant scattering mechanism in AlGaN, and that Al-rich AlGaN may outperform GaN on various figures of merit for power-electronics applications at high operating temperatures. Bellotti *et al.* performed a numerical study of alloy scattering in the ternary III-nitrides.[19] They used a modified nonlocal Empirical Pseudopotential Method (EPM), which includes corrections to the VCA via a disorder potential, and Monte-Carlo simulations to calculate transport parameters. Murphy-Armando and Fahy have investigated alloy scattering in $Si_{1-x}Ge_x$ alloys by directly evaluating the scattering matrix elements from first-principles calculations.[20] Sau and Cohen later used a similar method to calculate the rates of alloy scattering in a $Ge_{1-x}Sn_x$ alloy,[21] and Vaughan *et al.* extended the method to $Si_{1-x}C_x$.[22] However, there have been no first-principles reports on the effects of alloy scattering on the electron mobility of AlGaN.

4.1.3 Overview of our new method

Alloying breaks the symmetries of the parent crystals. The breaking of translational symmetry means that Bloch's theorem, and correspondingly the concept of bands, is not strictly valid. However, states in weakly perturbed alloys retain their Bloch-like character if the effects of localization are weak, and an effective band structure can be meaningfully constructed from the Bloch spectral function.[23] The appearance of non-diagonal terms in the Bloch-basis Hamiltonian, introduced by alloying, results in a broadening of wave vectors and energies in the effective band structure. The broadening of wave vectors corresponds to the lack of periodicity and the associated broadening of energies corresponds to the scattering lifetime or, equivalently, the imaginary part of the quasiparticle self energy.

In this work, we present a first-principles approach to study the effects of alloy scattering on the electron mobility of $Al_xGa_{1-x}N$. Our method is based on a first-principles effective band structure and a model alloy-scattering potential, with parameters in-

formed from first-principles calculations on disordered structures.[24, 25, 26, 27, 19] Our mobility results are in good agreement with experimental measurements and explain the composition and temperature dependence of the electron mobility.

4.2 Methodology

In our work, we employ special quasirandom structures (SQS) that accurately capture the effects of statistical disorder with a small simulation cell size by mimicking the correlation functions of a truly random alloy. [28, 29] In principle, the size of the supercell should be large to accurately capture the effects of statistical randomness. However, the computational cost of DFT calculations for large simulation cells is prohibitive. We thus limit ourselves to small SQS supercells, which accurately describe carrier scattering for the higher-energy Bloch states with wavelengths smaller than the supercell dimensions, and extrapolate the scattering rate to low-energy, long-wavelength states using analytical models.

For Al contents of $x = 0.25, 0.5$ and 0.75, we sampled eight SQSs, $4 \times 4 \times 2$ primitive-cells wide with 128 atoms each, as generated by the Monte-Carlo algorithm implemented in the Alloy Theoretic Automatic Toolkit.[30] We ran a special test case with a 300-atom supercell to verify the convergence of our calculations (see **??**). We minimized deviations in pairwise correlations up to 0.6 nm, and triplet correlations up to 0.5 nm. Figure 4.1 shows a sample SQS with 50% Al content. We performed DFT calculations in the local-density approximation (LDA) as implemented in the Quantum Espresso package.[31] We chose the LDA functional due to its low cost and its ability to predict band offsets in the III-nitrides within $\sim$0.1 eV accuracy of the HSE functional,[32] which is the important parameter for evaluating the scattering potential. We could have performed this study with another functional, such as the PBE functional, since it produces similar results as the LDA. To measure the level of agreement between LDA and HSE, we ran a special test case with the HSE functional (see section 4.5). We used norm-conserving pseudopotentials for the $3s^2p$ valence electrons of Al, $3d^{10}4s^2p$ valence electrons of Ga, and $2s^2p^3$ valence electrons

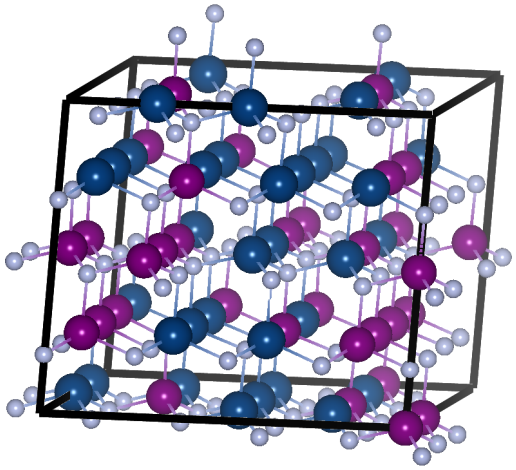

Figure 4.1: A characteristic $4 \times 4 \times 2$ special quasirandom supercell structure for $Al_{0.5}Ga_{0.5}N$ containing 128 atoms, used for DFT calculations of alloy disorder. Ga atoms are in blue, Al atoms are in purple and N atoms are in gray.

of N. For the self-consistent calculation, we used a converged plane-wave kinetic-energy cutoff of 100 Ry, and a converged $2 \times 2 \times 1$ Monkhorst-Pack Brillouin-zone sampling grid. We relaxed the structures until all forces were smaller than 0.001 Ry/a_B. and the pressure was below 0.5 Kbar. The spectral functions $A(\mathbf{k}, \varepsilon)$, as functions of wave vectors $\mathbf{k}$ and energy ε, were calculated by unfolding the supercell band structure onto the primitive wurtzite reciprocal cell using the BandUP code.[26, 27] The unfolded band structure is shown in Figure 4.2. We averaged the spectral functions for each composition after aligning the conduction-band edges to minimize contributions from variations in $\mathrm{Re}(\Sigma)$.

4.2.1 Alloy-scattering rates by unfolding the band structure

To calculate the spectral broadenings, we consider the finite integrated spectral moments,[29]

$$M_p(\mathbf{k}) = \int_{\varepsilon_1}^{\varepsilon_2} d\varepsilon \varepsilon^p A(\mathbf{k}, \varepsilon), \tag{4.1}$$

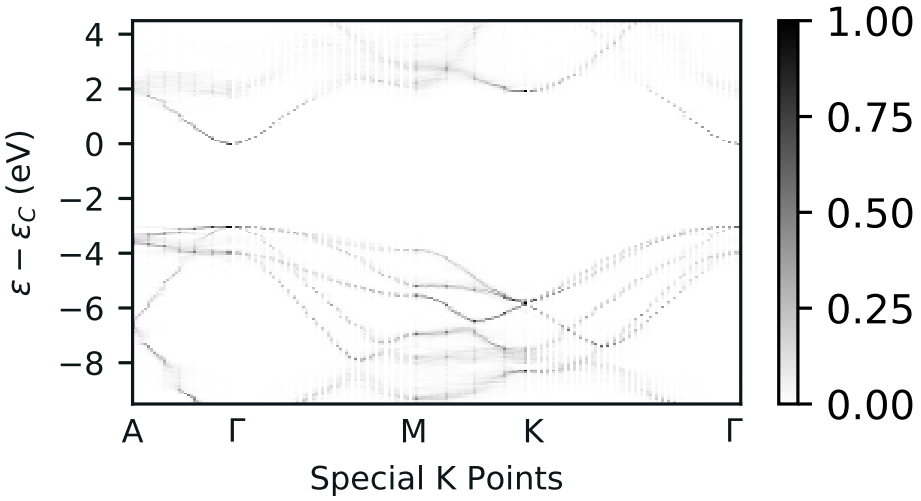

Figure 4.2: Unfolded band structure of a $4 \times 4 \times 2$ $Al_{0.5}Ga_{0.5}N$ special quasirandom supercell structure, obtained from BandUP. Note that LDA-DFT underestimates the band gap. Higher energy states exhibit moderate to significant energy-broadening due to alloy disorder.

where p is an integer, and the integration is performed over a finite range determined by the spectral peak positions. For a single-band Hamiltonian with site-diagonal disorder, Ehrenreich and Schwartz showed that the standard deviation of the spectral moment corresponds approximately to $\mathrm{Im}(\Sigma)$, calculated from perturbation theory.[33] We extend the result by Ehrenreich and Schwartz, and approximate the alloy-scattering rate $1/\tau = 2\,\mathrm{Im}(\Sigma)/\hbar$ by $1/\tau \approx \sqrt{\mu_2}/\hbar$, where $\mu_2 \equiv M_2(\mathbf{k}) - (M_1(\mathbf{k}))^2$ is the spectral variance, defined over a finite energy range. This approximation is reasonable because of the small lattice mismatch between AlN and GaN, which leads primarily to cation site-diagonal disorder.[29] As will be shown, our results agree with previous theoretical results and experiments, and therefore support the validity of our assumption. This method works only for those regions of the band structure where the band of interest is isolated from other bands, which is the case for the lowest conduction band of AlGaN. One advantage of our method, compared to previous *ab initio* approaches,[20, 21, 22] is the ability to treat full disorder and partial atomic ordering on equal footing, by simply changing the atomic coordinates. For the scope of this work, we only consider random alloys, which is known to be a good approximation for AlGaN.[34]

We constructed an averaged effective band structure for the conduction band by defining $M_1(\mathbf{k})$ as the band center and $\sqrt{\mu_2}$ as the width, as illustrated in Figure 4.3. A more complete effective band structure is shown in Figure 4.4. For energies up to approximately 1.5 eV from the minimum, the conduction band is well described by the Kane model for non-parabolic spherical bands,[35]

$$\varepsilon(1 + \alpha\varepsilon) = \frac{\hbar^2 k^2}{2m^*}, \tag{4.2}$$

where α is the non-parabolicity parameter and m^* is the effective mass. Electrons with small wave vectors (long wavelengths), near the Γ-point, experience the periodic potential of the repeating supercell, and therefore exhibit artificially decreased energy broadening. Because of this, we extrapolate the scattering rate for long-wavelength electron states by fitting a golden-rule expression[19] to the scattering

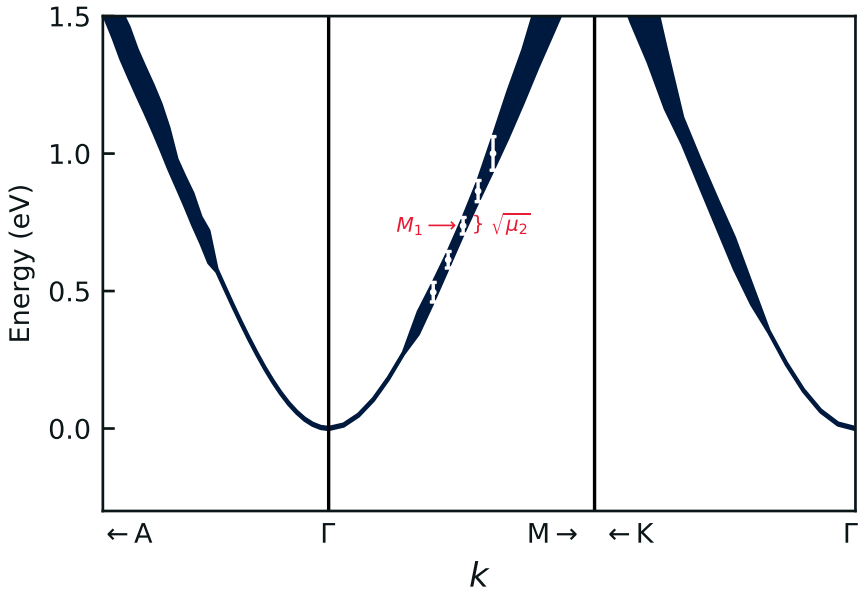

Figure 4.3: Averaged effective band structure of $Al_{0.5}Ga_{0.5}N$ partway along the high-symmetry directions of wurtzite. The energy broadening reflects the finite lifetime due to statistical disorder. The first spectral moment M_1 and the spectral width μ_2, used to construct the effective band structure, are shown as points with uncertainty bars. The supercell periodicity results in an artificial lack of broadening for the small wave vector (i.e., long wavelength) states near Γ.

rates of accurately described short-wavelength states with energies less than 1.5 eV. However, our method is general and can be extended to other plane-wave codes, such as atomistic tight-binding, that can handle much larger supercells thereby avoiding the need to extrapolate to long wavelengths.

4.2.2 Alloy-scattering potential from first principles

We treat alloy scattering perturbatively and assume that electrons collide with randomly positioned hard spheres. [35, 36, 37] Adopting the convention of using a statistically averaged disorder potential in the matrix element for Fermi's golden

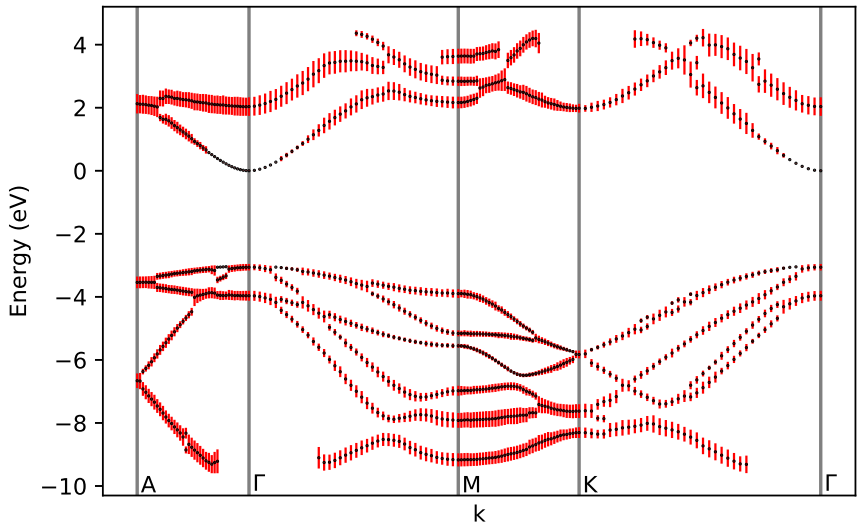

Figure 4.4: Effective band structure of a $4 \times 4 \times 2$ $Al_{0.5}Ga_{0.5}N$ special quasirandom supercell structure. The black points correspond to the discretely sampled band centers and the red uncertainty bars indicate the energy width $\sqrt{\mu_2}$, calculated from the spectral function. Our computational implementation, which uses peak detection and directed graphs to construct the effective band structure from the spectral function, works best for the lowest conduction band of AlGaN, which is isolated from other bands and for which the quasiparticle approximation is valid. Accurate determination of the band center and band broadenings for the valence band and very high (low) energy conduction (valence) band states is difficult due to significant spectral function overlap or ill-defined quasiparticle states.

rule, we can express the scattering lifetime $\tau(\varepsilon)$ as[19]

$$\begin{aligned}\frac{1}{\tau} =& \frac{2\pi}{\hbar} U_0^2 x(1-x)\Omega_0 \frac{m^{*3/2}}{\sqrt{2}\pi^2\hbar^3} \\ & \times (1+2\alpha\varepsilon)\sqrt{\varepsilon(1+\alpha\varepsilon)}I(\alpha,\varepsilon),\end{aligned} \tag{4.3}$$

where

$$I(\alpha,\varepsilon) \equiv \frac{1+2\alpha\varepsilon+(4/3)\alpha^2\varepsilon^2}{(1+2\alpha\varepsilon)^2}. \tag{4.4}$$

For materials with available experimental data, the strength of the effective scattering potential U_0, assumed to extend over the primitive-cell volume Ω_0, is typically estimated by fitting to experimental mobility measurements.[36, 7] In the absence of experiments, U_0 is taken to be the conduction-band offset $\Delta\varepsilon_c$ or band-gap difference $\Delta\varepsilon_G$ between GaN and AlN.[36, 19]

In contrast to empirical approaches, here we determine the scattering rate from the first-principles band-structure broadening data. To extract the scattering rates from the effective band structure of alloys, we sampled k-points along the Γ-A, Γ-M, and Γ-K directions. The resulting scattering rates, averaged across all eight SQS's for $x = 0.5$, are shown in Figure 4.5. As expected, the scattering rate is proportional to the density of states. As shown in Figure 4.5, we estimated the effective scattering potential U_0 by fitting equation (4.3) from approximately 0.5 eV to 1.5 eV. (See 4.4 for the error analysis.)

Table 4.1 lists the estimated scattering potentials as a function of Al content. We verified our estimates by comparing to the slope of the conduction band with respect to the composition[38] (see section 4.5) and by directly evaluating the long-wavelength scattering potential using a substitutional defect approach[21] (see section 4.5). The effective potentials evaluated by unfolding the band structure are smaller than the band-gap difference $\Delta\varepsilon_G = 2.6$ eV between GaN and AlN. They are also smaller than

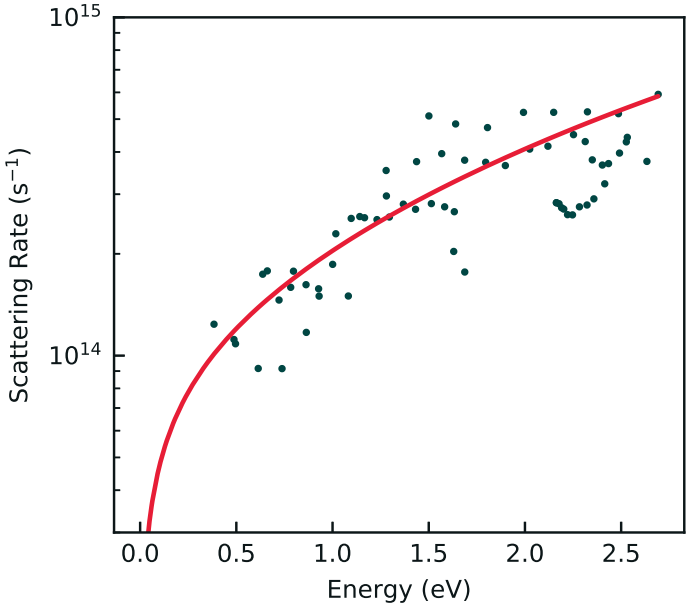

Figure 4.5: Scattering rates for $Al_{0.5}Ga_{0.5}N$ as a function of electron energy, referenced to the conduction-band minimum. Each point is the averaged electron scattering rate for a k-state sampled along one of Γ-A, Γ-M, and Γ-K directions. The rates were calculated from energy broadenings of the conduction band using the uncertainty principle. The solid line is the fit for the scattering rate expression derived from Fermi's golden rule for the hard sphere scattering potential.

Table 4.1: Scattering potentials U_0 of $Al_xGa_{1-x}N$ calculated for different Al contents using various methods.

Method	Al content	U_0 (eV)
Band unfolding (this work)	0.25	1.5 ± 0.2
	0.5	1.8 ± 0.2
	0.75	1.9 ± 0.3
Substitutional defect (this work)[1]	0.5	2.1
$\partial\varepsilon_c/\partial x$ from hybrid DFT[2]	0.25	1.5
	0.5	1.8
	0.75	2.1
Monte Carlo + EPM[3]	0.2	2.0
	0.5	1.7
	0.8	1.3
Experiment 1[4]	0 to 0.3	1.5
Experiment 2[5]	0 to 0.3	1.8

the conduction-band offset $\Delta\varepsilon_C = 2.2$ eV, obtained by slab calculations,[32, 39] but comparable to $\Delta\varepsilon_C = 1.8$ eV, obtained by referencing the branch-point energies of bulk structures.[38] The potentials are also similar to the values reported by Bellotti *et al.*, albeit with a different composition dependence.[19] Finally, they are similar to values reported by experiment (1.5 eV and 1.8 eV for graded $Al_xGa_{1-x}N$ alloys), with differences likely arising due to minor differences in the fitting models.[9, 7]

4.3 Electron mobility of AlGaN

To estimate the low-field electron mobility μ, we work in the framework of the Boltzmann transport equation in the relaxation-time approximation,

$$\langle\mu\rangle = -\frac{e}{3n}\int_0^\infty d\varepsilon v^2(\varepsilon)\tau(\varepsilon)\frac{\partial f^0}{\partial\varepsilon}d_c(\varepsilon). \tag{4.5}$$

We calculated the band velocity $v(\varepsilon)$ and conduction band density of states $d_c(\varepsilon)$ in the effective-mass approximation, assuming $\alpha = 0$ since most electrons occupy states

near the band edge where the parabolic band approximation is valid. We calculated the lifetime $\tau(\varepsilon)$ using equation (4.3), setting $\alpha = 0$. Moreover, we interpolated the primitive-cell volume and scattering potentials for different compositions from the sampled compositions using a quadratic fit to account for statistical bowing. We treated the carrier density n as an input parameter, from which we estimated the Fermi energy and the Fermi-Dirac distribution of electrons f^0 using the bisection method for root finding.

In general, the LDA does not predict accurate effective masses.[40] This problem is exacerbated in our case, since m^* appears with a 5/2 power in equation (4.5), through the velocity, lifetime, and density of states. This can be alleviated by using the hybrid HSE functional[41] or by applying many-body GW corrections.[40] Unfortunately, these methods are more computationally demanding, and accurate experimental and theoretical reports of effective masses in $Al_xGa_{1-x}N$ at different compositions are not available. We have thus used a linear interpolation of the HSE effective masses of GaN and AlN, which amounts to Vegard's law.[41] This assumption is supported by our LDA calculations, which show a near-linear dependence of effective mass on Al content (see section 4.6). We approximate the conduction band as being isotropic, and take the effective mass to be $m_e^* = (m_\perp^2 m_\parallel)^{1/3}$. We use the HSE values reported by Dreyer *et al.* to approximate $m_e^* = 0.209 m_e$ for GaN and $m_e^* = 0.313 m_e$ for AlN, which is well within the range of experimentally reported values.[41]

4.3.1 Alloy-disorder-limited mobility

Figure 4.6 shows our calculated low-field, alloy-scattering electron mobility as a function of composition, at a carrier density of 10^{18} cm^{-3} and temperatures of 10 K, 300 K, and 500 K. We have fixed the carrier density by tuning the Fermi level rather than introducing dopants. Experimentally measured electron mobilities are also shown for comparison. All samples exhibit carrier concentrations of approximately 10^{18} cm^{-3},[9, 8, 7, 11, 10] except the sample by Armstrong *et al.* which exhibits a carrier concentration of 5×10^{17} cm^{-3}.[42] The samples by Simon *et al.*[9], Jena *et*

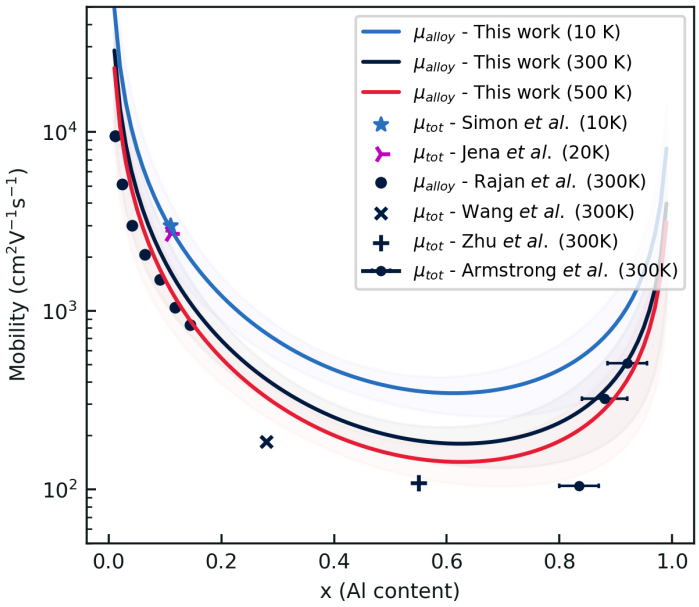

Figure 4.6: Alloy-scattering electron mobility of $Al_xGa_{1-x}N$ as a function of composition and temperature at $n = 10^{18}$ cm^{-3}. Blue corresponds to 10 K, black corresponds to 300 K and red corresponds to 500 K. The shaded regions correspond to the uncertainty in the mobility arising from the uncertainty in the scattering potentials. μ_{tot} denotes the directly measured total mobility, whereas μ_{alloy} denotes the alloy-scattering component of the total mobility.

al.[7], Rajan *et al.*[8], and Armstrong *et al.*[42] are polarization doped. The sample by Wang *et al.* is grown on a high-temperature AlN interlayer,[11] and the sample by Zhu *et al.* is delta-doped.[10] There is overall good agreement between theory and experiment. The discrepancy with some experimental values at room temperature could be due to electron-phonon scattering processes that are important at room temperature in addition to alloy scattering. The temperature dependence of the alloy-scattering mobility behaves as $T^{-1/2}$. At low temperatures, alloy and defect scattering become dominant due to freezing out of phonon scattering. Simon *et al.* and Jena *et al.* measured the mobilities of polarization-doped 3-dimensional electron slabs (3DES) with average compositions of $\langle x \rangle = 0.11$ at 10 K and 20 K, respectively.[9, 7] Their experimental values are plotted in Figure 4.6 and are in excellent agreement with our theoretical predictions, without accounting for defect scattering in our DFT calculations.

4.3.2 Total mobility

Besides alloy scattering, the mobility in AlGaN is reduced by dipole scattering, phonon scattering, and ionized-impurity scattering when dopants are present. Zhao *et al.* studied dipole scattering in AlGaN, in the relaxation-time approximation, and predicted the dipole-scattering electron mobility to be greater than 1300 $\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$ across the entire composition range, at $n = 10^{17}$ cm^{-3}.[43] In addition, Farahmand *et al.* investigated phonon and ionized-impurity scattering ($N_D^+ = 10^{17}$ cm^{-3}) in the III-nitrides using Monte-Carlo simulations on semi-empirical band structures.[44] They found the room-temperature electron mobility for GaN and AlN to be 990 and 533 $\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$, in agreement with experiments.[44, 45, 46] Using the virtual-crystal approximation, they calculated the electron mobility of $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}$ at $x = 0.2$, 0.5, and 0.8 to be equal to 978, 856, and 658 $\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$, respectively. [44] As shown in Figure 4.7, we calculated the total room-temperature electron mobility $1/\mu_{\mathrm{total}} = 1/\mu_{\mathrm{VCA}} + 1/\mu_{\mathrm{dipole}} + 1/\mu_{\mathrm{alloy}}$, by combining our μ_{alloy} with μ_{dipole} and μ_{VCA}, interpolated from the data of Zhao *et al.* and Farahmand *et al.* At $x \approx 0.6$, both the room-temperature alloy-scattering electron mobility and the total electron

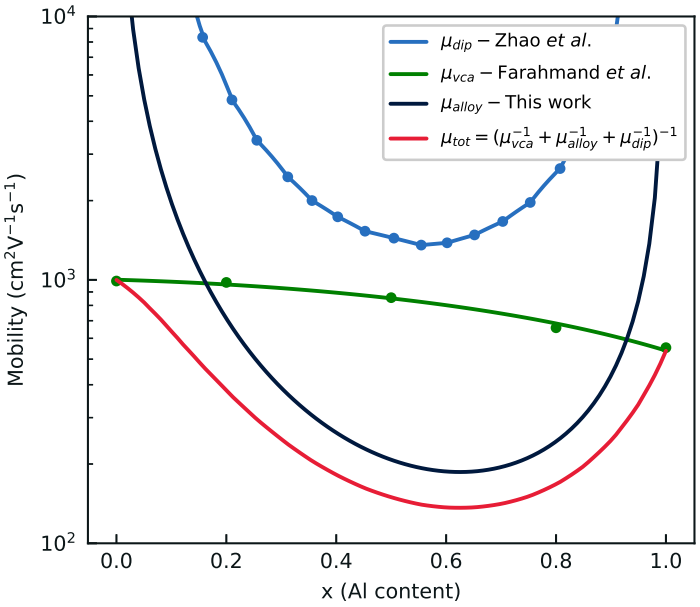

Figure 4.7: Total electron mobility of $Al_xGa_{1-x}N$ as a function of composition at room temperature and $n = 10^{17}$ cm^{-3}. The dipole-scattering mobility μ_{dip} by Zhao *et al.* and our alloy-scattering mobility μ_{alloy} assume $n = 10^{17}$ cm^{-3}. The VCA mobility μ_{vca}, which accounts for phonon and ionized-impurity scattering, by Farahmand *et al.* assumes $N_D^+ = 10^{17}$ cm^{-3}.

mobility reach their minima of 186 and 136 $\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$. For comparison, this is approximately equal to the *highest* electron mobility predicted in β-$(\mathrm{Al}_x\mathrm{Ga}_{1-x})_2\mathrm{O}_3$ at $x = 0$.[12, 13, 14, 15, 16] The comparable mobility in AlGaN for disordered compositions and superior mobility for Ga-rich and Al-rich compositions is due to stronger polar electron-phonon scattering in the III-oxides than in the III-nitrides.[15, 16] The difference in mobility is compounded at the high temperatures at which high-power devices operate,[18] due to the stronger temperature dependence of electron-phonon scattering compared to alloy scattering. Notably, Ma *et al.* showed that the phonon-scattering electron mobility in β-$\mathrm{Ga}_2\mathrm{O}_3$ decreases to 70 $\mathrm{cm}^2/\mathrm{V}\cdot\mathrm{s}$ at 500 K,[13] which is two times smaller than the lowest predicted alloy-scattering electron mobility in AlGaN at the same temperature. (See section 4.7 for a discussion of the advantages of AlGaN over β-$(\mathrm{Al}_x\mathrm{Ga}_{1-x})_2\mathrm{O}_3$ for high-power devices, considering other factors such as the band gap, ambipolar dopability, and thermal conductivity of the available substrates.)

4.4 Error analysis

In our work, we have averaged eight $4 \times 4 \times 2$ supercells for each composition and evaluated the scattering rate from first principles. Figure 4.5 shows the scattering rate that we have calculated from the energy broadening of the averaged spectral function for $\mathrm{Al}_{0.5}\mathrm{Ga}_{0.5}\mathrm{N}$.

There is a spread in the scattering rate data compared to the line of best fit. To identify the magnitude of the spread, we calculated the residuals of the first-principles scattering rates with respect to the line of best fit. We have converted the residuals back to units of energy to get the equivalent error in the spectral width, which is the quantity that we directly compute, and plotted it against the conduction-band energy on a linear scale (Figure 4.8). There is no correlation of the residuals with respect to the conduction-band energy, suggesting that the fitting model accurately describes the data. The average root mean squared error is 19 meV, calculated

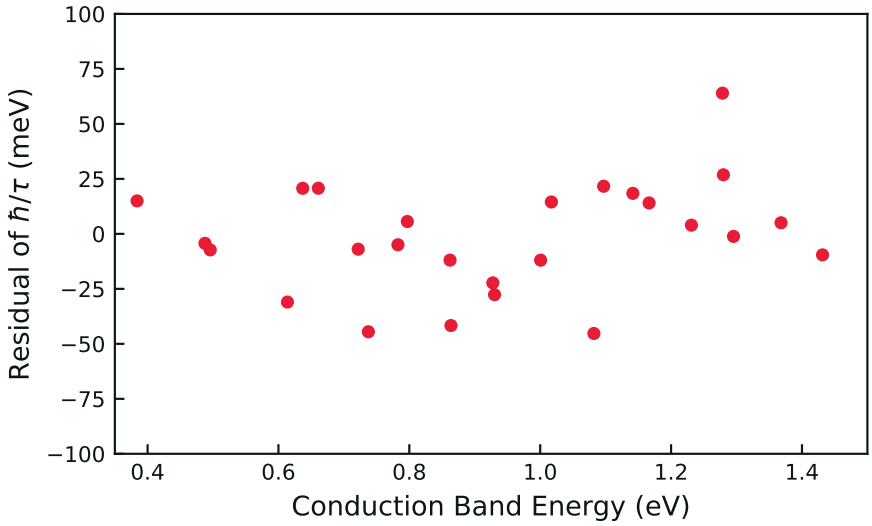

Figure 4.8: Residuals of the scattering rate compared to the line of best fit.. The scattering rates have been converted to energies to reflect the equivalent error in the spectral width. The conduction-band energy is referenced to the conduction-band minimum.

according to,

$$\sigma_{w_i} = \sqrt{\frac{1}{N-2}\sum_i \left(\hbar w_i - \frac{\hbar}{\hat{\tau}(\varepsilon_i)}\right)^2}, \tag{S6}$$

where w_i is the scattering rate, $1/\hat{\tau}(\varepsilon)$ is the line of best fit, and N is the number of scatter points. To get a better sense of the magnitude of this error, we convert it to the equivalent error in the scattering potential,

$$\sigma_{U_0} = \frac{\partial U_0}{\partial w_i}\sigma_{w_i}. \tag{S7}$$

For 50% Al content, the error in the scattering potential is 0.18 eV or 10%. We also calculated the errors in mobilities. For Al contents x = 0.25, 0.5, and 0.75, the mobility errors are 24%, 20% and 34% respectively. Therefore, the spread in the scattering rate data does not affect our mobility estimates to leading order, which is sufficient to support our comparisons with $(Al_xGa_{1-x})_2O_3$.

4.4.1 Source of the spread in the scattering rate

There are two main reasons for the spread in the scattering rate data compared to the line of best fit. First, the line of best fit assumes that the conduction band is isotropic. This is only approximately true, as shown in Figure 4.9, and explains part of the spread in the scattering rate, which is to be expected due to the directional anisotropy. Indeed, the effective scattering potential evaluated along Γ-A is 1.9 eV, which is slightly larger than the scattering potential evaluated along Γ-M (1.6 eV) and Γ-K (1.7 eV). We note that the exact value of the scattering potential can shift by ∼0.1 eV, depending on the parameters used for the peak-detection algorithm. Nonetheless, all three scattering potentials are within the error of the effective scattering potential of 1.8 eV ± 0.2 eV. Second, the periodicity of the finite-sized supercell gives rise to avoided crossings within the primitive-cell Brillouin zone. Haverkort *et al.* showed that these avoided crossings disappear after averaging the spectral functions of many structures.[47] However, since we have averaged only 8 structures, these avoided crossings remain. Our computational algorithm that

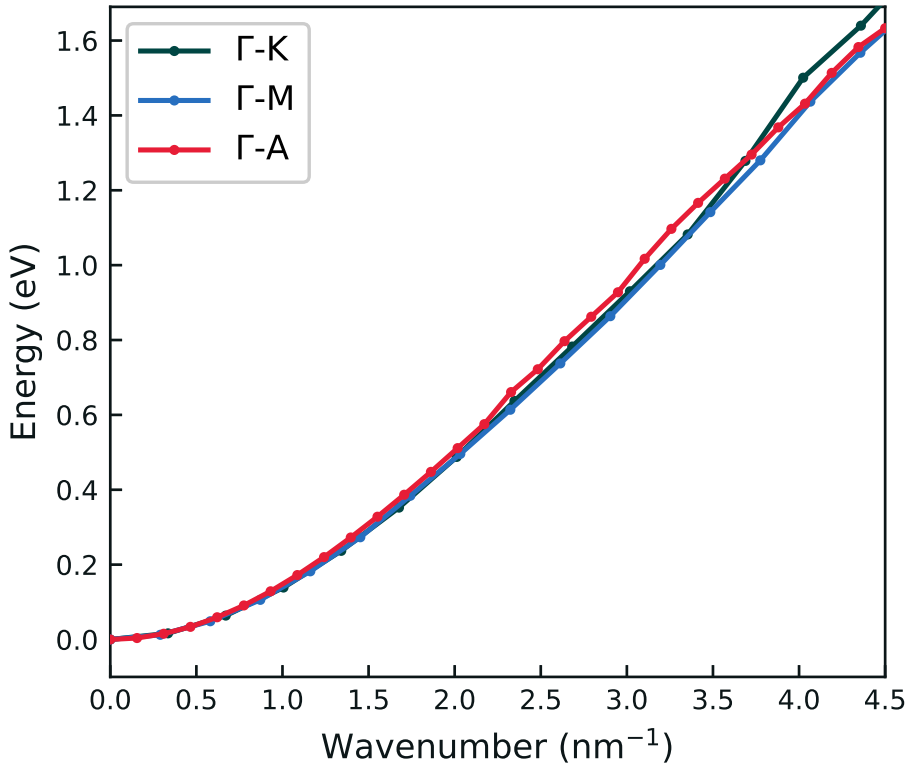

Figure 4.9: Energy dispersion of the conduction band of $Al_{0.5}Ga_{0.5}N$, plotted along the same axis for different wave-vector directions. The slight anisotropy of the conduction band gives rise to a spread in the scattering rate.

determines the spectral peak and width can be sensitive to the noise due to these avoided crossings, which accounts for part of the error in Figure 4.8.

4.5 Validation of the scattering potential

4.5.1 Comparison with a larger supercell

The finite size of the $4 \times 4 \times 2$ (128-atom) supercell that we have chosen limits the scattering rates at long wavelengths due to the periodicity of the supercell, and at short wavelengths by limiting the number of states that a carrier can scatter to. We

have overcome this problem by using special quasirandom structures (SQS's) whose site correlation functions closely match those of a perfectly random alloy.[28] We obtained the scattering rates by directly evaluating the spectral widths of the SQS spectral functions. The advantage of this approach is that the discrete SQS spectral function reproduces the statistics of a perfectly random alloy,[29] therefore the Bloch states of the supercell with wavelengths shorter than the SQS dimensions exhibit the correct energy broadenings in the primitive-cell basis. We have taken this further by averaging the spectral functions of eight SQS's since the site correlation functions are not perfect matches. We extrapolate the long-wavelength scattering rates by fitting the Golden-rule expression to the accurately described short-wavelength states. We do not expect that increasing the number of long-wavelength states (i.e., increasing the size of the supercell) would significantly affect our mobility estimates since we have already captured a majority of the (shorter-wavelength) states that contribute to the effective scattering potential.

To verify our assumptions, we evaluated the spectral function of a $5\times5\times3$ $Al_{0.5}Ga_{0.5}N$ supercell, consisting of 300 atoms. In the interest of computational efficiency, we excluded the *d*-orbitals of Ga in the valence shell since the atomic character of the lowest conduction band is predominantly *s* and *p* type. To further save computational cost, we evaluated three *k*-states as points of comparison between the 128-atom and 300-atom SQS's. In particular, we tested the ability of our fitting model to extrapolate long-wavelength scattering rates by considering a long-wavelength state, along the Γ-M direction, corresponding to an energy of 0.21 eV above the band minimum, which is not broadened in the 128-atom supercell. We also tested the convergence of the short-wavelength scattering rates by directly evaluating the spectral width of two states corresponding to energies of 0.7 eV and 1.3 eV, also along the Γ-M direction. Figure 4.10 shows that the scattering rates of the 300-atom supercell is well described by the line of best fit from the 128-atom supercell.

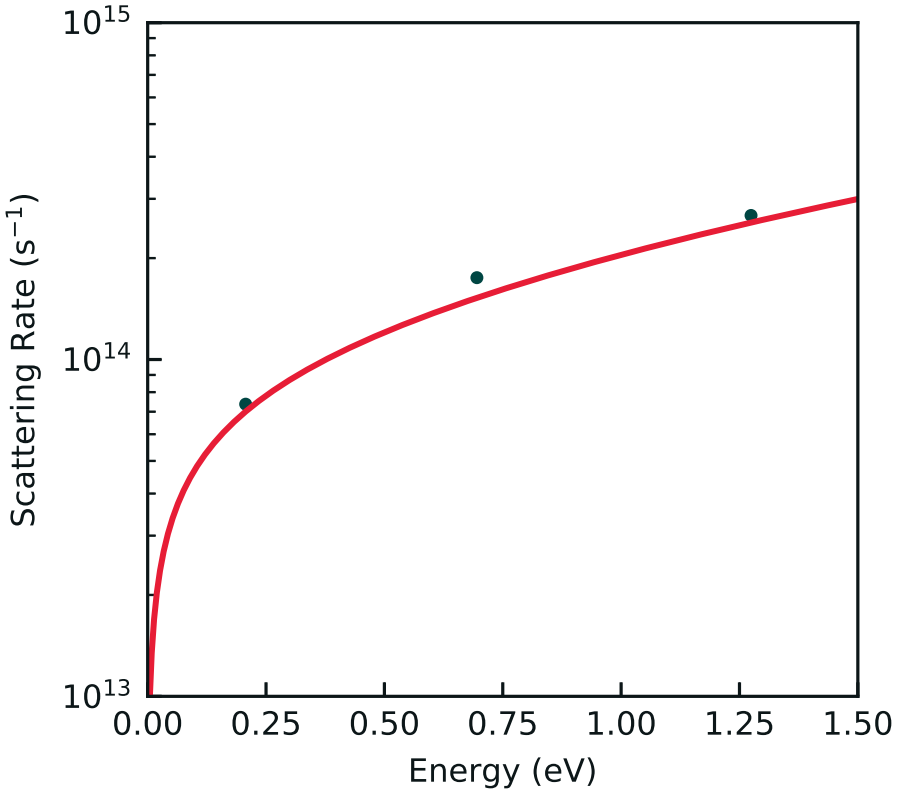

Figure 4.10: Agreement of the scattering rates (green points) evaluated directly from the spectral function of a 300-atom $Al_{0.5}Ga_{0.5}N$ supercell along the Γ-M direction to the line of best fit (solid curve) of the 128-atom supercell.

Table 4.2: Our scattering potentials U_0, calculated from DFT by unfolding the band structure for different Al contents, compared to $d\varepsilon_c/dx$ by Kyrstos *et al.*

x	U_0 (eV)	$d\varepsilon_c/dx$
0.25	1.5 ± 0.2	1.5
0.5	1.8 ± 0.2	1.8
0.75	1.9 ± 0.3	2.1

4.5.2 Energy shift due to an infinitesimal composition change

To check whether our estimate of the effective scattering potential is correct, we can directly calculate the long-wavelength scattering potential at the Γ-point by evaluating the eigenvalue shift of the conduction band minimum due to a small change in the composition.[21, 35] More precisely, for a perfectly random alloy, [35]

$$U_0 = \frac{\delta V}{(x' - x)}, \tag{S8}$$

where δV is the eigenvalue shift. In the limit $x' \to x$, this reduces to $U_0 = d\varepsilon_c/dx$ for the conduction band. Kyrstos *et al.* recently calculated the HSE conduction-band offsets for various compositions of $Al_xGa_{1-x}N$, and obtained the expression, $\varepsilon_c = 5.22x + 3.4(1 - x) - 0.55x(1 - x)$, which gives a slope of 1.8 eV at $x = 0.5$. This is precisely the value of the scattering potential that we obtained by unfolding the band structure. The slopes evaluated at $x = 0.25$ and 0.75 also agree with our scattering potentials within the margin of error. These are summarized in Table 4.2. [38] This strongly suggests the validity of the effective scattering potentials that we have derived, even at long wavelengths.

4.5.3 Substitutional-defect calculation

As a sanity check, we evaluated equation (S4.8) directly by calculating the energy shift of the Γ-point conduction-band minimum due to the substitution of a single Al-atom onto a Ga-site.[21, 35] We performed calculations on two 128-atom disordered

supercells: one with Al content $x = 64/128$, and the other with $x' = 65/128$. We relaxed the atomic coordinates after substituting the atoms. The potential of the supercell is arbitrary, thus we aligned the branch-point energies of the two SQS's, since each SQS can be thought of as a separate local micro-configuration of a larger alloy.[**?**] To calculate the branch-point energy, we included two conduction bands and four valence bands per primitive-cell in the supercell, for a total of 64 conduction bands and 128 valence bands.

Using this method, we obtained a scattering potential of 2.08 eV, which is roughly similar to the scattering potential of 1.8 eV that we derived by unfolding the band structure. The discrepancy is likely due to the fact that a single substitutional defect cannot fully account for structural disorder effects, such as internal relaxation and alloy fluctuations. These are critical in the III-nitrides since they give rise to strong piezoelectric and spontaneous polarization fields, which can significantly perturb the band structure. Our comparison of the mobility of $Al_xGa_{1-x}N$ and $(Al_xGa_{1-x})_2O_3$ are unaffected by this small discrepancy.

4.5.4 Hybrid-functional calculation

In our work, we have used the Local-Density Approximation (LDA) over the hybrid Heyd-Scuseria-Ernzerhof (HSE) functional to save computational cost. We justified this assumption by noting that both the LDA and the HSE functionals predict similar band offsets in the III-nitrides, within order 0.1 eV accuracy.[32] To verify this assumption, we evaluated the effective HSE band structure of a single 128-atom $Al_{0.5}Ga_{0.5}N$ supercell, and calculated the scattering rates, using a mixing parameter of $\alpha = 0.3$. We obtained an HSE band gap of 4.8 eV and effective mass of $0.25m_e$. The scattering rates, computed with the LDA and HSE functionals, are shown in Figure 4.11. The effective scattering potential obtained with LDA is 1.81 eV which is nearly identical to the HSE scattering potential of 1.83 eV, indicating that LDA is sufficient for this work.

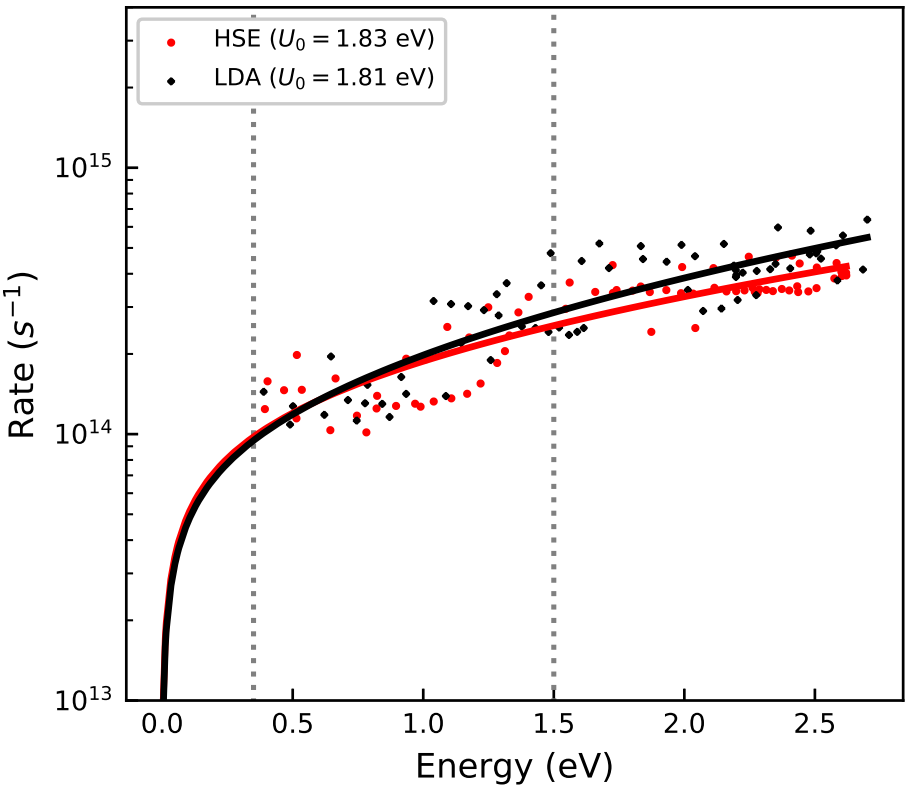

Figure 4.11: Scattering rates for a single 128-atom $Al_{0.5}Ga_{0.5}N$ SQS, computed with the LDA and HSE ($\alpha = 0.3$) functionals. The solid curves are lines of best fit, corresponding to the hard-sphere model Golden-Rule expression, with the domain of fit lying between the two vertical dotted lines. The effective scattering potentials, obtained by fitting the Golden-Rule expression to the scattering rates, are indicated in the legend.

4.6 Justification of Vegard's law for effective mass

The Local-Density Approximation (LDA) that we have used in this work does not accurately predict dispersion parameters, such as effective masses, for excited states.[40] To our knowledge, accurate experimental or theoretical reports of AlGaN effective masses for varying compositions are not available. Therefore, we have used Vegard's law (linear interpolation) to estimate the effective mass of AlGaN from the effective masses of GaN and AlN,

$$m^*_{\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{N}} = x m^*_{\mathrm{AlN}} + (1-x) m^*_{\mathrm{GaN}}. \tag{S9}$$

We obtain the effective masses of GaN and AlN from values reported by Dreyer *et al.*,[41] who used Density Functional Theory (DFT) with the Heyd–Scuseria–Ernzerhof (HSE) functional, which accurately predicts effective masses in the III-nitrides.[40, 41] The linear interpolation is justified because our LDA calculations show a near-linear dependence of effective mass on Al content, as shown in Figure 4.12.

4.7 Advantages of $\mathbf{Al_xGa_{1-x}N}$ over $\mathbf{(Al_xGa_{1-x})_2O_3}$ for high-power devices

We have shown that the low field electron mobility attainable with AlGaN is generally larger than that possible with $(\mathrm{Al}_x\mathrm{Ga}_{1-x})_2\mathrm{O}_3$. Beyond the low-field electron mobility, other considerations for high-power devices include the magnitude of the band gap, the possibility of ambipolar doping, and the thermal conductivity of the available substrates.

Wider band gaps are desirable for high-power applications because the breakdown electric field increases superlinearly with increasing band gap. The band gap of AlGaN at $x = 0.6$ is approximately 4.8 eV,[48] which is comparable to the band gap of β-$\mathrm{Ga}_2\mathrm{O}_3$.[49, 50, 51] In this regard, the advantage of AlGaN is clear, since increasing the Al content beyond $x = 0.6$ increases its band gap, thus the breakdown

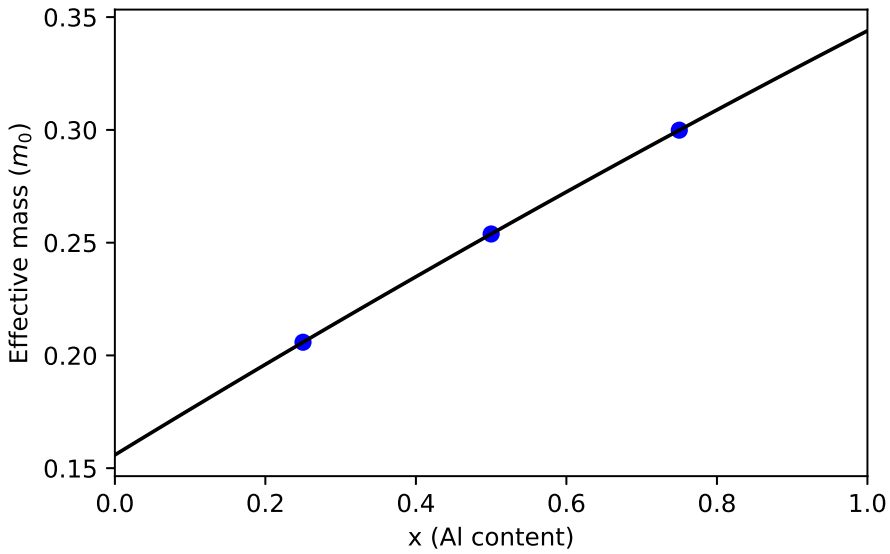

Figure 4.12: Near-linear dependence of the LDA electron effective mass of $Al_xGa_{1-x}N$ on Al content.

field, as well as its mobility to values higher than that of β-Ga_2O_3.

Although we have not considered the effect of alloying on hole mobility in this work, the ability to dope AlGaN p-type[52] further highlights AlGaN as an attractive platform for high-power complementary metal-oxide-semiconductor (CMOS) devices. Recently, Pandey *et al.* showed that Mg-doped Al-rich AlGaN, grown with metal-semiconductor junction assisted epitaxy, can exhibit hole concentrations as high as 4.5×10^{17} cm^{-3}, with resistivity $< 5\Omega \cdot$ cm and mobility of approximately $4 - 6$ cm^2/V $\cdot$ s.[52] This is in contrast to the absence of p-type doping in β-Ga_2O_3, to date.

Furthermore, strong anharmonic phonon-phonon coupling limits the thermal conductivity of β-Ga_2O_3 to values of approximately 20 W/m $\cdot$ K,[15] in contrast to GaN and AlN, which have thermal conductivities of 245 and 350 W/m$\cdot$K, respectively.[53,

54, 55, 56, 57, 58, 59, 60] The ability to grow AlGaN on thermally conductive substrates such as GaN and AlN enables efficient heat dissipation, necessary for optimal high-power performance.

These considerations elucidate the numerous advantages of AlGaN over β-$(Al_xGa_{1-x})_2O_3$ for high-performance, high-power devices, although the cost of GaN and AlN substrates remains an obstacle.

4.8 Conclusion

In summary, we have developed a first-principles approach to calculate the electron alloy-scattering rates and mobility of $Al_xGa_{1-x}N$, as a function of composition and temperature using the semiclassical Boltzmann transport equation. Our results are in agreement with experiments. Our work brings to focus the fundamental limits imposed by alloy disorder on the electron transport properties of random $Al_xGa_{1-x}N$ alloys. We confirm that experimentally measured mobilities in polarization doped samples are close to the intrinsic mobility limits. We have matched experimental results without including the effects of electron localization, which suggests that the weak electron localization in AlGaN may not be important for low-field electron transport. Finally, our analysis of the electron mobility underscores the viability of AlGaN for high-performance high-power applications. The simplicity of our computational technique makes it promising to screen the electronic properties of semiconductor alloys in a high-throughput fashion. We anticipate that our method can be further applied toward understanding the effects of alloy scattering on a broad range of alloyed semiconductors. Future work on the effects of alloy clustering and atomic ordering using the methods we have developed may lead to further insights on the mobility of materials that exhibit clustering.

Bibliography

[1] Umesh K. Mishra, Primit Parikh, and Yi Feng Wu. AlGaN/GaN HEMTs - An overview of device operation and applications. *Proceedings of the IEEE*, 90(6):1022–1031, 2002.

[2] Umesh K. Mishra, Likun Shen, Thomas E. Kazior, and Yi Feng Wu. GaN-based RF power devices and amplifiers. *Proceedings of the IEEE*, 96(2):287–305, 2008.

[3] Siddha Pimputkar, James S. Speck, Steven P. Denbaars, and Shuji Nakamura. Prospects for LED lighting. *Nature Photonics*, 3(4):180–182, apr 2009.

[4] S. Strite. GaN, AlN, and InN: A review. *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures*, 10(4):1237, 1992.

[5] Michael Kneissl, Tae Yeon Seong, Jung Han, and Hiroshi Amano. The emergence and prospects of deep-ultraviolet light-emitting diode technologies. *Nature Photonics*, 13(4):233–244, apr 2019.

[6] Debdeep Jena, Sten Heikman, Daniel Green, Dado Buttari, Robert Coffie, Huili Xing, Stacia Keller, Steve DenBaars, James S. Speck, Umesh K. Mishra, and Ioulia Smorchkova. Realization of wide electron slabs by polarization bulk doping in graded III-V nitride semiconductor alloys. *Applied Physics Letters*, 81(23):4395–4397, 2002.

[7] Debdeep Jena, Sten Heikman, James S. Speck, Arthur Gossard, Umesh K. Mishra, Angela Link, and Oliver Ambacher. Magnetotransport properties of a polarization-doped three-dimensional electron slab in graded AlGaN. *Physical Review B - Condensed Matter and Materials Physics*, 67(15):153306, 2003.

[8] Siddharth Rajan, Steven P. Denbaars, Umesh K. Mishra, Huili Xing, and Debdeep Jena. Electron mobility in graded AlGaN alloys. *Applied Physics Letters*, 88(4):042103, 2006.

[9] John Simon, Albert Wang, Huili Xing, Siddharth Rajan, and Debdeep Jena.

Carrier transport and confinement in polarization-induced three-dimensional electron slabs: Importance of alloy scattering in AlGaN. *Applied Physics Letters*, 88(4):042109, jan 2006.

[10] Shaoxin Zhu, Jianchang Yan, Yun Zhang, Jianping Zeng, Zhao Si, Peng Dong, Jinmin Li, and Junxi Wang. The effect of delta-doping on Si-doped Al rich n-AlGaN on AlN template grown by MOCVD. *Physica Status Solidi (C) Current Topics in Solid State Physics*, 11(3-4):466–468, feb 2014.

[11] Shuchang Wang, Xiong Zhang, Min Zhu, Fadi Li, and Yiping Cui. Crack-free Si-doped n-AlGaN film grown on sapphire substrate with high-temperature AlN interlayer. *Optik*, 126(23):3698–3702, 2015.

[12] K. Ghosh and U. Singisetti. Ab initio calculation of electron-phonon coupling in monoclinic β-Ga2O3. *Applied Physics Letters*, 109:072102, 2016.

[13] N Ma, N. Tanen, A. Verma, Z. Guo, T. Luo, H Xing, and D. Jena. Intrinsic electron mobility limits in β-Ga2O3. *Applied Physics Letters*, 109:212101, 2016.

[14] Y. Kang, K. Krishnaswamy, H. Peelaers, and C.G., Van de Walle. Fundamental limits on the electron mobility of β-Ga2O3. *Journal of Physics: Condensed Matter*, 29:234001, 2017.

[15] K.A. Mengle and E. Kioupakis. Vibrational and electron-phonon coupling properties of β-Ga2O3 from first-principles calculations: Impact on the mobility and breakdown field. *AIP Advances*, 9:015313, 2019.

[16] S. Ponce and F. Giustino. Structural, electronic, elastic, power, and transport properties of β-Ga2O3 from first principles. *Physical Review Research*, 2:033102, 2020.

[17] J. Bernard and A Zunger. Electronic structure of ZnS, ZnSe, ZnTe, and their pseudobinary alloys. *Physical Review B*, 36(6):3119–3228, 1987.

[18] Michael E. Coltrin and Robert J. Kaplar. Transport and breakdown analysis for

improved figure-of-merit for AlGaN power devices. *Journal of Applied Physics*, 121(5):55706, 2017.

[19] Enrico Bellotti, Francesco Bertazzi, and Michele Goano. Alloy scattering in AlGaN and InGaN: A numerical study. *Journal of Applied Physics*, 101(12):123706, 2007.

[20] F. Murphy-Armando and S. Fahy. First-principles calculation of alloy scattering in GexSi1-x. *Physical Review Letters*, 97(9):096606, 2006.

[21] J.D. Sau and Marvin Cohen. Possibility of increased mobility in Ge-Sn alloy system. *Physical Review B*, 75:045208, 2007.

[22] Vaughan M.P., Murphy-Armando F., and Fahy S. First-principles investigation of the alloy-scattering potential in dilute SixC1-x. *Physical Review B*, 85:165209, 2012.

[23] Voicu Popescu and Alex Zunger. Effective band structure of random alloys. *Physical Review Letters*, 104(23):236403, 2010.

[24] Wei Ku, Tom Berlijn, and Chi Cheng Lee. Unfolding first-principles band structures. *Physical Review Letters*, 104(21):216401, may 2010.

[25] Voicu Popescu and Alex Zunger. Extracting E versus k effective band structure from supercell calculations on alloys and impurities. *Physical Review B - Condensed Matter and Materials Physics*, 85(8):085201, feb 2012.

[26] Paulo V.C. Medeiros, Sven Stafström, and Jonas Björk. Effects of extrinsic and intrinsic perturbations on the electronic structure of graphene: Retaining an effective primitive cell band structure by band unfolding. *Physical Review B - Condensed Matter and Materials Physics*, 89(4):041407(R), jan 2014.

[27] Paulo V.C. Medeiros, Stepan S. Tsirkin, Sven Stafström, and Jonas Björk. Unfolding spinor wave functions and expectation values of general operators: Introducing the unfolding-density operator. *Physical Review B - Condensed Matter and Materials Physics*, 91(4):041116(R), 2015.

[28] Alex Zunger, S. H. Wei, L. G. Ferreira, and James E. Bernard. Special quasirandom structures. *Physical Review Letters*, 65(3):353–356, 1990.

[29] K. C. Hass, L. C. Davis, and Alex Zunger. Electronic structure of random Al0.5Ga0.5As alloys: Test of the special-quasirandom-structures description. *Physical Review B*, 42(6):3757–3760, 1990.

[30] A. Van De Walle, P. Tiwary, M. De Jong, D. L. Olmsted, M. Asta, A. Dick, D. Shin, Y. Wang, L. Q. Chen, and Z. K. Liu. Efficient stochastic generation of special quasirandom structures. *Calphad: Computer Coupling of Phase Diagrams and Thermochemistry*, 42:13–18, 2013.

[31] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. Buongiorno Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. Dal Corso, S. De Gironcoli, P. Delugas, R. A. Distasio, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H. Y. Ko, A. Kokalj, E. Kücükbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H. V. Nguyen, A. Otero-De-La-Roza, L. Paulatto, S. Poncé, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni. Advanced capabilities for materials modelling with Quantum ESPRESSO. *Journal of Physics Condensed Matter*, 29(46):465901, 2017.

[32] Poul Georg Moses, Maosheng Miao, Qimin Yan, and Chris G. Van De Walle. Hybrid functional investigations of band gaps and band alignments for AlN, GaN, InN, and InGaN. *Journal of Chemical Physics*, 134(8):84703, 2011.

[33] H. Ehrenreich and L. M. Schwartz. The Electronic Structure of Alloys. *Solid State Physics - Advances in Research and Applications*, 31(C):149–286, jan 1976.

[34] L. Fang, L.M. Porter, Davis R.F., and Schreiber D.K. Analysis of compositional uniformity in AlxGa1-xN thin films using atom probe tomography and electron microscopy. *Journal of Vacuum Science & Technology A*, 34:041510, 2016.

[35] B.K. Ridley. *Quantum Processes in Semiconductors 4e.* Oxford Press, Oxford, United Kingdom, 1999.

[36] J. W. Harrison and J. R. Hauser. Theoretical calculations of electron mobility in ternary III-V compounds. *Journal of Applied Physics*, 47(1):292–300, 1976.

[37] D.K. Ferry. Alloy scattering in ternary III-V compounds. *Physical Review B - Condensed Matter and Materials Physics*, 17(2):912–913, jan 1978.

[38] A. Kyrtsos, M. Matsubara, and E. Belloti. Band offsets of AlxGa1-xN alloys using first-principles calculations. *Journal of Physics: Condensed Matter*, 32:365504, 2020.

[39] A. Kyrtsos, M. Matsubara, and E. Belloti. First-principles study of the impact of the atomic configuration on the electronic properties of AlxGa1-xN alloys. *Physical Review B*, 99:035201, 2019.

[40] Patrick Rinke, M. Winkelnkemper, A. Qteish, D. Bimberg, J. Neugebauer, and M. Scheffler. Consistent set of band parameters for the group-III nitrides AlN, GaN, and InN. *Physical Review B - Condensed Matter and Materials Physics*, 77(7):075202, 2008.

[41] C. E. Dreyer, A. Janotti, and C. G. Van De Walle. Effects of strain on the electron effective mass in GaN and AlN. *Applied Physics Letters*, 102(14):142105, 2013.

[42] Andrew M. Armstrong and Andrew A. Allerman. Polarization-induced electrical conductivity in ultra-wide band gap AlGaN alloys. *Applied Physics Letters*, 109(22):222101, 2016.

[43] W. Zhao and D. Jena. Dipole scattering in highly polar semiconductor alloys. *Journal of Applied Physics*, 96(4):2095, 2004.

[44] M. Farahmand, C. Garetto, E. Bellotti, K. Brennan, M. Goana, E. Ghillino, G. Ghione, J.D. Albrecht, and P.P. Ruden. Monte Carlo Simulation of Electron

Transport in the III-Nitride Wurtzite Phase Materials System: Binaries and Ternaries. *IEEE Transactions on Electron Devices*, 48(3):535–452, 2001.

[45] Y. Taniyasu, M. Kasu, and T. Makimoto. Increased electron mobility in -type Si-doped AlN by reducing dislocation density. *Applied Physics Letters*, 89:182112, 2006.

[46] D.C. Look and J.R. Sizelove. Predicted maximum mobility in bulk GaN. *Applied Physics Letters*, 79:1133, 2001.

[47] M.W. Haverkort, I.S. Elfimov, and G.A. Sawatzky. Electronic structure and self energies of randomly substituted solids using density functional theory and model calculations. *arXiv [cond-mat.mtrl-sci]*, 1109.4036, 2018.

[48] I Vurgaftman and J.R. Meyer. Band parameters for nitrogen-containing semiconductors. *Journal of Applied Physics*, 94:3675, 2003.

[49] C. Janowitz, V. Scherer, M. Mohamed, A. Krapf, H. Dwelk, R. Manzke, Z. Galazka, R. Uecker, K. Irmscher, and R. Fornari. Experimental electronic structure of In2O3 and Ga2O3. *New Journal of Physics*, 13:085014, 2011.

[50] T. Onuma, S. Saito, T. Masui, T. Yamaguchi, T. Honda, and M. Higashiwaki. Valence band ordering in β-Ga2O3 studied by polarized transmittance and reflectance spectroscopy. *Japanese Journal of Applied Physics*, 54(11):112601, 2015.

[51] K.A. Mengle, G. Shi, D. Bayerl, and E. Kioupakis. First-principles calculations of the near-edge optical properties of β-Ga2O3. *Applied Physics Letters*, 109:212104, 2016.

[52] A. Pandey, X. Liu, Z. Deng, D.A. Laleyan, K. Mashooq, E.T. Reid, E. Kioupakis, P. Bhattacharya, and Z. Mi. Enhanced doping efficiency of ultrawide band gap semiconductors by metal-semiconductor junction assisted epitaxy. *Physical Review Materials*, 3:053401, 2019.

[53] G.A. Slack, R.A. Tanzilli, Pohl R.O., and Vandersande J.W. The intrinsic ther-

mal conductivity of AlN. *Journal of Physics and Chemistry of Solids*, 48(7):641–647, 1987.

[54] G.A. Slack, L.J. Schowalter, D. Morelli, and J. A. Freitas Jr. Some effects of oxygen impurities on AlN and GaN. *Journal of Crystal Growth*, 246:287–298, 2002.

[55] C. Mion, J.F. Muth, E.A. Preble, and D. Hanser. Accurate dependence of gallium nitride thermal conductivity on dislocation density. *Applied Physics Letters*, 89:092123, 2006.

[56] L. Lindsay, D.A. Broido, and T.L. Reinecke. Thermal conductivity and large isotope effect in GaN from first principles. *Physical Review Letters*, 109:095901, 2012.

[57] R. Rounds, B. Sarkar, T. Sochacki, M. Bockowski, M. Imanishi, Y. Mori, R. Kirste, R. Collazo, and Z. Sitar. Thermal conductivity of GaN single crystals: Influence of impurities incorporated in different growth processes. *Journal of Applied Physics*, 124:105106, 2018.

[58] S.R. Choi, D. Kim, S. Choa, S. Lee, and J. Kim. Thermal Conductivity of AlN and SiC Thin Films. *International Journal of Thermophysics*, 27(3):896–905, 2006.

[59] R.L. Xu, M.M. Rojo, S.M. Islam, A. Sood, B. Vareskic, A. Katre, N. Mingo, K. Goodson, H. Xing, D. Jena, and E. Pop. Thermal Conductivity of crystalline AlN and the influence of atomic-scale defects. *Journal of Applied Physics*, 126:185105, 2019.

[60] S. Dagli, K.A. Mengle, and E. Kioupakis. Thermal conductivity of AlN, GaN, and AlxGa1-xN alloys as a function of composition, temperature, crystallographic direction, and isotope disorder from first principles. 2019.

CHAPTER V

Computational Design of Atomically Ordered Superlattices of AlN and GaN for Power Electronics

Alloy scattering in random AlGaN alloys drastically reduces the electron mobility and therefore the power-electronics figure of merit. As a result, Al compositions greater than 75% are required to obtain even a two-fold increase of the Baliga figure of merit compared to GaN. However, beyond approximately 80% Al composition, donors in AlGaN undergo the DX transition which makes impurity doping increasingly more difficult. Moreover, the contact resistance increases exponentially with increasing Al content, and integration with dielectrics becomes difficult due to the upward shift of the conduction band. Atomically thin superlattices of AlN and GaN, also known as digital alloys, are known to grow experimentally under appropriate growth conditions. These chemically ordered nanostructures could offer significantly enhanced figure of merit compared to their random-alloy counterparts due to the absence of alloy scattering, as well as better integration with contact metals and dielectrics. In this work, we investigate the electronic structure and phonon-limited electron mobility of atomically thin AlN/GaN digital-alloy superlattices using first-principles calculations based on density-functional and many-body perturbation theory. The band gap of the atomically thin superlattices reaches 4.8 eV, and the in-plane (out-of-plane) mobility is 369 (452) cm^2 V^{-1} s^{-1}. Using the modified Baliga figure of

merit that accounts for the dopant ionization energy, we demonstrate that atomically thin AlN/GaN superlattices with a monolayer sublattice periodicity have the highest modified Baliga figure of merit among several technologically relevant ultra-wide band-gap materials, including random AlGaN, β-Ga_2O_3, cBN, and diamond. This chapter was reprinted (adapted) with permission from (Appl. Phys. Lett. **121**, 032105 (2022)).

5.1 Introduction

Power electronics that [1] will drive the future electrical grid, and electric rail and aviation infrastructure need semiconductors with ultra-wide band gaps, high carrier mobilities, and shallow dopants [1, 2]. Semiconductors with ultra-wide band gaps can tolerate high electric fields without electrical breakdown due to impact ionization. High carrier mobility ensures that electrical transport is energy efficient and does not generate unnecessary heat. Finally, shallow dopants with low ionization energies are necessary to efficiently introduce free electrons that conduct electricity. Using predictive first-principles calculations [3], we propose atomically thin superlattices of AlN and GaN as candidate semiconductors that satisfy all three criteria for the active region of next-generation power electronics. These semiconductors have been experimentally demonstrated for use in light-emitting diodes and are compatible with existing industrial manufacturing processes.

5.1.1 Figure of Merit for power electronics

The performance of semiconducting materials in power-electronics applications is quantified by the Baliga figure of merit (BFOM) [4] and its modified version that accounts for dopant ionization [5]. The BFOM quantifies conduction losses, and is given by the expression:

$$\text{BFOM} = \frac{\epsilon_s \mu F_{br}^3}{4}$$

where ϵ_s is the dielectric constant, μ is the carrier mobility, and F_{br} is the critical breakdown electric field, which scales superlinearly with the band gap [4]. The cubic

dependence of the BFOM on the breakdown field has led to intense research efforts in developing ultra-wide-band-gap semiconductors for power electronics. In this work, we use the BFOM and its modified version that accounts for dopant ionization to quantify the performance of semiconductors. For lateral power devices, the lateral figure of merit is an alternative metric to quantify conduction losses, and is given by LFOM = $en_s\mu F_{br}^2$, where n_s is the sheet carrier density [6]. Since the LFOM and BFOM depend very similarly on the mobility and breakdown field, we use the BFOM in our analysis for simplicity; however, our conclusions would hold equally well using the LFOM as well. Although many ultra-wide-band-gap semiconductors, e.g., AlN, diamond, and cubic boron nitride, exhibit promising BFOM, their lack of shallow dopants has hampered their adoption. Therefore, the modified BFOM [5], which is the BFOM multiplied by the dopant ionization ratio, is a more useful quantity for evaluating the performance of ultra-wide-band-gap semiconductors for power-electronics applications.

5.1.2 III-nitrides for power electronics

GaN and AlGaN are some of the most promising materials for highly efficient power-electronic devices. GaN technology is the state of the art for low to moderate power applications [7, 8], e.g., phone chargers, electric cars, and photovoltaic inverters, due to its wide band gap of 3.5 eV [9], high electron mobility of 800-1600 cm^2 V^{-1} s^{-1} [10, 11], and availability of shallow dopants [12]. The (modified) BFOM approximately scales with the band gap to the sixth power, therefore a promising approach for improving the figure of merit of GaN is increasing its band gap by alloying it with aluminum. The alloy $Al_xGa_{1-x}N$ is a solid solution of GaN and AlN, and has a band gap that can be tuned from 3.5 eV ($x = 0$) to 6.3 eV ($x = 1$) [13, 14].

However, AlGaN alloys face several challenges regarding their doping and conductivity. Despite two decades of intense research, the anticipated gain to the performance of AlGaN has not been fully realized because the electrical conductivity drops dra-

matically as the Al composition increases. Below ∼85% Al composition, the conductivity is limited by alloy scattering, which occurs due to the random occupation of Al and Ga in the lattice [15]. At the most disordered compositions of 50-60% Al, the electron mobility reaches a minimum that is seven times smaller than the electron mobility of GaN [16]. Consequently, Al compositions of ∼75% are required to obtain even a two-fold increase of the (modified) BFOM compared to GaN. Unfortunately, at compositions greater than ∼80%, the conductivity decreases again due to the donor DX transition [17, 18], which occurs when donors, e.g., Si or Ge, preferentially occupy interstitial sites rather than substitutional sites. This causes the donor transition level to lie deep within the band gap, which makes doping highly inefficient. Consequently, the modified BFOM decreases exponentially beyond an Al composition of ∼85%.

5.1.3 Atomically ordered superlattices of AlN and GaN

Electrons in atomically ordered compounds, such as superlattices, do not undergo alloy scattering. Therefore, superlattices could offer a viable route toward increasing the mobility and modified BFOM of AlGaN at an Al composition where impurity doping is efficient. Fortunately, atomically thin superlattices of alternating AlN and GaN layers have been demonstrated using common growth techniques, e.g., molecular-beam epitaxy [19, 20, 21, 22] and metalorganic-vapor-phase epitaxy [23, 24, 25]. In the limit of atomic sublattice thickness, such ordered digital alloys show significant promise for performance improvements in light-emitting diodes compared to conventional random AlGaN alloys [26, 27, 28]. In contrast to previous work, which explored increasing the alloy-scattering mobility of the two-dimensional electron gas at the GaN/AlGaN interface with the insertion of an ultra-thin AlN interlayer [29, 30, 31], we are interested in using atomically thin AlN/GaN digital-alloy superlattices as the active region for power electronics.

In this work, we use atomistic calculations based on density-functional theory (DFT), density-functional perturbation theory (DFPT), and many-body perturbation the-

ory (MBPT) to uncover the electronic and electron-transport properties of atomically thin AlN/GaN superlattices, periodically repeating along the c-axis. Such structures retain the ultra-wide band gap of AlGaN, while exhibiting an enhanced phonon-limited mobility that is 3-4x larger than the mobility of random AlGaN alloys due to the absence of alloy disorder. Most importantly, these favorable properties occur at an effective composition of 50%, where impurity doping is efficient and there is good integration with contact metals and dielectrics. As a result, the atomically thin superlattices have the highest modified BFOM of all known ultra-wide-band-gap semiconductors, and show great promise for high-performance power electronics.

5.2 Methodology

We investigated atomically thin AlN/GaN superlattices with two different stacking periods along the c-axis: one monolayer of AlN by one monolayer of GaN (1ML) stacking and two monolayers of AlN by two monolayers of GaN (2ML) stacking. We also calculated the electron transport properties of GaN and AlN to interpolate the phonon-limited mobility of random alloys. To simulate pseudomorphic strain on AlN substrates, we lattice-matched each semiconductor to the basal plane of AlN using the experimental lattice constant, while allowing the atomic positions and c-axis length to relax. We separately investigated the relaxation of the ground-state crystal structures by minimizing the total energy with respect to the atomic coordinates, and requiring all forces to be less than 10^{-3} Ry/Bohr and the total energy to be converged within 10^{-4} Ry. We performed band structure and phonon calculations using Quantum Espresso[32] in the local-density approximation (LDA)[33]. We used norm-conserving pseudopotentials for the $3s^2p^1$ valence electrons of Al, $3d^{10}4s^2p^1$ valence electrons of Ga, and $2s^2p^3$ valence electrons of N. We used a plane-wave kinetic energy cutoff of 130 Ry, and a converged $8\times8\times4$ ($8\times8\times2$) Monkhorst-Pack Brillouin-zone sampling grid for the self-consistent calculation of the 1ML (2ML) superlattice, GaN, and AlN. For the non-self-consistent calculation and the phonon calculation, we used a coarse $8\times8\times8$ ($8\times8\times4$) Monkhorst-Pack grid. We applied many-body quasiparticle corrections in the G_0W_0 approximation using BerkeleyGW

to obtain accurate band gaps and effective masses[34, 35].

5.2.1 Phonon-limited mobility calculation

To obtain the phonon-limited mobility, we iteratively solved the linearized Boltzmann transport equation using EPW[36]. This requires calculating the ab initio electron-phonon matrix elements from density-functional perturbation theory, which calculates the linear response of the Kohn-Sham potential to a collective atomic displacement through the linear response of the charge density. We included all interband and intraband scattering processes between thermally occupied electron $|n\mathbf{k}\rangle$ and phonon $|\nu\mathbf{q}\rangle$ states by integrating the electron-phonon matrix elements across the Brillouin zone. Additionally, we solved the alloy-scattering-limited mobility using an in-house code in the relaxation-time approximation, which is a valid approximation since alloy scattering is elastic and has no angular dependence.

The EPW code uses the following definition of the low-field mobility:

$$\mu_{\alpha\beta} = -\frac{1}{V_{PC} n_c} \sum_n \int \frac{d^3k}{\Omega_{BZ}} v_{nk,\alpha} \frac{\partial f_{nk}}{\partial E_\beta}, \tag{5.1}$$

where α, β are cartesian coordinates, V_{PC} is the volume of the primitive cell, n_c is the carrier density, n is the band index, k is a crystal wave vector in the first Brillouin zone, and Ω_{BZ} is the volume of the first Brillouin zone. The band velocity $v_{nk,\alpha}$ corresponds to the momentum-space gradient of the bands, $v_{nk,\alpha} = \frac{1}{\hbar}\nabla_{k,\alpha}\varepsilon_{nk}$. The quantity $\frac{\partial f_{nk}}{\partial E_\beta}$ is the linear response of the electronic occupation function to a small electric field E applied along the β direction, which we obtained by self-consistently

solving the linearized Boltzmann transport equation:

$$\frac{\partial f_{n\mathbf{k}}}{\partial E_\beta} = ev_{n\mathbf{k},\beta}\left(\frac{\partial f^0_{n\mathbf{k}}}{\partial \varepsilon_{n\mathbf{k}}}\right)\tau_{n\mathbf{k}} + \frac{2\pi\tau_{n\mathbf{k}}}{\hbar}\sum_{m\nu}\int\frac{d^3q}{\Omega_{BZ}}|g_{nm\nu}(\mathbf{k},\mathbf{q})|^2$$
$$\times[(n_{\nu q}+1-f^0_{m\mathbf{k}+\mathbf{q}})\delta(\varepsilon_{n\mathbf{k}}-\varepsilon_{m\mathbf{k}+\mathbf{q}}-\hbar\omega_{\nu\mathbf{q}})$$
$$+(n_{\nu\mathbf{q}}+f^0_{m\mathbf{k}+\mathbf{q}})\delta(\varepsilon_{n\mathbf{k}}-\varepsilon_{m\mathbf{k}+\mathbf{q}}+\hbar\omega_{\nu\mathbf{q}})]\frac{\partial f_{m\mathbf{k}+\mathbf{q}}}{\partial E_\beta}, \tag{5.2}$$

where:

$$1/\tau_{n\mathbf{k}} = \frac{2\pi}{\hbar}\sum_{m\nu}\int\frac{d^3\mathbf{q}}{\Omega_{BZ}}|g_{nm\nu}(\mathbf{k},\mathbf{q})|^2[(n_{\nu\mathbf{q}}+1-f^0_{m\mathbf{k}+\mathbf{q}})\delta(\varepsilon_{n\mathbf{k}}-\varepsilon_{m\mathbf{k}+\mathbf{q}}-\hbar\omega_{\nu\mathbf{q}})+$$
$$(n_{\nu\mathbf{q}}+f^0_{m\mathbf{k}+\mathbf{q}})\delta(\varepsilon_{n\mathbf{k}}-\varepsilon_{m\mathbf{k}+\mathbf{q}}+\hbar\omega_{\nu\mathbf{q}})] \tag{5.3}$$

In the equations above, the quantum numbers n and m are electronic band indices, k is the electronic crystal wave vector, ν is the phonon branch index, and q is the phonon wave vector in the first Brillouin zone. We calculated the electronic eigenvalues $\varepsilon_{n\mathbf{k}}$ in the G_0W_0 approximation, from which the Fermi-Dirac occupation factors $f_{n\mathbf{k}}$ are calculated at room temperature. For the mobility calculation, we included electronic states within 300 meV of the conduction-band edge since higher states do not contribute to low-field transport due to their small occupation. We calculated the phonon eigenvalues $\omega_{\nu\mathbf{q}}$ from density-functional perturbation theory in the local-density approximation, from which the Bose-Einstein occupation factors $n_{\nu\mathbf{q}}$ are calculated. Density-functional perturbation theory also produces the electron-phonon matrix elements $g_{nm\nu}(\mathbf{k},\mathbf{q})$, which give the probability amplitudes for the interband and intraband scattering processes from electronic states $|n\mathbf{k}\rangle$ to $|m\mathbf{k}+\mathbf{q}\rangle$ mediated by all $3\times N_{\text{atom}}$ phonon modes $|\nu\mathbf{q}\rangle$ throughout the Brillouin zone. The integrals over $\mathbf{k}$ and $\mathbf{q}$ in equations (1), (2), and (3) converge for very fine grids with $O(100^3)$ grid points, however density-functional and density-functional perturbation theory calculations are typically performed for coarser grids with $O(10^3)$ grid points due to their computational cost. This challenge is overcome by interpolat-

ing the coarse-grid electronic and phononic eigenvalues, velocity matrix elements, and electron-phonon scattering matrix elements to fine grids using the maximally localized Wannier-function method as implemented in the EPW code [2, 3]. In polar materials, the electron-phonon matrix elements of longitudinal-optical Fröhlich modes exhibit an $O(1/q)$ divergence as $q \to 0$, due to the dipole charge contribution to the electron-phonon interaction [4]. This presents a challenge for Wannier interpolation since divergent functions do not have well-behaved Fourier transforms. We overcame this challenge using the EPW code by analytically treating the long-range dipolar divergence while numerically treating the well-behaved short-range interaction [4]. Overall, we interpolated the necessary quantities to $160 \times 160 \times 110$ fine k and q grids for primitive-cell structures, i.e., GaN, AlN, and the 1ML AlN/GaN superlattice, and $160 \times 160 \times 55$ fine k and q grids for the 2ML AlN/GaN supercell structure to obtain converged phonon-limited mobilities.

5.2.2 Alloy-limited mobility calculation

Unlike electron-phonon scattering, electron-alloy scattering is elastic and has no angular dependence, thus the relaxation-time approximation can be more reliably used to calculate the alloy-scattering mobility. We calculated the alloy-scattering-limited mobility using an in-house code in the relaxation-time approximation, where the composition-dependent relaxation time is given by the formula,

$$\frac{1}{\tau(\varepsilon)} = \frac{2\pi}{\hbar} U_0^2 x(1-x) V_{\mathrm{PC}} \frac{m^*}{\sqrt{2\pi^2\hbar^3}} \sqrt{\varepsilon} \tag{5.4}$$

where x is the aluminum composition. In this equation, we used the geometrically averaged conduction-band effective mass $m^*(x)$, which we calculated for random alloys as a linear interpolation of the G_0W_0 effective masses of GaN and AlN. In our previous work [5], we found that the alloy-scattering potential $U_0(x)$ in AlGaN can be calculated by unfolding the band structure of special-quasirandom-structure supercells and fitting the alloy-scattering-rate expression to the energy-broadened effective band structure. In the same work, we showed that a one-to-one correspondence exists

between the alloy-scattering potential $U_0(x)$ and the slope of the conduction band versus composition curve $\partial\varepsilon_c/\partial x$, obtained by referencing the branch-point energies of the alloys. For materials where the conduction band is a quadratic function of composition, e.g., AlGaN, it turns out that the slope $\partial\varepsilon_c/\partial x$, and therefore the scattering potential, at $x = 0.5$ is equal to the conduction band offset $\Delta\varepsilon_C$ between the end binary compounds. By referencing the branch-point energies of GaN and AlN, we calculated a conduction-band offset $\Delta\varepsilon_C \approx 1.8\,\text{eV}$ at the G_0W_0 level, which is equal to the alloy-scattering potential $U_0(x = 0.5) = 1.8\,\text{eV}$ that we calculated in our previous work by unfolding the band structure at the LDA and hybrid-functional level. We excluded neutral defect and ionized impurity scattering in our calculation of the room-temperature mobility, therefore our mobility estimates serve as theoretical upper bounds.

5.3 Structural, electronic, and transport properties

By performing structural-relaxation calculations, we found that the atomically thin superlattices of AlN and GaN are well suited for epitaxial growth on bulk AlN substrates. In contrast to traditional multi-quantum-well structures, whose critical thickness is independently limited by the bulk lattice constant of each sublattice layer, the critical thickness of atomically thin superlattices is determined by a single lattice constant that describes the entire superlattice structure. In Table 5.1, we list the relaxed in-plane lattice constants a that we calculated for GaN, AlN, and the 1ML and 2ML AlN/GaN superlattices. Our calculated lattice constants for GaN and AlN are in good agreement with experiment, [37] which we also list in Table **??**. We additionally show the epitaxial strain ϵ of each material if coherently strained to the basal c-plane of AlN. The 1ML and 2ML superlattices exhibit a lattice mismatch of only 1.6% compared to AlN, which should enable thick pseudomorphic superlattice stacks on AlN substrates. To estimate the critical thickness t_{crit}, we can make use of the fact that the critical thickness scales inversely with the lattice mismatch, i.e., $t_{\text{crit}} \sim 1/|\epsilon|$[38]. Recently, 30 nm thick pseudomorphic GaN layers in AlN/GaN/AlN double heterostructures were demonstrated on AlN substrates[39, 40]. The lattice

Material	a [nm] (Theory)	a [nm] (Expt.)	ϵ (Theory)	ϵ (Expt.)
GaN	0.318	0.319	-0.035	-0.026
1ML AlN/GaN Superlattice	0.312		-0.016	
2ML AlN/GaN Superlattice	0.312		-0.016	
AlN	0.307	0.311	0	0

Table 5.1: The relaxed in-plane lattice constants a and the corresponding epitaxial strain ϵ if coherently grown on the basal c-plane of AlN. The experimental values are from Vurgaftman and Meyer. The lattice constants of the superlattices are well described by Vegard's law.

mismatch between GaN and AlN is two times greater than the lattice mismatch between the superlattices and AlN. Extrapolating from the experimentally demonstrated thickness of GaN on AlN, we roughly estimate superlattice stacks with thickness of $\sim$60 nm to be experimentally feasible. Overall, we expect that thick stacks of atomically thin AlN/GaN superlattices can be grown on AlN substrates while being nearly free of misfit dislocations that are harmful for device operation.

Our band-structure results demonstrate that the atomically thin superlattices retain the ultra-wide band gap of random AlGaN alloys and exhibit dispersive conduction bands, indicating their promise for high-power devices. Figure 6.1 shows the quasiparticle band structure of the 1ML and 2ML AlN/GaN superlattices. For both structures, the band gap is direct at the Γ-point and energetically isolated from other valleys. We calculated band gaps of 4.6 eV and 4.3 eV for the 1ML and 2ML structures. We verified these values by comparing to previous calculations [28] that explicitly included the computationally expensive semicore Ga $3s^2p^6$ electrons in addition to the $3d^{10}$ $4s^2p^1$ valence electrons that we considered in the present work. The band gaps of the 1ML and 2ML structures, calculated with the semicore pseudopotentials, are 4.8 eV and 4.6 eV. These values are in good agreement with the band gaps calculated in the present work with valence pseudopotentials, and justify the choice of treating the $3s^2p^6$ states as frozen core electrons. The band-gap results are

Superlattice Stacking Period (AlN/GaN)	**Theory (electronic, this work)**	**Theory (electronic, previous work)**	**Theory (optical, this work)**	**Experiment (optical)**
1ML / 1ML	4.6	4.8	4.7	
2ML / 2ML	4.3	4.6	4.5	
1ML / 2ML		5.0	4.9	4.9

Table 5.2: Theoretical quasiparticle band gaps (in eV) of atomically thin AlN/GaN superlattices. The experimental optical gap, measured by Wu et al., agrees with the theoretical predictions once excitonic effects are considered.

summarized in Table 5.2, and show excellent agreement with optical measurements by Wu et al. [22]. In Table **??**, we list the basic ab initio electronic and electron-transport properties of GaN, AlN, and the atomically thin superlattices, namely the effective masses (m^*), the room-temperature electron mobility (μ), the frequency of the highest longitudinal-optical (LO) mode ($\hbar\omega_{LO}$), and the static dielectric constant (ε_s). As input to our mobility calculation, we use the G_0W_0-corrected eigenvalues. We also use the electron-phonon matrix elements calculated using density-functional perturbation theory at the LDA level, which is a valid approximation since LDA wave functions are nearly identical to G_0W_0 wave functions in common semiconductors [34], and therefore should give accurate electron-phonon matrix elements. Our mobility results agree with Monte-Carlo simulations [41] and experimental measurements [42] of the mobility of AlN to within 25%, which is typical for first-principles calculations [43]. The effective mass, frequency of the highest LO mode, and dielectric constants of the 1ML superlattice are close to linear interpolations of the end binary compounds. Therefore, as the sublattice thickness decreases, the atomically thin superlattices (approximately) approach the virtual-crystal limit.

Our electron transport calculations show that the mobility of atomically thin AlN/GaN superlattices is significantly higher than the mobility of random AlGaN alloys. We calculated the total mobility of random AlGaN alloys by combining the alloy-scattering-limited mobility of disordered AlGaN with the phonon-limited mobility of a virtual

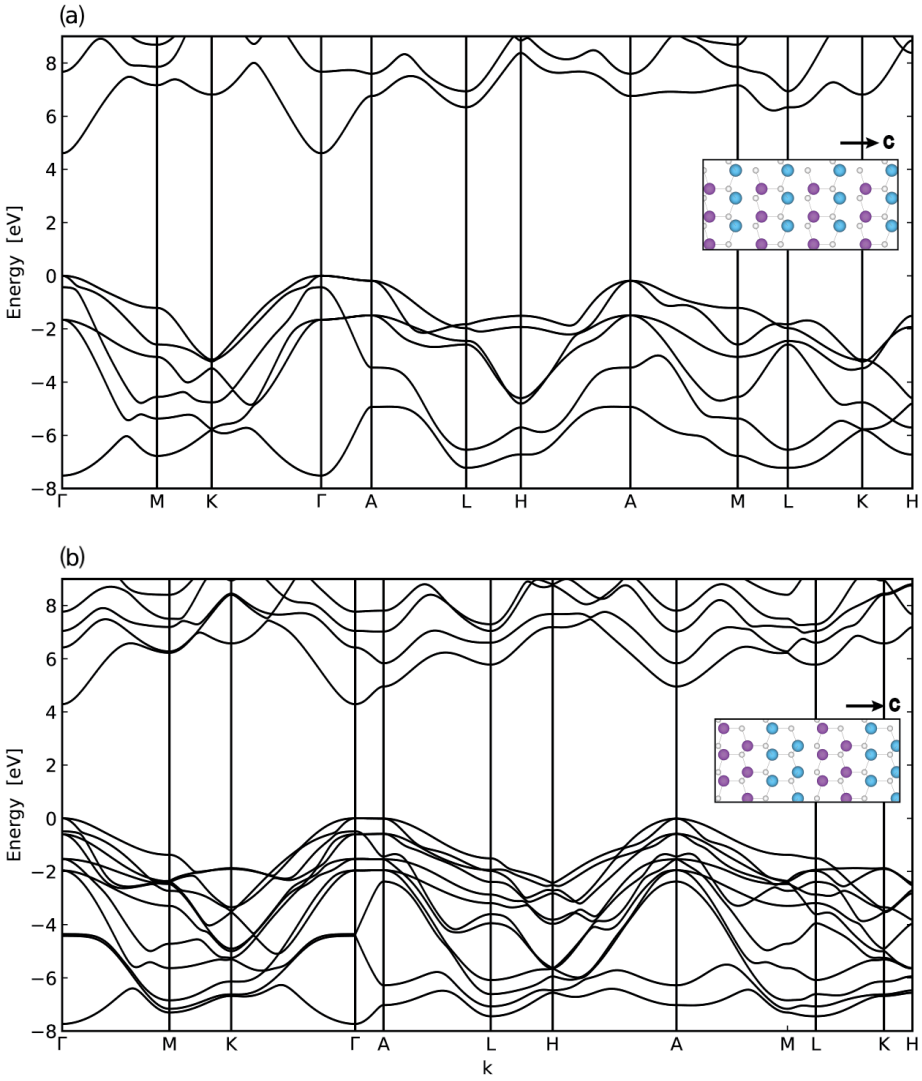

Figure 5.1: Quasiparticle band structure of (a) one-monolayer AlN / one-monolayer GaN superlattice and (b) two-monolayers AlN / two-monolayers GaN superlattice, periodically repeating along the c-axis. Both structures are pseudomorphically strained to AlN on the c-plane. The structural models for the superlattices are shown in the insets, with the wurtzite c-axis pointing to the right. The ultra-wide band gaps for both structures allow the materials to tolerate high electric fields without undergoing dielectric breakdown due to impact ionization.

Material	$m^*_\perp/m_0$	$m^*_\parallel/m_0$	$\mu_\perp$ [$\mathbf{cm^2/V{\cdot}s}$]	$\mu_\parallel$ [$\mathbf{cm^2/V{\cdot}s}$]	$\hbar\omega$ [**meV**]	$\epsilon_{s,\perp}/\epsilon_0$	$\epsilon_{s,\parallel}/\epsilon_0$
GaN	0.25	0.21	430	721	93.7	9.3	10.6
1ML/1ML AlN/GaN Superlattice	0.30	0.30	369	452	102.6	8.9	9.9
2ML/2ML AlN/GaN Superlattice	0.31	0.33	210	212	102.3	8.8	10.1
AlN	0.32	0.33	373	283	114.4	8.0	9.6

Table 5.3: Transport parameters (effective mass, room-temperature electron mobility, energy of the highest LO mode, and static dielectric constant) obtained from first-principles calculations. All materials are pseudomorphically lattice-matched to AlN on the c-plane, while the atoms and the c-axis length are allowed to relax.

crystal, using Matthiessen's rule. In our previous work, we calculated the alloy-limited mobility of disordered AlGaN alloys whose lattice constants were fully relaxed [16]. To facilitate comparisons with the superlattices, which are pseudomorphically strained to AlN, we recalculated the mobility of random AlGaN alloys that are also pseudomorphically strained to AlN. We found that strain does not change the alloy scattering potential to within 0.1 eV based on the conduction-band offset, but reduces the total mobility due to the increase of the effective mass. We calculated the virtual-crystal phonon-limited mobility by interpolating the mobility of GaN and AlN using an analytical model for piezoelectric scattering [44] that describes the functional dependence of the mobility on the effective mass, dielectric constant, and electromechanical coupling constant K, $\mu \propto \varepsilon/(K^2 m^{3/2})$. We found that the total mobility of the alloy is, to first order, independent of the electron-phonon interpolation model used because of the dominance of alloy scattering. Compared to the alloy scattering potential of 1.8 eV, the scattering potential due to monolayer fluctuations is only 0.1 eV, which is the energy difference between the conduction band of the 1ML and 2ML structures, evaluated by referencing their branch-point energies[45]. Therefore, we do not expect minor thickness fluctuations to significantly affect the mobility. In Figure 5.2, we compare the in-plane and out-of-plane

room-temperature mobility of AlN/GaN superlattices and random AlGaN alloys, at a typical electron density of 10^{18} cm^{-3}. The superlattices exhibit enhanced mobility compared to random AlGaN due to the absence of alloy disorder. In particular, the 1ML superlattice exhibits an in-plane (out-of-plane) mobility that is 3.1× (3.8×) larger than the mobility of random Al0.5Ga0.5N. As mentioned earlier, the mobility of the 1ML superlattice is close to the virtual-crystal phonon-limited mobility. The difference in the mobility between the 1ML and 2ML superlattices can be qualitatively understood in terms of the fact that there are more phonon modes that can scatter electrons in the 2ML superlattice compared to the 1ML superlattice since there are eight atoms in the primitive cell of the 2ML superlattice compared to four atoms in the 1ML superlattice. Indeed, the thermally averaged relaxation time is approximately 30% larger in the 1ML superlattice than in the 2ML superlattice. The mobility calculated in the self-energy-relaxation-time approximation (SERTA) is 35-40% larger in the 1ML superlattice ($\mu_\perp$ = 209 cm^2/Vs, $\mu_\parallel$ = 207 cm^2/Vs) than in the 2ML superlattice ($\mu_\perp$ = 154 cm^2/Vs, $\mu_\parallel$ = 146 cm^2/Vs), which additionally reflects the increased effective mass in the 2ML structure. Interestingly, self-consistently solving the iterative Boltzmann transport equation increases the mobility of the 1ML superlattice by a factor of $\sim$ 2, but the mobility of the 2ML superlattice increases only by $\sim$ 40%, compared to the SERTA mobility. This suggests that the additional electron-phonon scattering pathways in the 2ML structure contribute more strongly to backward scattering, as opposed to forward scattering, than in the 1ML structure[46]. Nevertheless, electrons in the 2ML superlattice still exhibit a 1.6× greater mobility than in random Al0.5Ga0.5N. Therefore, replacing disordered AlGaN alloys with atomically thin superlattices is a viable solution for increasing the electron mobility.

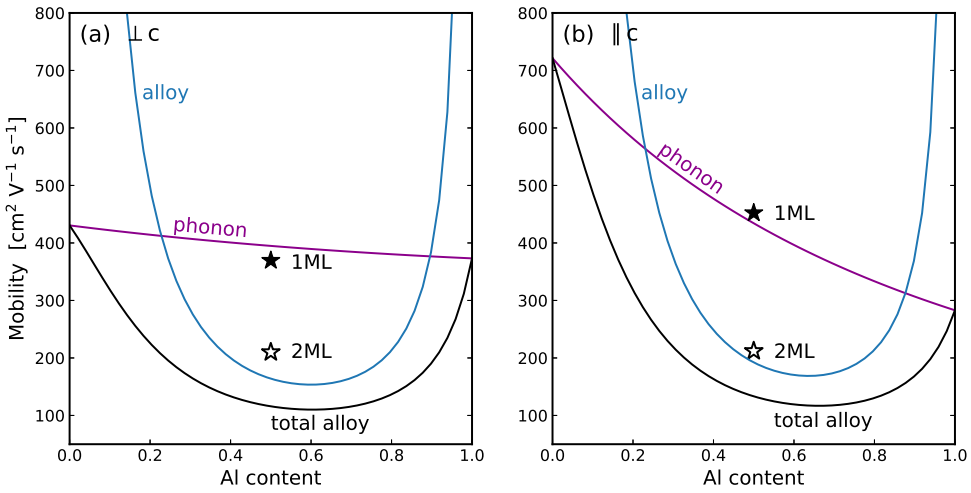

Figure 5.2: (a) In-plane ($\perp$c) and (b) out-of-plane ($\parallel$c) mobility of atomically thin AlN/GaN superlattices compared to AlGaN alloys. The semiconductors are pseudomorphically strained to AlN on the c-plane. The mobility of the superlattice with one-monolayer (1ML) sublattice periodicity is indicated by the filled star, and the mobility of the two-monolayers (2ML) superlattice is indicated by the unfilled star. The black curve is the total mobility of a random alloy, and the blue and purple curves show the alloy-scattering and phonon-scattering components, respectively. Both the in-plane and out-of-plane mobility of the superlattices exceed the mobility of random $Al_{0.5}Ga_{0.5}N$.

5.4 Prospects for power electronics

5.4.1 Power-electronics figure of merit

Our results show that the absence of alloy scattering in AlN/GaN superlattices increases the BFOM compared to both GaN and $Al_{0.5}Ga_{0.5}N$ alloys. To calculate the BFOM, we used the following formula to estimate the breakdown field, $F_{\rm br} = 3.3, {\rm MV\ cm^{-1}} \times (\varepsilon_G/3.5)^2$, where 3.3 MV cm^{-1} is the experimentally known breakdown field of GaN[47, 48]. The model proposed by Higashiwaki et al.[48] slightly overestimates the breakdown field of ultra-wide-band-gap semiconductors compared to the model that we have used; however, we verified that both models support the conclusions of our work. In Figure 5.3, we show that this simple phenomenological model properly describes the experimentally known breakdown fields in a wide range of semiconductors, including Si[48], GaAs[48], 4H-SiC[48, 49], AlGaN[50, 51, 52], diamond[53], and β-Ga_2O_3[54], although these experiments are subject to large uncertainties. This model also agrees with theoretical calculations of the breakdown field using the Von Hippel criterion[55, 56]. Accurate experimental measurements of the breakdown field do not yet exist for AlN and cBN. At a given (effective) composition, the breakdown field and dielectric constant in the superlattices are approximately equal to the breakdown field and dielectric constant in random AlGaN. However, the electron mobility is higher due to the absence of alloy scattering, thus the BFOM is also larger. In Figure 5.4, we show that the AlN/GaN superlattices exhibit greater BFOM than AlGaN alloys at an Al composition of 50% for both lateral and vertical transport. For reference, we have also shown the BFOM of relaxed GaN[10], i.e., GaN that has not been pseudomorphically strained to AlN, which is the state-of-the-art for power electronics. The advantage of the superlattices is highlighted by the fact that Al compositions of ∼75% is needed for random AlGaN alloys to obtain even a two-fold increase of its BFOM compared to GaN. For AlGaN alloys to be competitive with the 1ML superlattice, Al compositions greater than ∼85% is needed, at which point dopants undergo the DX transition. At a much lower effective composition of 50%, the 1ML superlattice has a lateral (vertical) BFOM of 15 (18) MW/cm^2, and the 2ML superlattice has a lateral (vertical) BFOM of 6.5

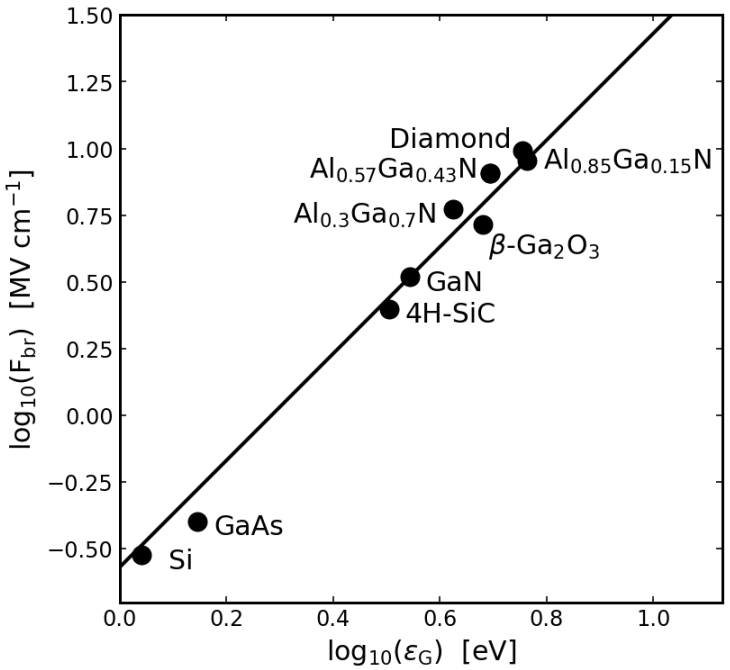

Figure 5.3: Breakdown field as a function of the band gap in a wide range of semiconductor. The scatter points are experimental, and the solid line is the phenomenological model, $F_{br} = 3.3MV/cm \times (\varepsilon_G/3.5)^2$.

(6.5) MW/cm^2, which are $\sim 4\times$ ($\sim 3\times$) and $\sim 1.5\times$ ($\sim 1.5\times$) times larger than in random $Al_{0}.5Ga_{0.5}N$.

We find that the modified BFOM, which accounts for dopant ionization, is higher in the atomically thin superlattices than in random AlGaN alloys throughout the entire composition range. We calculated the room-temperature dopant ionization ratio η using the formula, $\eta = \left(1 + g \exp\left(\frac{\varepsilon_F - \varepsilon_D}{k_B T}\right)\right)^{-1}$ where g is the degeneracy factor, ε_F is the electron quasi-Fermi level, ε_D is the dopant ionization energy, and $k_B T$ is the Boltzmann constant times the temperature. We assumed ultra-high purity of the materials, i.e., no charge compensation by impurities. We obtained ε_D by empirically fitting a sigmoid function to the experimental ionization energies of Si in AlGaN, measured by Collazo et al.[18] (see 5.5). We numerically calculated

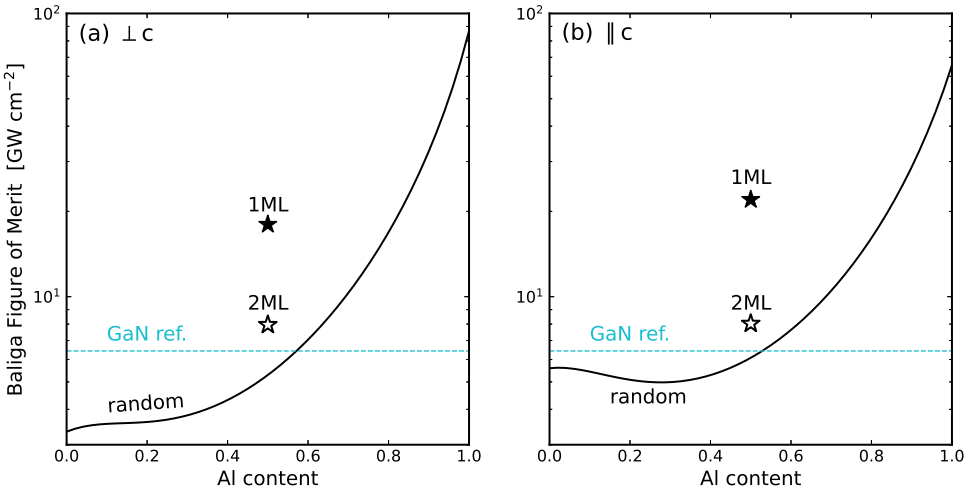

Figure 5.4: Baliga Figure of Merit for (a) lateral and (b) vertical transport in atomically thin AlN/GaN superlattices compared to AlGaN alloys. We assumed the breakdown field is related to the band gap according to, $F_{br} \propto \varepsilon_G^2$. The filled and unfilled stars show the BFOM of the one-monolayer (1ML) and two-monolayer (2ML) superlattices. The solid curve shows the BFOM of random AlGaN alloys. The dashed line shows the reference BFOM of relaxed GaN. All materials except the GaN reference are pseudomorphically lattice-matched to AlN.

the quasi-Fermi level for a fixed electron density of $10^{18}\,\mathrm{cm}^{-3}$ using the analytical 3D density-of-states expression. In Figure 5.6, we compare the modified BFOM of AlN/GaN superlattices and random AlGaN alloys. The modified BFOM of random AlGaN alloys reaches a maximum of 8.4 GW/cm^2 at an Al composition of 84%, very close to the DX transition. As we will show later in the text, this is higher than the modified BFOM of all known non-nitride semiconductors with experimentally demonstrated dopability. The modified BFOM of Al-rich AlGaN is exceeded only by the 1ML AlN/GaN superlattice, which exhibits a superior modified BFOM of 11.4 GW/cm^2 for vertical transport and 9.3 GW/cm^2 for lateral transport. Compared to random $Al_{0.5}Ga_{0.5}N$ and GaN, the modified BFOM of the 1ML superlattice is 300-400% greater. Although the modified BFOM of the 2ML superlattice is lower, it is still 65% greater than the modified BFOM of random $Al_{0.5}Ga_{0.5}N$ and 95% greater than GaN. These results underscore the advantage of nitride semiconductors for high-performance and high-power applications.

5.4.2 Contact resistance and ease of integration with dielectrics

In addition to their improved mobilities and figure of merit, the atomically thin superlattices offer lower specific contact resistance to metals and better integration with dielectrics compared to Al-rich AlGaN alloys. An additional consideration in this comparison, which is not reflected in our work, is the experimental fact that random alloys are easier to grow than atomically thin superlattices. We address this by highlighting the technological advantages that the superlattices offer compared to random alloys, beyond what is reflected in the modified BFOM. We believe that these benefits warrant experimental effort on the growth and characterization of the superlattices. Although our calculations show that Al-rich random AlGaN alloys with Al composition below ∼85% are promising in terms of their modified BFOM, their wider adoption has been hampered by the unfavorable position of their conduction band[2]. In particular, the large band offset between the conduction band of Al-rich AlGaN and the Fermi level of common ohmic-contact metals, e.g., Ti- or V/Zr-based contacts, leads to a large barrier for electron tunneling between the metal and the

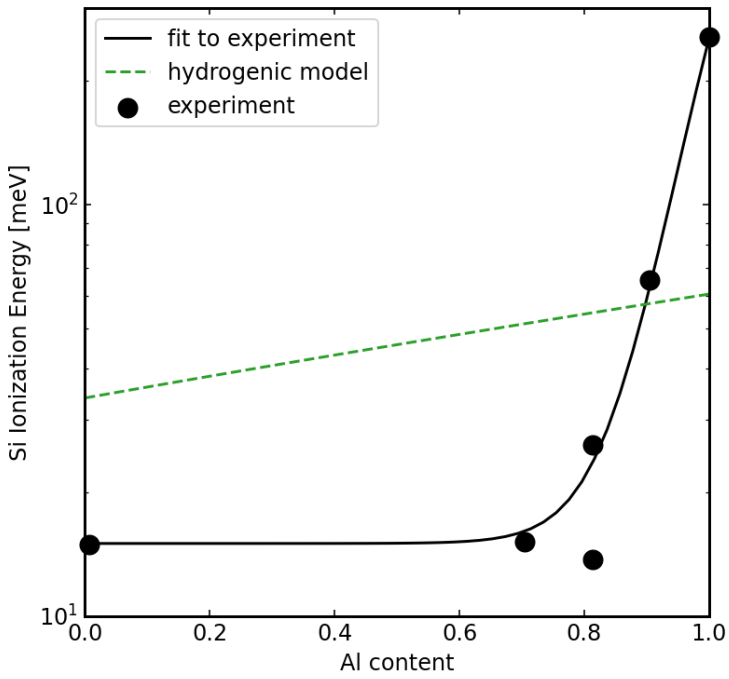

Figure 5.5: The ionization energy of Si in AlGaN as a function of Al composition. The experimental data points are obtained from Collazo et al.

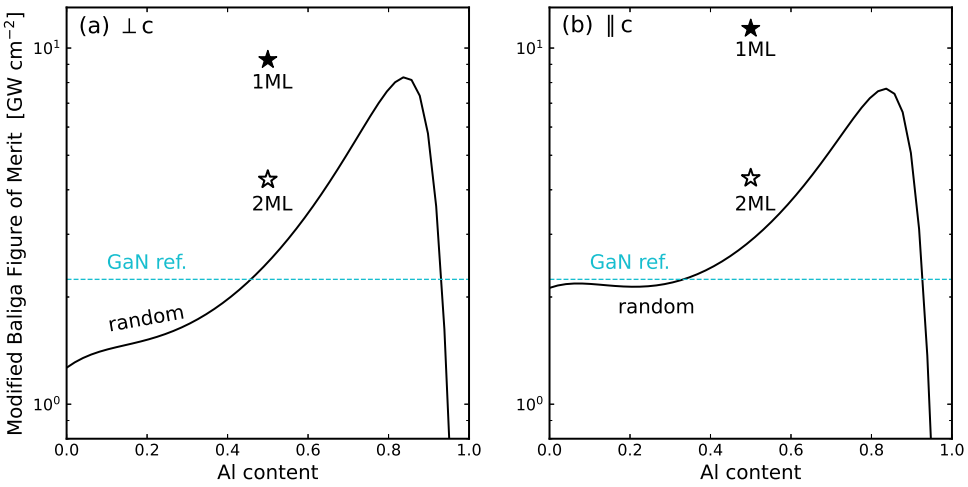

Figure 5.6: Modified Baliga Figure of Merit (BFOM) for (a) lateral transport and (b) vertical transport. The modified BFOM is the BFOM multiplied by the dopant ionization ratio, which we calculated using the dopant ionization energy measured by Collazo et al. The vertical-transport modified BFOM of the 1ML superlattice is superior to random AlGaN throughout its composition range. Compared to the current state-of-the-art GaN technology (blue line), AlN/GaN superlattices offer performance improvements of up to 400%.

semiconductor. This is problematic since the tunneling probability depends exponentially on the barrier height, i.e., $P \propto \exp\left(-\sqrt{\phi_B}\, L\right)$, where ϕ_B is the energetic barrier and L is the tunneling distance. Figure 5.7 shows the composition-dependent conduction-band position of random AlGaN alloys[14] and AlN/GaN superlattices, which we evaluated by referencing their branch point energies and used the bowing calculated by Kyrstos et al.[14] The conduction band in the 1ML (2ML) superlattice is lower by 0.43 (0.57) eV than in $Al_{0.75}Ga_{0.25}N$ and lower by 0.65 (0.79) eV than in $Al_{0.85}Ga_{0.15}N$, thus the barrier for electron tunneling is lower by the same amount. Further progress in compositionally graded AlGaN contacts[57, 58] is necessary for Al-rich random AlGaN alloys to be technologically viable. Related to the same problem, the small conduction band offset between Al-rich random AlGaN alloys and dielectrics, e.g., AlN, can lead to large leakage currents. For example, the band offset is only 0.58 eV in the $Al_{0.75}Ga_{0.25}N$/AlN system, and 0.44 eV in $Al_{0.8}Ga_{0.2}N$/AlN. In contrast, the band offset is 1.0 eV between the 1ML superlattice and AlN, and 1.15 eV between the 2ML superlattice and AlN. The more favorable conduction band position of the superlattices compared to random AlGaN alloys results in better integration with dielectrics. Hence, lower specific contact resistance and better integration with dielectrics is made possible for the atomically thin superlattices thanks to their lower effective composition and lower conduction-band position compared to Al-rich random AlGaN alloys.

5.4.3 Practical growth considerations

Although we have considered infinitely repeating periodic superlattices in this work, the structures that we have proposed can be experimentally realized by growing superlattices that are sufficiently thick. The electron thermal wavelength $\lambda_{th} = \sqrt{\frac{2\pi\hbar^2}{m^* k_B T}}$ is approximately 10 nm in AlGaN, and the scattering mean-free path $\lambda_{mfp} = \sqrt{\frac{3k_B T}{m^*}}\langle\tau\rangle$, which we estimated from our mobility calculations, is between 10 nm and 15 nm. For vertical transport, the superlattice stack thickness should exceed these length scales, with thicker stacks enabling higher breakdown voltages. For in-plane transport, we expect 30-nm-thick stacks to be sufficient, which is the typical thick-

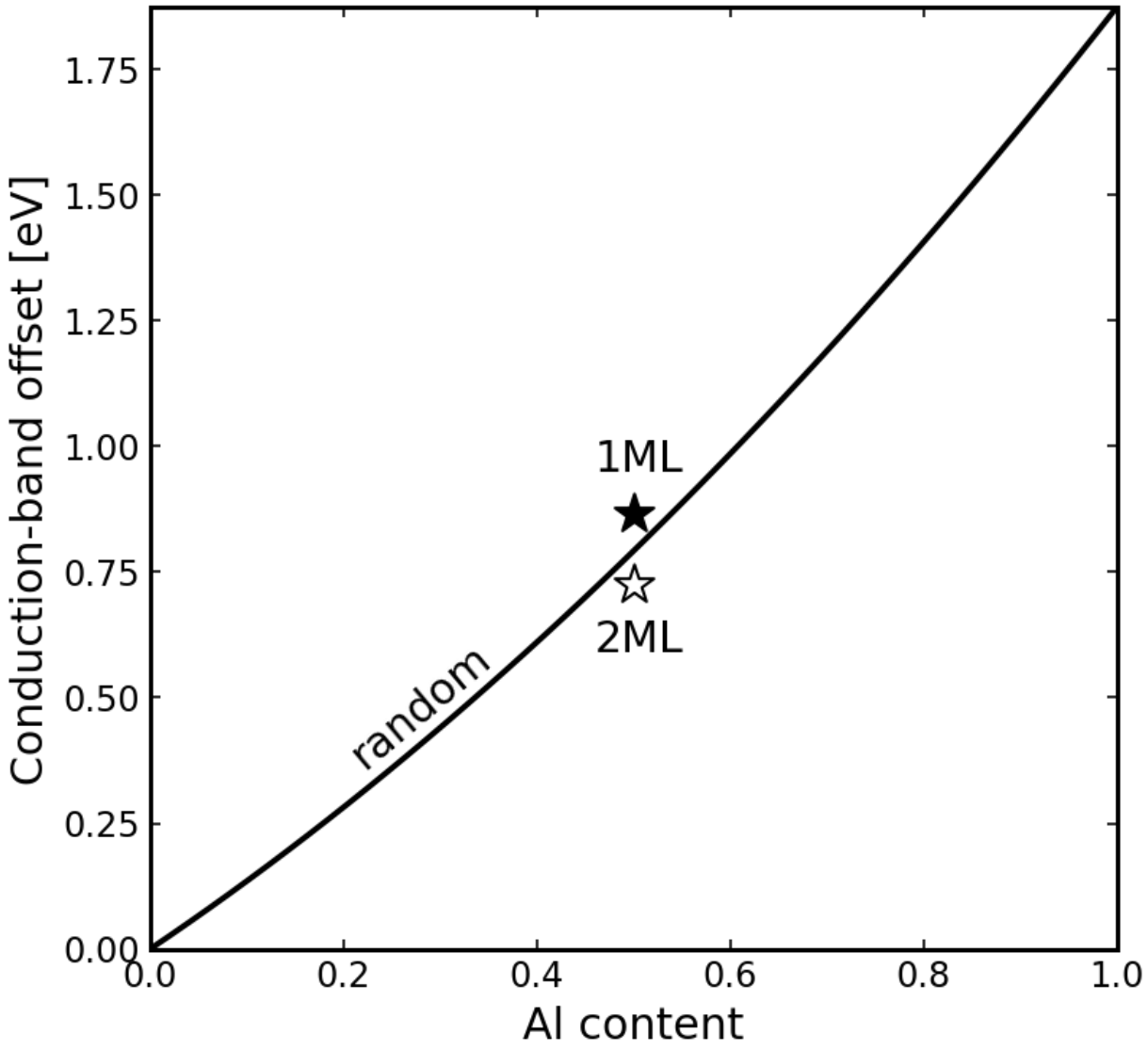

Figure 5.7: Conduction-band offset of random AlGaN alloys (solid curve) and atomically thin superlattices (stars) as a function of Al composition. The band offset is given relative to the conduction-band position of GaN, which we evaluated by referencing their branch-point energies. For random AlGaN alloys, we used the bowing parameter for the conduction band calculated by Kyrtsos et al.

ness used for GaN quantum-well high-electron mobility transistors[40]. In terms of growth, thermodynamic mixing may occur at high growth temperatures between the AlN and GaN sublattice layers, thereby producing ternary $Al_xGa_{1-x}N/Al_yGa_{1-y}N$ superlattices, with $x \approx 1$ and $y \approx 0$. Since the mobility of Al-rich and Ga-rich AlGaN alloys is phonon-limited rather than disorder-limited, we expect the performance of the atomically thin superlattices to be robust against minor ternary-cation mixing in the sublattice layers. Therefore, the superlattices that we have proposed should be experimentally feasible as long as good uniformity in the sublattice composition and thickness is maintained.

5.4.4 Comparison with other semiconductors

Overall, the 1ML AlN/GaN superlattice has the largest modified BFOM among all known semiconductors with experimentally demonstrated dopability. Its modified BFOM is larger than the modified BFOM of β-Ga_2O_3 by a factor of ∼3, 4H-SiC by a factor of ∼7, cBN by a factor of ∼12, Si by a factor of ∼1300, and diamond by a factor of ∼10,000. Table **??** lists the band gap, breakdown field, dielectric constant, dopant ionization energy, and carrier mobility that we used for the calculation of the BFOM and the modified BFOM for all semiconductors that we considered (References: [a][18], [b][5], [c][10], [d][59], [e][43]). We assume a carrier density of 10^{18} cm^{-3} for all materials when calculating the dopant-ionization fraction. For consistency in our comparison of the modified BFOM with other materials, we used first-principles band gaps and phonon-limited mobilities calculated with many-body corrections and the iterative Boltzmann transport equation with dipole-corrected ab-initio electron-phonon matrix elements[10, 43, 59]; if not available, we used values that are widely accepted in the literature[5, 56, 60, 61]. We calculated the breakdown fields using the model presented above, and obtained the dopant ionization energies from literature[5, 18, 62]. In addition to the 1ML superlattice, Al-rich AlGaN with Al composition below ∼85% shows great promise for high-power devices if the technological challenges associated with high specific contact resistance and integration with non-native dielectrics, e.g., MgO[63], can be resolved. However, these challenges are

fundamentally related to the unfavorable position of their conduction bands, and the extent to which progress can be made is unclear. In this regard, the superlattices offer a clear advantage since they have a lower effective composition and lower conduction band, which allows for better integration with metals and dielectrics. Random AlGaN alloys require a minimum Al composition of 61% for their modified BFOM to be competitive with their closest non-nitride competitor, β-Ga_2O_3, which exhibits a modified BFOM of 3.7 GW/cm^2. The 2ML AlN/GaN superlattice also exhibits a high modified BFOM of 4.3 GW/cm^2 that is comparable to the modified BFOM of β-Ga_2O_3. Unlike β-Ga_2O_3, which suffers from severe self-heating due to low thermal conductivity ($\sim$20 W m^{-1} K^{-1})[55], III-nitride semiconductors have higher thermal conductivity ($\sim$200-300 W m^{-1} K^{-1} for ordered compounds[64, 65, 66]) thanks to weaker anharmonic phonon-phonon coupling. This enables efficient cooling and, therefore, high performance since the phonon-limited mobility decreases sharply with temperature. Finally, an advantage of the III-nitrides is that they are among the few ultra-wide-band-gap semiconductors for which both n-type and p-type doping has been experimentally demonstrated, which is necessary for ambipolar high-power devices[67].

5.5 Conclusion

In summary, we propose an experimentally feasible design, i.e., atomically thin superlattices of AlN and GaN, that removes alloy scattering in AlGaN and, therefore, enhances its power-electronics figure of merit. Our calculations show that AlN/GaN superlattices are promising semiconductors for next-generation power electronics due to their ultra-wide band gap, high electron mobility, and availability of shallow dopants. They exhibit the largest modified BFOM among all the technologically relevant semiconductors that we have considered. Moreover, such superlattices offer lower specific contact resistance and better integration with dielectrics compared to Al-rich random AlGaN alloys. Most importantly, similar superlattices have already been demonstrated experimentally using industrial growth techniques. Similar theoretical characterization and materials prediction from first principles will enable the

Material	ε_G [eV]	F_{br} [MV/cm]	ϵ_s	ε_D [meV]	μ [cm^2/V·s]	BFOM [GW/cm^2]	MBFOM [GW/cm^2]
1ML/1ML AlN/GaN Superlattice	4.8	6.2	9.2	15[a]	452 (∥) 369 (⊥)	22 (∥) 18 (⊥)	11.4 (∥) 9.3 (⊥)
2ML/2ML AlN/GaN Superlattice	4.6	5.7	9.2	15[a]	210	8.0	
Random $Al_{0.75}Ga_{0.25}N$ (this work)	5.5	8.1	8.8	18[a]	125	13	6.4
Random $Al_{0.5}Ga_{0.5}N$ (this work)	4.8	6.2	9.1	15[a]	115	5.6	2.6
AlN (this work)	6.3	11	8.5	255[a]	373	87	3.5×10^{-3}
β-Ga_2O_3	4.8[b]	6.2	10[b]	30[b]	200[b]	11	3.7
GaN	3.5	3.8	9.7	15[a]	830[c]	6.4	2.2
4H-SiC	3.2[a]	3.1	9.7[b]	60[b]	900[b]	4.1	1.7
cBN	6.8[d]	12	7.1[d]	250[d]	1610[d]	490	0.95
Si	1.1[b]	0.3	11.7[b]	45[b]	1400[e]	1.2×10^{-2}	8.8×10^{-3}
Diamond	5.7[d]	8.8	5.7[d]	370[d]	1970[d]	170	1.1×10^{-3}

Table 5.4: Comparison of the Baliga Figure of Merit and Modified Baliga Figure of Merit for various semiconductors. The monolayer-thin AlN/GaN digital-alloy superlattice surpasses all known ultra-wide-band-gap semiconductors for power-electronics applications.

discovery of efficient semiconductors for a wide range of device applications.

Bibliography

[1] JY Tsao, S Chowdhury, MA Hollis, D Jena, NM Johnson, KA Jones, RJ Kaplar, S Rajan, CG Van de Walle, E Bellotti, CL Chua, R Collazo, ME Coltrin, JA Cooper, KR Evans, S Graham, TA Grotjohn, ER Heller, M Higashiwaki, MS Islam, PW Juodawlkis, MA Khan, AD Koehler, JH Leach, UK Mishra, RJ Nemanich, RCN Pilawa-Podgurski, JB Shealy, Z Sitar, MJ Tadjer, AF Witulski, M Wraback, and JA Simmons. Ultrawide-bandgap semiconductors: Research opportunities and challenges. *Adv. Electron. Mater.*, 4:1600501, 2018.

[2] MH Wong, O Bierwagen, RJ Kaplar, and H Umezawa. Ultrawide-bandgap semiconductors: An overview. *J. Mater. Res.*, 36:4601–4615, 2021.

[3] E Kioupakis, S Chae, K Bushick, N Pant, X Zhang, and W Lee. Theoretical characterization and computational discovery of ultra-wide-band-gap semiconductors with predictive atomistic calculations. *J. Mater. Res.*, 36:4616–4637, 2021.

[4] BJ Baliga. Power semiconductor device figure of merit for high-frequency applications. *IEEE Electron Device Lett.*, 10:455–457, 1989.

[5] Y Zhang and JS Speck. Importance of shallow hydrogenic dopants and material purity of ultra-wide bandgap semiconductors for vertical power electron devices. *Semicond. Sci. Technol.*, 35:125018, 2020.

[6] ME Coltrin, AG Baca, and RJ Kaplar. Analysis of 2d transport and performance characteristics for lateral power devices based on algan alloys. *ECS J. Solid State Sci. Technol.*, 6:S3114–S3118, 2017.

[7] MS Shur. Gan based transistors for high power applications. *Solid. State. Electron.*, 42:2131–2138, 1998.

[8] UK Mishra, L Shen, TE Kazior, and YF Wu. Gan-based rf power devices and amplifiers. *Proc. IEEE*, 96:287–305, 2008.

[9] PG Moses, M Miao, Q Yan, and CG Van De Walle. Hybrid functional investigations of band gaps and band alignments for aln, gan, inn, and ingan. *J. Chem. Phys.*, 134:084703, 2011.

[10] S Poncé, D Jena, and F Giustino. Hole mobility of strained gan from first principles. *Phys. Rev. B*, 100:085204, 2019.

[11] VA Jhalani, JJ Zhou, J Park, CE Dreyer, and M Bernardi. Piezoelectric electron-phonon interaction from ab initio dynamical quadrupoles: Impact on charge transport in wurtzite gan. *Phys. Rev. Lett.*, 125:136602, 2020.

[12] JL Lyons, D Wickramaratne, and CG Van De Walle. A first-principles understanding of point defects and impurities in gan. *J. Appl. Phys.*, 129:111101, 2021.

[13] A Kyrtsos, M Matsubara, and E Bellotti. First-principles study of the impact of the atomic configuration on the electronic properties of alxga1-x n alloys. *Phys. Rev. B*, 99, 2019.

[14] A Kyrtsos, M Matsubara, and E Bellotti. Band offsets of alxga1-xn alloys using first-principles calculations. *J. Phys. Condens. Matter*, 32:365504, 2020.

[15] J Simon, A Wang, H Xing, S Rajan, and D Jena. Carrier transport and confinement in polarization-induced three-dimensional electron slabs: Importance of alloy scattering in algan. *Appl. Phys. Lett.*, 88:013501–013503, 2006.

[16] N Pant, Z Deng, and E Kioupakis. High electron mobility of alxga1xn evaluated by unfolding the dft band structure. *Appl. Phys. Lett.*, 117:242105, 2020.

[17] L Gordon, JL Lyons, A Janotti, and CG Van De Walle. Hybrid functional calculations of d x centers in aln and gan. *Phys. Rev. B*, 89:085204, 2014.

[18] R Collazo, S Mita, J Xie, A Rice, J Tweedie, R Dalmau, and Z Sitar. Progress on n-type doping of algan alloys on aln single crystal substrates for uv optoelec-

tronic applications. *Phys. Status Solidi Curr. Top. Solid State Phys.*, 8:2031–2033, 2011.

[19] E Iliopoulos, KF Ludwig, TD Moustakas, and SNG Chu. Chemical ordering in algan alloys grown by molecular beam epitaxy. *Appl. Phys. Lett.*, 78:463–465, 2001.

[20] SM Islam, K Lee, J Verma, V Protasenko, S Rouvimov, S Bharadwaj, H Xing, and D Jena. Mbe-grown 232-270 nm deep-uv leds using monolayer thin binary gan/aln quantum heterostructures. *Appl. Phys. Lett.*, 110:041108, 2017.

[21] B Daudin, AM Siladie, M Gruart, M Den Hertog, C Bougerol, B Haas, JL Rouviere, E Robin, MJ Recio-Carretero, N Garro, and A Cros. The role of surface diffusion in the growth mechanism of iii-nitride nanowires and nanotubes. *Nanotechnology*, 32:085606, 2021.

[22] Y Wu, X Liu, P Wang, DA Laleyan, K Sun, Y Sun, C Ahn, M Kira, E Kioupakis, and Z Mi. Monolayer gan excitonic deep ultraviolet light emitting diodes. *Appl. Phys. Lett.*, 116:013101, 2020.

[23] Y Taniyasu and M Kasu. Polarization property of deep-ultraviolet light emission from c-plane aln/gan short-period superlattices. *Appl. Phys. Lett.*, 99:251112, 2011.

[24] V Jmerik, A Toropov, V Davydov, and S Ivanov. Monolayer-thick gan/aln multilayer heterostructures for deep-ultraviolet optoelectronics. *Phys. Status Solidi - Rapid Res. Lett.*, 15:2100242, 2021.

[25] M Asif Khan, JN Kuznia, DT Olson, T George, and WT Pike. Gan/aln digital alloy short-period superlattices by switched atomic layer metalorganic chemical vapor deposition. *Appl. Phys. Lett.*, 63:3470, 1998.

[26] X. Y. Cui, B. Delley, and C. Stampfl. Band gap engineering of wurtzite and zinc-blende gan/aln superlattices from first principles. *Journal of Applied Physics*, 108:103701, 2010.

[27] W. Sun, C. K. Tan, and N. Tansu. Aln/gan digital alloy for mid- and deep-ultraviolet optoelectronics. *Scientific Reports*, 7:1–8, 2017.

[28] D. Bayerl and E. Kioupakis. Room-temperature stability of excitons and transverse-electric polarized deep-ultraviolet luminescence in atomically thin gan quantum wells. *Applied Physics Letters*, 115:131101, 2019.

[29] K. Shinohara, D. Regan, I. Milosavljevic, A. L. Corrion, D. F. Brown, P. J. Willadsen, C. Butler, A. Schmitz, S. Kim, V. Lee, A. Ohoka, P. M. Asbeck, and M. Micovic. Electron velocity enhancement in laterally scaled gan dh-hemts with ft of 260 ghz. *IEEE Electron Device Letters*, 32:1074–1076, 2011.

[30] D. A. Deen, D. F. Storm, D. J. Meyer, R. Bass, S. C. Binari, T. Gougousi, and K. R. Evans. Impact of barrier thickness on transistor performance in aln/gan high electron mobility transistors grown on free-standing gan substrates. *Applied Physics Letters*, 105:093503, 2014.

[31] Y. Cao and D. Jena. High-mobility window for two-dimensional electron gases at ultrathin aln/gan heterojunctions. *Applied Physics Letters*, 90:182112, 2007.

[32] P. Giannozzi, O. Baseggio, P. Bonfà, D. Brunato, R. Car, I. Carnimeo, C. Cavazzoni, S. De Gironcoli, P. Delugas, F. Ferrari Ruffino, A. Ferretti, N. Marzari, I. Timrov, A. Urru, and S. Baroni. Quantum espresso toward the exascale. *Journal of Chemical Physics*, 152:154105, 2020.

[33] D. M. Ceperley and B. J. Alder. Ground state of the electron gas by a stochastic method. *Physical Review Letters*, 45:566–569, 1980.

[34] M. S. Hybertsen and S. G. Louie. Electron correlation in semiconductors and insulators: Band gaps and quasiparticle energies. *Physical Review B*, 34:5390–5413, 1986.

[35] J. Deslippe, G. Samsonidze, D. A. Strubbe, M. Jain, M. L. Cohen, and S. G. Louie. Berkeleygw: A massively parallel computer package for the calculation of the quasiparticle and optical properties of materials and nanostructures. *Computer Physics Communications*, 183:1269–1289, 2012.

[36] S. Ponce, E. R. Margine, C. Verdi, and F. Giustino. Epw: Electron-phonon coupling, transport and superconducting properties using maximally localized wannier functions. *Computer Physics Communications*, 209:116–133, 2016.

[37] I. Vurgaftman and J. R. Meyer. Band parameters for nitrogen-containing semiconductors. *Journal of Applied Physics*, 94:3675–3696, 2003.

[38] J. Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures*. 2003.

[39] M. Qi, G. Li, S. Ganguly, P. Zhao, X. Yan, J. Verma, B. Song, M. Zhu, K. Nomoto, H. Xing, and D. Jena. Strained gan quantum-well fets on single crystal bulk aln substrates. *Applied Physics Letters*, 110:063501, 2017.

[40] A. Hickman, R. Chaudhuri, S. J. Bader, K. Nomoto, K. Lee, H. G. Xing, and D. Jena. High breakdown voltage in rf aln/gan/aln quantum well hemts. *IEEE Electron Device Letters*, 40:1293–1296, 2019.

[41] J. Fang, M. V. Fischetti, R. D. Schrimpf, R. A. Reed, E. Bellotti, and S. T. Pantelides. Electron transport properties of alxga1-xn/gan transistors based on first-principles calculations and boltzmann-equation monte carlo simulations. *Physical Review Applied*, 11:044045, 2019.

[42] Y. Taniyasu, M. Kasu, and T. Makimoto. Increased electron mobility in n-type si-doped aln by reducing dislocation density. *Applied Physics Letters*, 89:182112, 2006.

[43] S. Poncé, E. R. Margine, and F. Giustino. Towards predictive many-body calculations of phonon-limited carrier mobilities in semiconductors. *Physical Review B*, 97:121201, 2018.

[44] S. Adachi. *Properties of Group-IV, III-V and II-VI Semiconductors*. 2005.

[45] A. Schleife, F. Fuchs, C. Rödl, J. Furthmüller, and F. Bechstedt. Branch-point energies and band discontinuities of iii-nitrides and iii-/ii-oxides from quasiparticle band-structure calculations. *Applied Physics Letters*, 94:012104, 2009.

[46] S. Poncé, F. Macheda, E. R. Margine, N. Marzari, N. Bonini, and F. Giustino. First-principles predictions of hall and drift mobilities in semiconductors. *Physical Review Research*, 3:043022, 2021.

[47] T. Kabemura, S. Ueda, Y. Kawada, and K. Horio. Enhancement of breakdown voltage in algan/gan hemts: Field plate plus high-k passivation layer and high acceptor density in buffer layer. *IEEE Transactions on Electron Devices*, 65:3848–3854, 2018.

[48] M. Higashiwaki, K. Sasaki, A. Kuramata, T. Masui, and S. Yamakoshi. Gallium oxide (ga2o3) metal-semiconductor field-effect transistors on single-crystal -ga2o 3 (010) substrates. *Applied Physics Letters*, 100:013504, 2012.

[49] H. Niwa, J. Suda, and T. Kimoto. 21.7 kv 4h-sic pin diode with a space-modulated junction termination extension. *Applied Physics Express*, 5:064001, 2012.

[50] D. Khachariya, S. Mita, P. Reddy, S. Dangi, P. Bagheri, M. Hayden Breckenridge, R. Sengupta, E. Kohn, Z. Sitar, R. Collazo, and S. Pavlidis. Al0.85ga0.15n/al0.6ga0.4n high electron mobility transistors on native aln substrates with ¿9 mv/cm mesa breakdown fields. In *Device Research Conference (DRC)*, pages 1–2, June 2021.

[51] A. A. Allerman, A. M. Armstrong, A. J. Fischer, J. R. Dickerson, M. H. Crawford, M. P. King, M. W. Moseley, J. J. Wierer, and R. J. Kaplar. Al0.3ga0.7n pn diode with breakdown voltage ¿1600 v. *Electron. Lett.*, 52(16):1319–1321, 2016.

[52] A. Nishikawa, K. Kumakura, and T. Makimoto. High critical electric field exceeding 8 mv/cm measured using an algan p-i-n vertical conducting diode on n-sic substrate. *Japanese J. Appl. Physics, Part 1 Regul. Pap. Short Notes Rev. Pap.*, 46:2316–2319, 2007.

[53] R. J. Kaplar, O. Slobodyan, J. D. Flicker, and M. A. Hollis. (invited) a new

analysis of the dependence of critical electric field on semiconductor bandgap. In *ECS Meet. Abstr. MA2019-02*, pages 1334–1334, 2019.

[54] X. Yan, I. S. Esqueda, J. Ma, J. Tice, and H. Wang. High breakdown electric field in -ga2o3/graphene vertical barristor heterostructure. *Appl. Phys. Lett.*, 112(3):032101, 2018.

[55] K. A. Mengle and E. Kioupakis. Vibrational and electron-phonon coupling properties of -ga2o3 from first-principles calculations: Impact on the mobility and breakdown field. *AIP Adv.*, 9:015313, 2019.

[56] S Poncé and F Giustino. Structural, electronic, elastic, power, and transport properties of -ga2o3 from first principles. *Phys. Rev. Res.*, 2:033102, 2020.

[57] S Bajaj, F Akyol, S Krishnamoorthy, Y Zhang, and S Rajan. Algan channel field effect transistors with graded heterostructure ohmic contacts. *Appl. Phys. Lett.*, 109:133508, 2016.

[58] T Razzak, S Hwang, A Coleman, H Xue, S H Sohel, S Bajaj, Y Zhang, W Lu, A Khan, and S Rajan. Design of compositionally graded contact layers for mocvd grown high al-content algan transistors. *Appl. Phys. Lett.*, 115:043502, 2019.

[59] N Sanders and E Kioupakis. Phonon- and defect-limited electron and hole mobility of diamond and cubic boron nitride: A critical comparison. *Appl. Phys. Lett.*, 119:062101, 2021.

[60] N Ma, N Tanen, A Verma, Z Guo, T Luo, H (Grace) Xing, and D Jena. Intrinsic electron mobility limits in -ga2o3. *Appl. Phys. Lett.*, 109:212101, 2016.

[61] Y Kang, K Krishnaswamy, H Peelaers, and C G Van De Walle. Fundamental limits on the electron mobility of -ga2o3. *J. Phys. Condens. Matter*, 29, 2017.

[62] O Mishima, J Tanaka, S Yamaoka, and O Fukunaga. High-temperature cubic boron nitride p-n junction diode made at high pressure. *Science (80-.)*, 238:181–3, 1987.

[63] E A Paisley, M Brumbach, A A Allerman, S Atcitty, A G Baca, A M Armstrong, R J Kaplar, and J F Ihlefeld. Spectroscopic investigations of band offsets of mgo—alxga1-xn epitaxial heterostructures with varying aln content. *Appl. Phys. Lett.*, 107:102101, 2015.

[64] S Dagli, K A Mengle, and E Kioupakis. Thermal conductivity of aln, gan, and alxga1-xn alloys as a function of composition, temperature, crystallographic direction, and isotope disorder from first principles. 2019.

[65] R Rounds, B Sarkar, T Sochacki, M Bockowski, M Imanishi, Y Mori, R Kirste, R Collazo, and Z Sitar. Thermal conductivity of gan single crystals: Influence of impurities incorporated in different growth processes. *J. Appl. Phys.*, 124:105106, 2018.

[66] R L Xu, M Munõz Rojo, S M Islam, A Sood, B Vareskic, A Katre, N Mingo, K E Goodson, H G Xing, D Jena, and E Pop. Thermal conductivity of crystalline aln and the influence of atomic-scale defects. *J. Appl. Phys.*, 126:185105, 2019.

[67] S J Bader, H Lee, R Chaudhuri, S Huang, A Hickman, A Molnar, H G Xing, D Jena, H W Then, N Chowdhury, and T Palacios. Prospects for wide bandgap and ultrawide bandgap cmos devices. *IEEE Trans. Electron Devices*, 67:4010–4020, 2020.

CHAPTER VI

Origin of the Injection Dependence of the Optical Spectrum of III-nitride Light-Emitting Diodes

III-nitride light-emitting diodes (LEDs) exhibit an injection-dependent emission blueshift and linewidth broadening that is severely detrimental to their color purity. Using first-principles multi-scale modelling that accurately captures the competition between polarization-charge screening, phase-space filling, and many-body plasma renormalization, we explain the current-dependent spectral characteristics of polar III-nitride LEDs fabricated with state-of-the-art quantum wells. Our analysis uncovers a fundamental connection between carrier dynamics and the injection-dependent spectral characteristics of light-emitting materials. For example, polar III-nitride LEDs offer poor control over their injection-dependent color purity due to their poor hole transport and slow carrier recombination dynamics, which forces them to operate at or near degenerate carrier densities. Designs that accelerate carrier recombination and transport and reduce the carrier density required to operate LEDs at a given current density lessen their injection-dependent wavelength shift and linewidth broadening. This chapter was reprinted (adapted) with permission from AIP Advances 12, 125020 (2022).

6.1 Introduction

Although III-nitride light-emitting diodes (LEDs) have been highly successful for producing blue light efficiently, they face several challenges for the longer green and red wavelengths [1]. Their wall-plug efficiency decreases as the emission wavelength increases and becomes worse for high-power operation, a phenomenon known as the green gap [2, 3, 4, 5, 6]. Another challenge is the blueshift of the emission wavelength and the broadening of the spectral linewidth with increasing carrier injection. These effects change the perceived hue, which severely deteriorates the color purity of LEDs at high operating powers [**?**]. In many cases, the perceived hue is blueshifted, and this worsens the efficiency gap by requiring even longer wavelength devices to compensate for the perceived blueshift. Despite the overwhelming technological importance of this problem, a quantitative understanding of the injection-dependent spectral blueshift and linewidth broadening has been missing.

6.1.1 Physics of band-to-band emission

The band-edge emission of polar InGaN quantum wells is determined by the interplay of competing mechanisms that contribute to the emission by shifting the band gap or by filling the bands (Figure 6.1). To date, the most widely accepted explanation of the injection-dependent blueshift is screening of polarization fields by free carriers, with a smaller role attributed to phase-space filling [**?** **?**]. Meanwhile, there is no widely accepted explanation for the origin of the linewidth broadening. III-nitride quantum wells exhibit strong piezoelectric and spontaneous polarization fields, which contribute to a quantum-confined Stark shift of the band gap [7, 8, 9]. As free carriers are injected into the quantum well, they screen the polarization charges, which results in a blueshift of the band gap as the bands flatten (Figure 6.1(a)). A competing, and often overlooked, effect that redshifts the energy is the renormalization of the band gap by many-body effects in the free-carrier plasma [10, 11, 12, 13], an effect that has been directly measured in bulk samples [14, 15]. At carrier densities exceeding 10^{18} cm^{-3} relevant for LED operation, excited carriers exist predominantly in the correlated plasma state rather than as bound excitons [16], due to Pauli blocking

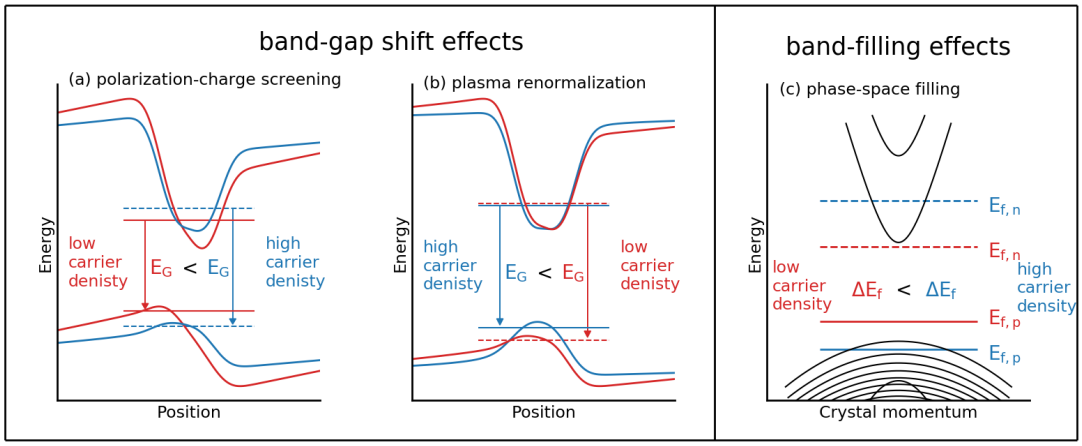

Figure 6.1: Schematic illustrations of the three primary effects that contribute to the band-edge emission of polar III-nitride quantum wells at carrier densities relevant for LED operation. Band-gap shift effects such as polarization-charge screening (panel (a)) and plasma renormalization (panel (b)) contribute to the emission spectrum by shifting the band gap E_G. Band-filling effects such as phase-space filling (panel (c)) contribute to the emission spectrum by changing the finite occupation of carriers (indicated in the figure by the electron and hole quasi-Fermi levels $E_{f,n}$ and $E_{f,p}$, and their difference ΔE_f), which in turn determines the region of phase-space from which carriers recombine to produce light.

and screening of the Coulomb interaction [**?**]. An electron (hole) in a plasma repels other electrons (holes), creating a surrounding region of positive (negative) charge, called the exchange-correlation hole [**?**]. The net result is an effective attractive potential for the carriers, which lowers the conduction band and raises the valence band as the carrier density increases (Figure 6.1(b)). In contrast to band-gap shift effects, phase-space filling contributes to a blueshift of the peak-emission energy by changing the occupancies of the bands [7, 8, 9]. As the carrier density increases and the quasi-Fermi levels penetrate deeper into the bands (Figure 6.1(c)), the emission occurs from states that are further away from the band edge. This effect becomes pronounced only if both carriers are degenerate. Therefore, the emission of InGaN quantum wells is influenced by the complex interplay of band-gap shift and band-filling effects in the free-carrier plasma.

6.1.2 Literature review

An experimental understanding of the band-edge emission of InGaN LEDs has been impeded by the difficulty in distinguishing the competing effects. For example, Kuokstis et al. compared the luminescence of bulk films against quantum wells to isolate the effects of phase-space filling from polarization-charge screening [17]. However, this approach assumes that polarization fields do not affect phase-space filling, which is not true as we will show later. On the other hand, several experimental works have attempted to explain the injection-dependent broadening of the high-energy tail of the luminescence spectrum in terms of carrier delocalization [18, 19, 20, 21, 22]. Although these works reveal interesting correlations, it is difficult to establish causation from their data. On the theoretical front, previous studies have not explained the experimentally observed injection dependence of the blueshift and linewidth broadening. Della Sala et al. used self-consistent tight-binding simulations in the virtual-crystal approximation to conclude that polarization-charge screening is responsible for the injection-dependent blueshift but they neglected phase-space filling, carrier localization, and many-body renormalization [23]. On the other hand, Peng et al. neglected alloy disorder, and it is unclear what simulation parameters they used to match experimental data since the work dates from a time when various fundamental parameters, e.g., the band gap of InN [24] and polarization constants [25], were not accurately known. Therefore, a theory of the injection dependence of the emission spectrum of III-nitride LEDs is entirely missing.

6.1.3 Overview of this work

In this work, we use first-principles multi-scale modelling to explain the carrier-injection dependence of the emission blueshift and linewidth broadening of III-nitride quantum wells. We benchmark our calculations against electroluminescence (EL) measurements of a polar InGaN quantum-well device, and show that our calculation explains the experimentally observed injection dependence of the EL spectrum. In context of these results, we identify design strategies that minimize the wavelength shift and linewidth broadening of III-nitride emitters.

6.2 Methods

6.2.1 Computational methods

We self-consistently solved the Schrödinger and Poisson equations using `nextnano++`[26] and an in-house code, with parameters determined from first-principles density-functional theory (DFT) calculations [25, 27, 28, 29, 30]. As input to our Schrödinger-Poisson calculations, we used elastic constants obtained in the local-density approximation [27] and improper polarization constants [25], deformation potentials [28], and band gaps and offsets calculated with hybrid-functional DFT [29]. To obtain room temperature values for the band gaps, we used empirical Varshni parameters [31], although the temperature-dependent band-gap narrowing is very weak in the III-nitrides. We used the two-band effective-mass model for the conduction and valence bands, which is justified since we are interested in the band-edge optical properties [32, 33, 34]. We used $m_e^* = 0.19$ ($\parallel$), $m_e^* = 0.21$ ($\perp$) and $m_h^* = 1.89$ for GaN, and $m_e^* = 0.07$ and $m_h^* = 1.81$ for InN, which are consistent with hybrid-functional [35] and many-body-perturbation-theory calculations [30].

6.2.1.1 Details of 3D calculation

For our 3D calculations with nextnano++, we simulated thirty supercells of size $18\,\text{nm} \times 18\,\text{nm} \times 21\,\text{nm}$ containing an InGaN quantum well with periodic boundaries, which is a valid approximation to the quantum well in an LED since the junction field is negligible if the device is fully turned on [12]. To account for alloy disorder, we randomly assigned the composition in each grid site as either InN or GaN, and did not perform any further compositional averaging. We used a grid-size spacing of 0.3 nm in all directions, which corresponds to the interaction distance in (In)GaN [36]. As input to our modeling, we used the out-of-plane composition profile of a commercial device, which we measured experimentally using energy-dispersive X-ray spectroscopy (EDS) and cross-validated with X-ray diffraction (XRD) measurements. Using this approach, we find that holes near the valence-band edge are localized within the plane due to alloy disorder, meanwhile electrons are extended within the

plane [12].

6.2.1.2 Details of 1D calculation

For our 1D calculations with our in-house code, we self-consistently solved the one-dimensional Schrödinger and Poisson equations with a grid spacing of 0.01 nm [37]. As mentioned in the main text, we accounted for many-body effects using the local-density approximation for the exchange-correlation potential [38]. We screened the local-density exchange-correlation potential with the low-frequency dielectric constant ϵ_0 [38]. We treated the electron and hole renormalization independently, in accordance with previous work on other semiconductors [12, 9]. This approach allows us to self-consistently calculate the band-gap shift effects due to polarization-charge screening and many-body band-gap renormalization [12, 9].

6.2.1.3 Calculation of spontaneous emission spectrum

We calculated the spontaneous-emission spectrum at first considering only band-filling effects in the disordered landscape of the quantum well, later shifting the spectrum energies to account for band-gap shift effects. We verified the validity of such a shift by checking that polarization-charge screening and plasma renormalization lead predominantly to a rigid shift of the bands (Figure 6.7). We calculated the spontaneous-emission spectrum with the equation,

$$R_{sp}(\hbar\omega) = \frac{e^2 n_r \omega}{\hbar m_0 \epsilon_0 c^3 V} |p_{cv}|^2 \frac{1}{3} \sum_{n,m} f_n f_m \left| \int_V d^3 r \psi_n(r) \psi_m(r) \right|^2 \delta(\varepsilon_n - \varepsilon_m - \hbar\omega) \quad (6.1)$$

where ω is the photon frequency, n_r is the refractive index, V is the recombination volume, p_{cv} is the bulk interband momentum matrix element between the conduction and valence bands, ψ_n and ψ_m are electron and hole envelope functions, f_n and f_m are electron and hole occupation factors, and ε_n and ε_m are electron and hole energies. We approximated the delta function with a Gaussian, and used a broadening parameter of 50 meV. We obtained the quasi-Fermi levels using the bisection method for

root finding, assuming Fermi-Dirac statistics. To account for phase-space filling, we calculated the spontaneous-emission spectrum in the rigid-band approximation. We separately calculated the change to the band gap due to polarization-charge screening and many-body renormalization at the level of the local-density approximation and the virtual-crystal approximation. We then combined these two calculations to obtain the net carrier-density dependence of the peak-emission energy, as described in the main text, thus accounting for phase-space filling, polarization-charge screening, and many-body renormalization.

6.2.1.4 Treatment of exchange and correlation

We treated many-body exchange-correlation effects of the free carriers in the local-density approximation, using the Perdew-Wang parameterization [39] of the Monte-Carlo calculation by Ceperley and Alder [40]. This treatment of exchange and correlation accurately describes the experimentally measured band-gap renormalization of bulk GaN (Figure 6.2).[14] Although the LDA works well for free-carrier plasmas in the virtual-crystal approximation, it cannot be faithfully applied to three-dimensional calculations with alloy disorder. As the plasma becomes more inhomogeneous, the LDA exchange becomes less effective in cancelling the spurious self-interaction of occupied carriers caused by the Hartree approximation [**?**]. Therefore, we have chosen to perform our calculations of polarization-charge screening and many-body renormalization in the virtual-crystal approximation, where the use of the LDA exchange-correlation is justified. Nevertheless, we do not expect the conclusions of our one-dimensional virtual-crystal calculations to change in the presence of carrier localization. Localized holes will screen the polarization charge less effectively compared to extended virtual-crystal states, but they will also contribute to a smaller band-gap renormalization due to reduced Coulomb matrix elements with other holes. Hence, we expect a cancellation of errors between polarization-charge screening and many-body renormalization in one-dimensional calculations, which justifies our virtual-crystal treatment of free-carrier screening. We note that this is an improvement over previous works that have neglected many-body exchange-correlation effects entirely

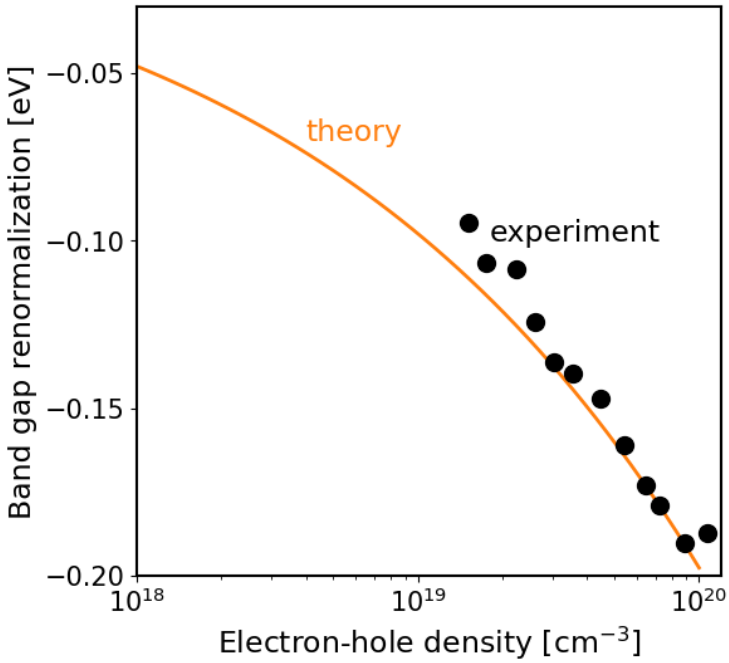

Figure 6.2: Theoretical band-gap renormalization by free carriers due to many-body exchange-correlation effects in bulk GaN (solid curve), compared to experimental measurements (scatter points) by Nagai et al.

[32, 34, 41, 42, 43].

6.2.2 Experimental methods

To validate the accuracy of our calculations, our collaborators at the University of New Mexico performed experimental measurements of the current-dependent electroluminescence (EL) spectrum of an InGaN LED packaged at Lumileds. These LEDs were designed so that practically all of the recombination occurs over a single quantum well, thus allowing us to determine the carrier density, which is needed to compare experiment with simulation. In state-of-the-art green LEDs, the growth of the active layers is optimized around V-defects in order to enable efficient hole injection into quantum wells farther away from the p-side of the device[38]. In such LEDs, recombination occurs in multiple quantum wells, resulting in improved EQE

droop; however, this also gives rise to large uncertainty in estimating the carrier density, which is needed for comparison to theory. To avoid the uncertainty in carrier density, a quasi-single-quantum-well LED of simplified epitaxial design was the focus of our experimental study. This simplified device is representative of the quantum-well recombination dynamics in state-of-the-art LEDs, but not the inter-well carrier transport. The active region is comprised of three 3 nm quantum wells but practically all of the recombination occurs in the well closest to the p-type AlGaN electron blocking layer. This conclusion is supported by analysis of the measured angular distribution of the far-field radiation of unencapsulated planar LEDs[44] and also by comparing the device characteristics to those of an otherwise equivalent LED with the two wells closer to the n-side of the junction modified to emit blue instead of green by reducing their indium concentrations. The latter LED shows an obvious difference in photoluminescence spectra but its current-dependent electroluminescence characteristics (spectra, EQE, and forward voltage) are practically identical to those of the studied LED having three green wells. The epitaxial wafers were fabricated into LEDs using established manufacturing processes at Lumileds, and packaged into LUXEON C packages for testing.

The electroluminescence measurements of the quasi-single-quantum-well LED were performed under pulsed operation to minimize Joule heating, while ensuring that the time-averaged current density is only 1% of the peak current density. Our measurements exhibit both a current-dependent blueshift of the peak emission energy and broadening of the spectral linewidth (Figure 6.3(a)). The injection-dependent broadening is stronger on the high-energy side of the luminescence spectrum, which other groups have observed as well [18, 19, 20, 21, 45]. In order to compare our measurements with theory, we measured the recombination lifetime and carrier density using a previously developed small-signal RF technique [46, 47], in which we acquired and simultaneously fit the input impedance and modulation response to an equivalent circuit model of the LED to obtain the differential carrier lifetime. We then integrated the differential carrier lifetime to obtain the full carrier lifetime [**?** 48]. Figure 6.3(b) shows the recombination lifetime as a function of the current density;

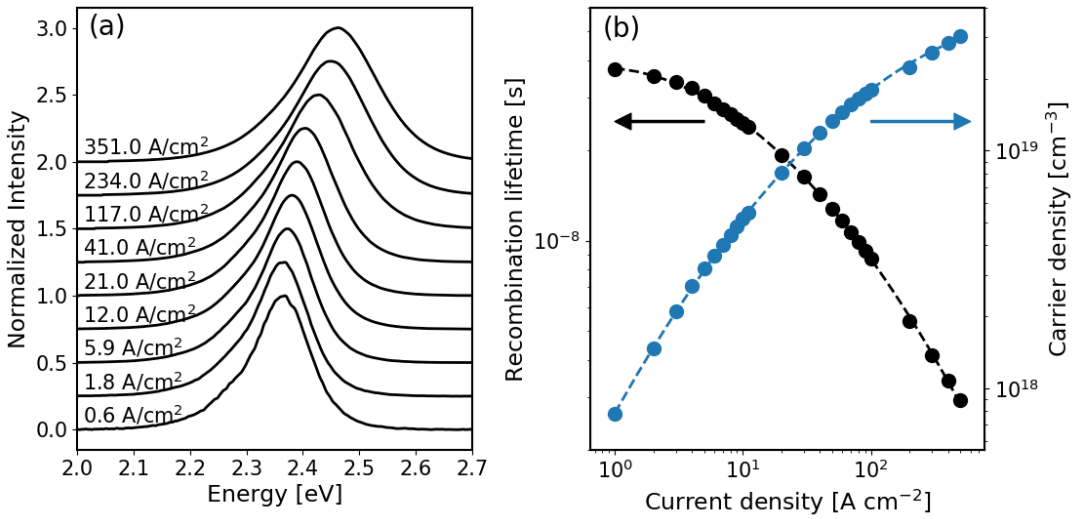

Figure 6.3: (a) Experimentally measured electroluminescence spectra of the InGaN quantum-well LED exhibiting a current-dependent blueshift and linewidth broadening. (b) Experimentally measured recombination lifetime (left axis) and the carrier density (right axis) calculated from the recombination lifetime, as a function of the injected current density.

we also show the equivalent carrier density calculated from the relation, $J = en_{2D}/\tau$, where J is the current density, n_{2D} is the two-dimensional carrier density, and τ is the recombination lifetime. By measuring the recombination lifetime at various current densities, we converted the current dependence of the EL spectra to a carrier-density dependence, which is directly accessible in our calculations.

6.3 Source of the blueshift and linewidth broadening

6.3.1 Peak-emission shift

Our modeling shows that we can accurately describe the carrier-density dependence of the peak-emission blueshift if we include the contributions of phase-space filling, polarization-charge screening, and many-body renormalization. In Figure 6.4, we show that our calculated carrier-density dependence of the peak-emission energy is in excellent agreement with experiment. We found that we needed to rigidly shift the

band gap by -0.4 eV to quantitatively match the experimental gap, which suggests the presence of a systematic band-gap error in the modified $k \cdot p$ model [49]. We calculated the relative contribution of phase-space filling in the rigid-band approximation by assuming that polarization-charge screening and plasma renormalization can be treated, to first order, as a rigid shift of the bands. We obtained the relative contribution of polarization-charge screening by calculating the band-gap shift in a one-dimensional calculation at the level of the mean-field Hartree approximation. Finally, we obtained the relative contribution of plasma renormalization by taking the difference of the band-gap shift between the Hartree and local-density approximations. We also find that for quantum wells with thicknesses of $\sim$3 nm, the band-gap blueshift due to polarization-charge screening is compensated by a redshift due to plasma renormalization. Importantly, we show that polarization screening, phase-space filling, and plasma renormalization do not independently describe the shape of the carrier-density dependence curve. Therefore, our results show that it is crucial to accurately capture the contribution of all three effects to correctly model the emission spectra of InGaN emitters.

6.3.2 Cancellation between polarization screening and plasma renormalization

The blueshift of the band gap due to polarization-charge screening is compensated by a redshift of the band gap due to plasma renormalization. We demonstrate that the quantum-well thickness influences how polarization-charge screening and many-body effects compete in shifting the band gap. In the range of carrier densities relevant for LED operation, the band gap remains approximately independent of the carrier density due to a cancellation of the blueshift due to polarization-charge screening by the redshift due to plasma renormalization, as shown in Figure 7.4(a). The cancellation of these two effects depends on the quantum-well thickness. As a point of comparison, we consider the carrier-density range between 310^{10} cm^{-2} and 310^{12} cm^{-2}. In Figure 7.4(b), we show that there is an approximate cancellation of the two effects for 3 nm quantum wells, which is the thickness typically used in

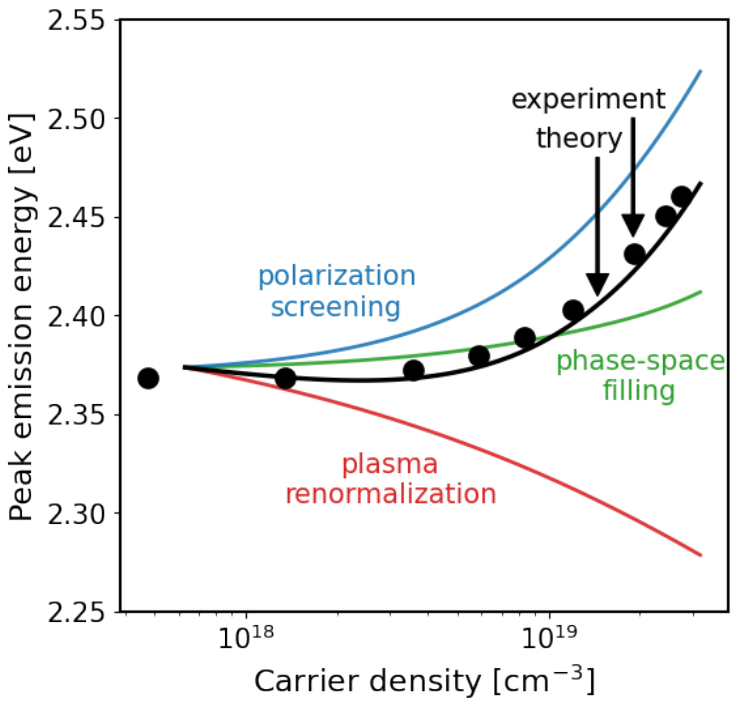

Figure 6.4: Theoretical carrier-density dependence of the peak emission energy of an InGaN quantum well (solid black curve) compared to experiment (scatter points). We show the relative contributions from polarization-charge screening (blue curve), phase-space filling (green curve) and plasma renormalization (red curve). There is excellent agreement between theory and experiment only if all three effects are included.

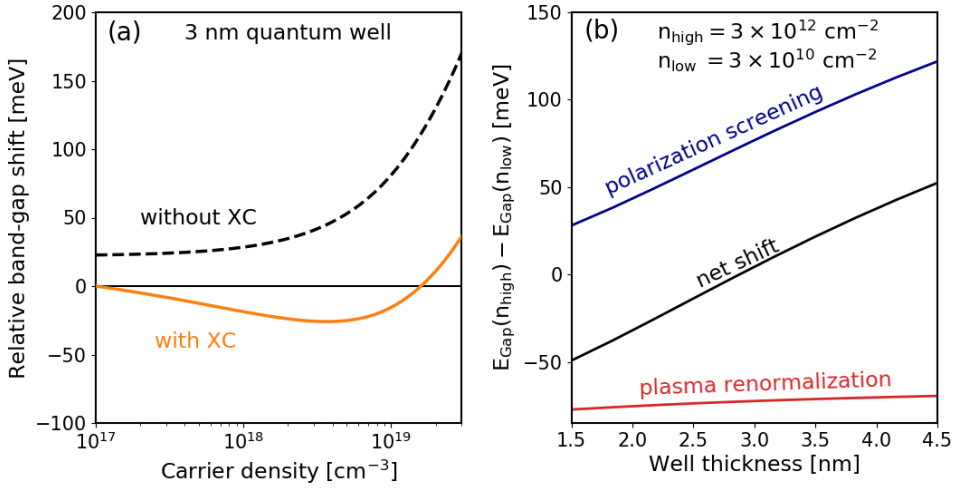

Figure 6.5: (a) Relative band-gap shift of a 3 nm InGaN quantum well, compared to the band gap at a carrier density of 10^{17} cm^{-3}, with (solid curve) and without (dashed curve) exchange-correlation (XC) effects, showing the importance of including many-body effects in calculations to describe the band-gap shift. (b) The band-gap shift of green-emitting quantum wells between carrier densities of $n_{low} = 3 \times 10^{10}$ cm^{-2} and $n_{high} = 3 \times 10^{12}$ cm-2, as a function of the quantum-well thickness. There is virtually no net band-gap shift from n_{low} to n_{high} for 3 nm quantum wells due to a fortuitous cancellation between polarization screening and plasma renormalization.

commercial LEDs, thus the blueshift in this range is predominantly due to band-filling effects rather than band-gap shift effects. Overall, thicker wells experience a net band-gap blueshift while thinner wells experience a net redshift, thus reflecting the challenge in fabricating long-wavelength emitters based on thick polar quantum wells.

6.3.3 Linewidth broadening

Furthermore, we find that phase-space filling of carriers in the disordered potential landscape of the InGaN quantum well accurately describes the experimentally measured linewidth broadening. In Figure 6.6(a), we show that our calculations of phase-space filling in the rigid-band approximation predict the relative increase of

the full-width at half-maximum (FWHM) of the EL spectrum as a function of the carrier density. We report only the relative change to the FWHM rather than the exact value since only the former is physically meaningful due to the use of a constant energy-broadening parameter in calculating the joint density of states. As shown in Figure 6.6(b), a signature of phase-space filling is broadening of the high-energy luminescence tail, which is visible in the experimental EL spectrum of Figure 6.3(a) as well. According to the van-Roosbroeck-Shockley relation [50], the low-energy tail of the luminescence spectrum corresponds to the shoulder of the joint density of states while the high-energy tail corresponds to the tail of the product of the electron and hole occupation functions. Since electrons are lighter than holes in the III-nitrides, the onset of hole degeneracy determines the onset of the broadening of the high-energy tail since both carriers need to be degenerate for phase-space filling to contribute to the peak wavelength blueshift and linewidth broadening. Since strongly localized carriers have smaller density of states than extended states, carrier localization exacerbates phase-space filling. However, localization is not a requirement for linewidth broadening, as previously conjectured [18, 20, 22], since broadening of the Fermi tail is a general feature of degenerate-carrier statistics. Our observation that polarization-charge screening and plasma renormalization lead predominantly to a rigid shift of the bands (see Figure 6.7) further supports the argument that these two effects are less important than phase-space filling in explaining the linewidth broadening.

We argue that the injection-dependent linewidth broadening must be predominantly determined by phase-space filling, rather than polarization-charge screening or many-body renormalization. Indeed, polarization-charge screening leads to the removal of the quantum-confined Stark effect, which lifts the level repulsion between states that were mixed by the electric field. This increases the density of states near the band edge, thus shrinking the region of phase-space that carriers occupy. Such an effect would lead to linewidth narrowing, which is qualitatively inconsistent with the experimentally observed broadening. Therefore, polarization-charge screening can be ruled out as a source of the broadening [36]. Polarization fields have an additional

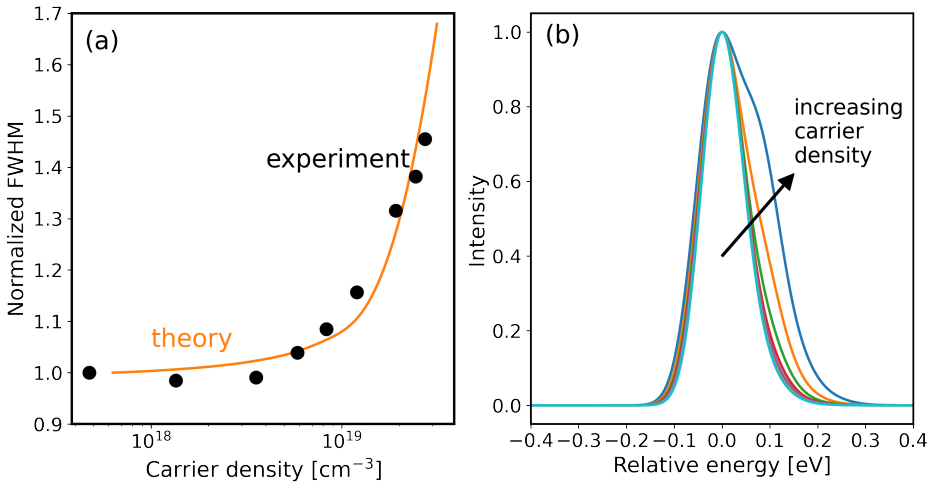

Figure 6.6: (a) Carrier-density dependence of the luminescence full-width at half-maximum due to phase-space filling of carriers in the disordered potential landscape of the InGaN quantum well. (b) Theoretical luminescence curve of a representative InGaN quantum well, with the peak-emission energy centered at zero. The signature of phase-space filling is broadening of the high-energy tail of the luminescence spectrum.

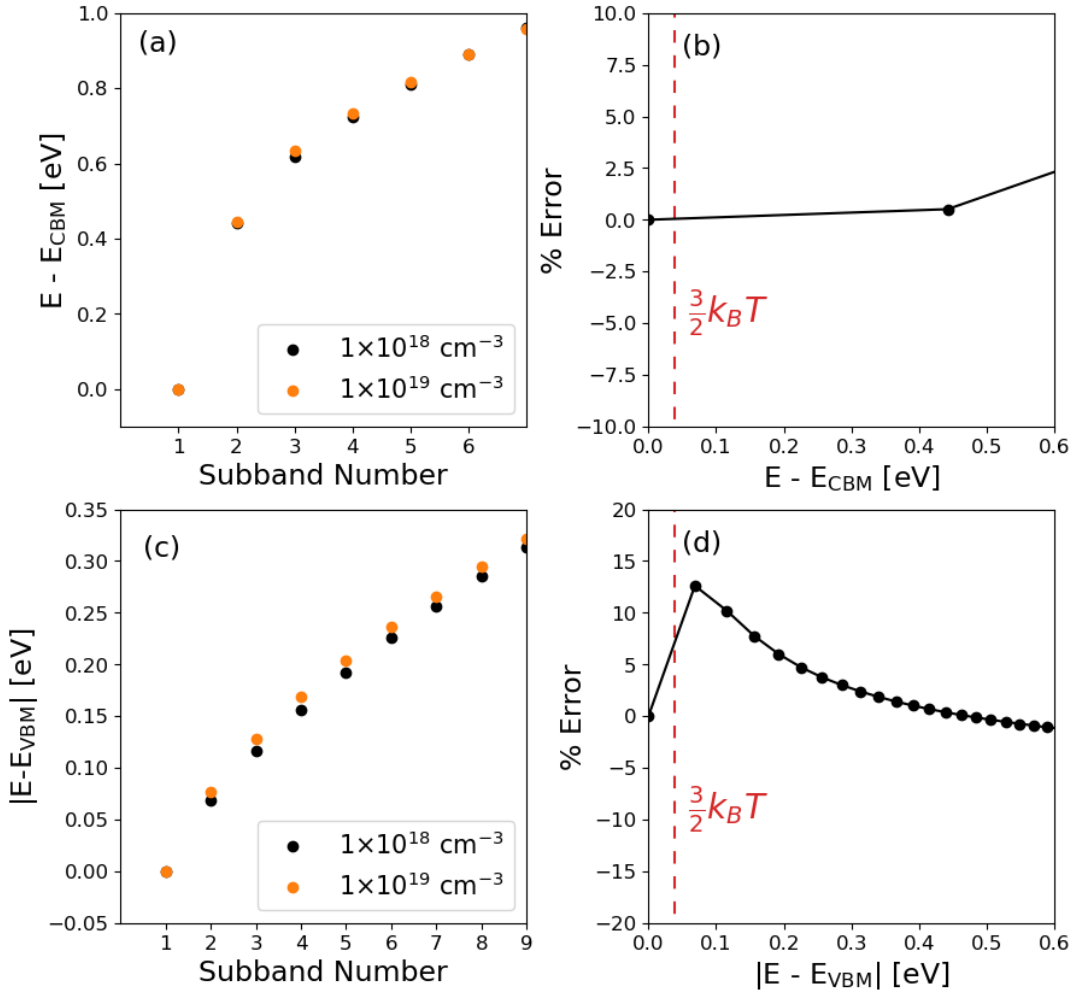

Figure 6.7: Evidence that polarization-charge screening and plasma renormalization lead predominantly to a rigid shift of the bands in the carrier-density range of interest for LED operation. Panel (a) compares the electron energy of the subbands in an InGaN quantum well with carrier densities of 10×18 cm^{-3} and 10^{19} cm^{-3}, and panel (b) shows the relative error accrued by assuming the conduction band is rigidly shifted due to screening effects. The error in the conduction band accrued by assuming a rigid shift of the bands is negligible. Panel (c) compares the hole energy of the subbands in an InGaN quantum well with carrier densities of 10^{18} cm^{-3} and 10^{19} cm^{-3}, and panel (d) shows the relative error accrued by assuming the valence band is rigidly shifted due to screening effects. The error in the valence band accrued by assuming a rigid shift of the bands is small; the largest error is for the first excited subband, however the error is small (less than 15%), which is further diminished by the fact that the thermal occupation of this subband is small.

second-order effect on the linewidth broadening since polarization fields modulate the B coefficient, which controls the carrier density required to operate the LED at a given current density ($J = Bn^2$). Weaker polarization fields, e.g., due to screening, increase the B coefficient, thus reducing the carrier density at a given current density, thereby lessening phase-space filling. This effect is still relatively small compared to first-order phase-space filling effects at the carrier densities that we have considered, and indeed would reduce the relative linewidth broadening as the carrier density increases. Moreover, in the carrier densities relevant for LED operation, many-body renormalization of the band structure can be treated, to first order, as a rigid shift of the band edges [36], thus it does not contribute strongly to linewidth broadening either. This is supported by our observation that the net result of polarization-charge screening and plasma renormalization is approximately a rigid shift of the bands (see Figure 6.7).Therefore, while the injection-dependence of the peak-emission energy is due to the interplay of various physical effects, the injection-dependent linewidth broadening is predominantly due to phase-space filling.

6.4 Which designs improve spectral characteristics?

6.4.1 Role of polarization field

One important question that remains to be answered is why III-nitride LEDs grown on polar planes suffer from more severe injection-dependent linewidth broadening than III-phosphide and semipolar/non-polar III-nitride LEDs even though phase-space filling is a universal phenomenon that is present in all materials. The answer is simply that polar III-nitride LEDs operate at higher carrier densities due to their weaker oscillator strengths and correspondingly smaller radiative recombination (B) coefficients [51], and are thus more susceptible to phase-space filling. In Figure 6.8, we show the carrier density required to operate 3 nm single-quantum-well LEDs at radiative current densities of 1A/cm^2, 50 A/cm^2, and 1000 A/cm^2 as a function of the B coefficient. We also show experimentally measured B coefficients for various (0001) polar [34] and ($20\overline{2}1$) semipolar [52] LEDs. Polar LEDs have low B coef-

ficients due to their strong polarization field, which separates electrons and holes to opposite sides of the quantum well and lowers the probability of recombination. The B coefficient of polar LEDs decreases with increasing emission wavelength (or indium content), therefore longer wavelength emitters undergo more severe injection-dependent spectral broadening. In contrast, semipolar LEDs have higher B coefficients due to their smaller polarization fields; consequently, they can operate at much lower carrier densities for a given current density. For this reason, semipolar LEDs exhibit less injection-dependent linewidth broadening than polar LEDs, a conclusion that is directly supported by optical measurements of semipolar LEDs in the literature [53, 54, 55, 56]. The B coefficient of III-phosphide LEDs tend to be even higher than semipolar III-nitride LEDs, with typical B coefficients of the order of $\sim 10^{-10}$ cm^3 s^{-1} [57]. In fact, such high radiative recombination coefficients mean that III-phosphide LEDs are more likely to experience stimulated emission before undergoing significant linewidth broadening, which may explain why luminescence broadening is typically not observed in the III-phosphide system. Our results also explain why some non-polar LEDs exhibit an (often small) injection-dependent blueshift and linewidth broadening despite the absence of a polarization field [21, 58, 59]. Because there is no quantum-confined Stark effect in non-polar LEDs, higher indium compositions are required to obtain a given wavelength. Carrier localization due to stronger alloy disorder reduces the density of states and lowers the B coefficient (if electrons and holes are not co-localized) [4, 60], which makes phase-space filling important in non-polar LEDs. Although our analysis has been for InGaN LEDs, it applies equally well to AlGaN quantum-well LEDs, which also have strong polarization fields [61] and carriers localized by alloy disorder [**?**]. Hence, we have shown that recombination coefficients, and in particular the B coefficient, are important parameters that determine the likelihood of a device undergoing phase-space filling and injection-dependent linewidth broadening.

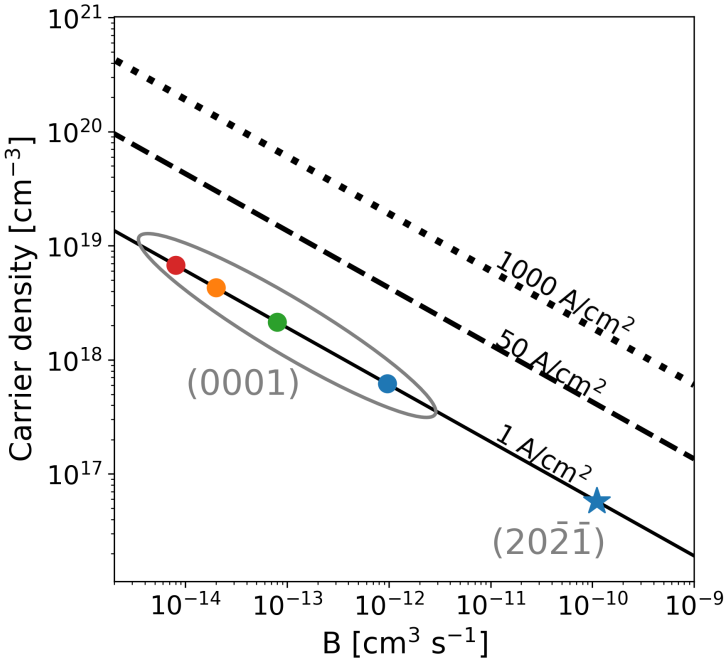

Figure 6.8: Effect of the B coefficient on the carrier density required to obtain a given radiative current density. The circles correspond to experimental B coefficients for polar "(0001)" LEDs measured by David et al. for blue (450 nm), green (535 nm), orange (600 nm), and red (645 nm) emitters. The star is the experimental B coefficient measured by Monavarian et al. for a semi-polar blue LED (430 nm). LEDs with lower B coefficients are more susceptible to phase-space filling, and consequently to stronger spectral broadening, because they operate at higher carrier densities for a given current density.

6.4.2 Importance of the steady-state carrier density

Our results demonstrate that device designs that reduce the carrier density required to operate the device at a given current density reduce the injection-dependent blueshift and linewidth broadening. Improving the inter-well hole transport and spreading the number of carriers over more quantum wells enables the same light-power output for a lower carrier density [38]. 3D engineering of the active region using V-pits has recently been shown to be a practical way of improving hole transport, as evidenced by state-of-the-art multi-quantum-well LEDs fabricated with 3D V-pit engineering that show improved efficiency droop as well as smaller wavelength shift and linewidth broadening compared to LEDs with poor inter-well hole transport [**?**]. Designs that minimize the polarization field, e.g., semi-polar, non-polar, and thinner polar LEDs, minimize the injection-dependent wavelength blueshift because they allow the device to be operated at a lower carrier density for a given current density. Such designs simultaneously reduce the injection-dependent linewidth broadening and reduce efficiency droop, albeit at the expense of also requiring higher indium concentrations, which may inadvertently lead to a broader linewidth at low carrier density. In contrast, inefficient designs with more defects also operate at lower carrier densities for a given current density due to their higher non-radiative recombination rates, and thus exhibit less linewidth broadening. In general, it is important to identify the origin of small injection-dependent linewidth broadening, particularly in devices that are more susceptible to defects, e.g., micro-LEDs, as it can be a reflection of their high non-radiative recombination rate, which is highly undesirable. We highlight that the designs that minimize efficiency droop by reducing the operating carrier density of LEDs also lead to better color purity.

6.4.3 Analysis of other designs

Some authors have suggested that adding indium to the barriers may enable better hole transport in multi-quantum-well LEDs [62]. Although such designs work well for blue emitters, it is unclear whether they work for green and longer wavelength emitters since adding indium to the barrier increases the emission energy. A similar

argument applies for doping the barriers to minimize the polarization fields in polar multi-quantum-well structures [63]. As discussed in the main text, thinner polar quantum wells are more desirable than thicker polar quantum wells due to their higher oscillator strengths and larger B coefficients, which allows them to operate at lower carrier densities for a given current density. However, higher indium concentrations are needed in thinner quantum wells to obtain a given wavelength, which may lead to a broader linewidth at low carrier density due to stronger alloy disorder. Moreover, polar quantum wells with thickness greater than 3 nm experience an additional net blueshift of the band gap for high carrier densities, due to an incomplete cancellation of the polarization-charge blueshift by the many-body redshift, therefore they are especially undesirable for applications that require good color purity. Conversely, thicker non-polar quantum wells are more desirable than thinner non-polar quantum wells since the current density achievable for a given carrier density scales linearly with the thickness if the B coefficient is held constant. Overall, any strategy that reduces the B coefficient by reducing the polarization field will inadvertently increase the indium concentration required to obtain a given wavelength. Therefore, such strategies will require a careful tradeoff between an increase of the B coefficient due to minimizing polarization fields versus stronger carrier localization and material degradation due to increasing indium concentrations.

6.5 Conclusion

In summary, we have calculated the carrier-density dependence of the emission spectrum of InGaN LEDs. In contrast to the widely accepted hypothesis that the injection-dependent emission blueshift in III-nitride LEDs is primarily due to polarization-charge screening, we have shown that the emission shift depends on a complex interplay between polarization-charge screening, exchange-correlation effects, and phase-space filling of carriers in the disordered potential landscape of the quantum well. We have also shown that the injection-dependent linewidth broadening is caused primarily by phase-space filling, which is exceptionally prominent in polar III-nitride quantum wells due to their weaker oscillator strengths and lower

radiative recombination coefficients. This emphasizes the innate connection between carrier dynamics and the current-dependent spectral characteristics of LEDs. Namely, emitters with poor transport and recombination dynamics offer poorer control over the injection-dependent color purity. Hence, designs that reduce the carrier density required to operate the LED at a given current density simultaneously reduce efficiency droop and improve the high-power color purity of III-nitride LEDs.

Bibliography

[1] S. Pimputkar, J.S. Speck, S.P. DenBaars, and S. Nakamura. Nature photon. 3:180, 2009.

[2] Y.C. Shen, G.O. Mueller, S. Watanabe, N.F. Gardner, A. Munkholm, and M.R. Krames. Appl. phys. lett. 91:141101, 2007.

[3] E. Kioupakis, P. Rinke, K.T. Delaney, and C.G. Van de Walle. Appl. phys. lett. 98:161107, 2011.

[4] S.Y. Karpov. Applied sciences. 8:818, 2018.

[5] A. David, N.G. Young, C. Lund, and M.D. Craven. Ecs j. solid state sci. technol. 9:016021, 2019.

[6] D.S.P. Tanner, P. Dawson, M.J. Kappers, R.A. Oliver, and S. Schulz. Phys. rev. applied. 13:044068, 2020.

[7] A. Hangleiter, J. s. Im, J. Off, and F. Scholz. Physica status solidi (b). 216:427, 1999.

[8] S.F. Chichibu, A.C. Abare, M.S. Minsky, S. Keller, S.B. Fleischer, J.E. Bowers, E. Hu, U.K. Mishra, L.A. Coldren, S.P. DenBaars, and T. Sota. Appl. phys. lett. 73:2006, 1998.

[9] R.A. Oliver, S.E. Bennett, T. Zhu, D.J. Beesley, M.J. Kappers, D.W. Saxey, A. Cerezo, and C.J. Humphreys. J. phys. d: Appl. phys. 43:354003, 2010.

[10] P. Vashishta and R.K. Kalia. Phys. rev. b. 25:6492, 1982.

[11] G. Tränkle, H. Leier, A. Forchel, H. Haug, C. Ell, and G. Weimann. Phys. rev. lett. 58:419, 1987.

[12] S. Das Sarma, R. Jalabert, and S.-R.E. Yang. Phys. rev. b. 41:8288, 1990.

[13] F. Caruso and F. Giustino. Phys. rev. b. 94:115208, 2016.

[14] T. Nagai, T.J. Inagaki, and Y. Kanemitsu. Appl. phys. lett. 84:1284, 2004.

[15] D. Hirano, T. Tayagaki, and Y. Kanemitsu. Phys. rev. b. 77:073201, 2008.

[16] A. David, N.G. Young, and M.D. Craven. Phys. rev. applied. 12:044059, 2019.

[17] E. Kuokstis, J.W. Yang, G. Simin, M.A. Khan, R. Gaska, and M.S. Shur. Appl. phys. lett. 80:977, 2002.

[18] N.I. Bochkareva, V.V. Voronenkov, R.I. Gorbunov, A.S. Zubrilov, P.E. Latyshev, Yu.S. Lelikov, Yu.T. Rebane, A.I. Tsyuk, and Yu.G. Shreter. Semiconductors. 46:1032, 2012.

[19] M.J. Davies, T.J. Badcock, P. Dawson, M.J. Kappers, R.A. Oliver, and C.J. Humphreys. Appl. phys. lett. 102:022106, 2013.

[20] M.J. Davies, T.J. Badcock, P. Dawson, R.A. Oliver, M.J. Kappers, and C.J. Humphreys. Physica status solidi c. 11:694, 2014.

[21] M.J. Davies, P. Dawson, S. Hammersley, T. Zhu, M.J. Kappers, C.J. Humphreys, and R.A. Oliver. Appl. phys. lett. 108:252101, 2016.

[22] G.M. Christian, S. Schulz, M.J. Kappers, C.J. Humphreys, R.A. Oliver, and P. Dawson. Phys. rev. b. 98:155301, 2018.

[23] F. Della Sala, A. Di Carlo, P. Lugli, F. Bernardini, V. Fiorentini, R. Scholz, and J.-M. Jancu. Appl. phys. lett. 74:2002, 1999.

[24] J. Wu, W. Walukiewicz, K.M. Yu, J.W. Ager, E.E. Haller, H. Lu, W.J. Schaff, Y. Saito, and Y. Nanishi. Appl. phys. lett. 80:3967, 2002.

[25] C.E. Dreyer, A. Janotti, C.G. Van de Walle, and D. Vanderbilt. Phys. rev. x. 6:021038, 2016.

[26] S. Birner, T. Zibold, T. Andlauer, T. Kubis, M. Sabathil, A. Trellakis, and P. Vogl. Ieee transactions on electron devices. 54:2137, 2007.

[27] A.F. Wright. Journal of applied physics. 82:2833, 1997.

[28] Q. Yan, P. Rinke, A. Janotti, M. Scheffler, and C.G. Van de Walle. Phys. rev. b. 90:125118, 2014.

[29] P.G. Moses, M. Miao, Q. Yan, and C.G. Van de Walle. J. chem. phys. 134:084703, 2011.

[30] P. Rinke, M. Winkelnkemper, A. Qteish, D. Bimberg, J. Neugebauer, and M. Scheffler. Phys. rev. b. 77:075202, 2008.

[31] I. Vurgaftman and J.R. Meyer. Journal of applied physics. 94:3675, 2003.

[32] T.-J. Yang, R. Shivaraman, J.S. Speck, and Y.-R. Wu. Journal of applied physics. 116:113104, 2014.

[33] C.M. Jones, C.-H. Teng, Q. Yan, P.-C. Ku, and E. Kioupakis. Applied physics letters. 111:113501, 2017.

[34] A. David, N.G. Young, C.A. Hurni, and M.D. Craven. Phys. rev. applied. 11:031001, 2019.

[35] C.E. Dreyer, A. Janotti, and C.G. Van de Walle. Appl. phys. lett. 102:142105, 2013.

[36] H. Haug and S. Schmitt-Rink. Progress in quantum electronics. 9:3, 1984.

[37] G. Tränkle, E. Lach, A. Forchel, F. Scholz, C. Ell, H. Haug, G. Weimann, G. Griffiths, H. Kroemer, and S. Subbanna. Phys. rev. b. 36:6712, 1987.

[38] C.-K. Li, C.-K. Wu, C.-C. Hsu, L.-S. Lu, H. Li, T.-C. Lu, and Y.-R. Wu. Aip advances. 6:055208, 2016.

[39] J.P. Perdew and Y. Wang. Phys. rev. b. 45:13244, 1992.

[40] D.M. Ceperley and B.J. Alder. Phys. rev. lett. 45:566, 1980.

[41] C.-K. Li, M. Piccardo, L.-S. Lu, S. Mayboroda, L. Martinelli, J. Peretti, J.S. Speck, C. Weisbuch, M. Filoche, and Y.-R. Wu. Phys. rev. b. 95:144206, 2017.

[42] M. O'Donovan, D. Chaudhuri, T. Streckenbach, P. Farrell, S. Schulz, and T. Koprucki. J. appl. phys. 130:065702, 2021.

[43] J.A. Gonzalez Montoya, A. Tibaldi, C. De Santi, M. Meneghini, M. Goano, and F. Bertazzi. Phys. rev. applied. 16:044023, 2021.

[44] A. David, M.J. Grundmann, J.F. Kaeding, N.F. Gardner, T.G. Mihopoulos, and M.R. Krames. Appl. phys. lett. 92:053502, 2008.

[45] C. Frankerl, F. Nippert, A. Gomez-Iglesias, M.P. Hoffmann, C. Brandl, H.-J. Lugauer, R. Zeisel, A. Hoffmann, and M.J. Davies. Appl. phys. lett. 117:102107, 2020.

[46] A. David, N.G. Young, C.A. Hurni, and M.D. Craven. Appl. phys. lett. 110:253504, 2017.

[47] A. Rashidi, M. Monavarian, A. Aragon, and D. Feezell. Appl. phys. lett. 113:031101, 2018.

[48] A. Rashidi, M. Monavarian, A. Aragon, and D. Feezell. Sci rep. 9:19921, 2019.

[49] D. Chaudhuri, M. O'Donovan, T. Streckenbach, O. Marquardt, P. Farrell, S.K. Patra, T. Koprucki, and S. Schulz. Journal of applied physics. 129:073104, 2021.

[50] R. Bhattacharya, B. Pal, and B. Bansal. Appl. phys. lett. 100:222103, 2012.

[51] E. Kioupakis, Q. Yan, and C.G. Van de Walle. Appl. phys. lett. 101:231107, 2012.

[52] M. Monavarian, A. Rashidi, A. Aragon, S.H. Oh, M. Nami, S.P. DenBaars, and D. Feezell. Opt. express, oe. 25:19343, 2017.

[53] Y. Zhao, S. Tanaka, C.-C. Pan, K. Fujito, D. Feezell, J.S. Speck, S.P. DenBaars, and S. Nakamura. Appl. phys. express. 4:082104, 2011.

[54] C.-C. Pan, S. Tanaka, F. Wu, Y. Zhao, J.S. Speck, S. Nakamura, S.P. DenBaars, and D. Feezell. Appl. phys. express. 5:062103, 2012.

[55] Y. Zhao, S.H. Oh, F. Wu, Y. Kawaguchi, S. Tanaka, K. Fujito, J.S. Speck, S.P. DenBaars, and S. Nakamura. Appl. phys. express. 6:062102, 2013.

[56] D.F. Feezell, J.S. Speck, S.P. DenBaars, and S. Nakamura. Journal of display technology. 9:190, 2013.

[57] O.A. Fedorova, K.A. Bulashevich, and S.Y. Karpov. Opt. express, oe. 29:35792, 2021.

[58] A. Chakraborty, B.A. Haskell, S. Keller, J.S. Speck, S.P. DenBaars, S. Nakamura, and U.K. Mishra. Appl. phys. lett. 85:5143, 2004.

[59] A. Chitnis, C. Chen, V. Adivarahan, M. Shatalov, E. Kuokstis, V. Mandavilli, J. Yang, and M.A. Khan. Appl. phys. lett. 84:3663, 2004.

[60] S. Schulz, M.A. Caro, C. Coughlan, and E.P. O'Reilly. Phys. rev. b. 91:035439, 2015.

[61] Q. Guo, R. Kirste, S. Mita, J. Tweedie, P. Reddy, S. Washiyama, M.H. Breckenridge, R. Collazo, and Z. Sitar. Jpn. j. appl. phys. 58:SCCC10, 2019.

[62] K.S. Qwah, M. Monavarian, W.Y. Ho, Y.-R. Wu, and J.S. Speck. Phys. rev. materials. 6:044602, 2022.

[63] N.G. Young, R.M. Farrell, S. Oh, M. Cantore, F. Wu, S. Nakamura, S.P. DenBaars, C. Weisbuch, and J.S. Speck. Appl. phys. lett. 108:061105, 2016.

CHAPTER VII

Mechanism for the Apparent Defect Tolerance of InGaN Emitters

The tolerance of InGaN emitters to defects is widely attributed to the suppression of diffusion by carrier localization. However, recent experiments have challenged this hypothesis by showing long diffusion lengths of up to ten microns at room temperature. Here, we examine the competition between radiative and Shockley-Read-Hall recombination in InGaN alloys. Without assuming that carrier diffusion is suppressed, we show that the interplay of carrier localization and polarization fields with carrier recombination enhances the quantum efficiency at low current densities, leading to an apparent defect tolerance. Our analysis demonstrates that decreasing the oscillator strength by promoting carrier localization or increasing the quantum-well thickness can enhance the quantum efficiency of light emitters for low-power applications, although it will exacerbate efficiency droop and impair the control of color purity at high operating powers.

7.1 Introduction

III-nitrides have greatly advanced solid-state lighting by enabling the invention of white light-emitting diodes (LED) [1, 2]. Although InGaN alloys have achieved remarkable success, a fundamental understanding of their performance remains incomplete, which has hindered the development of long-wavelength and UV emitters

[3]. For example, despite having dislocation densities greater than 10^{10} cm^{-2}, six orders of magnitude larger than those found in arsenide- or phosphide-based LEDs, early InGaN LEDs demonstrated bright luminescence [4]. Experiments showed that increasing the In mole fraction of InGaN emitters enhanced their luminescence intensity, leading to the interpretation that InGaN is tolerant to defects due to the localization of holes around In-N atomic condensates, which prevents carriers from diffusing to non-radiative centers [5, 6, 7, 8, 9]. However, recent experiments have cast doubt on the hypothesis that suppression of diffusion by localization is responsible for the defect tolerance of InGaN. These experiments have shown that diffusion lengths in InGaN alloys can reach tens of microns at room temperature [10]. At this temperature, a substantial fraction of carriers are completely extended, and even localized states can diffuse by coupling to lattice vibrations [11]. These inconsistencies demand a reexamination of the origin of defect tolerance in InGaN.

While the origin and even the existence of defect tolerance remains disputed [12, 13, 14, 15, 16, 17, 18], several mechanisms have been proposed that shed light on the issue. For instance, Hangleiter et al. proposed that V-shaped pits around threading dislocations in InGaN quantum wells produce an energetic barrier, which prevents carriers from reaching the dislocation center [19]. Although this explanation successfully accounts for the tolerance of InGaN quantum wells to high threading dislocation densities, it does not explain why increasing the In mole fraction increases the luminescence intensity of bulk films as well. Massabuau et al. proposed that phase segregation of cations in the vicinity of dislocations traps carriers away from dislocation centers [20]. However, this mechanism does not apply to point defects that occur away from dislocations, which account for a significant portion of non-radiative recombination.

7.1.1 Overview of this work

In this work, we develop a framework to evaluate the impact of carrier localization on non-radiative recombination by multi-phonon emission. Our analysis of the competition between radiative and non-radiative recombination suggests that carrier

localization limits the maximum achievable internal quantum efficient (IQE) and the IQE at a given carrier density. However, at low current densities, an interplay of carrier localization and recombination in InGaN alloys causes the IQE to increase with stronger carrier localization, leading to an apparent defect tolerance. Importantly, the mechanism that we are proposing does not invoke suppressed diffusion and is operational in both bulk alloys and in quantum wells. We propose that polarization fields in quantum wells have a similar effect in promoting radiative recombination over non-radiative recombination as well.

7.2 Methodology

7.2.1 Computational methods

To investigate the impact of carrier localization on the recombination rates, we used *nextnano++* to solve the Schrödinger and Poisson equations [21] and evaluated wavefunction overlaps using an in-house post-processing code. We used elastic constants [22], deformation potentials [23], polarization constants [24], and band gaps and offsets [25] determined from first-principles calculations based on density-functional theory. We employed the effective-mass approximation and used $m_e^* = 0.2$ and $m_h^* = 1.9$ for GaN, $m_e^* = 0.07$ and $m_h^* = 1.8$ for InN [26, 27]. We modeled the alloy disorder by randomly assigning the composition x in each grid site as either $x = 0$ or $x = 1$, and did not perform any further compositional averaging. We simulated supercells of size 18 nm $\times$ 18 nm $\times$ 18 nm with periodic boundary conditions, using a grid-size spacing of 0.3 nm in all directions, which approximately corresponds to the cation-cation distance in InGaN. Unless specified otherwise, we repeated every calculation for ten different configurations of the random alloy. Our results show that electrons are weakly localized while holes are strongly localized in the disordered potential landscape of InGaN alloys, consistent with atomistic tight-binding calculations [28]. We show the ground-state hole and electron wave functions of an $In_{0.15}Ga_{0.85}N$ alloy in Figure 7.1. The asymmetry in localization between electrons and holes, arising from the asymmetry of the effective masses in the III-nitrides ($m_e^* \approx 0.2m_0$ and $m_h^* \approx$

$1.8m_0$), has important implications for radiative and non-radiative recombination rates, which we discuss later.

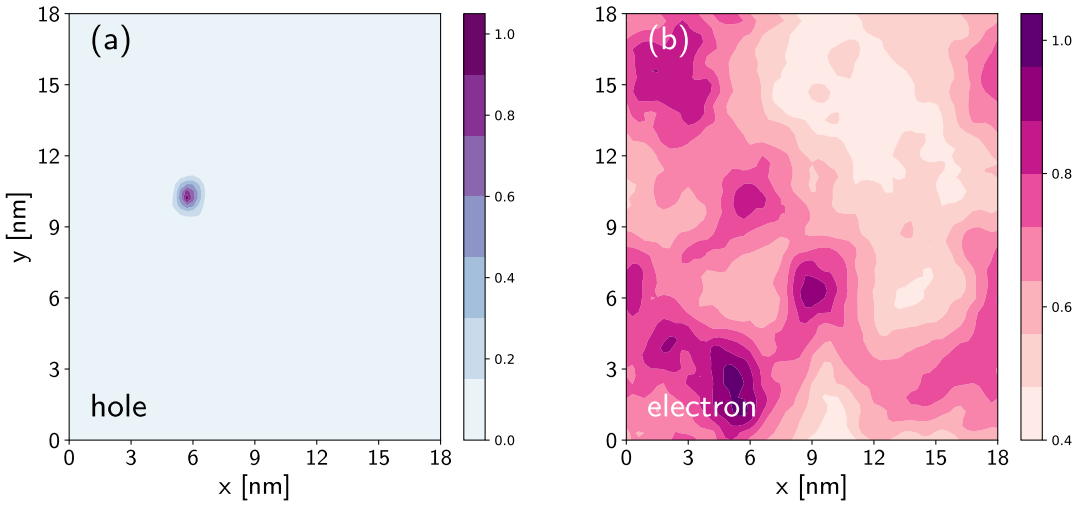

Figure 7.1: Squared modulus of the ground-state (a) hole and (b) electron envelope wave functions of an $In_{0.15}Ga_{0.85}N$ alloy, showing that holes are strongly localized with a characteristic length scale of ∼1 nm while electrons are extended. The wave functions are rescaled so that their peak value is one. We used a VB offset of 0.6 eV between GaN and InN, and a CB offset of 2.3 eV.

7.2.2 Recombination within k · p formalism

Central to our analysis is the relationship between recombination rates and $k \cdot p$ wave-function overlaps. The rate of radiative recombination is proportional to the squared overlap of the electron and hole wave functions $|F_{eh}|^2$ [1]:

$$\left|\tilde{F}_{eh}\right|^2 = \sum_{c\in CB, v\in VB} f_c(1-f_v)\left|\int dr\psi_c(r)\psi_v(r)\right|^2 \Big/ \sum_{c\in CB, v\in VB} f_c(1-f_v) \tag{7.1}$$

$$|F_{eh}|^2 = \left|\tilde{F}_{eh}\right|^2 \Big/ \left|\tilde{F}_{eh}^{VCA}\right|^2 \tag{7.2}$$

where f is the non-degenerate occupation probability at room temperature, and the indices c and v correspond to conduction band (CB) and valence band (VB)

states, respectively. Equation 7.2 rescales the overlap to ensure the limit $|F_{eh}|^2 = 1$ for systems with translational symmetry. $|F_{eh}|^2$ is proportional to the radiative recombination coefficient B, which relates to the rate of radiative recombination R_{rad} according to $R_{rad} = Bn^2$, where n is the carrier density. (In this work, we will interchangeably refer to $|F_{eh}|^2$ as the oscillator strength since it directly proportional to the oscillator strength up to some constant.) The rate of non-radiative recombination by multi-phonon emission (or Shockley-Read-Hall (SRH) recombination) is also proportional to an overlap term,

$$|F_{SRH}|^2 = V(1+\kappa)\int dr \frac{\delta n(r) \times \delta p(r)}{\delta n(r) + \kappa \delta p(r)}, \tag{7.3}$$

where $\kappa \equiv c_p/c_n$ is the ratio of a given defect's hole and electron capture coefficients, and $\delta n(r) \equiv n(r)/N$ and $\delta p(r) \equiv p(r)/N$ are the electron and hole densities divided by the total number of carriers in the simulation volume (N). To derive equation (7.3), we have assumed a random trap distribution, however we will later explicitly verify this assumption. The term $|F_{SRH}|^2$ is proportional to the SRH recombination coefficient A, which is related to the SRH rate according to $R_{SRH} = An$. It is straightforward to check that $|F_{SRH}|^2 = 1$ for systems with translational symmetry. We note that $|F_{SRH}|^2$ is completely independent of the details of the multi-phonon-emission physics, apart from the ratio of the defect capture coefficients κ. Very large or very small values of κ indicate that non-radiative recombination is limited by either electron or hole capture, meaning these two processes are not closely coupled in time. In contrast, values of κ near unity indicate that a defect captures electrons and holes successively and quickly, thus the probability of recombination depends strongly on the probability of both carriers being found at the defect site. The fact that $|F_{\mathrm{SRH}}|^2$ depends on the ratio κ rather than explicitly on c_p and c_n represents a convenient separation of physics by length scales. This allows us to evaluate the impact of localization on non-radiative recombination within $k \cdot p$ theory, without having to explicitly evaluate the capture coefficients at the level of density-functional theory. Thus, by studying how localization impacts $|F_{eh}|^2$ and $|F_{\mathrm{SRH}}|^2$, we can evaluate its influence on the radiative and SRH recombination rates.

7.2.3 Derivation of SRH overlap term

The rate of non-radiative recombination by multi-phonon emission (or Shockley-Read-Hall (SRH) recombination) is also proportional to an overlap term, which we term $|F_{SRH}|^2$. To derive this term, we start from the generalized SRH recombination rate that accounts for non-uniformities in the charge density,

$$R_{SRH} = n_T \int dr \frac{c_n n(r) \times c_p p(r)}{c_n n(r) + c_p p(r)}, \quad (S1) \tag{7.4}$$

where c_n and c_p are any given defect's electron and hole capture coefficients, n and p are the macroscopic electron and hole carrier density, and n_T is the trap density. For now, we have assumed that traps are uniformly distributed, however we will later explicitly verify our conclusions by relaxing this assumption. Under symmetric injection of electrons and holes, i.e., $\int dr, n(r) = \int dr, p(r) = N$, where N is the total number of carriers, we can rewrite the electron and hole densities as $n(r) = N\delta n(r)$ and $p(r) = N\delta p(r)$, where the spatially varying parts are given by $\delta n(r) \equiv \sum_{c\in CB} |\psi_c(r)|^2 f_c / \sum_{c\in CB} f_c$ and $\delta p(r) \equiv \sum_{v\in VB} |\psi_v(r)|^2 (1 - f_v) / \sum_{v\in VB} (1 - f_v)$. Factoring out c_n, we can rewrite the SRH rate as,

$$R_{SRH} = n_T N c_n \int dr \frac{\delta n(r) \times \delta p(r)}{\delta n(r) + (c_p/c_n)\delta p(r)}, \quad (S2) \tag{7.5}$$

For a system with translational symmetry, the macroscopic charge density is spatially uniform and $\delta n(r) = \delta p(r) = 1/V$, thus $R_{SRH}^{VCA} = n_T (N/V) c_n / (1 + c_p/c_n)$. Since we have defined $|F_{SRH}|^2$ to be the correction factor to the SRH recombination rate due to carrier localization effects, we obtain,

$$|F_{SRH}|^2 = \frac{R_{SRH}}{R_{SRH}^{VCA}} = V(1 + c_p/c_n) \int dr \frac{\delta n(r) \times \delta p(r)}{\delta n(r) + (c_p/c_n)\delta p(r)}, \quad (S3) \tag{7.6}$$

This is the expression that we provide in equation (3) of the main text, where we additionally defined $\kappa \equiv c_p/c_n$.

7.3 Impact of localization on recombination

7.3.1 Radiative recombination

We investigated the influence of hole localization on recombination by controlling the VB offset between InN and GaN. The VB offset has been previously determined experimentally to be between 0.5 eV and 1.1 eV [29, 30, 31, 32, 33, 34, 35], which agrees with the prediction of 0.6 eV from hybrid DFT [25]. We varied the VB offset from 0.0 eV to 1.0 eV, while fixing the CB offset to the theoretical value of 2.3 eV. To accurately quantify the degree of localization, we used the thermally averaged participation ratio, which measures the number of sites that the wave function "participates" in. We define the participation ratio as, $\left(\int dr\psi^2(r)\right)^2 / \int dr\psi^4(r)$, where ψ is the wave function. Generally, a smaller participation ratio indicates a more strongly localized state. Our results show that electrons are extended within the simulation cell but holes are strongly localized (Figure 7.2(a)). Moreover, larger VB offsets between GaN and InN lead to more strongly localized hole wave functions (Figure 7.2(b)).

Our findings demonstrate that strong hole localization decreases $|F_{eh}|^2$ in the III-nitrides due to the asymmetry of the carrier effective masses. Figure 7.3 illustrates how $|F_{eh}|^2$ changes with hole localization for $In_{0.15}Ga_{0.85}N$ alloys. Breaking translational symmetry by alloying initially increases $|F_{eh}|^2$ because of the relaxation of crystal-momentum conservation, explaining why $|F_{eh}|^2$ is larger than 1 for small ΔE_V (even if $\Delta E_V = 0$, strain and polarization fluctuations are sufficient to weakly localize holes). However, as holes become more localized, their spectral weight is transferred away from the Γ point, reducing their coupling strength with extended electrons whose spectral weight is highly concentrated near Γ. This result stands in contrast to our previous work that showed that carrier localization increases the B coefficient [36]. The previous work employed an averaging procedure to obtain local In composition, which resulted in coarser spatial resolution compared to atomistic models, and led to the incorrect prediction that electrons are also localized. In this work, we did not perform averaging and our localization results are closer to tight-

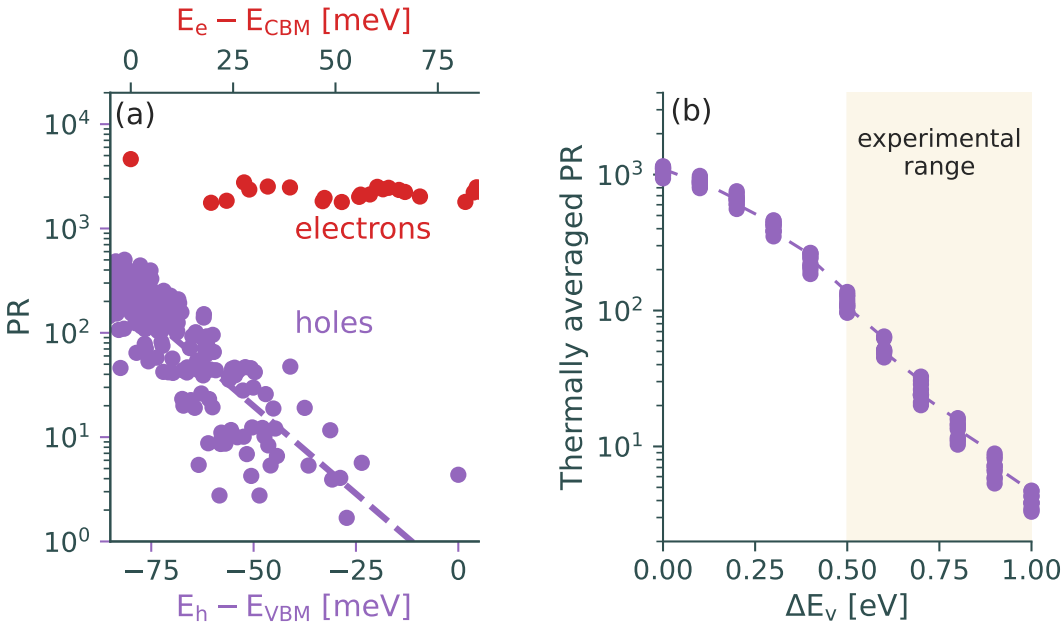

Figure 7.2: (a) Participation ratio of electron and hole wave functions in an $In_{0.15}Ga_{0.85}N$ alloy as a function of their energy with respect to the band edge. A smaller participation ratio indicates a more strongly localized wave function. (b) Larger VB offsets between InN and GaN lead to more strongly localized holes in InGaN. Experimental VB offsets range from 0.5 eV to 1.0 eV.

binding calculations [37]. In Figure 7.4, we show that artificially localizing electrons alongside holes increases the wave-function overlap. Therefore, the asymmetry of effective masses in the III-nitrides means that carrier localization reduces the rate of radiative recombination by reducing the spectral overlap of electron and hole wave functions.

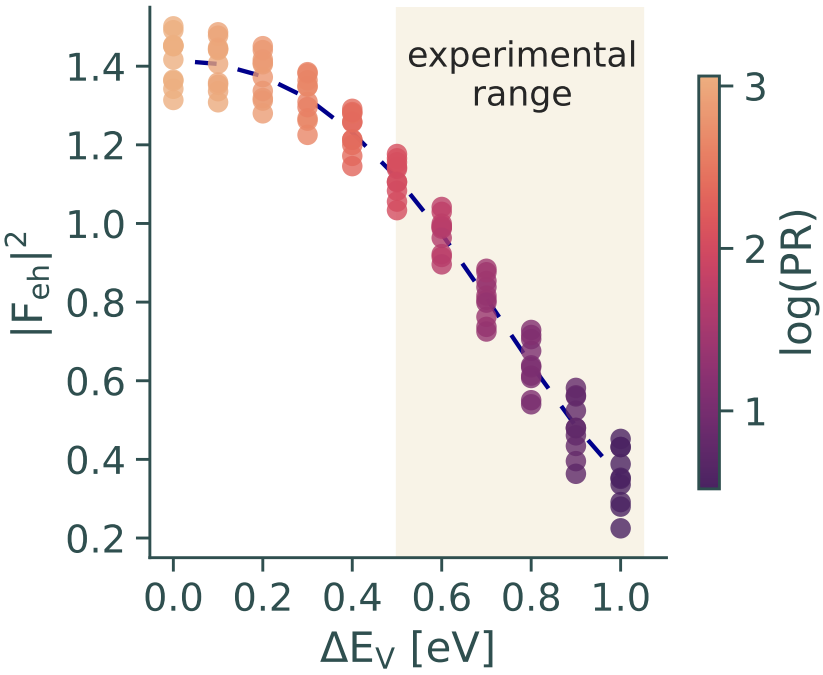

Figure 7.3: Strong hole localization in the absence of strong electron co-localization reduces the wave-function overlap in III-nitride alloys, thus reducing the rate of radiative recombination. The colors indicate the thermally averaged participation ratio of the hole wave functions; darker colors correspond to stronger localization. The shaded region shows the range of experimentally measured values of the InN/GaN VB offset. The CB offset is fixed at the natural value of 2.3 eV predicted by hybrid DFT.

7.3.2 Non-radiative recombination

On the other hand, the effect of carrier localization on non-radiative recombination depends on the κ of the defect over which recombination occurs. In Figure 7.5,

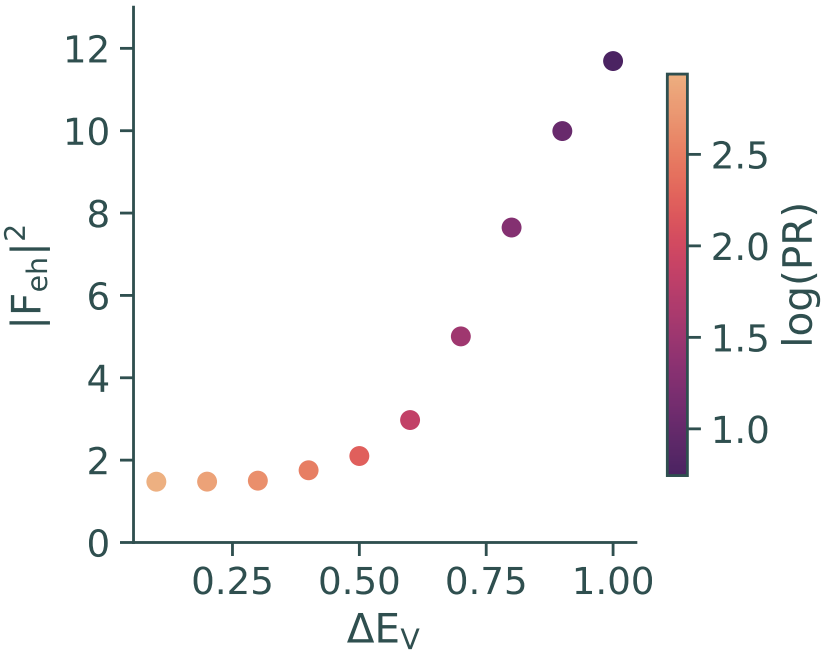

Figure 7.4: Strong hole localization in the presence of strong electron localization increases the wave-function overlap, thus increasing the rate of radiative recombination. The colors indicate the thermally averaged participation ratio of the hole wave function; darker colors correspond to stronger localization. We varied the CB offset alongside the VB offset according to the formula, $\Delta E_c = \Delta E_v(m_h^*/m_e^*)$, thus localizing electrons as well.

we show the dependence of $|F_{SRH}|^2$ on the participation ratio for different values of κ. For κ close to unity, hole localization reduces the SRH overlap by reducing the probability of finding an electron and hole at a defect site. In contrast, hole localization has no effect on the SRH overlap for extreme values of κ. This is because the SRH cycle is limited by multi-phonon emission rather than hole localization since the probability of finding a hole at a defect site is always finite due to the thermal occupation of extended states. Since defects with symmetric capture coefficients are typically the most efficient non-radiative centers, the overall effect of carrier localization is to reduce the rate of SRH recombination.

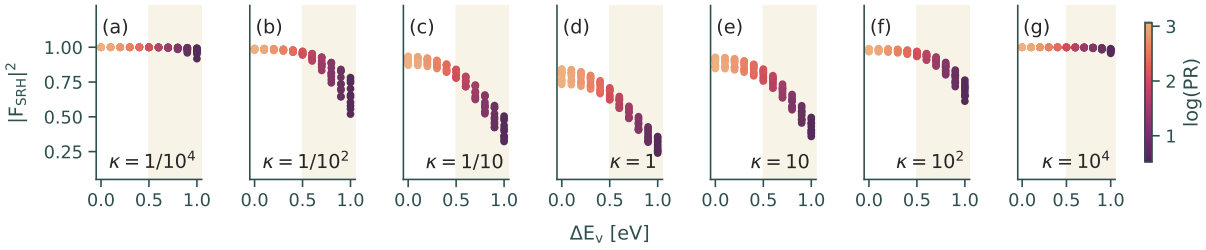

Figure 7.5: Increasing hole localization with increasing valence-band offset reduces the SRH wave-function overlap for recombination over defects with symmetric electron and hole capture coefficients ($\kappa \sim 1$). However, localization has little to no effect on recombination over defects with asymmetric capture coefficients ($\kappa \ll 1, \kappa \gg 1$). The colors indicate the thermally averaged participation ratio of the hole wave functions; darker colors correspond to stronger localization. The shaded region shows the range of experimentally measured values of the InN/GaN VB offset. The CB offset is fixed at the natural value of 2.3 eV.

7.3.3 Power-law scaling of recombination

The efficiency ultimately depends on the ratio of the radiative and non-radiative recombination rates. To capture the qualitative features of interest, we use a simple ABC model that accounts for wave-function overlap. For the current regime where SRH recombination dominates, we can neglect the C term corresponding to third-

order Auger-Meitner recombination [38]. The resulting expression is given by:

$$IQE = \left(\left| \frac{F_{SRH}}{F_{eh}} \right|^2 \frac{A_0}{B_0 n} + 1 \right)^{-1} = \left(s(\kappa) \left| F_{eh} \right|^{2p(\kappa)-2} \frac{A_0}{B_0 n} + 1 \right)^{-1}. \quad (7.7)$$

Here, A_0 and B_0 are the bulk virtual-crystal recombination coefficients, and $|F_{SRH}|^2$ and $|F_{eh}|^2$ introduce corrections due to localization (or polarization fields, if applicable). We simplified the IQE expression by expressing the SRH overlap in terms of the electron-hole overlap as a power law, $|F_{SRH}|^2 = s(\kappa) |F_{eh}|^{2p(\kappa)}$, where s and p are functions of κ. In Figure 7.6, we show this scaling relation explicitly for various values of κ. In 7.7, we show how the scaling constant s and power p depend on κ. By fitting $s(\kappa)$ and $p(\kappa)$ to Gaussian functions, we find the following approximate expressions: $s(\kappa) \approx -0.67 \exp\left(-\frac{\log_{10}^2(\kappa)}{4.0}\right) + 1$ and $p(\kappa) \approx 0.71 \exp\left(-\frac{\log_{10}^2(\kappa)}{4.6}\right)$. If κ is close to unity, the non-radiative rate is proportional to the probability of finding both an electron and hole at a defect site; therefore, $|F_{SRH}|^2$ is proportional to $|F_{eh}|^2$. In contrast, $|F_{SRH}|^2$ is independent of $|F_{eh}|^2$ for extremely large or extremely small values of κ because the non-radiative rate is limited by multi-phonon capture of one carrier. The power-law scaling relation between radiative and non-radiative recombination, which we have shown for the case of varying hole localization, appears to be a general feature of recombination in the III-nitrides. Other authors have observed it for varying quantum-well thickness and composition, showing that polarization fields have a similar effect [39, 40, 41].

7.3.3.1 Competition with radiative capture by defects

In the main text, we showed that $|F_{eh}|^2$ and $|F_{SRH}|^2$ are almost linearly correlated for $\kappa \sim 1$, and completely uncorrelated for $\kappa \gg 1$ and $\kappa \ll 1$. We also showed that the near-linear correlation between $|F_{eh}|^2$ and $|F_{SRH}|^2$ leads to an apparent defect tolerance at low current densities. In the extreme limits of κ, multi-phonon emission becomes prohibitively slow, in which case the dominant monomolecular recombination process that competes with band-to-band radiative recombination is radiative capture by point defects, such as the C_N impurity in GaN. This modifies

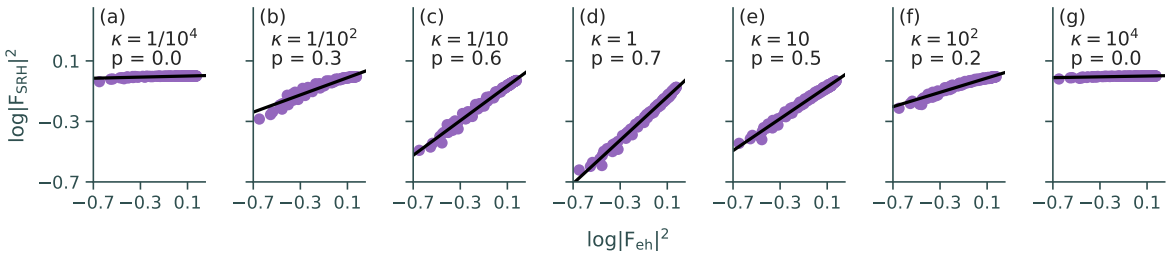

Figure 7.6: The wave-function overlap for SRH recombination is related to the electron-hole wave-function overlap probability as a power law of the form $|F_{SRH}|^2 \propto |F_{eh}|^{2p}$. The scaling exponent p (slope in log-log plot) depends on the κ of the defect over which recombination occurs. Each panel corresponds to a different value of κ. The SRH overlap and radiative overlap are strongly correlated for κ close to one.

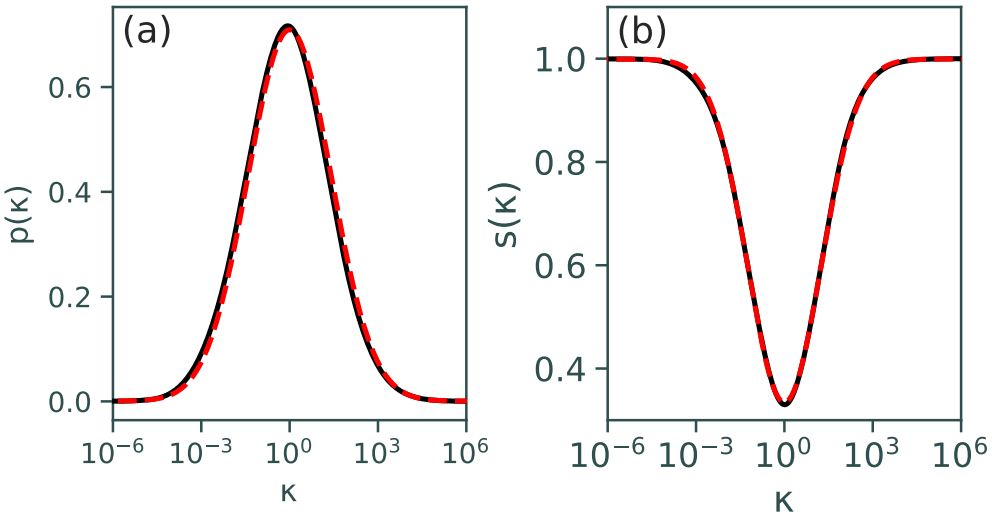

Figure 7.7: Figure S3. The (a) scaling power p and (b) scaling exponent s as a function of the capture-coefficient ratio κ, in the power law relation $|F_{eh}|^2 \propto s(\kappa)|F_{eh}|^{2p(\kappa)}$. The red dashed curves show Gaussian fits, with the fitting expressions provided in the main text.

the ABC model as,

$$IQE = \frac{Bn^2}{B'n + Bn^2 + Cn^3}, \tag{7.8}$$

where B' is the radiative-capture recombination coefficient. Similar to the B coefficient, the B' coefficient is also proportional to the electron-hole overlap $|F_{eh}|^2$ since it depends on the momentum-matrix element, $|p_i f|^2$. Therefore, we expect to observe apparent defect tolerance (at low current densities), even if band-to-band recombination competes with radiative capture by point defects rather than multi-phonon emission.

7.3.4 Quantum efficiency and defect tolerance

We propose that the defect tolerance of InGaN can be explained by the interplay of recombination dynamics with the increase in carrier density due to a reduction in the wave-function overlaps. To evaluate the IQE expression in equation (7.9), we varied $|F_{eh}|^2$ from 0.1 to 1.5, and used $B_0 = 6 \times 10^{-11}$ cm^3 s^{-1} and $A_0 = 10^7$ s^{-1}. Since defects with symmetric capture coefficients tend to be the most active recombination centers, we assumed $\kappa = 1$, but later we explicitly verify our conclusions with experimentally measured recombination coefficients, including the C coefficient. Figure 7.8a shows that carrier localization decreases the IQE at a given carrier density since $|F_{eh}|^2$ decreases more rapidly than $|F_{SRH}|^2$ as holes become more localized. Despite this decrease, carrier localization increases the IQE at a given current density, as seen in Figure 7.8b. By reducing the wave-function overlaps, carrier localization increases the carrier density required to operate an LED at a given current density, which in turn increases the relative rate of radiative recombination compared to SRH recombination. This is because SRH recombination scales linearly with the carrier density while radiative recombination scales quadratically. Therefore, although carrier localization decreases the quantum efficiency at a given carrier density, it increases the quantum efficiency at a given current density, which is the relevant quantity for experimental measurements. As we discuss later in the text, this is the opposite of what occurs at higher current densities in the efficiency-droop regime.

$$\text{IQE} = \left(\frac{|F_{SRH}|^2}{|F_{eh}|^2} \frac{A_0}{B_0 n} + 1 \right)^{-1} = \left(s(\kappa) |F_{eh}|^{2p(\kappa)-2} \frac{A_0}{B_0 n} + 1 \right)^{-1} \tag{7.9}$$

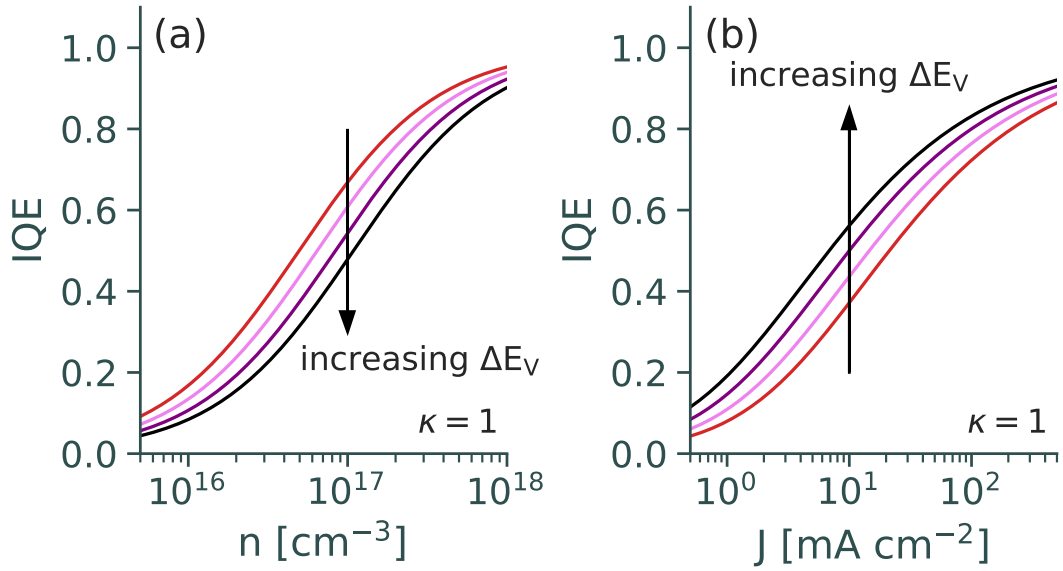

Figure 7.8: (The IQE as a function of the carrier density versus current density can be expressed using the scaling relation between $|F_{SRH}|^2$ and $|F_{eh}|^2$ that we calculated for $\kappa = 1$. Specifically, (a) stronger carrier localization due to larger valence-band offsets decreases $|F_{eh}|^2$ more quickly than $|F_{SRH}|^2$, thus reducing the IQE at a given carrier density; (b) however, at a fixed current density, carrier localization increases the IQE by increasing the carrier density n required to obtain a given current density J, which promotes radiative recombination over SRH recombination.

We now show that our proposed mechanism applies to commercial LEDs as well, despite the presence of additional factors such as carrier separation by polarization fields in quantum wells, Auger-Meitner recombination, and various types of defects contributing to non-radiative recombination [40]. To account for these factors in our analysis, we use the empirical scaling relation between the A, B, and C coefficients observed by David et al. for commercial LEDs of various thicknesses and compositions [40]. In Figure 7.9, we present the IQE calculated using a simple ABC model as a function of both carrier and current density, for B coefficients ranging from 10^{-13} cm^3 s^{-1} to 10^{-11} cm^3 s^{-1}. We observe that the IQE at a given carrier decreases as the B coefficient decreases, while the IQE at a given current density has the opposite behavior. This behavior is qualitatively similar to that observed in

Figure **??**, indicating that our proposed mechanism contributes to defect tolerance in commercial LEDs at low current densities.

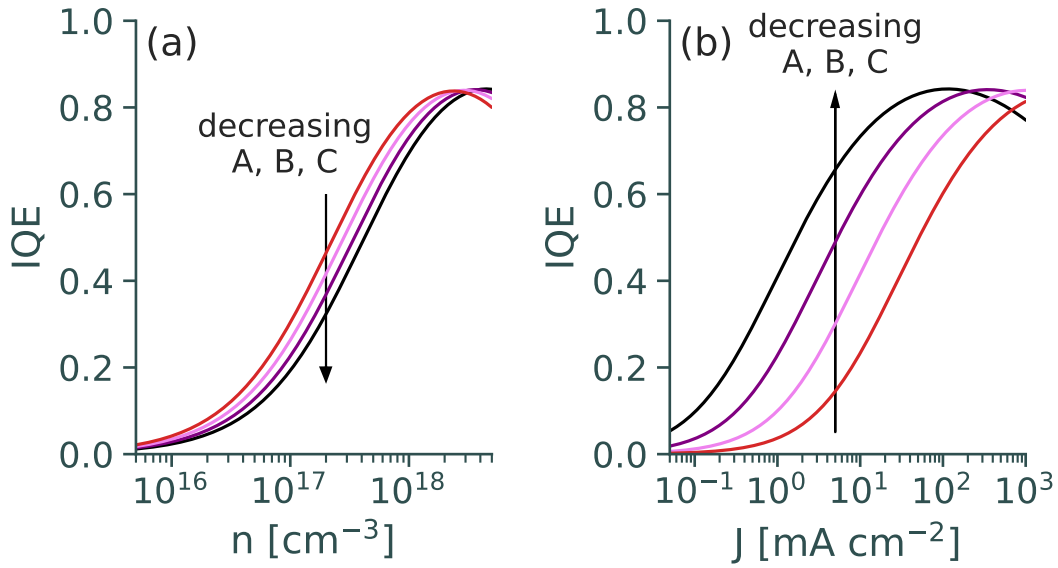

Figure 7.9: The IQE as a function of the carrier density versus current density, using the empirical scaling relation between the A, B, and C recombination coefficients measured by David et al. (a) Carrier localization and polarization fields decrease B more quickly than A , thus reducing the IQE at a given carrier density. (b) However, at a fixed current density, slower recombination dynamics increases the IQE by increasing the carrier density n required to obtain a given current density J, which promotes radiative recombination over SRH recombination.

We should note that although mechanisms that reduce the oscillator strength can enhance the IQE at low current densities, they are ultimately detrimental to high-power device performance. Slow recombination dynamics due to weak carrier overlaps increases the carrier density, which promotes bimolecular radiative recombination over monomolecular non-radiative recombination. However, it's important to recognize that this is only an apparent defect tolerance since reducing the carrier overlap overall decreases the IQE at a given carrier density. Our conclusions that InGaN alloys do not exhibit tolerance to point defects for fixed carrier density is consistent with the observation that In-containing underlayers improve the IQE by reducing the point-

defect density [16, 42]. At higher current densities, the same mechanism promotes third-order Auger-Meitner recombination over bimolecular radiative recombination [43], shifts the maximum efficiency to lower current densities, and impairs the ability to control the color purity [44]. Nevertheless, there may be some niche applications where reducing the oscillator strength, e.g., by promoting carrier localization or increasing the (polar) quantum-well thickness, could be advantageous for improving the low-current efficiency of LEDs, such as low-power micro-LEDs.

7.4 Conclusion

In summary, we have developed a method for calculating the impact of carrier localization on Shockley-Read-Hall recombination within the envelope-function approximation. Our proposed mechanism for the defect tolerance observed in InGaN LEDs does not invoke the suppression of carrier diffusion. Our analysis shows that reducing the carrier wave-function overlap can enhance the IQE at low current densities by increasing the carrier density, leading to an apparent defect tolerance. However, this comes at the expense of exacerbated efficiency droop at higher current densities and poor control of color purity. Nevertheless, reducing the oscillator strength by carrier localization or polarization fields may be beneficial for improving the low-current efficiency of LEDs for some niche low-power applications. Our proposed mechanism provides a theoretical framework for understanding the defect tolerance observed in InGaN LEDs and may guide the design of future high-performance devices.

Bibliography

[1] S. Nakamura, T. Mukai, and M. Senoh. Candela-class high-brightness InGaN/AlGaN double-heterostructure blue-light-emitting diodes. *Appl. Phys. Lett.*, 64:1687, 1994.

[2] S. Pimputkar, J. S. Speck, S. P. DenBaars, and S. Nakamura. Prospects for LED lighting. *Nature Photon*, 3:180–182, 2009.

[3] C. Weisbuch, S. Nakamura, Y.-R. Wu, and J. S. Speck. Disorder effects in nitride

semiconductors: Impact on fundamental and device properties. *Nanophotonics*, 10:565–594, 2021.

[4] S. F. Chichibu, A. Uedono, T. Onuma, T. Sota, B. A. Haskell, S. P. DenBaars, J. S. Speck, and S. Nakamura. Limiting factors of room-temperature nonradiative photoluminescence lifetime in polar and nonpolar gan studied by time-resolved photoluminescence and slow positron annihilation techniques. *Appl. Phys. Lett.*, 86:021914, 2005.

[5] S. F. Chichibu, T. Azuhata, T. Sota, and S. Nakamura. Origin of defect-insensitive emission probability in in-containing (Al,In,Ga)n alloy semiconductors. *Nature Mater*, 5:810–816, 2006.

[6] S. F. Chichibu. Review—defect-tolerant luminescent properties of low InN mole fraction InxGa1-XN quantum wells under the presence of polarization fields. *ECS J. Solid State Sci. Technol.*, 9(1):015016, 2019.

[7] S. Chichibu, K. Wada, and S. Nakamura. Spatially resolved cathodoluminescence spectra of ingan quantum wells. *Appl. Phys. Lett.*, 71(16):2346, 1997.

[8] K. P. O'Donnell, R. W. Martin, and P. G. Middleton. Origin of luminescence from InGaN diodes. *Phys. Rev. Lett.*, 82(2):237, 1999.

[9] Y. Narukawa, Y. Kawakami, M. Funato, S. Fujita, S. Fujita, and S. Nakamura. Role of self-formed InGaN quantum dots for exciton localization in the purple laser diode emitting at 420 nm. *Appl. Phys. Lett.*, 70(8):981, 1997.

[10] A. David. Long-range carrier diffusion in (In,Ga)N quantum wells and implications from fundamentals to devices. *Phys. Rev. Applied*, 15(5):054015, 2021.

[11] Jasprit Singh. *Electronic and Optoelectronic Properties of Semiconductor Structures*. Cambridge University Press, 2007.

[12] M. F. Schubert, S. Chhajed, J. K. Kim, E. F. Schubert, D. D. Koleske, M. H. Crawford, S. R. Lee, A. J. Fischer, G. Thaler, and M. A. Banas. Effect of

dislocation density on efficiency droop in GaInN/GaN light-emitting diodes. *Appl. Phys. Lett.*, 91:231114, 2007.

[13] J. Abell and T. D. Moustakas. The role of dislocations as nonradiative recombination centers in InGaN quantum wells. *Appl. Phys. Lett.*, 92:091901, 2008.

[14] B. Monemar and B. E. Sernelius. Defect related issues in the "current roll-off" in InGaN based light emitting diodes. *Appl. Phys. Lett.*, 91:181103, 2007.

[15] A. Armstrong, T. A. Henry, D. D. Koleske, M. H. Crawford, K. R. Westlake, and S. R. Lee. Dependence of radiative efficiency and deep level defect incorporation on threading dislocation density for InGaN/GaN light emitting diodes. *Appl. Phys. Lett.*, 101:162102, 2012.

[16] C. Haller, J.-F. Carlin, G. Jacopin, D. Martin, R. Butté, and N. Grandjean. Burying non-radiative defects in InGaN underlayer to increase InGaN/GaN quantum well efficiency. *Appl. Phys. Lett.*, 111:262101, 2017.

[17] T. Langer, A. Kruse, F. A. Ketzer, A. Schwiegel, L. Hoffmann, H. Jönen, H. Bremers, U. Rossow, and A. Hangleiter. Origin of the "green gap": Increasing nonradiative recombination in indium-rich GaInN/GaN quantum well structures. *Physica Status Solidi c*, 8:2170, 2011.

[18] T. J. Badcock, S. Hammersley, D. Watson-Parris, P. Dawson, M. J. Godfrey, M. J. Kappers, C. McAleese, R. A. Oliver, and C. J. Humphreys. Carrier density dependent localization and consequences for efficiency droop in InGaN/GaN quantum well structures. *Jpn. J. Appl. Phys.*, 52:08JK10, 2013.

[19] A. Hangleiter, F. Hitzel, C. Netzel, D. Fuhrmann, U. Rossow, G. Ade, and P. Hinze. Suppression of nonradiative recombination by v-shaped pits in GaInN/GaN quantum wells produces a large increase in the light emission efficiency. *Phys. Rev. Lett.*, 95:127402, 2005.

[20] F. C.-P. Massabuau et al. Carrier localization in the vicinity of dislocations in InGaN. *J. Appl. Phys.*, 121:013104, 2017.

[21] S. Birner, T. Zibold, T. Andlauer, T. Kubis, M. Sabathil, A. Trellakis, and P. Vogl. Nextnano: General purpose 3-d simulations. *IEEE Transactions on Electron Devices*, 54(9):2137–2142, 2007.

[22] A. F. Wright. Elastic properties of zinc-blende and wurtzite aln, gan, and inn. *Journal of Applied Physics*, 82(6):2833–2839, 1997.

[23] Q. Yan, P. Rinke, A. Janotti, M. Scheffler, and C. G. Van de Walle. Effects of strain on the band structure of group-iii nitrides. *Physical Review B*, 90(12):125118, 2014.

[24] C. E. Dreyer, A. Janotti, C. G. Van de Walle, and D. Vanderbilt. Correct implementation of polarization constants in wurtzite materials and impact on iii-nitrides. *Physical Review X*, 6(2):021038, 2016.

[25] P. G. Moses, M. Miao, Q. Yan, and C. G. Van de Walle. Hybrid functional investigations of band gaps and band alignments for aln, gan, inn, and ingan. *The Journal of Chemical Physics*, 134(8):084703, 2011.

[26] P. Rinke, M. Winkelnkemper, A. Qteish, D. Bimberg, J. Neugebauer, and M. Scheffler. Consistent set of band parameters for the group-iii nitrides aln, gan, and inn. *Phys. Rev. B*, 77:075202, 2008.

[27] C. E. Dreyer, A. Janotti, and C. G. Van de Walle. Effects of strain on the electron effective mass in gan and aln. *Appl. Phys. Lett.*, 102:142105, 2013.

[28] S. Schulz, M. A. Caro, C. Coughlan, and E. P. O'Reilly. Atomistic analysis of the impact of alloy and well-width fluctuations on the electronic and optical properties of ingan/gan quantum wells. *Phys. Rev. B*, 91:035439, 2015.

[29] G. Martin, A. Botchkarev, A. Rockett, and H. Morkoc. Valence-band discontinuities of wurtzite gan, aln, and inn heterojunctions measured by x-ray photoemission spectroscopy. *Applied Physics Letters*, 68(18):2541, 1996.

[30] Cheng-Fu Shih, Ning-Cheng Chen, Ping-Hsing Chang, and Kuo-Shen Liu. Band

offsets of inn/gan interface. *Japanese Journal of Applied Physics*, 44(11):7892, 2005.

[31] Takayuki Ohashi, Petter Holmstrom, Akira Kikuchi, and Katsumi Kishino. High structural quality inn/$in_{0.75}ga_{0.25}n$ multiple quantum wells grown by molecular beam epitaxy. *Applied Physics Letters*, 89(4):041907, 2006.

[32] C.-L. Wu, H.-M. Lee, C.-T. Kuo, C.-H. Chen, and S. Gwo. Cross-sectional scanning photoelectron microscopy and spectroscopy of wurtzite InN/GaN heterojunction: Measurement of "intrinsic" band lineup. *Appl. Phys. Lett.*, 92(16):162106, 2008.

[33] Z. H. Mahmood, A. P. Shah, A. Kadir, M. R. Gokhale, S. Ghosh, A. Bhattacharya, and B. M. Arora. Determination of InN-GaN heterostructure band offsets from internal photoemission measurements. *Appl. Phys. Lett.*, 91(15):152108, 2007.

[34] C.-L. Wu, H.-M. Lee, C.-T. Kuo, S. Gwo, and C.-H. Hsu. Polarization-induced valence-band alignments at cation- and anion-polar InN/GaN heterojunctions. *Appl. Phys. Lett.*, 91(4):042112, 2007.

[35] K. (Albert) Wang, C. Lian, N. Su, D. Jena, and J. Timler. Conduction band offset at the InN/GaN heterojunction. *Appl. Phys. Lett.*, 91(23):232117, 2007.

[36] C. M. Jones, C.-H. Teng, Q. Yan, P.-C. Ku, and E. Kioupakis. Impact of carrier localization on recombination in ingan quantum wells and the efficiency of nitride light-emitting diodes: Insights from theory and numerical simulations. *Appl. Phys. Lett.*, 111:113501, 2017.

[37] J. M. McMahon, D. S. P. Tanner, E. Kioupakis, and S. Schulz. Atomistic analysis of radiative recombination rate, stokes shift, and density of states in c-plane InGaN/GaN quantum wells. *Appl. Phys. Lett.*, 116(18):181104, 2020.

[38] D. Matsakis, A. Coster, B. Laster, and R. Sime. A renaming proposal: "the auger–meitner effect". *Physics Today*, 72(10):10, 2019.

[39] Aurelien David, Christophe A. Hurni, Nathan G. Young, and Michael D. Craven. Field-assisted shockley-read-hall recombinations in iii-nitride quantum wells. *Applied Physics Letters*, 111(23):233501, 2017.

[40] A. David, N. G. Young, C. Lund, and M. D. Craven. Review—the physics of recombinations in iii-nitride emitters. *ECS J. Solid State Sci. Technol.*, 9(1):016021, 2019.

[41] A. David, N. G. Young, C. Lund, and M. D. Craven. Compensation between radiative and auger recombinations in iii-nitrides: The scaling law of separated-wavefunction recombinations. *Appl. Phys. Lett.*, 115(19):193502, 2019.

[42] A. M. Armstrong, B. N. Bryant, M. H. Crawford, D. D. Koleske, S. R. Lee, and J. J. Wierer. Defect-reduction mechanism for improving radiative efficiency in ingan/gan light-emitting diodes using ingan underlayers. *Journal of Applied Physics*, 117(13):134501, 2015.

[43] E. Kioupakis, Q. Yan, and C. G. Van de Walle. Interplay of polarization fields and auger recombination in the efficiency droop of nitride light-emitting diodes. *Appl. Phys. Lett.*, 101(23):231107, 2012.

[44] N. Pant, X. Li, E. DeJong, D. Feezell, R. Armitage, and E. Kioupakis. Origin of the injection-dependent emission blueshift and linewidth broadening of iii-nitride light-emitting diodes. *AIP Advances*, 12(12):125020, 2022.

CHAPTER VIII

Summary and Future Work

8.1 Summary

In chapter I, I provide a comprehensive introduction to III-nitride semiconductors, highlighting the importance of alloy disorder and polarization fields in understanding their materials physics. I give a brief overview of the physics of light-emitting diodes, and survey the applications, as well as the role nitride semiconductors play in these applications. I also give a brief overview of the physics of power electronics, and survey power conversion and high-power radio-frequency technologies.

In chapter II, I present a comprehensive review of first-principles methods based on density-functional theory for predicting and describing materials phenomena. I introduce the Kohn-Sham equations and density-functional theory, and discuss exchange and correlation with an emphasis on physical intuition. I also cover many-body perturbation theory, the GW approximation, lattice vibrations, density-functional perturbation theory, electron-phonon interactions, Fermi's golden rule, and the Boltzmann transport equation. Finally, I discuss how alloy disorder can be modeled from first principles using the method of special quasirandom structures, which I connect to the lattice theory of cluster expansions, and demonstrate how to calculate the spectral function of alloys in any crystal basis by unfolding and refolding the Brillouin zone.

In chapter III, I review basic semi-empirical methods used in device simulations based

on the envelope-function and effective-mass approximations. I explain the relationship between $\mathbf{k} \cdot \mathbf{p}$ perturbation theory and the effective-mass approximation, derive the effective-mass equation, and discuss Schrödinger-Poisson modeling. I emphasize the importance of including many-body exchange-correlation effects in modeling and explain how to calculate material parameters from first-principles. Finally, I discuss the limitations of these methods and situations where more advanced approximations may be necessary.

In chapter IV, I develop a first-principles method to calculate the alloy-scattering potential and electron mobility of semiconductors with a focus on composition-dependent disorder. The scattering rates of AlGaN alloys are evaluated by unfolding the band structure and estimating the scattering potential and electron mobility using the hard-sphere model. Our results show that alloy disorder limits electron mobility for a wide range of compositions, and we compare the performance of $Al_xGa_{1-x}N$ and β-$(Al_xGa_{1-x})_2O_3$ for power-electronics applications.

In chapter V, I use predictive atomistic calculations to show that atomically thin superlattices of AlN and GaN have a phonon-limited mobility that is 3-4$\times$ larger than random AlGaN alloys due to the absence of alloy disorder. These superlattices also have better integration with dielectrics and contact metals and have the highest modified Baliga figure of merit among all known semiconductors.

In chapter VI, I develop a Schrödinger-Poisson model to investigate the injection-dependent emission spectra of III-nitride light-emitting diodes (LEDs). I show that the commonly accepted hypothesis, which attributes the blueshift and linewidth broadening solely to screening of internal electric fields, is incomplete. I find that plasma renormalization and phase-space filling also play important roles, with the latter being the predominant cause of spectral broadening. Our work provides new insights into the connection between the color purity of LEDs and their carrier recombination dynamics, and suggests that improving carrier transport and recombination lifetimes can enable better color over their purity.

Finally, in chapter VII, I investigate the notion of defect tolerance in InGaN emitters.

The widely accepted hypothesis of disorder-induced carrier localization hindering carrier diffusion has been challenged by recent experiments showing long diffusion lengths in InGaN. By developing a formalism for calculating non-radiative recombination rates, I find that an *apparent* defect tolerance emerges at low current densities due to an interplay of carrier localization, polarization fields, and recombination dynamics. However, the maximum quantum efficiency is *not* defect tolerant, emphasizing the need for defect management strategies for long wavelength emitters.

8.2 Future Work

The field of first-principles calculations of materials and devices is ripe with opportunities. There is a large community of researchers working on describing the physics of materials from first-principles, and another large community of researchers working on describing devices with empirical or semi-empirical approaches. However, the interface at which these two length scales meet is relatively unexplored.

8.2.1 Transport phenomena in materials and devices

Over the past decade, there have been significant advancements in the ability to describe and predict the low-field transport properties of materials. These developments have been primarily driven by the raw increase in supercomputing power and technical advancements in the field. One such advancement is the efficient interpolation of eigenvalues and matrix elements using maximally localized Wannier functions.

Moving forward, there are several low-hanging fruits in the field of low-field mobility calculations. This thesis proposes a novel method for calculating the mobility of semiconductor alloys that can be easily applied to other semiconductor alloys with similar band structures, such as $(Al_xGa_{1-x})_2O_3$ and $Ge_xSn_{1-x}O_2$, which exhibit isolated conduction or valence bands centered at the Γ point. These semiconductor alloys are promising for next-generation power electronics due to their ultra-wide band gaps and native bulk substrates, similar to III-nitrides.

The methodology developed in this thesis for calculating alloy mobility is limited to cases where the band of interest is not entangled with other bands in the manifold of valence or conduction states. In cases where entanglement occurs, it may be useful to combine this method with the disentanglement procedure implemented in the Wannier90 code. Another promising route is to explore the use of the Zacharias-Giustino special displacement configuration, which is a special supercell configuration that yiels the exact electron-phonon renormalization, to estimate low-field mobility by unfolding the band structure. This approach would require large-scale atomistic tight-binding simulations to obtain the broadening of small k-states, which are essential for low-field transport. This approach may be a way to surpass the limitations of the first Born approximation commonly used in modern calculations of electron-phonon mobility without the need for a many-body non-equilibrium Green's function description of transport. By coupling mobility calculations with large-scale atomistic tight-binding simulations, it should also be possible to calculate the mobility of graded alloys and of alloyed structures with partial atomic ordering.

In addition to the potential avenues for improving the methodology for calculating the mobility of semiconductor alloys, there are several open questions that need to be addressed. One such question is how to choose the primitive cell for the unfolding procedure. Does the choice of primitive cell affect the outcome of the unfolding procedure? For instance, superlattices exhibit periodicity, and the crystal-momentum eigenstates are eigenstates of the Hamiltonian. However, when the superlattice band structure is unfolded onto the primitive cell basis, the resulting band structure exhibits a broadening. It remains unclear how to interpret this broadening in terms of the scattering lifetime.

In this thesis, we demonstrated that atomically thin superlattices of AlN and GaN can exhibit many of the desirable properties of AlGaN without the alloy disorder that limits transport in random alloys. This raises the question of whether this approach could be a general strategy for enhancing the transport of other semiconductor alloys. One promising avenue for constructing novel superlattices is the use of van der Waals materials, which can be stacked without producing dislocations and defects. What

new phenomena can emerge in these superlattice systems, and how will electrons and holes transport in them? In addition, while calculations of low-field mobility of layered semiconductors are becoming more commonplace, these calculations are typically performed for isolated layered semiconductors without taking into account the effects of the dielectric environment of the substrate. Similarly, can we predict the mobility of 2D electron gases at semiconductor interfaces, such as the technologically important AlN/GaN interface, which are greatly affected by defects and disorder in the barrier environment? Interestingly, certain alloys, including some compositions of AlGaN, are predicted to exhibit chemical ordering due to kinetic or electrostatic driving forces. What is the fundamental limit of electron transport in these ordered systems?

Semiconductor superlattices, although not atomically thin, are already used in the drift regions of many important semiconductor devices in industry. However, the Boltzmann transport equation predicts these structures to be non-conducting, as it does not account for quantum-mechanical effects such as tunneling and field emission. Therefore, to capture these effects, one would need to go beyond the Boltzmann transport equation to a non-equilibrium Green's function description of transport. This approach may also be necessary for describing transport in LED structures, since current approaches such as the drift-diffusion model or Boltzmann transport formalism greatly overestimate the turn-on voltage. Thus, incorporating quantum-mechanical effects in transport modeling will be crucial for developing more accurate models and improving the design of semiconductor devices.

Furthermore, it is still a challenge to solve the Boltzmann transport equation for devices entirely from first principles, without any semi-empirical approximation. Phonon spectra calculations for common device geometries are not yet available due to the high computational cost. This is important because devices have different phonon modes than their bulk counterparts, which may lead to different electron-phonon interactions. First-principles device modeling tools would also need to account for phonon transport by solving the phonon Boltzmann transport equation. This is crucial for modeling modern technologies, such as GaN/AlGaN HEMTs,

which are limited by heat transport rather than electronic transport. Additionally, there is a lack of understanding of the effects of non-equilibrium phonons on electronic transport and relaxation at low and high electric fields. A comprehensive device modeling tool that describes electrons and phonons from first principles would be beneficial for addressing these challenges.

Many devices, such as RF transistors and high-power electronics, operate under high electric fields. Therefore, it is essential to have a better understanding of how materials behave under high fields, which can be addressed by solving the high-field Boltzmann transport equation. By solving the high-field Boltzmann transport equation, one can account for spatial variations and describe devices because the device geometry is often limiting rather than the material properties. The breakdown modes of materials and devices, such as impact ionization, which leads to Avalanche breakdown, and defect generation from Auger-Meitner recombination, are also poorly understood. There is a need to develop new methodologies to comprehensively describe these phenomena, not just for bulk compounds but also taking spatial variations into account. Systematic comparisons with non-adiabatic simulations that can capture these phenomena, albeit with severe restrictions on system sizes and time scales, would benefit these developments.

Finally, there are several long-standing questions in the field of transport that will likely take several years if not decades to answer in a satisfying manner. A central challenge is to unify the description of transport in localized and extended states. The transport of extended states is described by the Boltzmann transport equation, while the transport of localized states can be described using non-adiabatic time-dependent density-functional theory by calculating the auto-correlation function in the Kubo formula, but these calculations are limited to small cells and short times. An alternate approach is to use configuration coordinate diagrams to calculate hopping rates, but this approach is not general to all forms of localization and does not work well in the intermediate regime where the wave function has a large localization radius. Moreover, how should we account for charge trapping and emission by defects, which contribute to the resistivity but are currently not accounted for in cal-

culations of the mobility? In addition, the impact of disorder on phonons rather than electrons is still not well-understood. How does *phonon* disorder impact electron-phonon coupling? Can we use mass disorder, such as through isotope engineering, to enhance electron transport while suppressing phonon transport, which may be relevant for thermoelectric applications?

8.2.2 Optical phenomena in materials and devices

In the past decade, significant progress has been made in understanding the optical properties of materials. Researchers can now calculate phonon-assisted optical absorption rates, direct and indirect excitonic absorption spectra, and radiative and non-radiative recombination coefficients. Moreover, semi-empirical approaches that train or parameterize semi-empirical Hamiltonians using first-principles data have emerged as a cost-effective means of calculating device properties, albeit with lower accuracy. However, these calculations are currently only performed by a limited number of research groups worldwide. Additionally, the methods used to link microscale physics to macroscale phenomena are still in their early stages, providing ample opportunity for further advancement.

The main focus of this thesis has been on modeling InGaN devices for visible light emitters. The techniques used in this study can also be applied to investigate the performance limitations of AlGaN UV emitters and non-polar/semi-polar emitters, without requiring further method development. Furthermore, these methods can be employed to explore emerging III-nitride semiconductors, such as quaternary InAlGaN alloys and compounds containing B, Sc, or La. The knowledge of the fundamental material parameters of these compounds, even in their bulk form, is still incomplete, and future research could focus on calculating these basic parameters, such as polarization constants, bowing parameter, band offsets, etc. Once these parameters are determined, the semi-empirical methods developed in this thesis can be utilized to examine mesoscale structures that employ these new compounds. The approaches developed in this study are not restricted to III-nitrides, and other material systems can also be studied. In addition, the methods in this thesis could be utilized

to investigate device geometry, which is crucial for designing efficient quantum wells. Systematic studies of the impact of compositional profiles, well-width fluctuations, and short- and long-range ordering are still incomplete for the InGaN system and missing entirely for emerging nitrides and other semiconductor systems, representing a promising avenue for future short-term research.

In addition, it is worth noting that there is plenty of room for methodological developments to the semi-empirical methods employed in this thesis. Specifically, current methods for treating devices do not account for excitonic effects, which are only of secondary importance in polar quantum wells, as they suppress the wave-function overlap and binding energy. However, excitonic effects are crucial for non-polar semiconductors and ultra-thin emitters. It is essential to capture excitonic effects, either by propagating the time-dependent Schrodinger equation or by solving the Wannier exciton equation, to accurately describe these systems. Moreover, even for polar semiconductors, the accurate description of excitonic effects for recombination rates may be essential since it is still not possible to predict the IQE vs current curves exhibited by actual devices with accuracy. Furthermore, it is becoming increasingly clear that semi-empirical approaches may have systematic errors that are not necessarily well-controlled. This makes it challenging to use them predictively to study materials for which no experimental data exist. Therefore, there is considerable scope for the development of ab initio tight-binding approaches that are as close to first principles as possible to maximize their generalizability.

In addition to device-level excitonic effects, there is a need for methodological advancements to treat excitonic effects at the material level. Specifically, methods are needed to accurately calculate Shockley-Read-Hall, radiative, and Auger-Meitner recombination rates at the first-principles level, such as via the Bethe-Salpeter equation, that systematically treat higher-order correlations that give rise to excitons, trions, and biexcitons.

Overall, even calculations of recombination rates and functional optical properties in materials are in their infancy and have only been applied to a handful of semiconductors. Certain trends are already emerging in some semiconductors, while other

trends are well-known. For example, it is well-established that for conventional semiconductors, the Kane energy that describes the optical oscillator strength is a nearly universal value of 20 eV. It is worth investigating whether similar universal trends exist for non-radiative recombination rates in semiconductors.

The impact of polarization fields on non-radiative recombination rates is also not fully understood. While it is known that polarization fields can reduce carrier overlap, they can also introduce a spatial dependence to energy levels in the device, and this effect is often not considered in calculations. Similarly, the effect of band bending at interfaces in devices is also not fully accounted for. The ultimate goal of the field should be to develop predictive calculations that comprehensively include these effects at the device level without relying on empirical parameters.

Another challenge is connecting highly accurate transport calculations with device simulations, without resorting to empirical parameters. Currently, most device simulations use empirical parameters to match experimental results, compromising their reliability. A solution to this challenge could be the development of self-consistent first-principles models for transport and recombination by coupling $\mathbf{k} \cdot \mathbf{p}$ or tight-binding Hamiltonians, with the first-principles Boltzmann transport equation, and later with the non-equilibrium Green's function method. While some tools already exist that fill this niche, they rely on certain empirical parameters and make several simplifying approximations whose validity is not justified.

Although this thesis primarily focuses on Anderson-type localization, there are other forms of localization that can impact optical and recombination properties, such as polarons and Mott localization. There is a need for fundamental research to understand how these effects impact optical properties at both the material and device level. Such advancements would have broad applications across a range of materials, including organic semiconductors, Mott insulators, transition-metal oxides, and semiconductor alloys used in lighting, photo-catalysis, and solar cells. One challenge is to reliably calculate Urbach tails in materials using very large supercells, as their origin is still debated. Urbach tails in the joint density of states and Lifshitz tails in the density of states are known to emerge in the presence of disorder, whether it be static

disorder due to defects and alloying, or dynamic disorder due to phonons. Methodological advancements are needed to study systems that are as large as possible, while still retaining an accurate atomistic description that can capture the microscopic physics of the system. Advancements in modeling the functional properties of localized excitations would complement advancements in modeling their transport, as discussed in the previous section. While there are emerging studies on the impact of localization on emerging systems, these studies are often not systematic and often focus on one functional property while ignoring the effects of localization on other important properties such as transport, radiative and non-radiative recombination rates. There is a need for more comprehensive studies that consider the impact of localization on all relevant functional properties.

Finally, the methods presented in this thesis have mainly focused on studying light-emitting diodes, but there are still many other types of semiconductor devices that require investigation. These devices include laser diodes, photo-detectors, solar cells, and modulation and switching devices. While there are empirical or semi-empirical approaches available to describe these devices, they often require significant tuning of empirical parameters to match experimental results. This severely compromises their reliability, and sometimes leads to qualitatively incorrect conclusions. Therefore, obtaining a predictive description of various types of devices from first principles that comprehensively includes transport and optical phenomena, including excitonic and higher order correlation effects, remains an important long-term goal for the field.

APPENDICES

APPENDIX A

Adiabatic Approximation of the Many-Body Schrödinger Equation

A.1 Full many-body Schrödinger equation

The fundamental equation of matter is the Schrödinger equation,

$$\hat{H}\Psi = -i\frac{\partial}{\partial t}\Psi, \tag{A.1}$$

where Ψ is the wave function, and $\hat{H}$ is the Hamiltonian (energy) operator. If $\hat{H}$ is not a function of time, then this equation can be separated as,

$$\hat{H}\Psi = E\Psi, \tag{A.2}$$

$$-i\frac{\partial}{\partial t}\Psi = E\Psi. \tag{A.3}$$

The former, which we name the time-independent Schrödinger equation, is simply an eigenvalue problem, where Ψ is an *eigen*state or *stationary* state of the system that does not change with time, and E is the corresponding *eigen*energy. If Ψ is the ground-state wave function, then E is the ground state energy of the system, which includes both the electronic and nuclear contributions. The latter equation is

an ordinary differential equation, from which it is clear that complete knowledge of a system's stationary states and energies fully characterizes its dynamics. This is only true if $\hat{H}$ is independent of time, which turns out to be an excellent starting assumption for the study of real materials. Time dependence in $\hat{H}$ can be incorporated *a posteriori* using approximate methods such as perturbation theory.

The Hamiltonian H, which is an operator related to the total energy of the system, has contributions from the energy of electrons (e), nuclei (n), and their interaction ($e-n$),

$$H = H_e + H_n + H_{e-n} = T_e + V_e + T_n + V_n + H_{e-n}. \tag{A.4}$$

We have dropped the hats denoting that $\hat{H}$ is an operator for notational convenience. We have also split the H_e and H_n terms into their kinetic (T) and potential (V) contributions. Comparing this equation to $H\Psi = E\Psi$, it is clear that H couples the electronic and nuclear degrees of freedom, such that $\Psi = \Psi(r_1, r_2, ..., r_{N_e}, R_1, R_2, ..., R_{N_I})$ is a many-body wave function that simultaneously describes electrons and nuclei, and depends on both the electronic coordinates r_i and the nuclear coordinates R_I.

A.2 Adiabatic approximation

To facilitate the problem of solving the Schrödinger equation, we decouple the electronic and nuclear degrees of freedom. This is done using a procedure known as the Born-Oppenheimer approximation, which assumes that electrons move so much more quickly than the nuclei that their degrees of freedom are uncoupled. The decoupling of the electronic and nuclear degrees of freedom prevents the exchange of energy between them. This has very important consequences on how we deal with electron-nuclear interactions when we later consider functional properties such as the electron mobility.

We obtain more clarity on how to proceed by defining the operators of equation

(A.4). The electronic operators are defined as,

$$T_e := -\frac{1}{2}\sum_i \nabla_i^2, \tag{A.5}$$

$$V_e := \frac{1}{2}\sum_{i\neq j} \frac{4\pi}{|\mathbf{r}_i - \mathbf{r}_j|}, \tag{A.6}$$

where the sums run over the electron coordinates. Similarly, the nuclear operators are defined as,

$$T_n := -\frac{1}{2}\sum_I \nabla_I^2, \tag{A.7}$$

$$V_n := \frac{1}{2}\sum_{I\neq J} Z_I Z_J \frac{4\pi}{|\mathbf{R}_I - \mathbf{R}_J|}, \tag{A.8}$$

where Z is the nuclear charge, and the sums run over the nuclear coordinates. The problematic term, which is the electron-nuclear interaction, is given by,

$$H_{e-n} = -\sum_{i,I} Z_I \frac{4\pi}{|\mathbf{r}_i - \mathbf{R}_I|}. \tag{A.9}$$

To facilitate the problem of solving the Schrödinger equation, we decouple the electronic and nuclear degrees of freedom. This is done using a procedure known as the Born-Oppenheimer approximation, which assumes that electrons move so much more quickly than the nuclei that their degrees of freedom are uncoupled. The decoupling of the electronic and nuclear degrees of freedom prevents the exchange of energy between them. This has very important consequences on how we deal with electron-nuclear interactions when we later consider functional properties such as the electron mobility.

A.2.1 Frozen-nuclei approximation

Armed with the knowledge that nuclei are approximately frozen in the electrons' reference frame, we can try to substitute the nuclear coordinates with effective frozen

coordinates. In a crystal's ground state, the nuclei oscillate about their equilibrium positions in a harmonic potential. Therefore, we may make the substitution $R_I \to \langle R_I \rangle$, in which case the wave function is only a function of the electronic coordinates, $\Psi = \Psi(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e}, \langle \mathbf{R}_1 \rangle, \langle \mathbf{R}_2 \rangle, ..., \langle \mathbf{R}_N \rangle) \equiv \Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e})$. The Schrödinger equation for electrons with the nuclear coordinates frozen is,

$$[T_e + V_e + H_{e-n}]\, \Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e}) = (E - T_n - V_n)\Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e}) \quad \text{(A.10)}$$

$$\equiv E(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_{N_I})\Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e}), \quad \text{(A.11)}$$

where $E(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)$ is the total electronic energy that depends parametrically on the ionic coordinates.

A.2.2 Born-Oppenheimer approximation

Having determined how to solve the many-electron wave function for a fixed set of nuclear coordinates, the question remains how we can reconstruct the total wave function. In the Born-Oppenheimer approximation, we make the Ansatz,

$$\Psi \approx \Psi_{\{\mathbf{R}\}}(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_e})\chi(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N), \quad \text{(A.12)}$$

where χ is the nuclear wave function. Intuitively, this approximation is justified if we assume that nuclei remain close to their equilibrium position as their wave function evolves, such that the electronic wave function is unaffected. Since the nuclear timescale is much longer than the electronic timescale, nuclear evolution is adiabatic in the electron reference frame. According to the *adiabatic theorem*, if the electrons are in an instantaneous eigenstate, they will remain in the instantaneous eigenstate as the nuclei evolve.

The Schrödinger equation for the nuclear wave function is obtained by substituting

Ansatz (A.12) into the full Schrödinger equation, and using equation (A.11).

$$(T_e + V_e + T_n + V_n + H_{e-n})\,\Psi_{\{\mathbf{R}\}}\chi = E_{tot}\Psi_{\{\mathbf{R}\}}\chi \tag{A.13}$$

$$E(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)\Psi_{\{\mathbf{R}\}}\chi + \Psi_{\{\mathbf{R}\}}(T_n + V_n)\chi = E_{tot}\Psi_{\{\mathbf{R}\}}\chi \tag{A.14}$$

Multiplying by $\Psi^*_{\{\mathbf{R}\}}$ and integrating over the electronic coordinates, we obtain the Schrödinger equation for nuclear wave functions,

$$[T_n + V_n + E(\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_N)]\,\chi = E_{tot}\chi. \tag{A.15}$$

From this, it is clear that the effective potential felt by the nuclei is a sum of their Coulomb repulsion and the total energy of electrons at fixed nuclear positions. Thus, using the Born-Oppenheimer approximation, we have decoupled the many-body Schrödinger equation into two equations: one for the electrons and one for the nuclei. Within this thesis, we primarily focus on solving the many-body equation for electrons.

APPENDIX B

Derivation of the Kohn-Sham Equation

B.1 Derivation of the Kohn-Sham equations by variational minimization

To derive the Kohn-Sham equations, we start with the definition of the charge density, $n(\mathbf{r}) = \sum_i \phi_i^*(\mathbf{r})\phi_i(\mathbf{r})$, for Kohn-Sham orbitals ϕ. We make the Ansatz that the interacting system of electrons can be mapped onto a non-interacting system of electrons, which we call the Kohn-Sham electrons. The energy functional of the non-interacting system contains contributions from the non-interacting kinetic energy, electron-nuclear interaction, ion-ion interaction E_{II}, and an unknown contribution due to many-body effects $E_{ks}[n]$,

$$E[n] = -\frac{1}{2}\sum_i \langle\phi_i|\,\nabla^2\,|\phi_i\rangle + \sum_{i,I} Z_I\,\langle\phi_i|\,\frac{4\pi}{|\mathbf{r}-\mathbf{R}_I|}\,|\phi_i\rangle + E_{II} + E_{KS}[n]. \tag{B.1}$$

We are able to write the unknown Kohn-Sham energy term that contains many-body effects as a functional of the carrier density *because* of Hohenberg and Kohn's theorems. We know that $E_{KS}[n]$ maps the non-interacting system onto the interacting system exactly for the ground-start charge density. At this point, we use the *method of Lagrange multipliers* to minimize $E[n]$ with the constraint that the wave functions

are orthonormal. The orthonormality constraint can be written as, $\langle\phi_i|\phi_j\rangle - \delta_{ij} = 0$. Thus, the system of equations that we have to solve is,

$$\frac{\delta E}{\delta n} = 0 \tag{B.2}$$

$$\langle\phi_i|\phi_j\rangle = \delta_{ij} \tag{B.3}$$

The constrained optimization problem can be written as an unconstrained optimization problem by constructing the Lagrange functional $L \equiv E[n] - \sum_{ij}\lambda_{ij}[\langle\phi_i|\phi_j\rangle - \delta_{ij}]$, and demanding that $\delta L/\delta\phi_i^* = 0$ and $\delta L/\delta\lambda_{ij} = 0$. Note that the sum over i, j for the constraint is crucial since we require that all wave functions be orthonormal. Before we proceed, we make a few general observations using the properties of functional derivatives:

$$\frac{\delta n}{\delta\phi_i^*} = \phi_i \tag{B.4}$$

$$\frac{\delta}{\delta\phi_i^*}\langle\phi_i|\phi_j\rangle = \psi_j \tag{B.5}$$

$$\frac{\delta}{\delta\phi_i^*}E[n] = -\frac{1}{2}\nabla^2\phi_i + \sum_I Z_I\frac{4\pi}{|\mathbf{r}-\mathbf{R}_I|}\phi_i + \frac{\delta E_{KS}[n]}{\delta n}\frac{\delta n}{\delta\phi_i^*} \tag{B.6}$$

The method of Lagrange multipliers gives the equations,

$$\frac{\partial L}{\partial\phi_i^*} = -\frac{1}{2}\nabla^2\phi_i + \sum_I Z_I\frac{4\pi}{|\mathbf{r}-\mathbf{R}_I|}\phi_i + \frac{\delta E_{KS}[n]}{\delta n}\phi_i - \sum_{ij}\lambda_{ij}\phi_j = 0, \tag{B.7}$$

$$\frac{\partial L}{\partial\lambda_{ij}} = \langle\phi_i|\phi_j\rangle - \delta_{ij} = 0. \tag{B.8}$$

At this point, we define the matrix S as the matrix that diagonalizes $\Lambda\big|_{ij} = \lambda_{ij}$, such that $S^{-1}\Lambda S = E$, where $E_{ij} = \varepsilon_i\delta_{ij}$. It turns out that the matrix Λ is Hermitian because H is Hermitian. Because it is a property of Hermitian matrices that they are diagonalized by unitary matrices, we know that S is unitary, *i.e.*, $SS^\dagger = I$. Thus, defining new rotated wave function as $\psi_i \equiv \sum_{ij} S_{ij}\phi_j$, and keeping in mind that unitary transformations preserve inner products, we obtain the Kohn-Sham

equations,

$$-\frac{1}{2}\nabla^2\psi_i + \sum_I Z_I \frac{4\pi}{|\mathbf{r}-\mathbf{R}_I|}\psi_i + \frac{\delta E_{KS}[n]}{\delta n}\psi_i = \varepsilon_i\psi_i \tag{B.9}$$

$$\langle\psi_i|\psi_j\rangle = \delta_{ij}, \tag{B.10}$$

where we see that the mean-field Kohn-Sham potential is obtained by taking the functional derivative of the Kohn-Sham energy functional, $V_{KS}[n] \equiv \delta E_{KS}/\delta n$. Importantly, it can be seen in the equation that the Kohn-Sham potential does not depend on details of where the nuclei are, and thus the Kohn-Sham potential that maps the Kohn-Sham equation onto the many-body Schrödinger equation is universal and depends only on the ground-state charge density.

APPENDIX C

Background for Green's Functions

This section sets up the math for the Green's function formulation of quantum mechanics. In particular, the focus is on zero-temperature Green's functions.

C.1 Creation and annihilation operators

Consider a complete set of single distinguishable particle wave functions $|\psi(\mathbf{r})\rangle$. We denote the indistinguishable *antisymmetric* ("A") many-body state of N single-particle states as,

$$\frac{1}{\sqrt{N!}} det(|\psi(x)\rangle) \equiv |k_1, k_2, k_3, ..., k_N\rangle_A \tag{C.1}$$

In this abbreviated tensor-product notation, the particle at position 1 is in state k_1, the particle at position 2 is in state k_2, and so on. (As an aside, note that $|k_1, k_2, ..., k_N\rangle_A$ lives in a Hilbert space of N particle dimensions, whereas $|k_1, k_2, ..., k_{N+1}\rangle_A$ lives in a Hilbert space of $N+1$ particle dimensions. Recall that our normal rules of operator algebra are only defined for kets within the same Hilbert space.) For the simplest case of two particles,

$$|k_1, k_2\rangle_A = \psi_{k_1}(\mathbf{r}_1)\psi_{k_2}(\mathbf{r}_2) - \psi_{k_2}(\mathbf{r}_1)\psi_{k_1}(\mathbf{r}_2) \tag{C.2}$$

$$|k_2, k_1\rangle_A = \psi_{k_2}(\mathbf{r}_1)\psi_{k_1}(\mathbf{r}_2) - \psi_{k_1}(\mathbf{r}_1)\psi_{k_1}(\mathbf{r}_2) \tag{C.3}$$

Immediately, we see that interchanging the order of the particles produces a negative sign, *i.e.*, $|k_1, k_2\rangle_A = -|k_2, k_1\rangle_A$. The *creation operator* $c_k^\dagger$ and the *annihilation operator* c_k are *defined* as,

$$c_k^\dagger |k_1, k_2, ..., k_N\rangle_A = |k, k_1, k_2, ..., k_N\rangle_A , \tag{C.4}$$

$$c_k |k, k_1, k_2, ..., k_N\rangle_A = |k_1, k_2, ..., k_N\rangle_A \tag{C.5}$$

$$c_k c_k |k, k_1, k_2, ..., k_N\rangle_A = 0 \tag{C.6}$$

$$c_k^\dagger c_k^\dagger |k_1, k_2, ..., k_N\rangle_A = 0 \tag{C.7}$$

Both c_k and $c_k^\dagger$ are defined such that they only operate on kets with index k corresponding to particle 1. In other words, k has to be on the left-most position in order for c_k and $c_k^\dagger$ to operate on the ket. If k is not at the left-most position, one must interchange the order of *neighbouring* particles, incurring a negative sign each time a particle is exchanged. Use these rules for the creation and annihilation operators, it is quite straightforward to show,

$$\{c_k, c_{k'}^\dagger\} = \delta_{k,k'} \tag{C.8}$$

$$\{c_k, c_{k'}\} = 0 \tag{C.9}$$

$$\{c_k^\dagger, c_{k'}^\dagger\} = 0 \tag{C.10}$$

As an example of how to prove these relations, we consider the question: is $c_k c_{k'}^\dagger = c_{k'}^\dagger c_k$? Consider the case where $k \neq k'$,

$$\begin{aligned} &c_{k_1} c_k^\dagger |k_1, k_2, ..., k_N\rangle_A = c_{k_1} |k, k_1, k_2, ..., k_N\rangle_A = -c_{k_1} |k_1, k, k_2, ..., k_N\rangle_A = -|k, k_2, ..., k_N\rangle_A \\ &c_k^\dagger c_{k_1} |k_1, k_2, ..., k_N\rangle_A = c_k^\dagger |k_2, ..., k_N\rangle_A = |k, k_2, ..., k_N\rangle_A \end{aligned} \tag{C.11}$$

Clearly, $c_{k_1} c_k^\dagger = -c_k^\dagger c_{k_1}$. Now consider the case where $k = k'$,

$$\begin{aligned} &c_k c_k^\dagger |k_1, k_2, ..., k_N\rangle_A = c_k |k, k_1, k_2, ..., k_N\rangle_A = |k_1, k_2, ..., k_N\rangle_A \\ &c_k^\dagger c_k |k_1, k_2, ..., k_N\rangle_A = 0 \end{aligned} \tag{C.12}$$

Clearly, $c_k c_k^\dagger = \mathbf{1} - c_k^\dagger c_k$. Thus, $\{c_k, c_k^\dagger\} = \delta_{k,k'}$.

C.2 Field operators

Rather than working with creation and annihilation operators directly, we typically work with *field operators* that act on all space,

$$\hat{\psi}(\mathbf{r}) \equiv \sum_k \psi_k(\mathbf{r}) c_k, \tag{C.13}$$

$$\hat{\psi}^\dagger(\mathbf{r}) \equiv \sum_k \psi_k^*(\mathbf{r}) c_k^\dagger. \tag{C.14}$$

These field operators act on *Fock space*, where rather than keeping tracking of all the wave functions in the system, we simply keep track of the number of particles in each state. We define the vacuum state $|0\rangle$ as the state which satisfies $\hat{\psi}(\mathbf{r})\,|0\rangle = 0$. The effect of $\psi(\mathbf{r})$ is to annihilate a particle at position $\mathbf{r}$ and the effect of $\psi^\dagger(\mathbf{r})$ is to create a particle at position $\mathbf{r}$.

To see why this definition of the field operator is useful, consider the Hamiltonian in second quantization,

$$H = \sum_k \langle \psi_k |\, h(\mathbf{r})\, | \psi_l \rangle\, c_k^\dagger c_l \; + \frac{1}{2} \sum_{k_1,k_2,k_3,k_4} \langle \psi_{k_1}, \psi_{k_2} |\, v(\mathbf{r},\mathbf{r}')\, | \psi_{k_3}, \psi_{k_4} \rangle\, c_{k_1}^\dagger c_{k_2}^\dagger c_{k_3} c_{k_4} + V_{nucl}, \tag{C.15}$$

where $h(\mathbf{r})$ is the non-interacting part of the Hamiltonian and v is the Coulomb interaction. With field operators, the Hamiltonian becomes,

$$H = \int d^3\mathbf{r} \hat{\psi}^\dagger(\mathbf{r}) h(\mathbf{r}) \hat{\psi}(\mathbf{r}) + \int d\mathbf{r} d\mathbf{r}' \hat{\psi}^\dagger(\mathbf{r}) \hat{\psi}^\dagger(\mathbf{r}') v(\mathbf{r},\mathbf{r}') \hat{\psi}(\mathbf{r}') \hat{\psi}(\mathbf{r}) + V_{nucl}. \tag{C.16}$$

The field operators satisfy,

$$\{\psi(\mathbf{r}), \psi^\dagger(\mathbf{r}')\} = \delta(\mathbf{r} - \mathbf{r}'), \tag{C.17}$$

$$\{\psi(\mathbf{r}), \psi(\mathbf{r}')\} = \{\psi^\dagger(\mathbf{r}), \psi^\dagger(\mathbf{r}')\} = 0. \tag{C.18}$$

C.3 Green's functions

Using creation and annihilation operators, we can track the transition probability *between* states. In additional to tracking the position variables, we will also track time. To do this, we introduce the space-time coordinates $1 \equiv (\mathbf{r}_1, t_1)$. Consider the quantity below, which we will call the single-particle Green's function,

$$G(1,2) := -i \left\langle N^0 \right| \hat{T} \hat{\psi}(1) \hat{\psi}^\dagger(2) \left| N^0 \right\rangle , \tag{C.19}$$

where the ground-state ket $|N^0\rangle$ has been explicitly written out in the expectation. For the moment, ignore the imaginary factor, which is due to historical notation. $\hat{T}$ is the time-ordering operator, which ensures that particles travel forward in time by time-ordering the operators from right (past) to left (future); the time ordering operator introduces as many negative signs as needed to ensure causality is obeyed. (Time-ordering is the mathematical trick that introduces Dirac delta functions in the equation of motion for the Green's function.) The physical interpretation of the Green's function is as follows: G is (the square root of) the transition probability of a particle being created at a spacetime point $2 \equiv (\mathbf{r}_2, t_2)$ and being annihilated at a different spacetime point $1 \equiv (\mathbf{r}_1, t_1)$. One way of seeing this is directly writing out the ground state of a fictitious two-level system as $|00\rangle$. The transition probability of a particle propagating from 2 to 1 would then be given by the $|\langle 10|01\rangle|^2$. Consider the definitions, $\hat{\psi}^\dagger(2)\,|00\rangle = |01\rangle$ and $\hat{\psi}^\dagger(1)\,|00\rangle = |10\rangle$. Substituting these definitions into $\langle 10|01\rangle$ gives $\langle 00|\,\hat{\psi}(1)\hat{\psi}^\dagger(2)\,|00\rangle$, which is identically the definition of the Green's function for the fictitious system (apart from the imaginary factor). For this reason, the Green's function is often called the "propagator" since it propagates a particle in spacetime. It is also called the single-particle correlation function since it describes how a single-particle's occupation number is correlated in space and time.

The Green's function is the central quantity of interest. Once we have obtained the Green's function, we have solved the many-body problem. Note that the definition above does not give a prescription of *how* to solve for the Green's function, only *what* it is. Neither does it tell us *why* obtaining the Green's function means that we

have solved the many-body problem. To make progress in this regard, we need to write out the equation of motion for the Green's function. In the Heisenberg picture, operators evolve according to,

$$\frac{d}{dt}\hat{O}(\mathbf{r},t) = [\hat{O}(\mathbf{r}), H]. \tag{C.20}$$

By plugging the annihilation and creation operators for $\hat{O}$, and using some clever mathematical tricks, it is possible to show that the equation of motion for the Green's function is,

$$\left[i\frac{\partial}{\partial t} - h_0(1)\right] G(1,2) = i\int d3v(1,3)G_2(1,2,3,3^+) + \delta(1,2). \tag{C.21}$$

Here, $\delta(1,2) = \delta(\mathbf{r}_1 - \mathbf{r}_2)\delta(t_1 - t_2)$. The quantity,

$$G_2(1,2,3,3^+) \equiv -\left\langle \hat{T}\hat{\psi}^\dagger(\mathbf{r}_3, t_3^+)\hat{\psi}(\mathbf{r}_3,t_3)\hat{\psi}(\mathbf{r}_1,t_1)\hat{\psi}^\dagger(\mathbf{r}_2,t_2)\right\rangle, \tag{C.22}$$

is the "two-particle Green's function" or "four-point correlation function" (since four field operators - two creation and two annihilation operators - appear in the expectation value). To evaluate G_2, we can substitute it into the Heisenberg equation of motion, and we will obtain another partial differential equation, which contains a six-point correlation function $G_3 \times O(v^2)$. G_3 in turn will depend on $G_4 \times O(v^3)$, and so on. The recursive dependence on higher order terms is a feature of the many-body Coulomb interaction. If the interaction energy v is smaller than the kinetic energy of the electrons, then perturbation theory can be applied to truncate higher order terms by treating the single-particle Hamiltonian as the starting-point of the perturbative expansion. Often, it is sufficient to solve the expansion up to $O(v)$. Typically, the expansion is performed in terms of Feynman diagrams; the interested reader should refer to the text by R. D. Mattuck, "A Guide to Feynman Diagrams in the Many-Body Problem."

C.4 Self-Energy

If we perform the many-body diagrammatic perturbative expansion, it turns out that the perturbative expansion repeats as a geometric series. Freeman Dyson remarked that if we define a term Σ in the series, we can write out the entire perturbative expansion as a recursive relation,

$$G = G_0 + G_0 \Sigma G = G_0 + G_0 \Sigma G_0 + G_0 \Sigma G_0 \Sigma G_0 + ... \tag{C.23}$$

Then, the full Green's function can be written in terms of the free-particle Green's function,

$$G = \frac{G_0}{1 - G_0 \Sigma} = \frac{1}{G_0^{-1} - \Sigma} = \frac{1}{\omega - \varepsilon_n(\mathbf{k}) - \Sigma}, \tag{C.24}$$

where we note that Σ has units of energy. From this, we see that if the interaction is small, i.e., perturbation theory applies, the full Green's function looks exactly like the free-particle Green's function with the energy shifted by Σ. Therefore, Σ is the correction to the free-particle energy due to the many-body interaction, and is called the "self-energy." In general, Σ is non-Hermitian (but symmetric); the real-part of Σ gives the energy renormalization of the electronic state due to interactions, and the imaginary-part of Σ is related to the finite lifetime of the electronic state in consideration.

C.5 Green's functions in terms of Dyson orbitals

Thus far, we have written the Green's function in terms of an expectation value of creation and annihilation operators. It would be convenient to rewrite the Green's function in terms of *wave functions*, because density-functional theory gives us the Kohn-Sham wave functions. To this end, we introduce the concept of *Dyson orbitals*.

C.5.1 Dyson orbitals

Consider two states in the Heisenberg representation: a ground state of N electrons, $|N\rangle$, and some eigenstate s of $N-1$ electrons, $|N-1,s\rangle$. We define a Dyson orbital as the expectation value,

$$\begin{aligned} f_s(\mathbf{r},t) &\equiv \langle N-1,s|\,\hat{\psi}(\mathbf{r},t)\,|N\rangle \\ &= \langle N-1,s|\,e^{iHt}\hat{\psi}(\mathbf{r})e^{-iHt}\,|N\rangle \\ &= \langle N-1,s|\,e^{iE_{N-1,s}t}\hat{\psi}(\mathbf{r})e^{-iE_Nt}\,|N\rangle \\ &= \langle N-1,s|\,\hat{\psi}(\mathbf{r})\,|N\rangle\,e^{-i(E_N-E_{N-1,s})t} \\ &\equiv f_s(\mathbf{r})e^{-i\varepsilon_s t}, \end{aligned} \tag{C.25}$$

where we see that this orbital oscillates with frequency $\varepsilon_s = E_N - E_{N-1,s}$. If $|N\rangle$ and $|N-1,s\rangle$ are anti-symmetric Slater determinants composed of non-interacting states, ψ_k, then we can identify the label s with a non-interacting state, and $\hat{c}_s\,|N\rangle = |N-1,s\rangle$. Thus, we can write the Dyson orbital with the wave function corresponding to the s independent-particle wave function,

$$\begin{aligned} \langle N-1,s|\,\hat{\psi}(\mathbf{r})\,|N\rangle &= \sum_k \langle N-1,s|\,\hat{c}_k\psi_k(\mathbf{r})\,|N\rangle \\ &= \sum_k \psi_k(\mathbf{r})\,\langle N-1,s|N-1,k\rangle \\ &= \sum_k \psi_k(\mathbf{r})\delta_{k,s} \\ &= \psi_s(\mathbf{r}). \end{aligned} \tag{C.26}$$

Here, s is evidently a filled state, e.g., in the valence manifold. (In general, $|N\rangle$ and $|N-1,s\rangle$ are interacting kets and the interpretation of the Dyson orbitals with single-particle wave functions is not always possible. In particular, $\langle N-1,s|\,\hat{c}_k\,|N\rangle \neq \delta_{k,s}$ for a general interacting system.)

We can similarly define Dyson orbitals corresponding to empty states. Consider two states, again in the Heisenberg representation: a ground state of N electrons, and an

eigenstate s of $N+1$ electrons, $|N+1,s\rangle$. We again define a Dyson orbital as,

$$\begin{aligned} f_s(\mathbf{r},t) &= \langle N|\,\hat{\psi}(\mathbf{r},t)\,|N+1,s\rangle \\ &= \langle N|\,\hat{\psi}(\mathbf{r})\,|N+1,s\rangle\, e^{-i(E_{N+1}-E_N)t} \\ &= f_s(\mathbf{r})e^{i\varepsilon_s t}. \end{aligned} \tag{C.27}$$

Some simple math shows that for a non-interaction system, $\langle N|\,\hat{\psi}(\mathbf{r})\,|N+1,s\rangle = \psi_s(\mathbf{r})$, where s corresponds to an empty state, e.g., in the conduction manifold.

C.5.2 Lesser and greater Green's function

We recall the definition of the time-ordered Green's function in the Heisenberg picture,

$$G(\mathbf{r},\mathbf{r}';t,t') = -i\,\langle N|\,\hat{T}\hat{\psi}(\mathbf{r},t)\hat{\psi}^\dagger(\mathbf{r}',t')\,|N\rangle \tag{C.28}$$

If $t > t'$ then $\hat{T}$ leaves the ordering of the field operators unchanged. If $t < t'$ then $\hat{T}$ interchanges the order of the field operators, incurring a negative sign. Thus, the Green's function can be rewritten with the Heaviside step function $\Theta(x)$,

$$\begin{aligned} G(\mathbf{r},\mathbf{r}';t,t') &= -i\,\langle N|\,\hat{\psi}(\mathbf{r},t)\hat{\psi}^\dagger(\mathbf{r}',t')\,|N\rangle\,\Theta(t-t') \\ &\quad + i\,\langle N|\,\hat{\psi}^\dagger(\mathbf{r}',t')\hat{\psi}(\mathbf{r},t)\,|N\rangle\,\Theta(t'-t) \\ &= -i\,\langle N|\,e^{iHt}\hat{\psi}(\mathbf{r})e^{-iHt}e^{iHt'}\hat{\psi}^\dagger(\mathbf{r}')e^{-iHt'}\,|N\rangle\,\Theta(t-t') \\ &\quad + i\,\langle N|\,e^{iHt'}\hat{\psi}^\dagger(\mathbf{r}')e^{-iHt'}e^{iHt}\hat{\psi}(\mathbf{r})e^{-iHt}\,|N\rangle\,\Theta(t'-t) \\ &= -i\,\langle N|\,e^{iE_Nt}\hat{\psi}(\mathbf{r})e^{-iHt}e^{iHt'}\hat{\psi}^\dagger(\mathbf{r}')e^{-iE_Nt'}\,|N\rangle\,\Theta(t-t') \\ &\quad + i\,\langle N|\,e^{iE_Nt'}\hat{\psi}^\dagger(\mathbf{r}')e^{-iHt'}e^{iHt}\hat{\psi}(\mathbf{r})e^{-iE_Nt}\,|N\rangle\,\Theta(t'-t) \\ &= -i\,\langle N|\,\hat{\psi}(\mathbf{r})e^{-i(H-E_N)(t-t')}\hat{\psi}^\dagger(\mathbf{r}')\,|N\rangle\,\Theta(t-t') \\ &\quad + i\,\langle N|\,\hat{\psi}^\dagger(\mathbf{r}')e^{i(H-E_N)(t-t')}\hat{\psi}(\mathbf{r})\,|N\rangle\,\Theta(t'-t) \end{aligned} \tag{C.29}$$

We can simplify this notation by defining $\tau \equiv t - t'$. τ is therefore the time at which a particle is annihilated, relative to the time that it was created. We can also

define the *greater Green's function* G_R as the Green's function for the case $t > t'$ or $\tau > 0$.

$$G^{>}(\mathbf{r}, \mathbf{r}', \tau) = -i \langle N | \hat{\psi}(\mathbf{r}) e^{-i(H-E_N)\tau} \hat{\psi}^{\dagger}(\mathbf{r}') | N \rangle \tag{C.30}$$

The greater Green's function corresponds to the probability of creating an electron at (r', t'), and later finding it at (r, t).

We can similarly define the *lesser Green's function* as the Green's function for the case $t < t'$ or $\tau < 0$. The negative time is simplify an artefact of having defined $\tau \equiv t - t'$, where t is the time of annihilation and t' is the time of creation.

$$G^{<}(\mathbf{r}, \mathbf{r}', \tau) = +i \langle N | \hat{\psi}^{\dagger}(\mathbf{r}') e^{i(H-E_N)\tau} \hat{\psi}(\mathbf{r}) | N \rangle \tag{C.31}$$

The lesser Green's function corresponds to the probability of creating creating a *hole* at (r, t) and later finding it at (r', t').

Thus, the time-ordered Green's function reduces to either the lesser or greater Green's function, depending on whether $\tau > 0$ (electrons) or $\tau < 0$ (holes), $G(\mathbf{r}, \mathbf{r}', \tau) \equiv G^{>}(\mathbf{r}, \mathbf{r}', \tau)\Theta(\tau) + G^{<}\mathbf{r}, \mathbf{r}', \tau)\Theta(-\tau)$.

C.5.3 Lehmmann representation of the Green's function

We can now represent the Green's functions in terms of Dyson orbitals by making use of the completeness relation,

$$\sum_s |N+1, s\rangle \langle N+1, s| = \mathbf{1}_{N+1} \tag{C.32}$$

$$\sum_s |N-1, s\rangle \langle N-1, s| = \mathbf{1}_{N-1} \tag{C.33}$$

where the label s denotes an eigenstate of the $N-1$ or $N+1$ many-body state, and $\mathbf{1}_N$ is the identity operator in the N-dimensional Hilbert space. (The label s for the eigenstates may make more sense if we recall that $|N\rangle$ corresponds to the ground state (lowest energy eigenstate) of the many-body state with N particles.)

Consider the greater Green's function $G^{>}$. What would happen if we inserted $\mathbf{1}_{N+1}$ into its definition? Making use of the fact that $\mathbf{1}_{N+1}$ commutes with $\hat{\psi}$ and using the definition of Dyson orbitals,

$$\begin{aligned} G^{>}(\mathbf{r},\mathbf{r}',\tau) &= -i\left\langle N\right|\hat{\psi}(\mathbf{r})\mathbf{1}_{N+1}e^{i(H-E_N)\tau}\hat{\psi}^{\dagger}(\mathbf{r}')\left|N\right\rangle \\ &= -i\sum_s \left\langle N\right|\hat{\psi}(\mathbf{r})\left|N+1,s\right\rangle\left\langle N+1,s\right|e^{-i(H-E_N)\tau}\hat{\psi}^{\dagger}(\mathbf{r}')\left|N\right\rangle \\ &= -i\sum_s f_s(\mathbf{r})f_s^*(\mathbf{r}')e^{-i(E_{N+1,s}-E_N)\tau} \qquad \text{(C.34)} \\ &= -i\sum_s f_s(\mathbf{r})f_s^*(\mathbf{r}')e^{-i\varepsilon_s\tau} \qquad \text{(C.35)} \end{aligned}$$

Clearly, taking the Fourier transform of $G(\mathbf{r},\mathbf{r}',\tau)$ gives the Lehmann representation,

$$G(\mathbf{r},\mathbf{r}',\varepsilon) = \sum_s \frac{f_s(\mathbf{r})f_s^*(\mathbf{r}')}{\varepsilon-\varepsilon_s\pm i\eta}, \qquad \text{(C.36)}$$

where $\eta\to 0$ is a regularization parameter.