University of California

Los Angeles

Physics and Modeling of Tunneling in Low Power Transistors

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Electrical Engineering

by

Andrew Samuel Pan

2015

Abstract of the Dissertation

Physics and Modeling of Tunneling in Low Power Transistors

by

Andrew Samuel Pan

Doctor of Philosophy in Electrical Engineering

University of California, Los Angeles, 2015

Professor Chi On Chui, Chair

As CMOS electronics grow ever more ubiquitous and essential to modern life, managing and reducing power dissipation becomes essential. At the device level, this requires new transistors with reduced leakage currents and operating voltages. Opportunities and challenges in this regard arise from quantum transport effects. For instance, novel tunneling field-effect transistors (TFETs) can potentially operate at substantially lower voltages than MOSFETs by utilizing interband tunneling as the conduction process. Conversely, the Moore's law-driven scaling of MOSFETs down to the nanometer regime increases source-drain intraband tunneling, which may limit leakage power in future CMOS. Conventional device models and simulations based on semiclassical concepts are inadequate for describing such effects. In this dissertation, we develop new theoretical models to study tunneling and apply the resulting insights to MOSFET and TFET device design.

To this end, we develop a complete device simulator that uses non-equilibrium Green's functions (NEGF) to rigorously model quantum transport. We utilize a combination of NEGF, full band structure calculations, and analytical derivations to study the physics of interband tunneling in semiconductors. We clarify and improve the accuracy of commonly used analytical tunneling models and extend them to quantum confined structures, which include present and future scaled devices like ultra-thin body (UTB) transistors, FinFETs, and nanowire devices. We merge our findings with electrostatic analyses to derive the first

general quasi-analytical current model for TFETs that provides device insight and is easily used for compact modeling. We show that existing TFETs are performance limited by the chemical source doping profiles, a particularly profound problem for III-V p-type TFETs. To overcome these limitations, we propose a new device design, the gate-induced source tunneling FET (GISTFET), which utilizes electrostatic doping to define the tunneling junction and allow for high performing complementary TFET systems. Finally, we derive the first model of source-drain tunneling in MOSFETs and study the effect of contact doping on leakage in scaled devices.

The dissertation of Andrew Samuel Pan is approved.

Vidvuds Ozolins

Oscar Stafsudd

Chi On Chui, Committee Chair

University of California, Los Angeles

2015

To the happy few.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I have been very blessed to pursue my PhD studies under the guidance of Prof. Chi On Chui. His insights into semiconductor technology, his excellent advice and encouragement, and the trust and the freedom he has consistently given me to explore a heterogeneous array of topics in engineering and physics have all been vital to my professional and personal development over the last few years. The lessons I've learned from him about performing research (and just as importantly, picking the right problems to attack) and communicating with others will remain with me the rest of my life. For all this, and a multitude of other reasons, I am truly grateful to him.

I deeply value the insights and advice I've received from Prof. Puneet Gupta on a wide variety of topics, particularly in helping me understand the broader context of device engineering within real circuits and systems; my conversations with him and his students through our research groups' collaboration have been a fantastic learning experience for me. I'd also like to thank him for agreeing to be on my dissertation committee on short notice. I thank Prof. Oscar Stafsudd and Prof. Vidvuds Ozolins for their help and for taking the time to serve on my dissertation committee. I also thank Prof. Kang Wang for being on my qualifying exam committee.

I have been fortunate to be part of a lab filled with fellow graduate students who have been great teachers (and co-learners), colleagues, and friends. I have learned much from and enjoyed getting to know all of my labmates over the years, including Ablai, Dingkun, Greg, Hyung Suk, Jorge, Kaveh, Kyeong-Sik, Kun-Huan, Raghav, Rowan, Song, Wuran, Xin, and Yufei. My graduate experience would be unimaginable without the deep discussions and time frittered away in their company. Greg has patiently endured my many maniacal monologues about the Spurs and unstintingly shared his common sense and knowledge of transistors, simulations, and the Chargers. Kaveh has been a constant sounding board and debate partner on just about every topic imaginable; I am continually stimulated and inspired by his creativity and curiosity. His and Yufei's efforts to educate me on the realities of biosensing have, I hope, not been completely in vain. Songtao Chen

was an able and willing helper during his summer in our lab. Outside our group, I have similarly benefited and drawn inspiration from continual discussions with many other fellow students. Out of so many educational and entertaining conversation partners, I've benefited in particular from knowing, and talking the ears off of, Ben, Zoe, Shaodi, and Kyoungwon.

Many of the Green's function calculations performed throughout this work were carried out using the Hoffman2 shared cluster provided by the UCLA Institute for Digital Research and Education (IDRE)'s Research Technology Group. I'd like to thank their staff, particularly Tajendra Singh, for assistance in getting my parallelized calculations off the ground. Most or all of the materials presented in Chapters 3, 4, 5, and 6 have been previously published in assorted IEEE and AIP journals[7, 8, 9, 10, 11, 12]; they are reproduced here with minor modifications in accord with the author rights agreements for the respective publications.

Thanking my family is both essential and inadequate in the face of all the experiences we've been through together. My father has been an example of integrity, intelligence, and Christ-centered living in many ways; his insights into scientific thinking and analysis are just the tip of the iceberg. My mother has taught me much about the importance and impact of a well-centered life. My sister Ruth has been friend, bully, compatriot, and teacher for many years and will be very disappointed at the quality of my writing in this thesis. My brother-in-law Stephen has always been supportive and somehow fit in with our kooky ways.

I know that all things work together for the good of those who love God. May my life and work be in accord with His will and glorify His name.

Vita

2008 B.S. (Physics), UCLA, Los Angeles, California.

2010 M.S. (Electrical Engineering), UCLA, Los Angeles, California.

Publications

Greg Leung, Andrew Pan, and Chi On Chui, "Junctionless Silicon and $In_{0.53}Ga_{0.47}As$ Transistors – Part II: Device Variability From Random Dopant Fluctuation," *IEEE Trans. Electron Devices*, to be published, 2015.

Andrew Pan, Greg Leung, and Chi On Chui, "Junctionless Silicon and $In_{0.53}Ga_{0.47}As$ Transistors – Part I: Nominal Device Evaluation With Quantum Simulations," *IEEE Trans. Electron Devices*, to be published, 2015.

Andrew Pan and Chi On Chui, "Gate-Induced Source Tunneling FET (GISTFET)," *IEEE Trans. Electron Devices*, vol. 62, pp. 2390, 2015.

Andrew Pan and Chi On Chui, "Modeling Source-Drain Tunneling in Ultimately Scaled III-V Transistors," *Appl. Phys. Lett.*, vol. 106, pp. 243505, 2015.

Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta, "PROCEED: A Pareto Optimization-Based Circuit-Level Evaluator for Emerging Devices," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, to be published, 2015.

Andrew Pan and Chi On Chui, "Modeling Direct Interband Tunneling. II. Lower-Dimensional Structures," *J. Appl. Phys.*, vol. 116, pp. 54509, 2014.

Andrew Pan and Chi On Chui, "Modeling Direct Interband Tunneling. I. Bulk Semiconductors," *J. Appl. Phys.*, vol. 116, pp. 54508, 2014.

Shaodi Wang, Andrew Pan, Chi On Chui, and Puneet Gupta, "PROCEED: A Pareto Optimization-based Circuit-level Evaluator for Emerging Devices," in *Proc. ASP-DAC 2014*, pp. 818-824.

Andrew Pan and Chi On Chui, "RF Performance Limits of Ballistic Si Field-Effect Transistors," in *Proc. SiRF 2014*, pp. 68-70 [invited talk].

Andrew Pan, Songtao Chen, and Chi On Chui, "Electrostatic Modeling and Insights Regarding Multigate Lateral Tunneling Transistors," *IEEE Trans. Electron Devices*, vol. 90, pp. 2712-2720, Sept. 2013.

Kun-Huan Shih, Andrew Pan, Yang Liu, and Chi On Chui, "A systematic approach for hydrodynamic model calibration in the quasi-ballistic regime," *Solid-State Electron.*, vol. 87, pp. 90-97, Sept. 2013.

Shaodi Wang, Greg Leung, Andrew Pan, Chi On Chui, and Puneet Gupta, "Evaluation of Digital Circuit-Level Variability in Inversion-Mode and Junctionless FinFET Technologies," *IEEE Trans. Electron Devices*, vol. 60, pp. 2186-2193, July 2013.

Andrew Pan and Chi On Chui, "A Quasi-Analytical Model for Double-Gate Tunneling Field-Effect Transistors," *IEEE Electron Device Lett.*, vol. 33, pp. 1468-1470, Oct. 2012.

Kyeong-Sik Shin, Andrew Pan, and Chi On Chui, "Channel Length Dependent Sensitivity

of Schottky Contacted Silicon Nanowire Field-Effect Transistor Sensors," *Appl. Phys. Lett.*, vol. 100, pp. 123504, 2012.

Andrew Pan, Dee-Son Pan, and Chi On Chui, "New Mechanism for Excess Noise in Mixed Tunneling and Avalanche Breakdown of Silicon," *Appl. Phys. Lett.*, vol. 96, pp. 263503, 2010.

CHAPTER 1

Introduction

> It's all right, Colin. Sit down. We're going to tunnel.
>
> *The Great Escape*

The enormous unprecedented scope of modern semiconductor devices enables a wide scale of technologies such as the vast data centers powering the Internet, ubiquitous personal electronic devices like smartphones, and the continuously expanding range of "smart" appliances and sensors used in personal, healthcare, and industrial settings. The remarkable capabilities of electronics, however, are accompanied by an equally remarkable level of energy consumption, leading to major environmental and economic costs. There is an urgent need to develop "green" technologies to sustain the existing (and expanding) technological infrastructure while reducing energy consumption. At the most basic level this requires electronic devices, and in particular transistors, which consume less power by minimizing operating voltages and leakage currents. Conventional devices like the metal-oxide-semiconductor field-effect transistor (MOSFET) face fundamental limitations on their ability to reduce these quantities, so new device physics and engineering designs are necessary to overcome this challenge. As a result, much recent activity has been devoted to devices not limited by thermal processes, among which the tunneling FET (TFET) has emerged as particularly promising. While experimental progress has occurred for this concept, many theoretical aspects of the device operation and design remain under investigation.

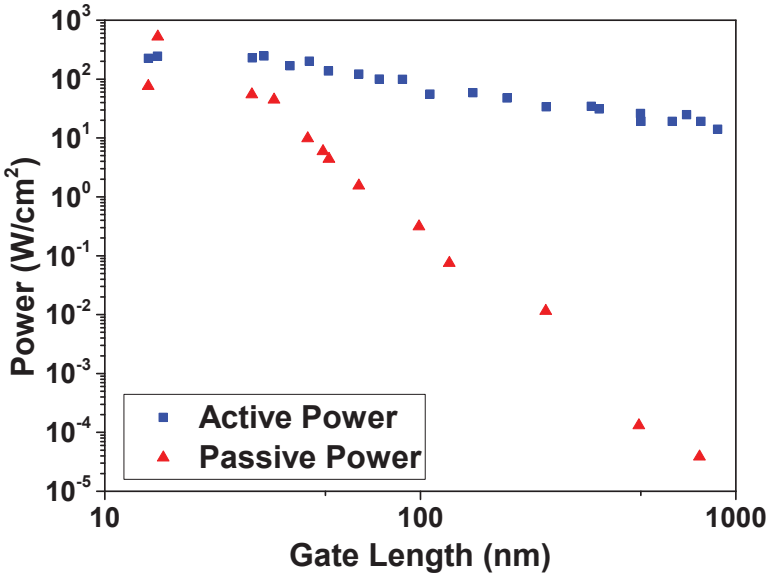

Figure 1.1: Scaling trends of active and passive power density with gate length in modern CMOS technologies. Data points taken from Ref. [1].

1.1 Power Consumption with Scaling

The technological trend for power dissipation in modern CMOS systems is shown in Fig. 1.1 and can be basically understood from the simple formula

$$P_{diss} = \alpha C_{eff} V_{dd}^2 f + I_{leak} V_{dd} \tag{1.1}$$

where α is the activity factor, C_{eff} is the load capacitance, V_{dd} is the supply voltage, f is the operating frequency, and I_{leak} is the leakage current in the off-state[1]. For circuits dominated by active dissipation during switching, the quadratic dependence of power on V_{dd} promises significant power savings via voltage scaling. In fact, because the system frequency and capacitance also have some voltage dependence, the power-voltage relationship may be cubic or greater in practice[2]. In low activity circuits such as ultralow power or remote systems, passive Joule heating represented by the second term of Eq. 1.1 may play a more significant role; here again, reducing V_{dd} and I_{leak} (which also scales with voltage) is key.

In traditional Dennard scaling, supply voltages and device dimensions (among other

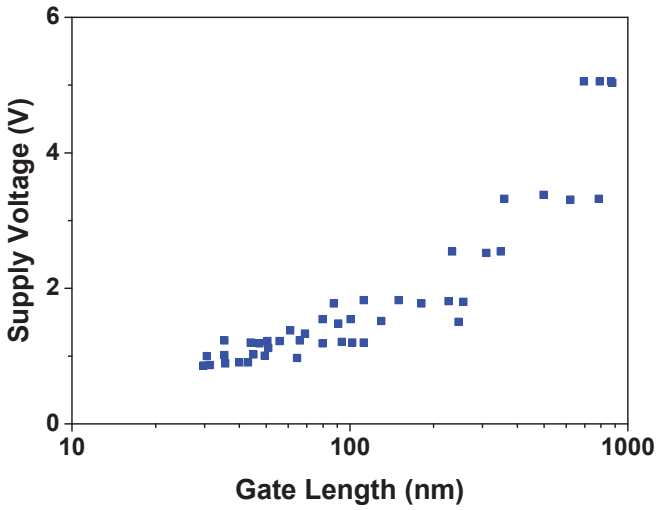

Figure 1.2: Operating voltage dependence on gate length in modern CMOS technologies. Data points taken from Ref. [2].

parameters) are each reduced in tandem by a factor κ, allowing the electric field magnitudes to remain constant inside the device[13]. In the ideal case, this leads to a reduction of power dissipation per circuit by κ^2 and hence maintains constant power density across the chip, independent of technology generation. Unfortunately, as Fig. 1.1 shows, this is not the case in real systems, where both active and passive power are increasing at an unsustainable rate as device dimensions shrink. This is because threshold voltages and subthreshold currents do not scale linearly in accord with other device parameters, constraining voltage scaling; in fact, operating voltages has scarcely decreased in high performance CMOS since technology nodes reached 100 nm, as illustrated in Fig. 1.2[1, 2]. This basic bottleneck in MOSFET-based technologies is due to the non-scalability of the subthreshold swing (SS). SS is equal to the change in gate voltage required to increase the subthreshold current by an order of magnitude; in traditional MOSFETs it is fundamentally limited to be $ln(10)kT/q = 60$ mV/decade or greater[14]. This is because the subthreshold diffusion current is driven by the "Boltzmann tail" of the source carrier population $\sim \exp(E/kT)$ with energies greater than the top of the barrier in the channel. The end result is that $SS = \dfrac{dV_{gs}}{d\log_{10} I_d}\ \ln(10)mkT/q = 60m$ mV/dec at room tempera-

ture, where $m = 1 + \frac{C_{dep}}{C_{ox}}$ describes the effective voltage drop across the channel. This basic limitation on the SS prevents the operating voltage of MOSFETs from being scaled in tandem with other device properties without either drastically increasing off-state current or reducing on-state current. In combination with other device factors, including increased variability in scaled devices due to random dopant fluctuations, this inhibits voltage scaling as devices shrink[2].

To overcome this fundamental limitation, new transport mechanisms must be employed besides carrier drift and diffusion. This has led to much research in so-called "steep swing" transistors which can have SS below 60 mV/decade and hence achieve good I_{on}/I_{off} ratios at lower voltages. Many such device candidates have been proposed, including the I-MOS (impact ionization MOS)[15], NEMS (nanoelectromechanical) switches[16], and negative capacitance FETs[17], which use avalanche breakdown, mechanical contact, and ferroelectric free energy instabilities, respectively, to achieve steep swings. Among a plethora of such candidates, interband tunneling field-effect transistors (TFETs) have emerged as especially promising because of their potential for low voltage operation and compatibility with common semiconductor technologies[18].

1.2 The Promise and Status of Tunneling FETs as a Low Power Device Alternative

In its most common incarnation, the TFET operates as a gated p-i-n diode where carrier transport occurs by interband tunneling between the source and channel. As illustrated in Fig. 1.3, the channel potential is controlled by the gate and forms an effective p-n junction with the source. For an n-type device in the off-state, the channel conduction band is higher than the valence band in the source, preventing tunneling from occurring. With increased gate bias, the channel conduction band is pulled below the source valence band edge, such that interband tunneling occurs and the device turns on. P-type operation can be realized by reversing the polarity of the device doping. Because the tunneling

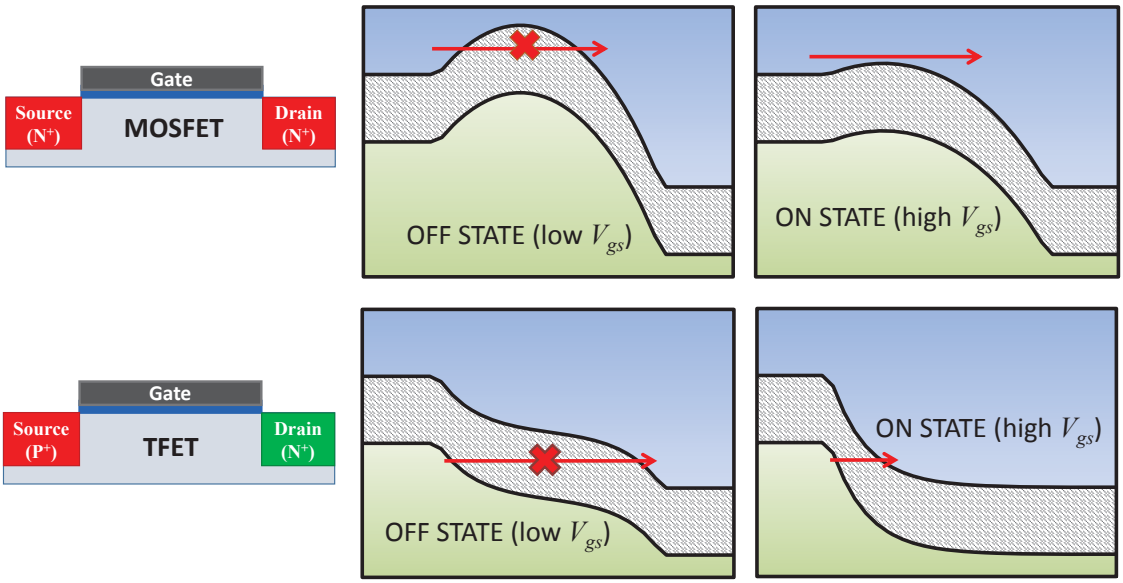

Figure 1.3: Schematic device structure and representative band diagrams in the off and on states for MOSFETs (top) and TFETs (bottom).

probability is a sharp function of electric field and the high-energy Boltzmann tail of the source distribution does not contribute to the tunneling current at low bias, it is possible for TFETs to attain subthreshold swings significantly below $\ln 10kT/q$ mV/decade[19].

The concept of tunneling transistors has a long history, dating back to early efforts to create tunable Esaki triodes using the field effect[20]. The idea of modulating interband tunneling using a gate seems to have been proposed and rediscovered several times, usually as a curiosity or with the goal of realizing negative differential conductance (NDC)[21, 22, 23, 24]. Perhaps the first indication of the true potential of TFETs came in 2004, with the experimental observation of steep SS in carbon nanotube (CNT) FETs biased to induce interband tunneling[25]. Around the same time, simulation studies reported the possibility of sub-60 mV/dec swing in silicon TFETs[26]. These findings generated intense excitement and extensive research into TFETs, leading to a number of experimental reports of steep SS in silicon- and germanium-based devices[27, 28, 29, 30, 31, 32, 33]. Unfortunately, thus far the steep SS in such devices has invariably been accompanied by low currents. Amusingly (and disappointingly), to the author's knowledge the highest

experimental on-current in a steep SS group IV TFET to date was reported in the first demonstration of such a device[27]. In group IV semiconductors, the relatively large indirect band gaps and effective masses in these materials are likely to reduce the tunneling probability and hence limit the achievable current drive, obviating any obvious technology advantages in terms of CMOS compatibility.

To overcome these limitations, TFETs based on III-V materials and/or heterojunctions have become more widely explored owing to their smaller and adjustable direct band gaps and effective masses, which should substantially increase the tunneling probability. Theoretical predictions of improved performance in such devices abound. Experimentally, many studies have shown larger on-currents in all-III-V TFETs compared to silicon, but most have failed to observe notably steep SS [34, 35, 36, 37]. Possibly the most promising experimental reports have come in InGaAs-based TFETs which have achieved minimum subthreshold swing of 60 mV/dec[6] and 64 mV/dec[5], the latter in a planar CMOS-like process. Several InGaAs/Si or InAs/Si heterojunction TFETs with steep SS (as low as 30 mV/decade at very low currents) have also appeared, though invariably with minuscule on-currents[38, 39]. A InP-GaAs heterostructure nanowire TFET has also been measured to have SS below 50 mV/dec, though the strong temperature dependence of SS in the device suggests that the subthreshold current may be due to something other than direct interband tunneling[40].

The underwhelming results of experimental TFET studies thus far has been attributed to a variety of complications, such as large leakage currents stemming from material quality issues, band tails due to random dopant fluctuations in the heavily doped source, or insufficiently optimized tunnel junction electrostatics due to dopant profile and oxide and channel geometric scaling limitations. Experimental evidence for interface trap-associated limitations has come via pulsed biasing $I - V$ measurements of silicon[41] and III-V devices[37] where subthreshold swings steeper than the DC values have been measured, presumably because the measuring frequency is faster than the response time of the trap states. By contrast, a study of silicon nanowire TFETs found that leakage currents scale with the device cross-sectional area rather than circumference, suggesting that bulk traps

dominate the parasitics[42]. The actual mechanisms limiting each experimental device are likely to be heavily dependent on design and processing. In principle, it should be possible to minimize these effects and realize the full potential of TFETs, yet it is clear that substantial progress remains to be made experimentally before the promise of these devices can be fulfilled.

It is also important to note that whereas n- and p-type devices have been demonstrated in silicon and germanium, in III-V materials only n-type TFETs have been experimentally demonstrated. This is likely due to the requirement for heavily n-doped regions in the source of p-type devices, which is difficult to realize due to the limited donor solid solubilities in common III-V semiconductors[43]. In addition to different materials, a variety of structural designs have been explored for TFETs beyond the lateral tunneling structure illustrated in Fig. 1.3. For instance, to boost current, vertical TFETs have been proposed wherein tunneling occurs perpendicular to the oxide-semiconductor interface[44, 45, 46, 47]; in such devices, the effective tunneling area is enlarged, increasing the drive current. Although some experimental devices have adopted this configuration, surface field-induced quantization and fabrication complexity pose possible significant challenges[48].

At present, experimental TFETs are still heavily constrained by performance compared to conventional devices. We have shown this in our collaborative work with Prof. Puneet Gupta, where we developed a device-circuit evaluation methodology, Pareto-Optimization-Based Circuit-Level Evaluator of Emerging Devices (PROCEED), which can cross-compare the performance of systems constructed using alternative device technologies [3]. For TFETs this draws on the analytical device modeling work discussed later in this thesis. In Fig. 1.4, when we compare the microprocessor-level performance of current state-of-the-art silicon TFET data[32] with commercial conventional SOI technology. It is evident that current TFETs can only outperform conventional devices in low performance (high delay) settings. This may be adequate for applications in certain ultralow power systems, such as remote sensors. If the challenges outlined above can be addressed, however, much greater advances in performance and a concomitantly wider

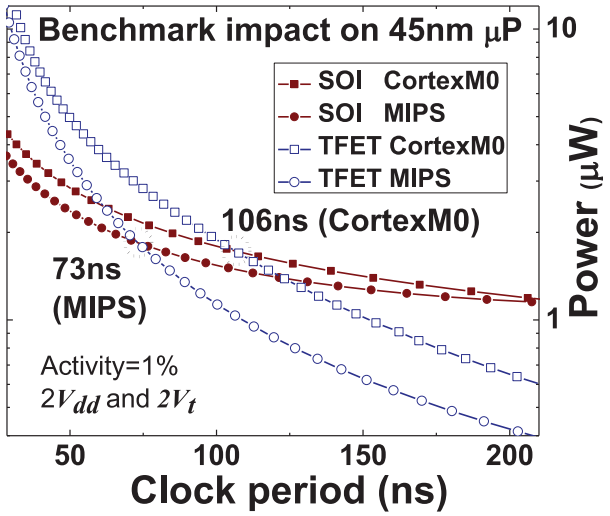

Figure 1.4: Optimized power-delay Pareto curves for MIPS and CortexM0 microprocessor designs implemented using current silicon experimental TFET and SOI characteristics, respectively, computed using PROCEED[3].

range of applications can be expected. Indeed, at present TFETs are still considered the most likely alternative device paradigm to conventional CMOS in the near future[49].

1.3 Outline of Work in this Dissertation

Despite the outpouring of experimental and simulation work in the field, many important questions remain about the fundamental operation of TFETs and the tunneling process. A solid theoretical understanding of these issues is critical for the development of optimized tunneling-based devices for low power and is the focus of this dissertation. Many simple models of tunneling currents are semiclassical in origin; however, the phenomenon is fundamentally quantum mechanical, and therefore we must adopt a formalism that reflects this fact. This would provide us with a rigorous basis for evaluating tunneling phenomena, as well as a standard by which simpler, approximate methods may be judged. The method of non-equilibrium Green's functions (NEGF) provides a natural starting point since it is the most general nonequilibrium quantum theory available. NEGF has also emerged as an increasingly popular method for studying device physics as quantum effects have

become more apparent in the latter. In Chapter 2, we discuss the basic theory of NEGF and the approximations and algorithms which are used to make it suitable for device simulations. Based on this methodology, we create a self-consistent quantum transport simulator program with which we will study various tunneling problems.

In Chapter 3 we analyze and validate traditional direct interband tunneling theories, which were derived via perturbation or semiclassical Wentzel-Kramers-Brillouin (WKB) arguments, by comparing their predictions with those given by NEGF calculations. The effects of material band structure and quantum size confinement (for tunneling in quantum wells and wires) are considered in detail. We clarify confusion in the literature regarding the range of validity of the semiclassical theory (which is widely used in device modeling and commercial simulation tools), and show that the main sources of error in the model lies in neglect of spin-orbit coupling and the nonparabolicity of transverse states. We derive simple new analytical formulas which incorporate these effects and yield good quantitative agreement with NEGF calculations for different materials and a wide range of electric fields. We also present a simple procedure, band gap scaling (BGS), which enables simple, quantitatively accurate evaluation of tunneling in confined structures. The BGS scheme is conceptually buttressed by arguments regarding the nature of the wave functions and complex band structure of confined materials. Taken together, these results provide a set of physically justified and quantitatively accurate analytical models for interband tunneling in direct gap materials. These results enable predictive calculations of tunneling currents in analytical models and widely-used device simulation tools. The work in this chapter is adapted from our prior journal publications[10, 9].

As TFET devices evolve, physically motivated design insights as well as sophisticated circuit-level understanding of the potential role of TFETs in real systems will be important in making them commercially useful technologies. Developing simple and accurate TFET current models is critical for both purposes, and our work in this area is documented in Chapter 4. We use analytical approximations to develop a general bias-dependent model of the electrostatic potential in lateral TFETs. The general formalism applies equally to ultrathin body, double-gate, and nanowire TFETs, among other structures, and can be

extended to consider various effects like graded doping junctions. We verify the accuracy of the model via extensive comparisons with numerical simulations and use it to derive the first quantitatively accurate analytical TFET $I-V$ model, showing the dependence of the electrical characteristics on device geometry, doping, and band structure. The resulting model elucidates the device physics and provides a foundation for compact models used in SPICE circuit-level simulations. Much of the work in this chapter is adapted from our prior journal publications[9, 7, 8].

Our device analysis helps demonstrate the critical importance of the TFET source doping concentration and abruptness for high device performance. Unfortunately, such chemical doping profiles are difficult to achieve repeatedly, particularly for high donor concentrations needed in p-type TFETs made of III-V materials. This helps account for the near-total absence of the latter in experimental TFET demonstrations to date. In Chapter 5, we propose a possible solution to this problem by introducing a new device concept, the gate-induced source TFET (GISTFET). Rather than relying on a chemical doping profile to achieve short tunneling distances, the GISTFET utilizes electrostatic doping from a metal gate heterojunction to form abrupt tunnel junctions. We explain the basic principle of the device and demonstrate its performance advantages for p-type TFETs using NEGF simulations. Our findings show that the GISTFET is a promising new direction for low power complementary logic devices.

The methods of analysis we have developed are not constrained to interband tunneling devices, and in Chapter 6 we apply our models and insights to study source-drain tunneling currents in MOSFETs, which is expected to become the limiting leakage source in sub-10 nm transistors. Using multiband NEGF simulations of III-V double gate FETs, we demonstrate that band nonparabolicity is quantitatively important even in intraband tunneling. We generalize the BGS procedure to describe this effect and present the first analytical model for source-drain tunneling, including nonparabolic corrections. We validate our model using NEGF simulations and show that the source/drain doping concentration in future MOSFETs must be carefully optimized to control parasitic tunneling leakage currents. Our results are important for analytical design and modeling of sub-10

nm MOSFETs.

Finally, in Chapter 7 we summarize the main findings of this dissertation and comment on some possible directions for future exploration.

CHAPTER 2

Quantum Transport Modeling using Non-Equilibrium Green's Functions

I do not care for dirty greens
By any means.

Gilbert and Sullivan, *Patience*

Traditionally, electronic transport in semiconductors has been explored within a semiclassical framework based on the Boltzmann transport equation (BTE), where a local distribution function in real and momentum space is assumed and the electron dynamics and scattering rates reflect the material band structure. Depending on the problem, the BTE can be solved exactly using Monte Carlo methods or approximately in the form of drift-diffusion or hydrodynamic equations[50]. In this scheme, quantum effects are negligible (beyond specifying the single-particle band structure) when relevant length scales exceed the mean free path of carriers and intraband transport dominates. This method has had great success qualitatively and quantitatively explaining a variety of semiconductor-based devices. However, as device dimensions scale down and phenomena like quantum tunneling or size confinement become important, the assumptions underlying the semiclassical approach begin to break down, and a fundamentally quantum mechanical approach to transport is desirable.

Quantum mechanics provides the foundation for our microscopic understanding of the electronic properties of semiconductors. The development of effective Hamiltonians that accurately describe the electronic band structure of different materials have enabled analysis of the relevant ground state and linear response properties using the well-known

methods of quantum mechanics for closed systems[51]. However, active semiconductor devices like transistors are generally operated in highly nonequilibrium conditions and coupled to external "reservoirs" or contacts which are in equilibrium. Treatment of such systems requires a quantum kinetic theory of open systems. A variety of methods like the Pauli master equation, the Lindblad formalism, or the Wigner function have been used to applied to this problem[52]. Of these, the non-equilibrium Green's function (NEGF) formalism is particularly attractive because it is the most general known theory of non-equilibrium in interacting systems and therefore an excellent starting point for rigorous analysis. Practically, the formalism provides a well-defined way to include various kinds of interactions, including those with external contacts, within the system. Therefore we will construct a complete semiconductor device simulation program using NEGF which will be subsequently used throughout our work on quantum tunneling in devices.

The formalism of non-equilibrium Green's functions (NEGF) initially developed out of work on equilibrium statistical quantum mechanics and many-body theory. The nonequilibrium extension of these techniques was pioneered in the 1960s by Kadanoff and Baym[53] and Keldysh[54], building on Schwinger's work on real-time Green's functions[55]. A formulation suitable for finite spatially nonuniform systems was developed by Caroli and coworkers in the early 1970s[56, 57], motivated by the desire to provide a rigorous basis for the theory of tunneling in metal-insulator-metal (MIM) structures. Initial application of the technique to semiconductors focused on formalism-heavy attempts to derive corrections to the BTE in the case of high, uniform electric fields[58]. In the 1990s, the experimental development of mesoscopic physics, nanoscale semiconductor devices (particularly resonant tunneling diodes), and increase in computational capabilities led to the practical use of NEGF in simulating realistic structures[59, 60]. As a result, in recent years NEGF has become an increasingly popular methodology for studying a variety of electronic and photonic devices[61, 62]. In this chapter, we provide a brief introduction to the NEGF equations for steady state transport and explain the calculation framework for evaluating the transport quantities of interest.

2.1 Steady State NEGF Kinetic Equations

At the formal level, NEGF is the nonequilibrium generalization of the powerful techniques of equilibrium Green's function theory; a detailed account of the latter is beyond the scope of this work but can be found in many excellent references, such as [63, 64, 65]. The method is important in quantum many-body and field theories because it allows for a systematic approach to dealing with interacting systems which cannot otherwise be solved exactly. In general, the Hamiltonian H for such a system can be written in the form $H = H_0 + H_{int}$, where H_0 is the non-interacting Hamiltonian (which is exactly solvable analytically or numerically) and H_{int} contains the interacting terms whose effects are not easily diagonalized. The power of the Green's function formalism lies in its ability to systematically approximate the effects of H_{int}. For electrons, the key quantity is the Green's function in the Heisenberg representation

$$G(r_1, r_2, t_1, t_2) = -i\langle T(\psi(r_1, t_1)\psi^\dagger(r_2, t_2))\rangle \tag{2.1}$$

where T is the time-ordering operator, r_1 and r_2 are position coordinates, t_1 and t_2 are time coordinates, and $\psi^\dagger$ and ψ are the creation and annihilation operators, respectively. The solution of G leads to integration over real time at zero temperature, carefully accounting for the time-ordering operator. At finite temperature T, an ensemble average of the Green's function is required; formally this leads to an integration over imaginary time with respect to the density matrix $\rho(H) = \dfrac{\exp(-\beta H)}{Tr[\exp(-\beta H)]}$ where $\beta = 1/k_B T$ defines the temperature.

In non-equilibrium conditions, by contrast, temperature is undefined and the density matrix is in general unknown, making the conventional Green's function theory untenable. Kadanoff and Baym[53] and Keldysh[54] analyzed the problem by assuming that before some time t_0 the system was in equilibrium with inverse temperature β, and that subsequently the interaction is turned on and the system is non-equilibrium. This leads to an integral in time for the evaluation of the Green's function which is defined over a contour that runs from $i\beta$ at t_0 to the subsequent real times t_1 and t_2 at which observables are evaluated. The resulting Green's functions are structurally equivalent to their equilib-

rium counterparts, allowing the diagrammatic techniques of field theory to be used. (For steady state transport problems, correlations due to the initial state β can be neglected (i.e., $t_0 \to -\infty$) and we need only worry about proper ordering of the operators along the contour.) This leads to the Green's function

$$G(r_1, r_2, t_1, t_2) = -i\langle T_c(\psi(r_1, t_1)\psi^\dagger(r_2, t_2))\rangle \tag{2.2}$$

where T_c is the contour-ordering operator and $\psi^\dagger$ and ψ are the creation and annihilation operators, respectively, in the Heisenberg representation. The different possibilities for operator ordering on the time contour lead to six Green's functions (which are not all independent, but each of which have distinct advantages and meanings), which are called retarded (G^r), advanced (G^a), time-ordered (G^t), anti-time-ordered ($G^{\bar{t}}$), "lesser-than" ($G^<$), and "greater-than" ($G^<$) and are defined (analogously to their equilibrium counterparts)

$$G^<(r_1, r_2, t_1, t_2) = i\langle\psi^\dagger(r_2, t_2)\psi(r_1, t_1)\rangle \tag{2.3}$$

$$G^>(r_1, r_2, t_1, t_2) = -i\langle\psi(r_1, t_1)\psi^\dagger(r_2, t_2)\rangle \tag{2.4}$$

$$G^t(r_1, r_2, t_1, t_2) = \theta(t_1 - t_2)G^>(r_1, r_2, t_1, t_2) + \theta(t_2 - t_1)G^<(r_1, r_2, t_1, t_2) \tag{2.5}$$

$$G^{\bar{t}}(r_1, r_2, t_1, t_2) = \theta(t_2 - t_1)G^>(r_1, r_2, t_1, t_2) + \theta(t_1 - t_2)G^<(r_1, r_2, t_1, t_2) \tag{2.6}$$

$$G^r(r_1, r_2, t_1, t_2) = -i\Theta(t_1 - t_2)\langle\{\psi(r_1, t_1), \psi^\dagger(r_2, t_2)\}\rangle \tag{2.7}$$

$$G^a(r_1, r_2, t_1, t_2) = i\Theta(t_2 - t_1)\langle\{\psi(r_1, t_1), \psi^\dagger(r_2, t_2)\}\rangle \tag{2.8}$$

It can be verified that these Green's functions (and the corresponding self-energies) obey various relations with each other, such as $G^t + G^{\bar{t}} = G^< + G^>$ and $G^r - G^a = G^> - G^<$[65].

One of the most useful relations in Green's function theory is Dyson's equation, which links the bare (non-interacting) and dressed (interacting) Green's functions. Keldysh presented a convenient general matrix form of this equation which is valid both in and out of equilibrium:

$$\begin{aligned}\tilde{G}(r_1, r_2, t_1, t_2) = \tilde{G}_0(r_1, r_2, t_1, t_2) + \\ \int dr_3 dt_3 \int dr_4 dt_4 \tilde{G}_0(r_1, r_3, t_1, t_3)\tilde{\Sigma}(r_3, r_4, t_3, t_4)\tilde{G}(r_4, r_2, t_4, t_2)\end{aligned} \tag{2.9}$$

where

$$\tilde{G} = \begin{bmatrix} G^t & -G^< \\ G^> & -G^{\bar{t}} \end{bmatrix} \tag{2.10}$$

and the corresponding self-energies (which contain the interactions) are organized via

$$\tilde{\Sigma} = \begin{bmatrix} \Sigma^t & -\Sigma^< \\ \Sigma^> & -\Sigma^{\bar{t}} \end{bmatrix} \tag{2.11}$$

We use the convention that every G represents a dressed Green's function and G_0 its bare counterpart. For convenience, and also because such notation will lead naturally to the type of matrix equations that we seek to solve numerically, we will suppress the arguments and integrals of the various functions with the understanding that any product of two functions (G or Σ) involves integration over shared interior coordinates r and t. In this form, Eq. 2.9 becomes

$$\tilde{G} = \tilde{G}_0 + \tilde{G}_0 \tilde{\Sigma} \tilde{G} \tag{2.12}$$

Keldysh also pointed out that the interdependence of the Green's functions allows a canonical transformation of $\tilde{G}$ (the so-called Keldysh rotation[66]) into a form involving only three independent quantities, G^r, G^a, and the Keldysh Green's function $F \equiv G^< + G^>$ (with an analogous transform for $\tilde{\Sigma}$)[54]. The Keldysh rotation is more convenient in certain situations, but is of course fundamentally equivalent to using $\tilde{G}$ and $\tilde{\Sigma}$[66]. We will retain the current form of $\tilde{G}$ but individually transform G^t and $G^{\bar{t}}$ as necessary.

The equations of motion of the Green's functions describe the dynamics and kinetics of the system, so we examine their time evolution by operating on the Green's functions with $[i\hbar\partial/\partial t_1 - H_0]$[67]. Recalling that t_1 only appears in $\tilde{G}_0$, we obtain

$$\begin{aligned} \left[i\hbar \frac{\partial}{\partial t_1} - H_0(r_1) \right] \tilde{G}_0 &= \tilde{I} \\ \left[i\hbar \frac{\partial}{\partial t_1} - H_0(r_1) \right] \tilde{G} &= \tilde{I} + \tilde{\Sigma}\tilde{G} \end{aligned} \tag{2.13}$$

where $\tilde{I}$ is the 2×2 identity operator which also operates like a delta function $\delta(r_1 - r_2)\delta(t_1 - t_2)$. Let us concentrate on the time evolution of the retarded and lesser-than Green's functions. Eq. 2.13 for $G^<$ becomes

$$\left[i\hbar \frac{\partial}{\partial t_1} - H_0(r_1) \right] G^< = \Sigma^t G^< - \Sigma^< G^{\bar{t}}. \tag{2.14}$$

Substituting the known relations $\Sigma^t = \Sigma^r + \Sigma^<$ and $G^{\bar{t}} = -G^a + G^<$, we can rewrite this as

$$\left[i\hbar\frac{\partial}{\partial t_1} - H_0(r_1)\right] G^< = \Sigma^r G^< - \Sigma^< G^a. \tag{2.15}$$

The time evolution of the retarded Green's function can be obtained from that of G^t

$$\left[i\hbar\frac{\partial}{\partial t_1} - H_0(r_1)\right] G^t = I + \Sigma^t G^t + \Sigma^< G^>. \tag{2.16}$$

We use the identity $G^r = G^t - G^<$ and Eq. 2.15 to find

$$\left[i\hbar\frac{\partial}{\partial t_1} - H_0(r_1) - \Sigma_r\right] G^r = I. \tag{2.17}$$

We recall that the various Gs and Σs are functions of two time coordinates t_1 and t_2. If we restrict ourselves to steady-state problems, these functions depend only on the relative time interval $t_1 - t_2$, not the absolute times. Therefore we can Fourier transform such functions f to the energy domain (equivalent to frequency via $E = \hbar\omega$) to obtain

$$f(x, x', E) = \int \frac{d(t_1 - t_2)}{\hbar} e^{\frac{iE}{\hbar}(t_1 - t_2)} f(x, x', t_1 - t_2). \tag{2.18}$$

Integrals that involve the convolution of two functions in time become products of the corresponding Fourier transforms in energy. Since semiconductor devices are spatially inhomogeneous, transforming to momentum space offers no further advantages and we will remain in real space. Substituting the time evolution of the bare retarded Green's function G_0^r obtained from Eq. 2.13 in Eq. 2.17 and Fourier transforming, we obtain Dyson's equation of the retarded Green's function in a form familiar from equilibrium many-body theory

$$G^r = G_0^r + G_0^r \Sigma^r G^r \tag{2.19}$$

with

$$G_0^r = (E + i\delta - H_0)^{-1} \tag{2.20}$$

where δ is an infinitesimal quantity whose sign reflects the temporal retarded response of G_r. (The advanced Green's function G_0^a is given by the same equation with a negative sign in front of $i\delta$. In practice, it can be directly obtained from the retarded Green's

function via $G^a = G^{r\dagger}$.) For the lesser-than function, if we move $\Sigma_r G^<$ to the left-hand side of Eq. 2.15 and Fourier transform, we obtain

$$G^< = G^r \Sigma^< G^a. \tag{2.21}$$

The Dyson equation for the greater-than function can be similarly analyzed to find

$$G^> = G^r \Sigma^> G^a. \tag{2.22}$$

Eqs. 2.19 and 2.21 lie at the heart of the NEGF method. We can obtain a rough physical picture of their significance by considering the meaning of the various functions. If we refer to the definition of G^r in Eq. 2.3, we see that it describes the response of the system at (r_1, t_1) to an excitation at (r_2, t_2). Meanwhile, when $t_1 \to t_2$, $G^<$ and $G^>$ become the expectation values of the electron and hole density operators, respectively. In some sense, therefore, the retarded and advanced Green's functions encode the dynamics of the system while the lesser-than and greater-than functions describe the electron and hole distributions[54]. This partitioning should not be taken too literally, since the various Green's functions are implicitly coupled by the self-energies, but it can be useful in providing intuition. The self-energies then represent external interactions, such as with phonons or spatial electrodes, which inject or remove particles (physically in the case of contacts and between specific energies and momenta in the case of phonons). Eq. 2.21 indicates that changes in the distribution occur via excitations ($\Sigma^<$) which propagate through the system as determined by G^r and G^a. In this sense, Eq. 2.21 is a kinetic equation. Indeed, under certain approximations it may be used to derive the Boltzmann transport equation[53].

2.2 Evaluating the NEGF Equations for Semiconductor Devices

To apply Eqs. 2.19 and 2.21 to realistic semiconductor devices, we must first decide how to partition the system into bare and interacting sections[61]. This breakdown is illustrated in Fig. 2.1. At the simplest level, we may separate any device into an active region, where non-equilibrium electronic effects are important, and reservoirs or contacts to which the

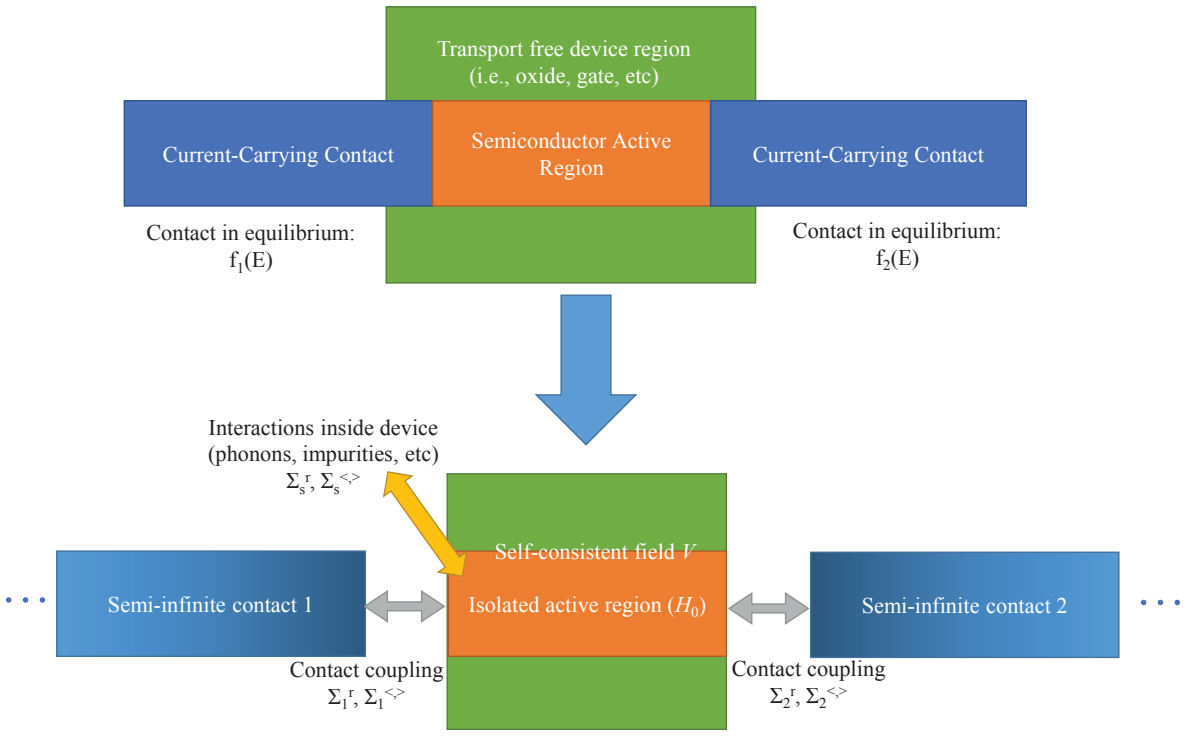

Figure 2.1: Conceptual division of a semiconductor device into constituent bare and interacting sections for NEGF analysis.

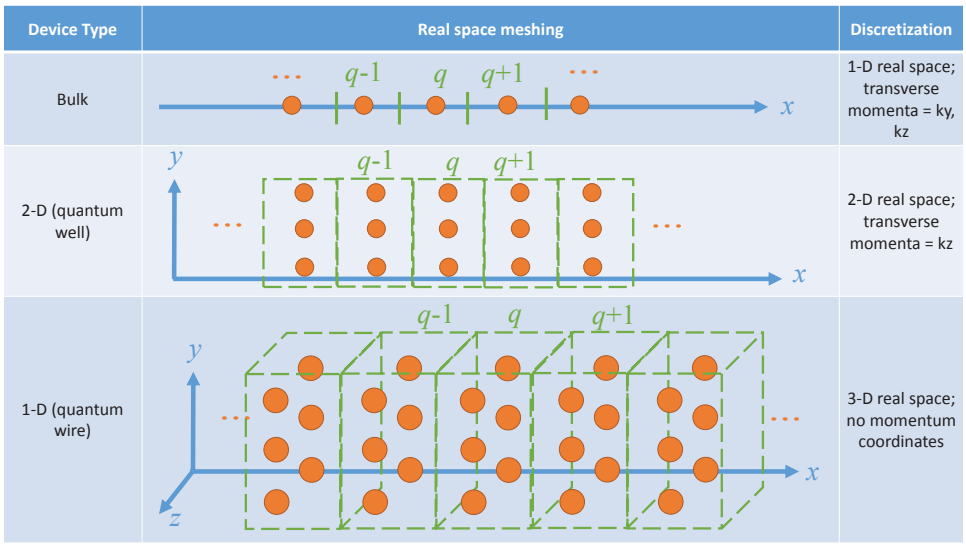

Figure 2.2: Schematic of real space gridding and layer partitioning definition for bulk, 2-D, and 1-D devices.

device is connected and through which carriers and applied biases are supplied. In the absence of dissipation or coupling to the contacts, the active region is treated as a closed system which can be modeled using a suitable Hamiltonian such as an effective mass, $\mathbf{k}\cdot\mathbf{p}$, or tight-binding (TB) model, along with a self-consistent potential V. Such a model can be solved exactly numerically and is thus treated as the non-interacting Hamiltonian H_0 from which the bare Green's function can be calculated. In this scheme, the coupling to the contacts (via hopping terms between the active and contact regions) is treated as an interaction which can be incorporated into the system via an appropriate self-energy[56]. Conceptually the contact may be regarded as a semi-infinite homogeneous strip of semiconductor with a well-defined distribution function (typically a fixed Fermi energy and temperature). If the contacts are assumed to be non-interacting, the coupling to the device can be included exactly using the surface Green's function of the semi-infinite contact. Any other interactions within the active region, via electron-phonon coupling, etc., can also be included perturbatively to various orders as additional self-energies[57, 60].

In practical situations, we will discretize the Hamiltonian in real space in the active

region and divide it into layers which are coupled with each other. Note that each layer may be a single discretized point or plane depending on the dimensions of the device; this is illustrated in Fig. 2.2. For instance, for a 2-D device structure, each layer spans the discretized region in the x axis and corresponds to a successive point along the y axis. In general, if the resulting device Hamiltonian couples only nearest neighbor layers, it will take the block tridiagonal matrix form

$$\begin{bmatrix} \ddots & & & & \\ \cdots & H_{q-1,q-1} & H_{q-1,q} & 0 & \cdots \\ \cdots & H_{q,q-1} & H_{q,q} & H_{q,q+1} & \cdots \\ \cdots & 0 & H_{q+1,q} & H_{q+1,q+1} & \cdots \\ & & & & \ddots \end{bmatrix}$$

where q denotes the layer index, $H_{q,q}$ is the "on-site" term for each layer (which may be a single element or a matrix block, depending on the dimensionality of the device and the band Hamiltonian), $H_{q,q\pm1}$ and $H_{q\pm1,q}$ are the coupling terms between adjacent layers, and all other terms are zero. Note that the general form of this matrix may be extended to n-nearest neighbor Hamiltonians by expanding the size of each Hamiltonian block $H_{q,q'}$ to span n mesh points. The size of each block depends on the number of basis states for each layer, as determined by the device dimension as well as the choice of band structure model. In the example of the 2-D system in Fig. 2.2, for N_x mesh points in the x direction, N_y mesh points in the y direction, and N_b basis states per lattice site in the chosen Hamiltonian, $H_{q,q'}$ will be of size $N_xN_b \times N_xN_b$ for $q, q' = 1, ...N_y$. An obvious numerical advantage of a block tridiagonal Hamiltonian is that the number of nonzero matrix elements which must be stored is reduced from $N_b^2N_x^2N_y^2$ to $3N_b^2N_x^2N_y$.

Given such a form for the device Hamiltonian, the Green's functions of interest are determined via the matrix equations derived in the previous section

$$((E + i\delta)I - H_0 - \Sigma^r)G^r = I \tag{2.23}$$

$$G^< = G^r\Sigma^<G^{r\dagger} \tag{2.24}$$

where I is the identity matrix. In general a block tridiagonal form for the Hamiltonian

does not lead to a corresponding structure for the Green's functions and the solution of Eq. 2.23, for example, requires matrix inversion of the term in parentheses (including the full H_0), a computationally costly task. If we impose the condition that the self-energy matrices Σ be block diagonal, however, we will find that most of the practically relevant quantities (DOS, current, electron density, etc.) can be extracted using only the block tridiagonal elements of the various Green's functions. In fact, these elements can be computed recursively using only matrices of the size of each block, which can lead to significant savings in computational processing and memory requirements. We will therefore use the so-called recursive Green's function algorithm and explain it in more detail below.

2.2.1 Defining the Contact Self-Energies

At the spatial boundaries of the active region (i.e., for layer index $q = 1, N$ for the left-most and right-most layers in the geometries of Fig. 2.2), we include the effects of carrier contacts by self-energies. To show this we follow the argument in Ref. [68]. This can be seen by considering the retarded Green's function for the infinite system including the contacts

$$((E+i\delta)I - H_0 - \Sigma^r)G^r = I \rightarrow \begin{bmatrix} A_{LL} & A_{LD} & O \\ A_{DL} & A_{DD} & A_{DR} \\ O & A_{RD} & A_{RR} \end{bmatrix} \begin{bmatrix} G^r_{LL} & G^r_{LD} & G^r_{LR} \\ G^r_{DL} & G^r_{DD} & G^r_{DR} \\ G^r_{RL} & G^r_{RD} & G^r_{RR} \end{bmatrix} = \begin{bmatrix} I & O & O \\ O & I & O \\ O & O & I \end{bmatrix} \tag{2.25}$$

where the subscripts L, D, R represent the semi-infinite left contact, active device region, and semi-infinite right contact regions, respectively, and the matrix blocks A represent the real-space discretized form of $(E + i\delta)I - H_0 - \Sigma^r$. The only assumption here is that the finite range of A leads to no direct coupling between the left and right regions. Writing

out the Green's function blocks directly we obtain

$$G^r_{LD} = -A^{-1}_{LL} A_{LD} G^r_{DD} \tag{2.26}$$

$$G^r_{RD} = -A^{-1}_{RR} A_{RD} G^r_{DD} \tag{2.27}$$

$$A_{DL} G^r_{LD} + A_{DD} G^r_{DD} + A_{DR} G^r_{RD} = I \tag{2.28}$$

from which the Green's function in the active region, which is the quantity we are interested in, becomes

$$[A_{DD} - A_{DL} A^{-1}_{LL} A_{LD} - A_{DR} A^{-1}_{RR} A_{RD}] G^r_{DD} = I \tag{2.29}$$

We note that $A^{-1}_{LL} = G^r_{LL,0}$ and $A^{-1}_{RR} = G^r_{RR,0}$, the Green's functions of the semi-infinite left and right contact regions in the absence of coupling to the device. Furthermore, if we assume that the self-energies are local and the Hamiltonian in A only couples nearest neighbors, the only nonzero terms in A_{DL}, A_{LD}, and their counterparts at the right side will be between the adjacent layers on the edge of the contacts and device, respectively. Hence, we only need to know $G^r_{LL,0}$ and $G^r_{RR,0}$ at the boundary of the contacts, i.e., the surface Green's functions $G^r_{s,L/R}$ for the isolated left and right contacts L and R. Once these are obtained, we can rewrite the second and third terms in the bracket in Eq. 2.29 as self-energies

$$\Sigma^r_{surf,L} = A_{10} G^r_{s,L} A_{01} \tag{2.30}$$

$$\Sigma^r_{surf,R} = A_{N,N+1} G^r_{s,R} A_{N,N+1} \tag{2.31}$$

where $A_{q,q'} = -H_{q,q'}$ are the off-diagonal coupling blocks of the leftmost and rightmost layers of the device Hamiltonian. Σ^r_{surf} only couple to the edge layers of the device, i.e., are nonzero only for $q = 1, N$.

This demonstrates that the interaction between the reservoir and active region can be calculated exactly assuming a nearest neighbor Hamiltonian describes the contact region. The isolated surface Green's function for each contact can be calculated exactly using the Sancho Lopez decimation scheme[69], which will now be described. Assume that the contact region is described by a Hamiltonian with nearest-neighbor interactions such that

H_{00} is the on-site Hamiltonian term and H_{10} and H_{01} are the couplings to adjacent layers on the left and right, respectively (note that the mathematical structure is unchanged if H_{00} includes a local self-energy). For specificity, let us calculate the surface Green's function for an semi-infinite slab in the right direction that is terminated on the left at $n = 0$; this corresponds to a right-side contact in the geometry of Fig. 2.2. We may then write

$$\begin{aligned}(E + i\delta - H_{00})G^r_{00} &= I + H_{01}G^r_{10} \\ (E + i\delta - H_{00})G^r_{10} &= H_{10}G^r_{00} + H_{01}G^r_{20} \\ &\vdots \\ (E + i\delta - H_{00})G^r_{n0} &= H_{10}G^r_{n-1,0} + H_{01}G^r_{n+1,0}\end{aligned} \tag{2.32}$$

for the retarded Green's function G^r of the isolated contact region. Rewriting the last line as

$$G^r_{n0} = (E + i\delta - H_{00})^{-1}(H_{10}G^r_{n-1,0} + H_{01}G^r_{n+1,0}), n \geq 1 \tag{2.33}$$

and substituting recursively for $G^r_{n+1,0}$ and $G^r_{n-1,0}$ we find

$$\begin{aligned}\left[E + i\delta - H_{00} - H_{01}(E + i\delta - H_{00})^{-1}H_{10} - H_{10}(E + i\delta - H_{00})^{-1}H_{01}\right] G^r_{n0} = \\ H_{10}(E + i\delta - H_{00})^{-1}H_{10}G^r_{n-2,0} + H_{01}(E + i\delta - H_{00})^{-1}H_{01}G^r_{n+2,0}\end{aligned} \tag{2.34}$$

for $n \geq 2$. We observe that the coupling is now only between next nearest sites ($n \pm 2$). If we define the parameters

$$\alpha_1 = H_{01}(E + i\delta - H_{00})^{-1}H_{01} \tag{2.35}$$

$$\beta_1 = H_{10}(E + i\delta - H_{00})^{-1}H_{10} \tag{2.36}$$

$$\epsilon_{1s} = H_{00} + H_{01}(E + i\delta - H_{00})^{-1}H_{10} \tag{2.37}$$

$$\epsilon_1 = H_{00} + H_{01}(E + i\delta - H_{00})^{-1}H_{10} + H_{10}(E + i\delta - H_{00})^{-1}H_{01} \tag{2.38}$$

we can now rewrite the relationship between evenly separated Green's functions as

$$\begin{aligned}(E - \epsilon_{1s})G^r_{00} &= I + \alpha_1 G^r_{20} \\ (E - \epsilon_1)G^r_{n0} &= \beta_1 G^r_{n-2,0} + \alpha_1 G^r_{n+2,0} \\ (E - \epsilon_1)G^r_{nn} &= I + \beta_1 G^r_{n-2,n} + \alpha_1 G^r_{n+2,n}\end{aligned} \tag{2.39}$$

If we consider only even values of n, we can further rewrite the chain coupling equations as

$$\begin{aligned}(E-\epsilon_{1s})G^r_{00} &= I+\alpha_1 G^r_{20}\\ (E-\epsilon_1)G^r_{2n,0} &= \beta_1 G^r_{2(n-1),0}+\alpha_1 G^r_{2(n+1),0}\\ (E-\epsilon_1)G^r_{2n,2n} &= I+\beta_1 G^r_{2(n-1),2n}+\alpha_1 G^r_{2)(n+1),2n}\end{aligned} \tag{2.40}$$

Formally, these equations are identical to Eqs. 2.32 with on-site and coupling Hamiltonian matrix elements renormalized via Eqs. 2.35; since only next nearest neighbors are coupled, the effective lattice spacing of the Hamiltonian has doubled. We can iterate this procedure to obtain after the i^{th} renormalization

$$\alpha_i = \alpha_{i-1}(E-\epsilon_{i-1})^{-1}\alpha_{i-1} \tag{2.41}$$

$$\beta_i = \beta_{i-1}(E-\epsilon_{i-1})^{-1}\beta_{i-1} \tag{2.42}$$

$$\epsilon_{is} = \epsilon_{i-1,s}+\alpha_{i-1}(E-\epsilon_{i-1})^{-1}\beta_{i-1} \tag{2.43}$$

$$\epsilon_i = \epsilon_{i-1}+\alpha_{i-1}(E-\epsilon_{i-1})^{-1}\beta_{i-1}+\beta_{i-1}(E-\epsilon_{i-1})^{-1}\alpha_{i-1} \tag{2.44}$$

where $\epsilon_0 = H_{00}-i\delta$, $\alpha_0 = H_{01}$, and $\beta_0 - H_{10}$. These coupling coefficients solve the equations

$$\begin{aligned}(E-\epsilon_{i,s})G^r_{00} &= I+\alpha_i G^r_{2^i n,0}\\ (E-\epsilon_i)G^r_{2^i n,0} &= \beta_i G^r_{2^i(n-1),0}+\alpha_i G^r_{2^i(n+1),0}\\ (E-\epsilon_i)G^r_{2^i n,2^i n} &= I+\beta_i G^r_{2^i(n-1),2n}+\alpha_i G^r_{2^i(n+1),2n}\end{aligned} \tag{2.45}$$

In the limit of large i, α_i and β_i approach zero and the final solutions for the on-site surface and bulk Green's functions G^r_s and G^r_b become

$$G^r_{00} = (E-\epsilon_{i,s})^{-1} = G^r_s \tag{2.46}$$

$$G^r_{2^i n,2^i n} = (E-\epsilon_i)^{-1} = G^r_b \tag{2.47}$$

For the opposite case of a slab terminated on the left at $n=0$, the corresponding surface Green's function is simply shown to be obtained using the identical procedure with the substitution

$$\epsilon_{is} = \epsilon_{i-1,s}+\beta_i(E-\epsilon_i)^{-1}\alpha_i. \tag{2.48}$$

In general we take H_{00}, H_{01}, and H_{10} from the left-most and right-most blocks of the device Hamiltonian and compute Eqs. 2.41 until the Euclidean norms of α_i and β_i fall below a convergence criteria (for example, 10 neV), whereupon the surface Green's functions are computed using Eq. 2.46. The surface retarded self-energies are then found via

$$\Sigma^r_{surf} = (-H_{01})G^r_s(-H_{10}) \qquad \text{left-facing surface on right contact} \tag{2.49}$$

$$= (-H_{10})G^r_s(-H_{01}) \qquad \text{right-facing surface on left contact} \tag{2.50}$$

The lesser-than and greater-than self-energies can be formulated in terms of the retarded self-energy if we assume that the region is in equilibrium with a distribution function $f(E)$ to obtain

$$\Sigma^<_{surf}(E) = -(\Sigma^r_{surf}(E) - \Sigma^a_{surf}(E))f(E) = -(\Sigma^r_{surf}(E) - \Sigma^{r\dagger}_{surf}(E))f(E) \tag{2.51}$$

$$\Sigma^>_{surf}(E) = (\Sigma^r_{surf}(E) - \Sigma^a_{surf}(E))(1 - f(E)) = (\Sigma^r_{surf}(E) - \Sigma^{r\dagger}_{surf}(E))(1 - f(E)). \tag{2.52}$$

The Lopez-Sancho decimation scheme is analogous to the renormalization group technique [66], where an effective Lagrangian or Hamiltonian is repeatedly rescaled by integrating over short-distance/high-momenta degrees of freedom and the scaling behavior of the coupling coefficients in the self-similar system is studied for insight into critical phenomena. In this case we repeatedly subsume nearest-neighbor couplings into a rescaled Hamiltonian with a greater lattice constant and asymptotically obtain the decoupled bulk and surface properties. With the contact self-energy defined and the active region Hamiltonian chosen, we can then exactly solve for the various Green's functions in the device and obtain physical observables of interest.

2.2.2 Computing the Green's Functions Recursively

We will follow the discussion in [70] to illustrate the recursive Green's function algorithm. We again take advantage of the matrix representation to rewrite Eq. 2.23 for the device

retarded Green's function G^r as

$$\begin{bmatrix} A_{Z,Z} & A_{Z,Z'} \\ A_{Z',Z} & A_{Z',Z'} \end{bmatrix} G^r = \begin{bmatrix} I & O \\ O & I \end{bmatrix} \tag{2.53}$$

where $A = (E + i\delta)I - H_0 - \Sigma^r$ and we divide the full matrices in two sections spanned by Z and Z' (which do not have to be the same size). The solution to this matrix equation is given by the Dyson equation

$$G^r = G^{r0} + G^{r0}UG^r = G^{r0} + G^rUG^{r0} \tag{2.54}$$

where

$$G^r = \begin{bmatrix} G^r_{Z,Z} & G^r_{Z,Z'} \\ G^r_{Z',Z} & G^r_{Z',Z'} \end{bmatrix}, \tag{2.55}$$

$$G^0 = \begin{bmatrix} G^{r0}_{Z,Z} & G^{r0}_{Z,Z'} \\ G^{r0}_{Z',Z} & G^{r0}_{Z',Z'} \end{bmatrix} = \begin{bmatrix} A^{-1}_{Z,Z} & O \\ O & A^{-1}_{Z',Z'} \end{bmatrix}, \tag{2.56}$$

and

$$U = \begin{bmatrix} O & -A_{Z,Z'} \\ -A_{Z',Z} & O \end{bmatrix}. \tag{2.57}$$

We effectively treat the isolated Hamiltonians in Z and Z' as "bare" systems which are coupled by a perturbation U which is solved exactly via Dyson's equation. Now let q be the index of a given block. Then we define the left-connected retarded Green's function g^{rLq}

$$A_{1:q,1:q}g^{rLq} = I_{1:q,1:q} \tag{2.58}$$

Dyson's equation gives us

$$g^{rLq+1}_{q+1,q+1} = (A_{q+1,q+1} - A_{q+1,q}g^{rLq}_{q,q}A_{q,q+1})^{-1} \tag{2.59}$$

Using this relation, the diagonal component of the full retarded Green's function can be written as

$$G^r_{q,q} = g^{rLq}_{q,q} - g^{rLq}_{q,q}A_{q,q+1}G^r_{q+1,q} \tag{2.60}$$

and the off-diagonal components are found from the Dyson equation to be

$$G^r_{q+1,q} = -G^r_{q+1,q+1} A_{q+1,q} g^{rLq}_{q,q} \tag{2.61}$$

$$G^r_{q,q+1} = -g^{rLq}_{q,q} A_{q,q+1} G^r_{q+1,q+1} \tag{2.62}$$

Putting these relationships together, we see that we can calculate the tridiagonal blocks of the retarded Green's function in increasing order from $q = 1$ to N_y starting from the initial left-connected Green's function $g^{rL1}_{11} = A^{-1}_{1,1}$. The advanced Green's function can be directly obtained via the relation $G^a = G^{r\dagger}$.

Similar arguments can be applied to the steady-state Keldysh equation for the lesser-than Green's function once G^r is known. In particular, a left-connected lesser-than Green's function $g^{<Lq}$ can be defined which obeys

$$g^{<Lq+1}_{q+1,q+1} = g^{rLq+1}_{q+1,q+1}[\Sigma^<_{q+1,q+1} + A_{q+1,q} g^{<Lq}_{q,q} A^\dagger_{q,q+1} - \Sigma^<_{q+1,q} g^{aLq}_{q,q} A^\dagger_{q,q+1} - A_{q+1,q} g^{rLq}_{q,q} \Sigma^<_{q,q+1}] g^{aLq+1}_{q+1,q+1} \tag{2.63}$$

from which the relevant blocks of $G^<$ are obtained

$$\begin{aligned} G^<_{q,q} =& g^{<Lq}_{q,q} + g^{rLq}_{q,q} A_{q,q+1} G^<_{q+1,q+1} A^\dagger_{q+1,q} g^{aLq}_{q,q} - g^{<Lq}_{q,q} A^\dagger_{q,q+1} G^a_{q+1,q} - \\ & G^r_{q,q+1} A_{q+1,q} g^{<Lq}_{q,q} - g^{<0}_{q,q+1} A^\dagger_{q+1,q} G^a_{q,q} - G^r_{q,q} A_{q,q+1} g^{<0}_{q+1,q} \end{aligned} \tag{2.64}$$

$$\begin{aligned} G^<_{q+1,q} =& g^{r0}_{q+1,q+1} \Sigma^<_{q+1,q} g^{a0}_{q,q} - G^r_{q+1,q} A_{q,q+1} g^{<0}_{q+1,q} - G^r_{q+1,q+1} A_{q+1,q} g^{<Lq}_{q,q} - \\ & G^<_{q+1,q+1} A^\dagger_{q+1,q} g^{aLq}_{q,q} \end{aligned} \tag{2.65}$$

2.2.3 Scattering Self-Energies

Interactions not included in the structure Hamiltonian (which can be solved exactly) may be incorporated via appropriate self-energies constructed from the diagrams of various interaction processes. This may include scattering off phonons, impurities, or imperfections like alloy disorder or surface roughness, as well as electron-electron interactions beyond the mean field (Hartree) solution of the electrostatic potential in H_0. Typically it is impossible to include these interactions to infinite order and some approximation must be made; frequently only first-order proper self-energies are included, which amounts to the Born approximation. The exact form of each self-energy is determined by the nature

of the interaction; in general, it will involve convolution over the electron Green's functions. It can be shown that current conservation requires that the Green's functions and self-energies must be solved self-consistently with each other, i.e., any Green's function lines appearing in the interior of self-energy diagrams must be dressed; this leads to the so-called self-consistent Born approximation[60]. As an example, the first-order retarded self-energy for electron-phonon scattering is[71]

$$\begin{aligned}\Sigma^r_{e-ph}(\vec{k},E) = i\int \frac{d\omega}{2\pi}\sum_{\vec{q}} |M_{\vec{q}}|^2 [G^<(\vec{k}-\vec{q},E-\omega)D^r(\vec{q},\omega) \\ + G^r(\vec{k}-\vec{q},E-\omega)D^<(\vec{q},\omega) + G^r(\vec{k}-\vec{q},E-\omega)D^r(\vec{q},\omega)]\end{aligned} \tag{2.66}$$

where $M_{\vec{q}}$ is the $\vec{q}$-dependent first-order electron-phonon matrix element and $D^{r,<}$ are the retarded and lesser-than phonon Green's functions, respectively. A similar equation can be written for the lesser-than and greater-than self-energies, and if we assume that the phonon system stays in equilibrium D reduces to the bare phonon Green's functions with a Bose-Einstein distribution. The point we emphasize here is that $\Sigma^r_{e,ph}$ and its counterparts depend on the dressed G^r and $G^<$, so that the solution to this equation must be performed iteratively. In practice this means we first solve for the bare Green's functions and use them to obtain the Born approximation $\Sigma_{e,ph}$, which are then fed back to obtain the perturbed Gs, etc., until the whole process converges.

We note that in general the self-energy may be spatially nonlocal; for instance, if we Fourier transform Eq. 2.66 into real space we will find in general that the self-energy at a lattice site r involves a convolution over Green's functions spanning the entire device domain. Locality is maintained only if $M_{\vec{q}}$ is $\vec{q}$-independent, i.e., the interaction is spatially localized (this is evident if we recall that the Fourier transform of a delta function in real space is a constant in momentum space). As noted before, the solution of off-diagonal elements in G beyond nearest neighbors leads to substantially greater computational costs. For this reason, in practice NEGF-based simulations often enforce locality by ignoring the nonlocal terms in self-energies, even though the physical validity of this approximation is uncertain and depends on the interaction in question[62].

In our device simulator, (local) scattering self-energies can be specified and calculated

self-consistently to include decoherence and dissipation effects. These have been used, for instance, to study transport limitations in silicon and junctionless transistors. Scattering also cannot be neglected for incoherent tunneling processes such as interband transitions in indirect band gap semiconductors where a phonon or impurity interaction is required to supply the momentum difference between the initial and final states. However, the tunneling problems we will focus on in this thesis are coherent in nature and depend only on the band coupling of the structure Hamiltonian; therefore we will mostly neglect scattering and the construction of appropriate self-energies in what follows.

2.2.4 Computing Observables from the Green's functions

We obtain the local density of states (LDOS) from the retarded Green's function via

$$\mathrm{LDOS}(E,\vec{r}) = \sum_{\vec{k}} -\frac{ig}{2\pi V}\mathrm{Tr}(\mathrm{Im}(G^r(E,\vec{r},\vec{k}))) \tag{2.67}$$

The electron density within a band is similarly obtained by

$$n(\vec{r}) = -\frac{ig}{2\pi V}\int_{E_{min}}^{E_{max}} dE \sum_{\vec{k}} \mathrm{Tr}(\mathrm{Im}(G^<(E,\vec{r},\vec{k}))) \tag{2.68}$$

and the hole density is obtained by substituting $-G^>$ for $G^<$ in the equation above. Here E_{min} and E_{max} denote the energy bounds of the band in question, V is the generalized volume of the mesh point at position $\vec{r}$, and g is the band degeneracy.

Assuming nearest-neighbor coupling in the device Hamiltonian, we can obtain the device current flowing between adjacent layers q and $q+1$. By enforcing current continuity and integrating over energy, transverse momenta, and degeneracies[56, 60], it can be shown that the current flowing between layers q and $q+1$ is given by

$$J_{q,q+1} = \frac{2eg}{h}\int dE \sum_{k} \mathrm{Tr}[H_{q,q+1}(k)G^<_{q+1,q}(k,E) - H_{q+1,q}(k)G^<_{q,q+1}(k,E)] \tag{2.69}$$

We thus observe that the carrier and current densities depend only on the tridiagonal blocks of the various Green's functions, validating the use of the recursive algorithm.

2.2.5 Boundary Conditions

The boundary conditions at the edges of the defined structure are important when computing the solution to the Hamiltonian. In regions connecting the active region to the electrodes, the inclusion of contact self-energies allows current to flow and acts as a kind of boundary condition for particle-exchanging reservoirs. In 2-D or 3-D structures, there are also regions where the active region terminates in an interface with an insulator, for instance, along the channel/oxide interface of a FET. In general, such interfaces form a finite potential barrier which serves to confine the electron wave functions within the device, although there may still be some finite tail of the wave function penetrating into the barrier (and potentially through it to form a tunneling current, if the insulator is thin enough). In principle this effect can be included by extending the device Hamiltonian with an appropriate form for the insulating regions (effectively forming a heterojunction and extending the active region of the device). This is possible in our program, though it increases the size of the system and slows computational runtime. In addition, in some cases physically appropriate Hamiltonians for the barrier material may not be as well developed as those for the device semiconductor(s). Hence, for simplicity and unless otherwise noted, the spatial edges of device structures (at oxide interfaces or the boundaries of standalone quantum wells or wires) are approximated as infinitely high potential barriers. In practice this means that the off-diagonal components of the Hamiltonian which couple to the barrier region are simply dropped. Phenomena that depend on this coupling, like gate tunneling currents in FETs, for instance, will not be included in this approximation.

Boundary conditions are also of paramount importance when solving the Poisson equation. In this case, structure edges which do not terminate in contacts are treated using Neumann boundary conditions where the normal electric field is zero. Voltage contacts (for instance the gates of FETs) are treated as Dirichlet boundaries constraining the local potential to be equal to the applied bias. In semiclassical transport calculations, current-carrying voltage contacts (such as the source and drain in transistors) are typically also treated using Dirichlet boundary conditions with voltages fixed by the applied

bias. However, what is fixed by the contact is actually the electrochemical potential. The mismatch between quasi-equilibrium contacts (described by a Fermi-Dirac distribution function) and the strongly non-equilibrium distribution in the device means that the precise electrostatic potential along the boundary must float to a value that satisfies charge conservation in the device[72, 73]. Therefore, at the contacts we fix the quasi-Fermi potential using the bias and local doping and we apply Neumann boundary conditions in the Poisson equation, which allow the electrostatic potential to adjust itself to satisfy this condition. These approximations are expected to become exact for low current densities. We note that the choice of proper boundary conditions in quantum simulations continues to be debated in the literature and that much work remains to be done in this area[52, 62].

2.3 Discretizing Semiconductor Band Structure Models

The choice of Hamiltonian for the device active region is critical for the accuracy and level of detail for the simulation, encompassing the electronic properties of the device material(s). The calculations in this dissertation are typically performed using one of three methods: the effective mass approximation (EMA), $\mathbf{k}\cdot\mathbf{p}$ theory, and empirical tight-binding (TB). In this section we discuss the basic implementation of these models in numerical calculations. The underlying microscopic justification and theory of these methods can be found in any number of treatises on solid state or semiconductor physics, such as [51]. The resulting real-space Hamiltonian H_{band} is then combined with the self-consistent electrostatic potential V calculated from the Poisson equation and the electron and hole densities taken from $G^<$ and $G^>$ to obtain the total device Hamiltonian, $H_0 = H_{band} + V$.

2.3.1 Effective Mass Approximation

In the effective mass approximation (EMA), only a single band is relevant and is described by effective masses m_x, m_y, and m_z. The resulting Hamiltonian in terms of momentum

k takes the form

$$H_{em} = \frac{\hbar^2}{2m_x}k_x^2 + \frac{\hbar^2}{2m_y}k_y^2 + \frac{\hbar^2}{2m_z}k_z^2. \tag{2.70}$$

In effective mass or **k·p** type Hamiltonians, the transformation to real space is performed by drawing on the relationship between the momentum and position operators

$$p_x = -i\hbar\frac{\partial}{\partial x} \tag{2.71}$$

and we can use $p = \hbar k$ to obtain $k_x = -i\dfrac{\partial}{\partial x}$. We can then discretize the resulting spatial or spatial gradient operators using conventional central difference schemes. As an example, if we work on a lattice with spacing Δx and use a discrete basis where the wave function at lattice site i is denoted Ψ_i, we can transform

$$k_x\Psi \rightarrow -i\frac{\partial}{\partial x}\Psi(x) \rightarrow -\frac{i}{2\Delta x}\left[\Psi_{i+1} - \Psi_{i-1}\right]. \tag{2.72}$$

Similarly

$$k_x^2\Psi \rightarrow -\frac{\partial^2}{\partial x^2}\Psi(x) \rightarrow -\frac{1}{\Delta x^2}\left[\Psi_{i+1} + \Psi_{i-1} - 2\Psi_i\right]. \tag{2.73}$$

The detailed justification of this procedure is found in envelope function theory[74].

In a 1-D simulation, only the x direction may be discretized in real space and translational invariance is maintained in the y and z directions. The effective mass equation then becomes

$$H_{em} = -\frac{\hbar^2}{2m_x}\frac{\partial^2}{\partial x^2} + \frac{\hbar^2}{2m_y}k_y^2 + \frac{\hbar^2}{2m_z}k_z^2. \tag{2.74}$$

The on-site and off-diagonal terms in the real space Hamiltonian are

$$H_{ii} = \frac{\hbar^2}{2m_x}\frac{2}{\Delta x^2} + \frac{\hbar^2}{2m_y}k_y^2 + \frac{\hbar^2}{2m_z}k_z^2 \tag{2.75}$$

$$H_{i,i+1} = H_{i,i-1} = -\frac{\hbar^2}{2m_x}\frac{1}{\Delta x^2}. \tag{2.76}$$

The generalization to 2-D and 3-D systems is straightforward.

2.3.2 Multiband k·p

The effective mass approximation is useful for studying low energy excitations at the edge of an isolated band, for example electrons in the Γ conduction band valley in direct-gap III-V semiconductors. However, at higher energies or when coupling between bands becomes

relevant, the EMA inevitably breaks down. In these situations $\mathbf{k}\cdot\mathbf{p}$ theory provides a powerful tool for more quantitatively and qualitatively accurate study. This approach, whose rigorous foundations were largely laid during work in the 1950s by Luttinger and Kohn[75] and Kane[76, 77], starts from the microscopic Bloch Hamiltonian and assumes that the wave functions and energies at a specific high-symmetry k value, usually the $k = 0$ Γ point, are known. The band structure and wave functions around this point are then constructed perturbatively to varying orders in k. The matrix elements between basis states are treated as fitting parameters which are adjusted against experimental data. Spin-orbit coupling can also be incorporated by including spin-dependent basis states. The accuracy of the method is limited primarily by the number of basis states and the necessity for sufficient data to independently fix the adjustable parameters. In this sense the EMA is the single-band limit of the $\mathbf{k}\cdot\mathbf{p}$ method. Particularly useful choices for diamond and zincblende semiconductors, which are the main materials of interest in this thesis, include 6-band methods for describing the valence band extrema (including the light hole, heavy hole, and split-off bands with spin), 8-band models including the conduction band, and 15- or 30-band models (without or with spin) which tend to yield good agreement throughout the entire Brillouin zone for the low lying conduction and valence bands of a material.

Many of the numerical $\mathbf{k}\cdot\mathbf{p}$ calculations in this dissertation will be performed using the 8-band Hamiltonian since it includes the coupling between conduction and valence bands critical for interband phenomena. The corresponding Hamiltonian H_8 is given as

$$H_8 = \begin{bmatrix} H_4 & 0 \\ 0 & H_4 \end{bmatrix} + H_{so} \tag{2.77}$$

where H_4 is the 4-band matrix

$$\begin{bmatrix} E_g + A_c k^2 & iPk_x & iPk_y & iPk_z \\ -iPk_x & -\frac{\Delta}{3} + L'k_x^2 + M(k_y^2 + k_z^2) & N'k_xk_y & N'k_xk_z \\ -iPk_y & N'k_yk_x & -\frac{\Delta}{3} + L'k_y^2 + M(k_x^2 + k_z^2) & N'k_yk_z \\ -iPk_z & N'k_zk_x & N'k_zk_y & -\frac{\Delta}{3} + L'k_z^2 + M(k_x^2 + k_y^2) \end{bmatrix} \tag{2.78}$$

and H_{so} is the spin-orbit coupling matrix given by

$$H_{so} = \frac{\Delta}{3}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -i & 0 & 0 & 0 & 0 & 1 \\ 0 & i & 0 & 0 & 0 & 0 & 0 & -i \\ 0 & 0 & 0 & 0 & 0 & -1 & i & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & i & 0 \\ 0 & 0 & 0 & -i & 0 & -i & 0 & 0 \\ 0 & 1 & i & 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2.79}$$

In these equations, Δ is the spin-orbit splitting, the assorted P are the interband momentum matrix elements, and L', M, and N' are the valence band Kane parameters[77], all of which depend on the choice of material.

The discretization of the k-dependent Hamiltonian Eq. 2.77 proceeds similarly to the EMA case. A complication in the 8-band model is provided by the presence of terms like $k_x k_y$ which couple orthogonal directions. The ordering of the operator during discretization is non-obvious and should be derived from the microscopic underpinnings of the Hamiltonian[78]; such an analysis shows that terms like

$$N' k_x k_y \tag{2.80}$$

should be more properly split into

$$k_x N'_+ k_y + k_y N_- k_x \tag{2.81}$$

where N'_+ and N_- represent the contributions to the matrix element N' from Γ_1, Γ_{12} and Γ_{15}, Γ_{25} bands, respectively. These constants can be approximated as $N_- = M - \hbar^2/2m_0$ and $N'_+ = N' - N_-$[78]. If we assume a homogeneous material, we obtain in real space (assuming wave functions discretized as $\Psi_{i,j}$ where i and j represent the x and y direction indices, respectively)

$$N' k_x k_y \Psi \rightarrow N' \frac{1}{4\Delta x \Delta y}\left[\Psi_{i-1,j+1} + \Psi_{i+1,j-1} - \Psi_{i-1,j-1} - \Psi_{i+1,j+1}\right]. \tag{2.82}$$

Table 2.1: Eight-band $\mathbf{k}\cdot\mathbf{p}$ parameters for materials in this study. Except where otherwise indicated, values are in units of $\hbar^2/2m_0$.

	$In_{0.53}Ga_{0.47}As$	InAs	InSb
E_g (eV)	0.74	0.37	0.237
Δ (eV)	0.329	0.393	0.81
A_c	1.43	3.6	1.74
E_P (eV)	18	17.4	23.1
$\tilde{L}'$	-3.406	-5.193	-0.172
$\tilde{M}$	-2.65	-2.84	-3.8
$\tilde{N}'$	-4.716	-6.613	-3.99

The corresponding discretization for k_xk_z and k_yk_z terms are obtained by cyclic permutation.

Most of our calculations will be done using three technologically important semiconductors: $In_{0.53}Ga_{0.47}As$, InAs, and InSb. We follow the guidelines in Ref. [79] for adjusting experimentally fitted $\mathbf{k}\cdot\mathbf{p}$ parameters[80, 81] to eliminate spurious gap states caused by discretization[78, 82]. Our final parameter sets are given in Table 2.1 and are free of spurious states while correctly reproducing the bulk band gaps and effective masses.

2.3.3 Empirical Tight-Binding

If very small dimensions or very high energies are relevant, atomistic Hamiltonians which accurately describe the band structure over the entire bulk Brillouin zone are preferable. This can be generated using *ab initio* methods such as density functional theory (DFT) calculations or semi-empirical Hamiltonians based on tight-binding (TB) or pseudopotential concepts. The use of the TB basis for empirical fitting of the band structure was first proposed by Slater and Koster[83] and subsequently became an important method for understanding semiconductor properties. The method consists in choosing a suitable number of atomic orbital basis states within the unit cell and constructing the form of

the Bloch Hamiltonian coupling these states in agreement with the symmetry properties of the material. As with **k**·**p** theory, the matrix elements are treated as fitting parameters adjusted to reach agreement with experimental energy gaps, effective masses, etc.

The overall accuracy and range of the resulting Hamiltonian is dictated by the basis size. Many TB parameterizations exist in the literature. Early work focused on sp^3 models to describe the conduction and valence band extrema, with limited success unless longer-range couplings were considered. For nearest neighbor coupling schemes, which are computationally favorable for the reasons discussed above, good success has been found using $spds^*$ Hamiltonians, which have forty states per unit cell (1 s, 3 p, 5 d, and one "excited" s^* orbital per spin per atom) in diamond and zincblende semiconductors[84]. Each layer in the device structure then corresponds to a cation or anion plane in the device, where the on-site energy levels of each orbital comprise the diagonal on-site Hamiltonian. Coupling between adjacent layers is given by the hopping parameters between different orbitals on neighboring cations and anions. We use the $spds^*$ model with parameters in Ref. [84] for InSb and Ref. [81] for InAs. For $In_{0.53}Ga_{0.47}As$ we obtain the TB matrix elements by interpolating InAs and GaAs parameters[81], accounting for alloy bowing effects[85]. In quantum confined structures, surface layers of the TB Hamiltonian are passivated to remove spurious gap states by shifting the dangling bond energies[86].

The use of TB provides atomistic detail in the device and is hence more accurate at small length scales or high energies. However, the large number of basis states, as well as the requirement that the mesh spacing between layers must be equal to the actual atomic spacing, renders TB calculations very computationally expensive, particularly in self-consistent NEGF evaluation. For this reason, in the work presented here, TB calculations are only used to examine the equilibrium band structure of quantum confined materials and to validate the use of simpler **k**·**p** models in actual NEGF transport calculations.

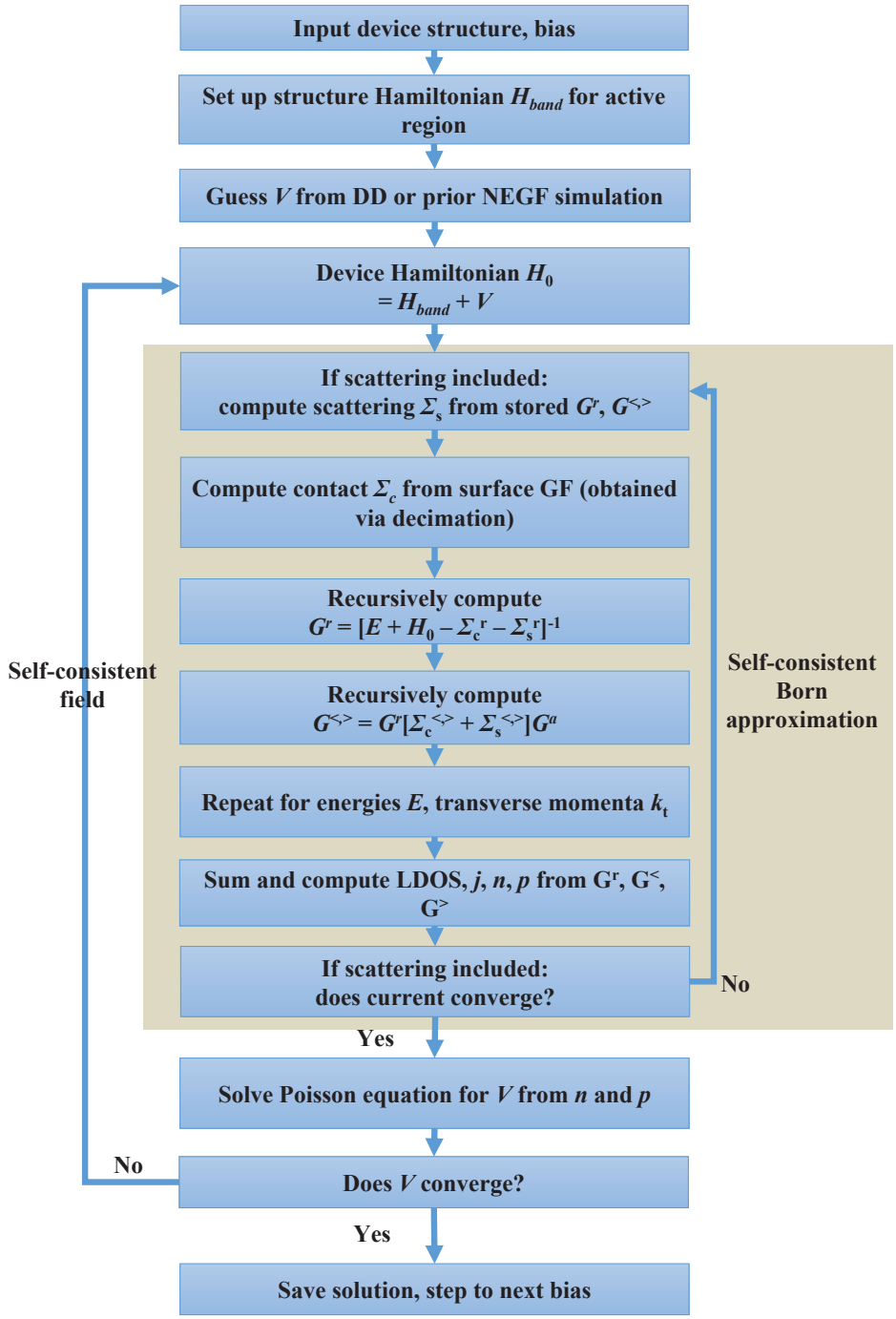

Figure 2.3: Computational methodology for self-consistent NEGF device simulations. The evaluation of the Poisson equation can be skipped if a self-consistent field calculation is not required.

2.4 Putting it All Together

The final program we develop allows for self-consistent simulation of 1-D, 2-D, or 3-D devices with arbitrary structure. The algorithmic structure is illustrated in Fig. 2.3. The input to the program (which is written entirely in MATLAB) is a script specifying the physical structure of the device, which is discretized using a finite difference scheme. The active region of the device (including all semiconductor regions contiguously connected to a current-carrying contact) is identified and the structure Hamiltonian is constructed from the mesh and material properties. Bias steps are specified and are solved sequentially. The device electrostatic profile at each step can either be fixed analytically (for non-self-consistent calculations where a fixed potential profile is assumed) or through a self-consistent computation of the Poisson equation, where the solution for the prior step serves as the initial guess for each new calculation. We implement Anderson mixing in our Poisson solver[87], though in practice the mixing parameters have had little impact on convergence in most of the structures simulated here. To provide an reasonable first guess for the electrostatic potential, we also created an auxiliary drift-diffusion simulator which solves the electron and hole continuity equations on the device grid assuming semiclassical mobilities and carrier concentrations[50].

Once the structure Hamiltonian and electrostatic potential are specified, we calculate the contact self-energies by applying the Lopez-Sancho decimation scheme to the left-most and right-most blocks of the device Hamiltonian. If scattering is enabled and a previous Green's function solution is available, the appropriate scattering self-energies are also computed. We then compute the tridiagonal blocks of the electron Green's functions using the recursive form of the Dyson and Keldysh equations. For 1-D or 2-D devices, spatial uniformity is assumed in the perpendicular, un-discretized dimension(s) and Green's functions are therefore solved and stored as a function of k_t, the transverse momenta. The energy interval over which the Green's functions are calculated is an initial input, but can change during the calculation if the program detects that a relevant energy band edge in the structure falls below or above the interval.

Having computed the Green's functions, we calculate the relevant quantities (LDOS, electron and hole densities, current density) as a function of space. The quantum carrier densities in the active region are used as an input to solve the Poisson equation via Newton's method, using a semiclassical Fermi-Dirac or Boltzmann form for the charge density Jacobian[50, 88]. For simulations without scattering, current conservation is guaranteed for each Green's function evaluation and convergence is assessed by comparing the change in electrostatic potential within each iteration. If scattering is included, we also check if the device current converges to a fixed value after each iteration, sufficiently satisfying the self-consistent Born approximation. If the error in the potential or current exceeds a specified value, we iterate the procedure, using the obtained solution to the Poisson equation (and Green's functions in the case of scattering) as the new initial condition.

2.5 Conclusion

In this chapter we introduce the basic structure of quantum device simulation using NEGF and outline the computational algorithms needed to calculate the device Hamiltonian, Green's functions, self-energies, and important observables. The computational framework we develop here provides a physically rigorous and numerically robust foundation for our investigation into the physics and design of quantum transport. Our program's modular structure makes it flexible and easy to use as a general simulator for semiconductor devices, not only transistors.

CHAPTER 3

Band Structure and Quantum Confinement in Direct Interband Tunneling

> Never calculate unless you already know the answer.
>
> John Wheeler

Direct interband or band-to-band tunneling has been heavily researched for decades because of its wide practical ramifications, for instance in enabling Zener and Esaki tunnel diodes[89], contributing unwanted gate-induced drain leakage (GIDL) to transistors[90] or parasitic dark currents in photodetectors[91], contacting multijunction solar cells[92], and potentially facilitating low power electronics using tunnel field-effect transistors (TFETs)[18]. Most early theoretical work[93, 94] used simple two-band Hamiltonian, constant field, and Wentzel-Kramers-Brillouin (WKB) or perturbation approximations to derive analytical semiclassical models such as the widely used Kane formula[95, 96]. As quantum transport theory and computational capabilities have evolved, these limitations have been relaxed in numerical studies of tunneling using more realistic $\mathbf{k}\cdot\mathbf{p}$[75, 77, 74] or tight-binding (TB)[83] band structures and quantum kinetic formalisms like scattering matrices[97], Wigner functions[98], or non-equilibrium Green's functions (NEGF)[60, 99]. However, the more convenient semiclassical formulas remain popular for providing physical understanding and developing device compact modeling or technology computer-aided design (TCAD) simulations. Therefore, it is of both theoretical and practical importance to determine, and if possible improve, their reliability.

While comparison to experiment would be ideal, and the Kane formula is sometimes

"calibrated" by fitting to tunnel junction measurements, the quantitative value of such exercises is often limited by uncertainties in the fabricated doping profiles and hence electrostatics, as well as concomitant disorder and many-body effects such as band gap narrowing (though p-i-n diodes can provide better controlled conditions[100, 101]). Alternatively, semiclassical predictions may be benchmarked against quantum calculations. Only a few such comparisons have been reported, with seemingly inconsistent conclusions. Di Carlo *et al.*[97] found that the Kane formula substantially underestimates scattering matrix calculations for GaAs but gives better agreement for InSb. Vandenberghe *et al.*[102] analyzed the two-band model for InSb using the envelope function approximation and found the Kane formula overestimates tunneling for small bias or very high fields. Schenk *et al.*[103] claimed good agreement between semiclassical and NEGF simulations for heavily doped InAs p-n junctions (i.e., high fields), though the former underestimates current at lower fields. By contrast, Ganapathi *et al.*[104] reported that semiclassical methods overestimate current at low fields but underestimate high field currents by orders of magnitude in their NEGF InAs p-n junction calculations. The disparity in these reports may be due to a number of factors, including the different materials, band structure models, and operating conditions considered.

Furthermore, the interplay of interband tunneling with quantum confinement (QC) are of increasing importance as semiconductor devices approach the nanometer scale. For instance, tunneling can be a limiting leakage pathway in modern transistors[105] or the operating principle in emerging devices like tunnel field-effect transistors (TFETs)[18], which in turn are being realized in multigate QC configurations like double-gate (DG) or nanowire (NW) structures. Inherently 2-D and 1-D materials such as graphene, MoS_2, and carbon nanotubes are gaining attention for a variety of electronic devices, including TFETs[106, 107]. Bulk concepts and models are still often used to study these and other devices, even though dimensional reduction can significantly modify physical properties. It is therefore important to establish whether and how such models can be extended to describe lower-dimensional devices.

The usefulness of the Kane formula and similar semiclassical tunneling models remains

an important question because of their continued popularity. To clarify this situation, we perform a detailed comparison of the semiclassical and quantum approaches to interband tunneling, identifying and correcting shortcomings in the former where possible. We extend our analysis to study how they should be modified when applied to QC devices. For bulk materials, we demonstrate that the primary corrections come from nonparabolicity and spin-orbit coupling effects. In QC structures we show that a simple band gap scaling (BGS) modification of the bulk Kane formula and its lower-dimensional counterparts allows the model to extended to QC devices. The work in this chapter has been published in Refs. [10, 9].

3.1 Semiclassical Bulk Tunneling Models

3.1.1 Two-Band Approach

Analytical studies of tunneling has been mostly based on the two-band Hamiltonian because of its simplicity. In this subsection, we briefly review some of the key prior work in the literature in this area. This provides background for understanding the new models we derive below and also calls attention to some useful previous results which perhaps have not been fully appreciated. Our focus throughout will be on homojunction direct tunneling; we do not discuss indirect tunneling as it requires consideration of the scattering (e.g., electron-phonon) matrix elements and involves transitions between band extrema at different k for which the two-band $\mathbf{k}\cdot\mathbf{p}$ model is not directly applicable, though related arguments are expected to apply[96]. The two-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian can be derived from the eight-band theory in the limit of large spin-orbit coupling strength Δ and describes coupling between the conduction (CB) and light hole (LH) bands[76, 77]

$$H = \begin{bmatrix} E_g + \dfrac{\hbar^2 k^2}{2m} & P_2 k \\ P_2 k & \dfrac{\hbar^2 k^2}{2m} \end{bmatrix} \tag{3.1}$$

where E_g is the band gap, P_2 is the two-band momentum matrix element, and m is a parameter equal to the free electron mass m_0 in the original theory, though its exact value

does not impact tunneling to lowest order as shown below. We therefore adjust it and P_2 to fit the experimental CB and LH effective masses, which are given by

$$m_{CB,LH} = \left(\pm \frac{1}{m} + \frac{2P_2^2}{\hbar^2 E_g} \right)^{-1}. \tag{3.2}$$

Note that interband coupling in the two-band model occurs along a set k orientation which is taken to be the direction of the electric field for tunneling. We will assume throughout that the basis is chosen such that all matrix elements, including P_2, are real[77].

Kane applied perturbation theory to this model under a constant electric field F to find the tunneling transmission coefficient[95]

$$T(E, k_\perp) = \frac{\pi^2}{9} \exp\left(-\frac{B}{F} \right) \exp\left(-\frac{2E_\perp}{\bar{E}_\perp} \right) \tag{3.3}$$

where $E_\perp$ is the energy associated with the transverse momentum $k_\perp$ perpendicular to the field and B and $\bar{E}_\perp$ are constants equal to

$$B = \frac{\pi m_r^{1/2} E_g^{3/2}}{2q\hbar} \tag{3.4a}$$

$$\bar{E}_\perp = \frac{2q\hbar F}{\pi m_r^{1/2} E_g^{1/2}}. \tag{3.4b}$$

The reduced mass m_r is related to the CB and LH effective masses through

$$m_r = \left(m_{CB}^{-1} + m_{LH}^{-1} \right)^{-1} = \frac{\hbar^2 E_g}{4P_2^2}. \tag{3.5}$$

We see that while the CB and LH effective masses are in terms of m, E_g, and P_2, the reduced mass (and hence tunneling probability) does not depend on m. For arbitrary dimension d and a tunnel junction like that in Fig. 3.1, the current density is obtained by integrating over energy E and $k_\perp$[96]

$$\begin{aligned} J &= \frac{q}{\pi\hbar} \int \frac{dk_\perp^{d-1}}{(2\pi)^{d-1}} \int dE\ T(E, k_\perp) \left[f_L(E) - f_R(E) \right] \\ &= A(F) \int_0^{\Delta E} dE\ \exp\left(-\frac{B}{F} \right) \left[f_L(E) - f_R(E) \right] \end{aligned} \tag{3.6}$$

where $f_{L,R}$ are the Fermi distribution functions on the left and right sides of the tunnel junction. In bulk structures ($d = 3$) and assuming parabolic bands in the perpendicular

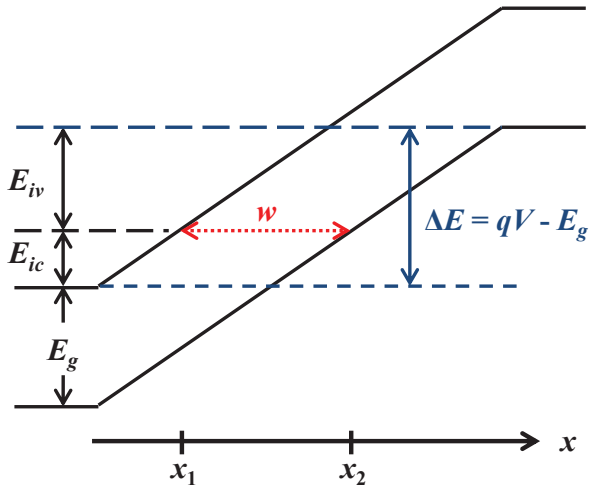

Figure 3.1: Definitions for tunnel junction with total potential difference V. A tunneling electron's energy relative to the conduction and valence band edges on either side of the junction is E_{ic} and E_{iv}, respectively, for which w is the tunneling distance between the classical turning points x_1 and x_2.

directions so $E_\perp = \hbar^2 k_\perp^2 / 2m_r$, the integration over $k_\perp$ gives[95]

$$A_{\text{Kane}} = \frac{q^3 m_r^{1/2} F}{18\pi\hbar^2 E_g^{1/2}} = \frac{q^2 BF}{9\pi^2 \hbar E_g^2}. \tag{3.7}$$

In Kane's work, the unperturbed basis was chosen to be the (localized) stationary states of individual bands under an unbounded linear potential, which leads to ambiguities in the boundary conditions. Fredkin and Wannier performed an alternative analysis of the problem using scattering theory and a finite bounded potential[108], with only minor differences in their result that are mostly due to the use of a different band model[109]. Shuey carefully analyzed the boundary conditions using this approach[110] and showed that tunneling is reduced for states with incident energy $E_{ic,iv}$ close to a band edge as defined in Fig. 3.1; this is related to the breakdown of the WKB approximation at small wavevectors. Some of the discrepancies observed in the Kane formula for small ΔE[102] may be attributed to this correction. The connection of these approaches to the WKB approximation and the parametrization of the two-band model are discussed next.

3.1.2 WKB Approach to Interband Tunneling

In the WKB approach to tunneling[94], the normal incident transmission probability is given by the action integral in the forbidden region between the classical turning points x_1 and x_2 in Fig. 3.1

$$\begin{aligned} T_{WKB} &= \exp\left(-2\int_{x_1}^{x_2}\kappa(x)dx\right) \\ &= \exp\left(-2\int_0^{E_g}\frac{\kappa(E)}{qF}dE\right), \end{aligned} \tag{3.8}$$

where the second line holds for constant field F. κ is the magnitude of the imaginary wavevector for energies E in the band gap and for the two-band Hamiltonian becomes

$$\kappa(E) = \frac{\sqrt{2}m}{\hbar^2}\sqrt{-\frac{\hbar^2 E}{m} + \beta + \sqrt{\frac{2\hbar^2 P_2^2 E}{m} + \beta^2}} \tag{3.9}$$

with $\beta = (E_g\hbar^2/2m) - P_2^2$. The action integral can be performed using this result and gives $T_{WKB} = \exp(-B/F)$, demonstrating the basic equivalence of the WKB and Kane models. This can be traced to Kane's use of a semiclassical phase and hence WKB-like wave functions as the unperturbed basis in his calculations. Similarly, while the Fredkin-Wannier scattering approach to tunneling is quite general, the two-band Green's function and WKB wave functions are used in the forbidden regions when obtaining analytical results[108, 109].

We observe again that while m affects the shape of the complex dispersion in Eq. 3.9 (which is why we use it to fit the experimental effective masses in Section 3.2), it disappears from the (constant field) integrated action, as can be seen most clearly in Kane's formulation where constant energy differences drop out[95]. Physically, this is because interband tunneling is concerned with the coupling between the conduction and valence band states, which is described entirely by P in the two-band model of Eq. 3.1. Under a nonuniform field, m does impact the transmission since κ will be weighted asymmetrically within the forbidden region, though we find numerically that these effects are negligible unless $m \ll m_0$, which is generally not the case for realistic material parameters. Similarly, we note that if the k dependence of the diagonal terms in the two-band Hamiltonian

is made asymmetric by imposing different masses for the upper (CB) and lower (LH) diagonals, then the asymmetry of the masses will enter the action integral. This is because the different masses, though they do not directly couple the CB-LH branches, change the curvature of the imaginary bands. Physically, different renormalized masses correspond to asymmetric corrections from remote bands; in eight-band $\mathbf{k}\cdot\mathbf{p}$ theory, for instance, this is reflected in the differing parameters A_c, $\tilde{L}'$, and $\tilde{M}$. These effects are neglected in the simple two-band Hamiltonian Eq. 3.1, which assumes a single mass m, but are partly recovered when we set P_2 using the experimental CB and LH effective masses through Eq. 3.2. This approximation is ultimately vindicated by the good agreement with full band structure and transport calculations documented in Sections 3.2-3.3.

3.1.3 Multiband Corrections

While the two-band Hamiltonian is easy to use and physically transparent, it neglects coupling with other bands, particularly the heavy hole (HH) and split-off (SO) valence bands. For instance, the use of the effective mass approximation in the transverse directions underestimates tunneling because $\mathbf{k}\cdot\mathbf{p}$ terms tend to flatten the dispersion and reduce the effective energy gap $E_g + E_\perp$. Using the eight-band Hamiltonian accounts for this but leads to difficult integrations over $k_\perp$ for Eq. 3.6. To obtain a tractable result, we assume the CB and LH dispersions along $k_\perp$ are also of the two-band form in Eq. 3.1, i.e.,

$$E_{CB,LH}(k_\perp) = \frac{E_g}{2} + \frac{\hbar^2 k_\perp^2}{2m} \pm \sqrt{\frac{E_g^2}{4} + P_2^2 k_\perp^2} \tag{3.10}$$

and hence the transverse energy difference

$$E_{\perp,NP} = E_{CB} - E_{LH} = E_g \left(\sqrt{1 + \frac{4P_2^2 k_\perp^2}{E_g^2}} - 1 \right). \tag{3.11}$$

This expression should be fairly accurate in capturing the nonparabolicity of the CB branch, though it neglects the warping of the LH band and cannot describe effects of anisotropic band mixing on tunneling. Additionally, for large $k_\perp$ the CB and LH bands may not be simply connected in imaginary space, such that the two-band WKB picture fails altogether[99]; we will assume such contributions are negligible at experimentally

accessible fields. Integrating Eq. 3.11 over $k_\perp$, we obtain a new nonparabolic counterpart to A_{Kane}

$$A_{\text{NP}} = \frac{q^2 BF}{9\pi^2 \hbar E_g^2} + \frac{q^3 F^2}{18\pi^2 \hbar E_g^2}, \tag{3.12}$$

where the new term quadratic in F shows that these effects are more important at high fields.

The inclusion of additional bands also impacts the normal incident tunneling probability. Krieger studied these effects using a four-band Hamiltonian[111]; despite some theoretical[97] and experimental[100] evidence of his model's success for GaAs, it is rarely used in the literature, perhaps because of its seeming complexity. We adapt his method to derive a new value of B for use within the exponential and prefactor A of the Kane and nonparabolic tunneling formulas

$$B_{\text{4band}} = \frac{\pi\sqrt{m_r} E_g^{3/2}}{2q\hbar} \sqrt{\frac{(5+4\alpha)(1+2\alpha)}{(2+2\alpha)(3+4\alpha)}}, \tag{3.13}$$

where $\alpha = \Delta/E_g$ and m_r is still in terms of the CB and LH effective masses. The derivation of this equation is outlined below. We see this new B_{4band} equals the two-band Kane version Eq. 3.4a in the limit of large α but is smaller by a factor of $\sqrt{5/6}$ for $\alpha = 0$, leading to an enhancement in tunneling current. The HH band does not enter this expression because it does not mix with the other bands to first order in the $\mathbf{k}\cdot\mathbf{p}$ interaction[76].

3.1.4 Derivation of Four-Band Tunneling Probability

Using spin degeneracy and neglecting remote band interactions, the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian can be transformed via a basis rotation into two copies of a four-band model with the eigenvalues[76]

$$\begin{aligned} E' &= 0 \\ E'(E'-E_g)(E'+\Delta) - k^2P^2(E'+2\Delta/3) &= 0 \end{aligned} \tag{3.14}$$

where $E' = E - (\hbar^2 k^2/2m_0)$. The lone band in the first line gives the uncoupled HH branch, so we focus on the three solutions in the second expression corresponding to the

CB, LH, and SO branches which enter the tunneling problem. Dropping the bare mass terms $\sim k^2/m_0$ for the same reasons as before, we can write the imaginary wavevector

$$\kappa_{\text{4band}}(E, k_\perp) = \frac{1}{P}\sqrt{\frac{E(E-E_g)(E+\Delta)}{E+2\Delta/3} - k_\perp^2 P^2}. \tag{3.15}$$

The WKB transmission can be evaluated using the action integral Eq. 3.8 of κ_{4band} generalized for arbitrary $k_\perp$. For $\alpha = \Delta/E_g \geq 0$, Krieger carried out this integration approximately using Taylor series expansions, leading to[111]

$$T_{\text{4band}} = \frac{\pi^2}{9}\exp\left[-\frac{\pi E_g^2}{4qFP}\sqrt{\frac{3+6\alpha}{3+4\alpha}}\left(1+\frac{4P^2k_\perp^2}{E_g^2}\right)\right]. \tag{3.16}$$

We have verified numerically that this approximation is accurate to within 1% error. Krieger then used perturbation theory to include remote bands, leading to a final form of the transmission coefficient summarized in Eqs. 5.39, 5.46-5.48, and 5.52 of Ref. [111] and which is in terms of E_g, Δ, and the effective masses of all four conduction and valence bands (as the remote bands reintroduce coupling to the HH branch)[1].

Instead, we choose to ignore the remote band effects and work with Eq. 3.16 directly. We express it in more familiar form using the CB and LH effective masses, which can be found from Eq. 3.14 to be

$$\begin{aligned} m_{CB} &= \left[\frac{1}{m_0} + \frac{2P^2}{3\hbar^2}\left(\frac{2}{E_g} + \frac{1}{E_g+\Delta}\right)\right]^{-1} \\ m_{LH} &= \left(-\frac{1}{m_0} + \frac{4P^2}{3\hbar^2 E_g}\right)^{-1}, \end{aligned} \tag{3.17}$$

from whence

$$P = \hbar\sqrt{\frac{3E_g}{2m_r\left(4+\frac{E_g}{E_g+\Delta}\right)}}. \tag{3.18}$$

With this identity, the integration over transverse modes in Eq. 3.6 yields a result identical in form to the Kane formula, but with the altered coefficient B given in Eq. 3.13. Interestingly, we have found for a wide range of realistic material parameters that this model is nearly indistinguishable numerically from the more complicated expressions in

[1]Note that while Eq. 5.16 in Ref. [111] is correct, there is an error of $\sqrt{2}$ in the coefficients of the remote-band-corrected transmission in Eq. 5.52 of that work.

Ref. [111]. This is because we use the experimental CB and LH masses when evaluating m_r, which is equivalent to lumping all remote band corrections into P (as is also the case when using m_r to define P_2 in the two-band model).

Eqs. 3.12 and 3.13 are our primary original analytical results for bulk tunneling; we will refer to them as the "multiband model" to distinguish them from the two-band Kane formula. Substituting them in Eq. 3.6 we observe that transverse nonparabolicity enhances the tunneling current at high fields, while the inclusion of the SO band leads to larger relative increases in tunneling at low fields because of the exponential dependence on B.

3.2 Complex Band Structure in Bulk Semiconductors

3.2.1 Band Structure Calculation Methods

To determine if the semiclassical Kane or multiband models are truly accurate, comparison with exact quantum calculations is necessary. It is also of interest to examine the complex band structure features relevant for tunneling; in particular, the two-band dispersion is important when modeling either bulk or quantum confined tunneling. Therefore we compare the two-band Hamiltonian with the more complete band structures predicted by eight-band $\mathbf{k}\cdot\mathbf{p}$[77] and *spds** empirical TB[83, 84] including spin-orbit coupling. To ensure that our results are not skewed by any accidental peculiarities of band structure for a particular material, we study three technologically important semiconductors: $In_{0.53}Ga_{0.47}As$, InAs, and InSb, corresponding to $\alpha < 1$, $\alpha \sim 1$, and $\alpha > 1$, respectively. We follow the guidelines in Ref. [79] for adjusting experimentally fitted $\mathbf{k}\cdot\mathbf{p}$ parameters[80, 81] to eliminate spurious gap states caused by discretization[78, 82]. Our final parameter sets are given in Table 2.1 and are free of spurious states while correctly reproducing the bulk band gaps and effective masses. For TB calculations, we use the parameters in Ref. [84] for InSb and Ref. [81] for InAs and obtain the TB matrix elements for $In_{0.53}Ga_{0.47}As$ by interpolating InAs and GaAs parameters[81], accounting for alloy bowing effects[85].

3.2.2 Bulk Band Modeling

In Fig. 3.2, we show the band structures of InGaAs, InAs, and InSb for real and imaginary k along the [100] direction calculated using the different Hamiltonians. In the two-band model of Eq. 3.1, m and P_2 are chosen such that Eq. 3.2 reproduces the bulk [100] LH and CB effective masses for each material. We see that the imaginary dispersion connecting the conduction and valence bands in the two-band model is close to those predicted by the other, more complete methods, showing it is indeed a useful starting point for analyzing bulk interband tunneling. The slight deviations can be attributed to the neglect of spin-orbit coupling and perturbations from remote bands in the two-band model. The eight-band and TB band structures are also quite similar, with the latter generally having slightly less area under the k-E curve, i.e., reduced action. This may be due to corrections from higher bands/additional orbitals included in the $spds^*$ basis, as well as the fact that the literature TB parametrizations yield effective masses somewhat smaller than the experimental values to which the $\mathbf{k}\cdot\mathbf{p}$ parameters in Table 2.1 are fitted. Since there is still good agreement overall between the TB and eight-band Hamiltonians, we will focus on the latter in our subsequent calculations and analysis due to its lighter computational load.

For intuitive understanding, we examine the relative contributions of the $\mathbf{k}\cdot\mathbf{p}$ basis states to the complex dispersion. Summing over spins for clarity and ignoring relative phases (as we will deal exclusively with the real projected weight of each state), the basis states of the eight-band model are $|S\rangle$, $|X\rangle$, $|Y\rangle$, and $|Z\rangle$, where the notation indicates the orbital-like symmetries of the wave functions[76]. In Fig. 3.3 we show the projections of the wave functions of the real and imaginary InAs conduction and valence bands (CB and LH) onto the eight-band basis states (similar results hold for InGaAs and InSb). The CB wave functions in Fig. 3.3a consist mainly of $|S\rangle$ as expected, with the increasing admixture of $|X\rangle$ leading to nonparabolic effects at larger k. For the LH band along the [100] direction, the zone center wave function weights[77] are 2/3 for $|X\rangle$ and 1/6 for $|Y\rangle$ and $|Z\rangle$, as seen in Fig. 3.3b. Interestingly, the "transverse" states $|Y\rangle$ and $|Z\rangle$ make

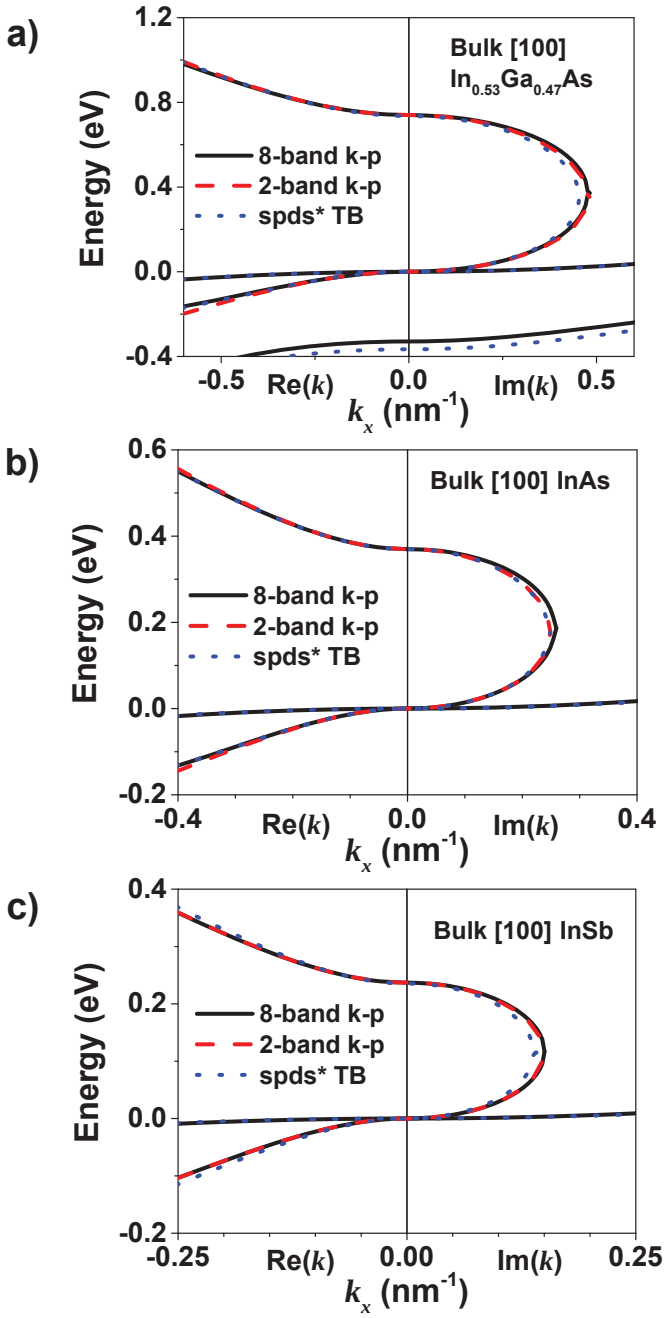

Figure 3.2: Bulk band structures for (a) InGaAs, (b) InAs, and (c) InSb calculated using two-band and eight-band $\mathbf{k}\cdot\mathbf{p}$ and *spds** TB. The left side of each plot (negative k) corresponds to real k and the right side to imaginary k.

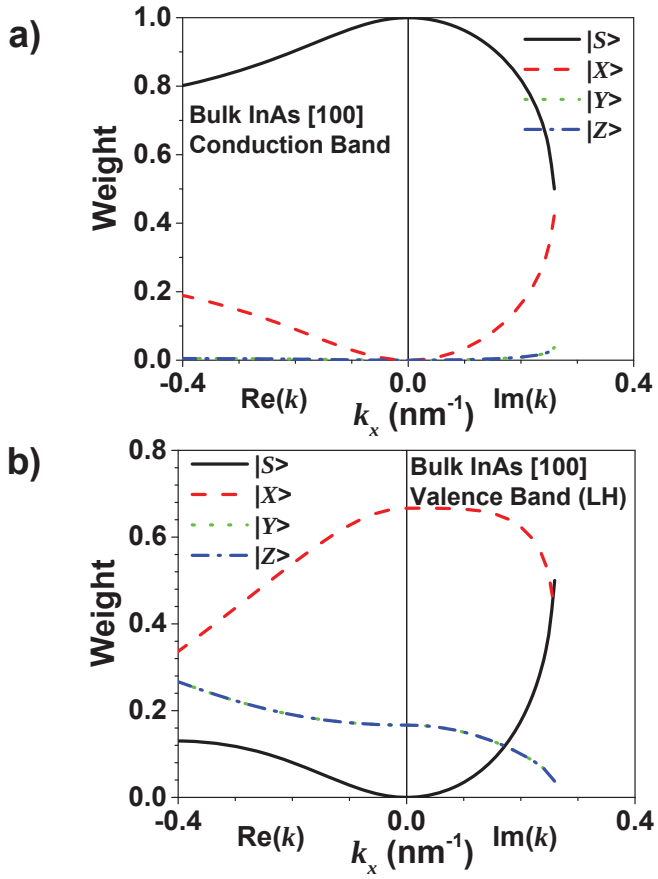

Figure 3.3: Weighted contributions of the spin-summed eight-band $\mathbf{k}\cdot\mathbf{p}$ basis states to the (a) conduction band and (b) valence band wave functions in InAs as the wavevector k_x is varied, corresponding to the dispersion in Fig. 3.2a.

increasing contributions to the real dispersion with k, reflecting the effects of band mixing. By contrast, the large ratio of $|X\rangle$ relative to $|Y\rangle$ and $|Z\rangle$ is roughly maintained for the imaginary k valence wave functions, while interband coupling comes from the increasing admixture of $|S\rangle$. This indicates that the CB and LH zone center states dominate the imaginary branch point. Therefore, in general the two-band model should be more useful for describing the complex CB-LH branch point than the real band dispersion. This is to be expected for bulk bands from $\mathbf{k}\cdot\mathbf{p}$ theory, but also has particular relevance for lower-dimensional quantum confined subbands.

The results in Figs. 3.2-3.3 are for states with no momentum $k_\perp$ perpendicular to the [100] orientation. While these make the largest contribution to the tunneling current

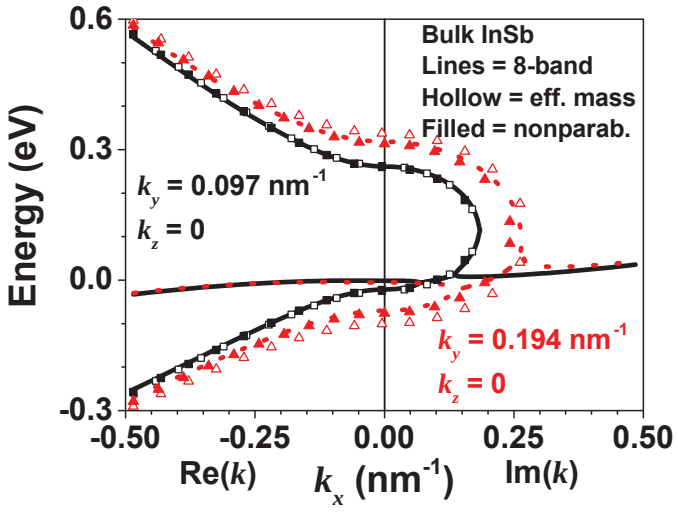

Figure 3.4: Complex dispersions for InSb with $k_y = 0.097$ nm^{-1} (black solid lines for eight-band and squares for two-band) and 0.194 nm^{-1} (red dashed lines for eight-band and triangles for two-band). Hollow symbols correspond to the use of the effective mass transverse gap $E_\perp$ and filled symbols to the nonparabolic $E_\perp$ in Eq. 3.11.

in that direction since $k_\perp$ increases the size of the energy gap to be surmounted, it is still important to examine the complex dispersion of wavevectors with lower symmetry for signs of band mixing effects. Since nonparabolicity generally increases for smaller gaps[77], we expect such effects to be large for InSb. In Fig. 3.4, we show the two-band and eight-band E-k in InSb for nonzero transverse momenta. It is evident that as $k_\perp$ increases, the transverse effective mass approximation overestimates the energy gap, indicating it will underestimate tunneling for these states. Using the nonparabolic correction Eq. 3.11 gives better agreement, particularly for the size of the energy gap.

3.3 Comparison of Interband Tunneling Models

3.3.1 NEGF Transport Modeling

Having examined the bulk band structure, we proceed to study the tunneling current. To do this, we perform numerical quantum transport simulations using the NEGF technique and the eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian as discussed in Chapter 2. Since we study only direct

Table 3.1: Bulk reduced masses and tunneling coefficients. B and $B_{4\text{band}}$ correspond to Eqs. 3.4a and 3.13, respectively. The coefficient A can be computed directly from B and E_g.

	$In_{0.53}Ga_{0.47}As$	InAs	InSb
m_r	$0.0235m_0$	$0.0128m_0$	$0.007m_0$
B (V/cm)	5.54×10^6	1.45×10^6	5.54×10^5
$B_{4\text{band}}$ (V/cm)	5.34×10^6	1.42×10^6	5.52×10^5

(i.e., elastic) tunneling, no scattering mechanisms are included in our calculations.

3.3.2 Constant Field Tunneling

For direct comparison of the analytical models with NEGF simulations, we first examine tunneling under constant electric field. The analytical currents are calculated using Eq. 3.6 with coefficients in Eqs. 3.4a and 3.7 for the Kane formula and Eqs. 3.12-3.13 for the multiband model. The model parameters are shown in Table 3.1 and are "uncalibrated" in the sense that they are based on material properties, not adjusted for best fit with quantum calculations. The NEGF simulations are performed using a linear potential drop V over a finite junction width, assuming nondegenerate statistics; to facilitate comparison and remove effects of the finite bias, we divide all current densities by the tunneling energy window ΔE defined in Fig. 3.1. In Fig. 3.5a, we observe good qualitative agreement between the normalized analytical and numerical current densities for InGaAs, InAs, and InSb over several orders of magnitude, with the multiband model being more accurate than the Kane formula.

To quantify our comparison and approximately normalize for differences in the materials, we plot the percentage difference between the NEGF and analytical model currents in Fig. 3.5b as a function of the tunneling transmission $T = \exp(-B/F)$. We observe that the underestimation of current by the Kane formula (open symbols) is largest for weak tunneling (low fields) and most significant for InGaAs. A similar, though weaker, trend

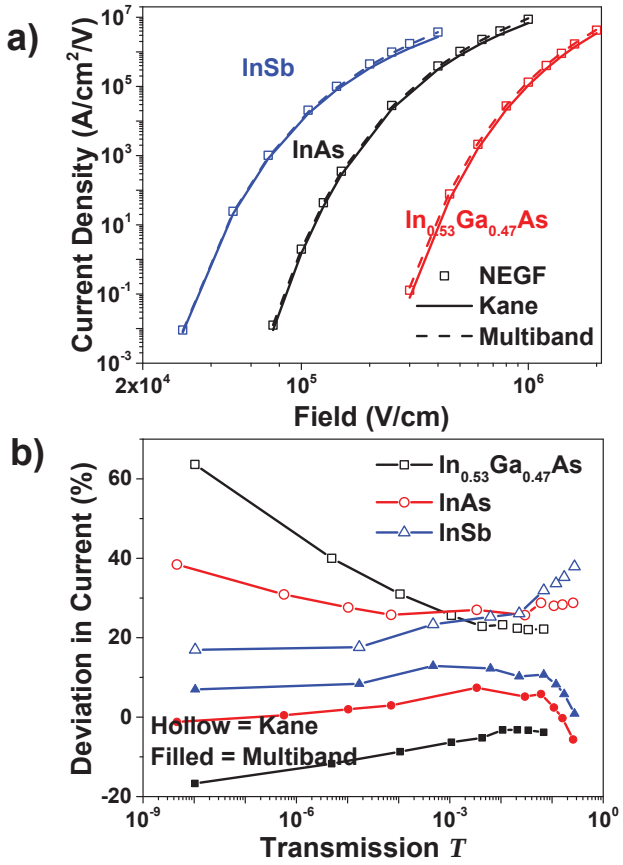

Figure 3.5: (a) Normalized current densities ($J/\Delta E$) versus applied field for InGaAs, InAs, and InSb using eight-band $\mathbf{k}\cdot\mathbf{p}$ NEGF (symbols), Kane formula (solid lines), and multi-band model (dashed lines) calculations. (b) Percentage difference between the NEGF and semiclassical model currents $(J_{\mathrm{NEGF}}/J_{\mathrm{model}})-1$ for InGaAs, InAs, and InSb as a function of transmission $T = \exp(-B/F)$. Hollow and filled symbols represent the NEGF-Kane and NEGF-multiband comparisons, respectively.

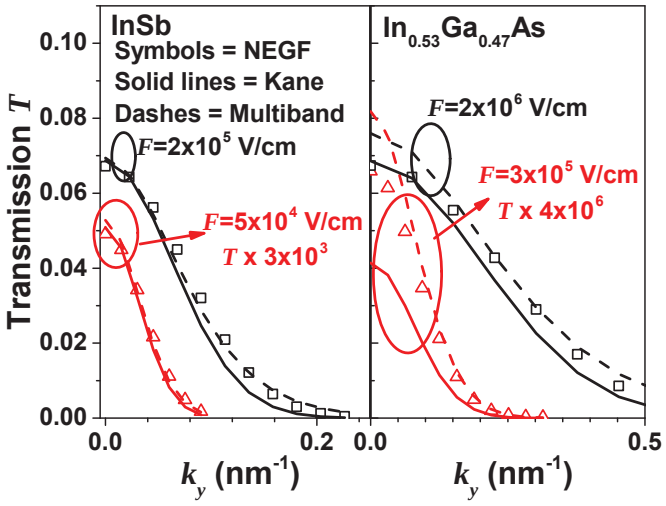

Figure 3.6: Energy-averaged transmission coefficients as a function of transverse momentum k_y ($k_z = 0$) for InSb (left) and InGaAs (right). Low (red triangles) and high (black squares) fields are considered in each case; for visibility, the transmissions for the smaller fields are multiplied by 3×10^3 for InSb and 4×10^6 for InGaAs. The Kane (solid lines) and multiband (dashed lines) model transmissions are calculated using Eq. 3.3 with corresponding choices of B and $E_\perp$.

is present in InAs. By contrast, the error increases with field for InSb. The multiband tunneling model (solid symbols) reduces the magnitude of the current deviation for all cases, with maximum errors of less than 10% for InSb and InAs and less than 20% for InGaAs over almost eight orders of magnitude in the transmission and total current.

To explain these trends, we directly inspect the dependence of the transmission coefficient on F and $k_\perp$ for InSb and InGaAs in Fig. 3.6. It is clear that in InSb, the Kane model (solid lines) is accurate at normal incidence ($k_\perp = 0$) but underestimates the transmission at large $k_\perp$, explaining why the error increases at higher fields. The impact of the 4-band correction is negligible at $k_\perp = 0$ (as seen by the near identical transmissions for the Kane and multiband models), but the nonparabolic correction leads to better agreement with the NEGF results for large $k_\perp$. Conversely, in InGaAs at low fields the normal incidence Kane transmission is about 60% smaller than the numerical result but converges towards the latter at high fields, while the multiband formula slightly

overestimates but is much closer to the NEGF values. These findings can be understood by recalling that InSb has large $\alpha = \Delta/E_g$, so we expect the two-band approximation to be good for low fields. However, nonparabolicity is more pronounced in small gap materials, leading to increased transverse momentum effects at high fields. Therefore the error in the Kane formula current increases with field for small gap materials. By contrast, α is small for InGaAs, implying that contributions from the SO band are important, especially at low fields where small changes in B lead to large differences in transmission. InAs falls in between these cases ($\alpha \sim 1$), such that the competing effects partly cancel and the relative error of the Kane model is flatter as a function of field, as seen in Fig. 3.5b.

These results demonstrate that the Kane formula is qualitatively useful but prone to underestimation of tunneling current, with the magnitude of the error dependent on the material α and applied field, while better quantitative results are provided by the multiband tunneling formula. The overall success of the semiclassical formulas in Fig. 3.5 might seem surprising since the WKB approximation (and first order perturbation theory) is expected to be applicable only for small fields. Some support for this may be found by recalling that in intraband (effective mass) tunneling, the solutions of the Schrodinger equation for exactly solvable cases such as a parabolic or exponential barrier are equal to the WKB result for large incident wavevectors[112]. Similarly, direct analysis of the two-band model shows that the Kane formula essentially holds for arbitrary field strength, at least in the limit of large bias [102]. However, the WKB approximation will break down in materials or nanostructures where an unambiguous path of least action connecting the conduction and valence branches cannot be identified[4].

Comparing our findings with other quantum transport studies, our results are consistent with those in Ref. [103] as well as Ref. [97], where the Kane model gave large errors for GaAs (large E_g, small α), but was better for InSb; the latter work also found that Krieger's 4-band model was more accurate in GaAs, though it overestimated the magnitude of the tunneling current, similar to our InGaAs results. Additionally, the theoretically calculated coefficients in Table 3.1 are within the range of experimentally extracted values for $In_{0.53}Ga_{0.47}As$ p-i-n diodes[101]. However, our qualitative and quan-

titative trends are different from those reported in Ref. [104] using NEGF for InAs, which may perhaps be partly attributed to that work's use of nonuniform fields as discussed in Section 3.3.3.

In Figs. 3.5-3.6, the tunneling currents are evaluated over energy windows ΔE on the order of 0.2-4 V (corresponding to the large bias condition in Ref. [102]), averaging out any energy dependent effects of quantum oscillations and junction edges. To illustrate how these effects impact tunneling at low bias, in Fig. 3.7 we show the energy dependent InAs transmission coefficient for $\Delta E = 0.02, 0.05,$ and 0.3 eV. For energies near the junction edges, i.e., small $E_{ic,iv}$, the transmission is strongly reduced because of the breakdown of the WKB approximation[110, 94]. This reduction is not due to the vanishing transverse density of states (DOS), since we examine the transmission for $k_\perp = 0$. The energy range over which these corrections are significant depends on the shape of the barrier and the value of ΔE[94], but is generally on the order of several to tens of millivolts. Hence, at small band overlap ($\Delta E = 0.02$ eV), this effect is especially noticeable and attenuates the average transmission below the Kane value. This, along with the reduced transverse DOS, explains the tendency of analytical models to overestimate current near the onset of tunneling, for instance around the peak voltage in forward biased Esaki diodes[102, 103], and can impact quantitative arguments on the possible steepness of ideal tunnel junctions[113]. In experimental situations, these effects might be partially obscured by leakage currents.

3.3.3 Tunneling in Nonuniform Fields

Since the electric fields in realistic structures are often nonuniform, we also investigate the tunneling current under these conditions. For direct comparison we examine abrupt p-n junctions using the depletion approximation with Fermi-Dirac statistics; in real devices the self-consistent electrostatic potential can be modified by quantum effects and band structure details, which might lead to additional quantitative differences between semiclassical and quantum calculations[104]. In the WKB approach (see Section 3.1.2)

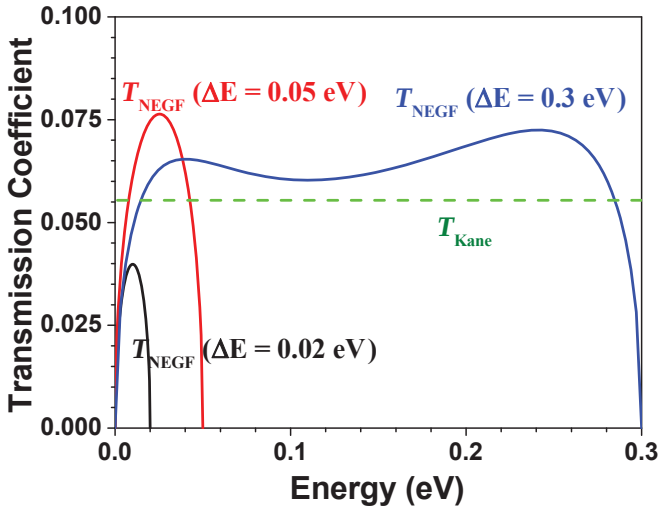

Figure 3.7: NEGF transmission coefficients $T(k_\perp = 0)$ for $\Delta E = 0.02, 0.05,$ and 0.3 eV in InAs at $F = 5 \times 10^5$ V/cm. The Kane transmission coefficient $\exp(-B/F)$ is independent of ΔE.

the transmission for an arbitrary potential may be evaluated numerically by integrating the action, as is done in some nonlocal TCAD tunneling models[114]; however, for convenience the Kane formula is often empirically extended to spatially varying fields by replacing the uniform field F with some energy-dependent F_{eff}. Different choices of F_{eff} have been suggested[115, 116], including the maximum field in the junction (which should overestimate current), the field at the midpoint $(x_1 + x_2)/2$ between the classical turning points (band edges) at a given energy illustrated in Fig. 3.1, or the average field E_g/w, where $w = x_2 - x_1$ is the tunneling distance.

In Fig. 3.8a we compare the p-n junction tunneling currents computed by integrating Eq. 3.6 using the various effective fields with two-band WKB action integration using Eq. 3.9 and NEGF calculations. The maximum field method is unsurprisingly inaccurate while the midpoint field curve exceeds the NEGF current at low bias and falls below it at high bias, similar to what was observed in Ref. [104]. The latter work took this trend to be a feature of the Kane model, but Fig. 3.8a demonstrates it is actually an artifact of the authors' choice of the midpoint effective field and the particular junction potential. Using the average field $F_{\text{eff}} = E_g/w$ gives better overall agreement with both the WKB and

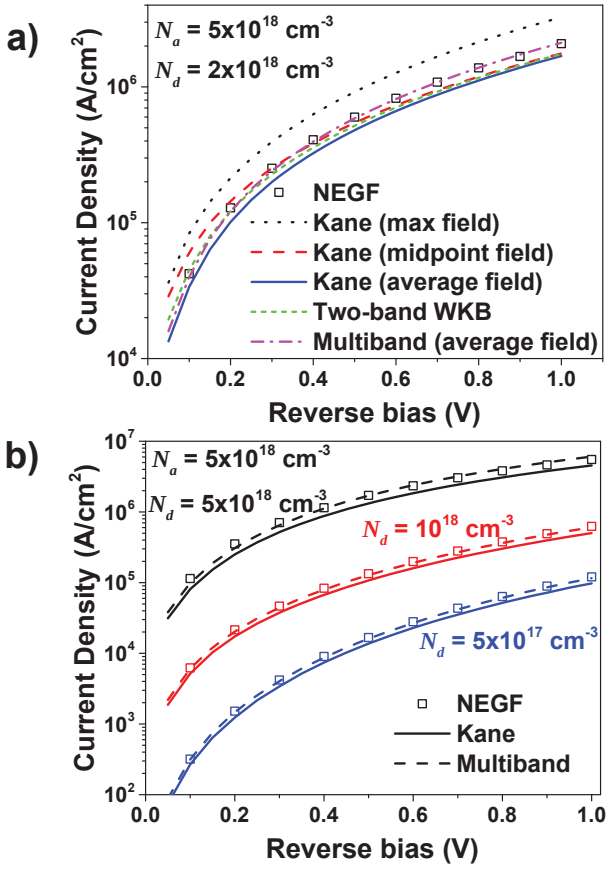

Figure 3.8: (a) Reverse bias tunneling current for InAs abrupt p-n junction calculated from NEGF, the integrated Kane formula using maximum junction field, midpoint field, and average field approximations, the two-band WKB action integral Eq. 3.8, and the multiband model using the average field approximation. (b) NEGF and average field Kane and multiband model current densities versus reverse bias for different InAs p-n junctions. N_a and N_d are the p- and n-doping levels in the junctions, respectively.

NEGF calculations, though the tunneling current is still quantitatively underestimated. Evaluating the multiband model with the average field leads to the best agreement with quantum simulations. While the use of the average field is heuristic, it relates the tunneling probability directly to the semiclassical tunneling distance w, lending it some physical credence; therefore, it is the best choice for analytically estimating tunneling when the WKB action integral cannot be performed. Finally in Fig. 3.8b we show the tunneling current for other junction profiles as computed from NEGF and the Kane and multiband

models with the average field approximation. In general we find that the Kane formula is qualitatively useful but tends to underestimate the current, especially at smaller bias, while the multiband model is the most accurate in general, similar to the uniform field results.

3.4 Complex Band Structure in QC Materials

Since tunneling depends sensitively on the imaginary band structure, it is important to examine the complex dispersion of QC materials like 2-D quantum wells and 1-D quantum wires. The subbands in these configurations can be quite complicated and require numerical evaluation because of band mixing effects. By contrast, most analytical tunneling models use a simple two-band Hamiltonian to describe interband coupling between the conduction (CB) and light hole (LH) valence bands, i.e., Eq. 3.3. It is unclear *a priori* whether or how the two-band model can be adapted to describe the lowest conduction and valence subbands (which should dominate tunneling in direct gap materials when confinement is significant[4]). To address this problem, we use eight-band $\mathbf{k}\cdot\mathbf{p}$[77] and *spds** empirical tight-binding (TB)[84] calculations to evaluate QC structures made of $In_{0.53}Ga_{0.47}As$, InAs, and InSb, similar to our study of bulk materials.

In Fig. 3.9, we show the material- and thickness-dependent quantum well band gaps extracted from our $\mathbf{k}\cdot\mathbf{p}$ and TB calculations; we see that deviations between the two methods occur in very thin structures, though they agree well for thicker wells. It is well known that under strong confinement, $\mathbf{k}\cdot\mathbf{p}$ and full-band calculations using methods like TB, empirical pseudopotentials, or density functional theory (DFT) can give different subband structures, with the full-band approaches expected to be more accurate due to their larger basis sets and reproduction of atomistic symmetries[117]. For computational efficiency we will still perform much of our analysis and transport calculations using eight-band $\mathbf{k}\cdot\mathbf{p}$, keeping in mind these limitations. Since our primary goal is to demonstrate the dependence of tunneling on the QC band gap, we expect and show where possible that our arguments can be extended to more precise theoretical or experimental data by

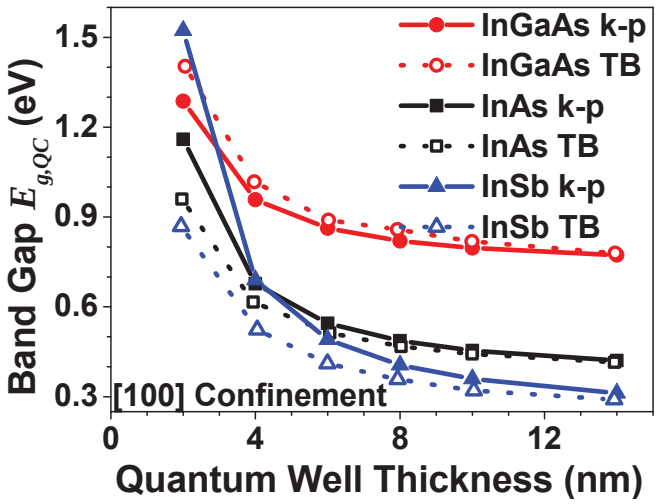

Figure 3.9: Eight-band **k**·**p** and *spds** TB energy gaps between the lowest conduction and valence subbands for quantum wells with varying thickness and material.

using the corresponding value of the band gap.

In the two-band Hamiltonian, there are essentially three parameters we can adjust to match the QC dispersion: the band gap E_g, the momentum matrix element P_2, and the mass m. In bulk materials, the latter two can be fitted to the experimental effective masses m_{CB} and m_{LH} through

$$P_2 = \hbar\sqrt{\frac{E_g}{4}\left[m_{CB}^{-1} + m_{LH}^{-1}\right]} \tag{3.19a}$$

$$m = \frac{2}{m_{CB}^{-1} - m_{LH}^{-1}}. \tag{3.19b}$$

Under confinement the gap between the lowest conduction and valence subbands becomes $E_{g,QC}$. The simplest approach to extend the bulk two-band Hamiltonian is to substitute $E_{g,QC}$ for the bulk E_g directly while keeping the bulk values of P_2 and m; we will call this the band gap scaling (BGS) method. Alternatively, since the lowest subband effective masses $m_{CB,QC}$ and $m_{LH,QC}$ also change from their bulk values, we can attempt to use them in Eq. 3.19 to adjust P_2 and m (which we will refer to as "subband scaling"). In Fig. 3.10, we compare the complex dispersion of a 9 nm InAs quantum well calculated using eight-band **k**·**p** and TB with the BGS and subband scaled two-band Hamiltonians (with all scaling being performed to the eight-band results). We observe that BGS provides

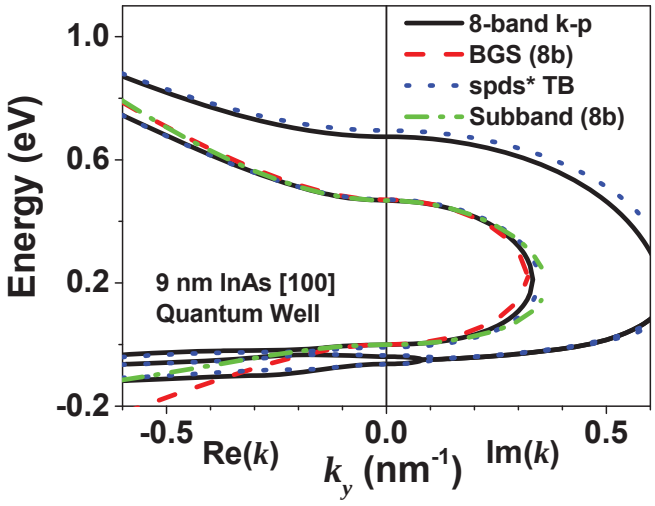

Figure 3.10: Band structure along the [010] direction of a 9 nm thick InAs quantum well calculated using eight-band $\mathbf{k}\cdot\mathbf{p}$, TB $spds^*$, and the two-band Hamiltonian with BGS or subband scaling (both fitted to the eight-band results). The left side of the plot (negative k) corresponds to real k and the right side to imaginary k.

better agreement for the eight-band imaginary dispersion; this is further borne out for the tunneling current, as shown in Section 3.5.

To better understand this finding, we decompose the wave functions of the lowest conduction and valence subbands in Fig. 3.11 in terms of the spin-degenerate eight-band basis states $|S\rangle$, $|X\rangle$, $|Y\rangle$, and $|Z\rangle$[77]. The effects of confinement can be best understood by comparison with the wave functions for bulk InAs shown in Fig. 3.3; note that the well is confined in the x direction and we examine the dispersion for k_y, so the dominant p-like state here is $|Y\rangle$ rather than $|X\rangle$. For real k and especially the valence subband, band mixing effects are larger and more complicated for the well states compared to bulk. For instance, the perpendicular ($|X\rangle$ and $|Z\rangle$) basis states are very unevenly weighted in Fig. 3.11b in contrast to their exact equality in the bulk case, signifying significant mixing of the SO and HH bands in the valence subband. The change of the subband effective masses from the bulk values is a product of this band mixing. By contrast, as the magnitude of imaginary k increases, the subband wave functions tend to converge towards their bulk counterparts; the increased ratio of $|Y\rangle$ to $|X\rangle$ and $|Z\rangle$, which in

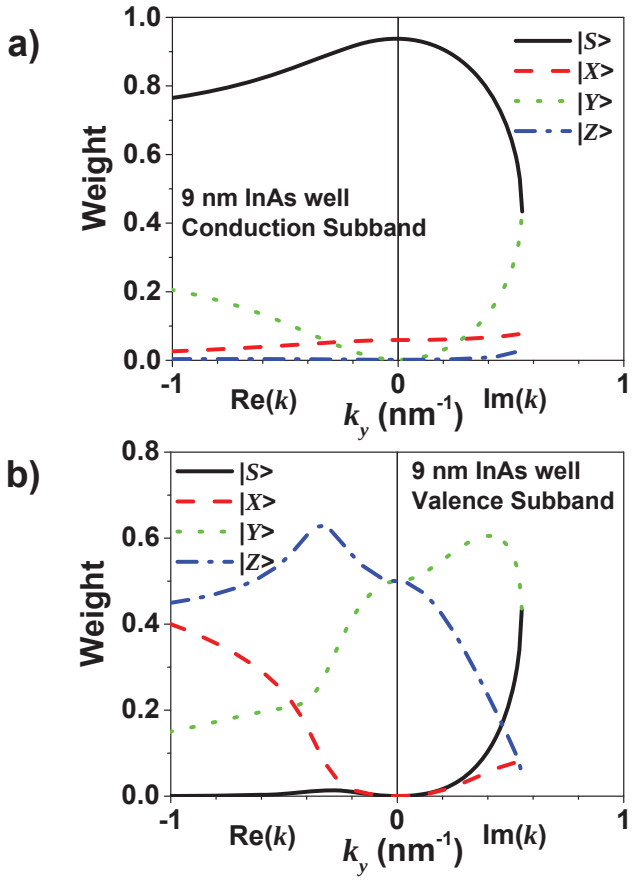

Figure 3.11: Projections onto the spin-summed eight-band basis states of the lowest (a) conduction and (b) valence subband wave functions for the quantum well in Fig. 3.10 as the wavevector k_y is varied for real and imaginary values.

turn become approximately equal, is in agreement with the form of the bulk LH wave functions[77]. This indicates that the imaginary dispersion is dominated by CB- and LH-like components, justifying the two-band Hamiltonian. Since P_2 describes the interaction between the CB and LH bands in this model, fitting it to the subband effective masses incorporates HH and SO contributions that are less relevant for the imaginary band structure. Retaining the bulk P_2 in the BGS method implies that the interband matrix element is basically unaltered by confinement.

Similar results hold for narrower quantum wells and different materials, as shown in Figs. 3.12a-b. In Fig. 3.12c, we also compare BGS with the subband structure for a cylindrical NW using TB[4] and a rectangular NW using eight-band $\mathbf{k}\cdot\mathbf{p}$. In all cases

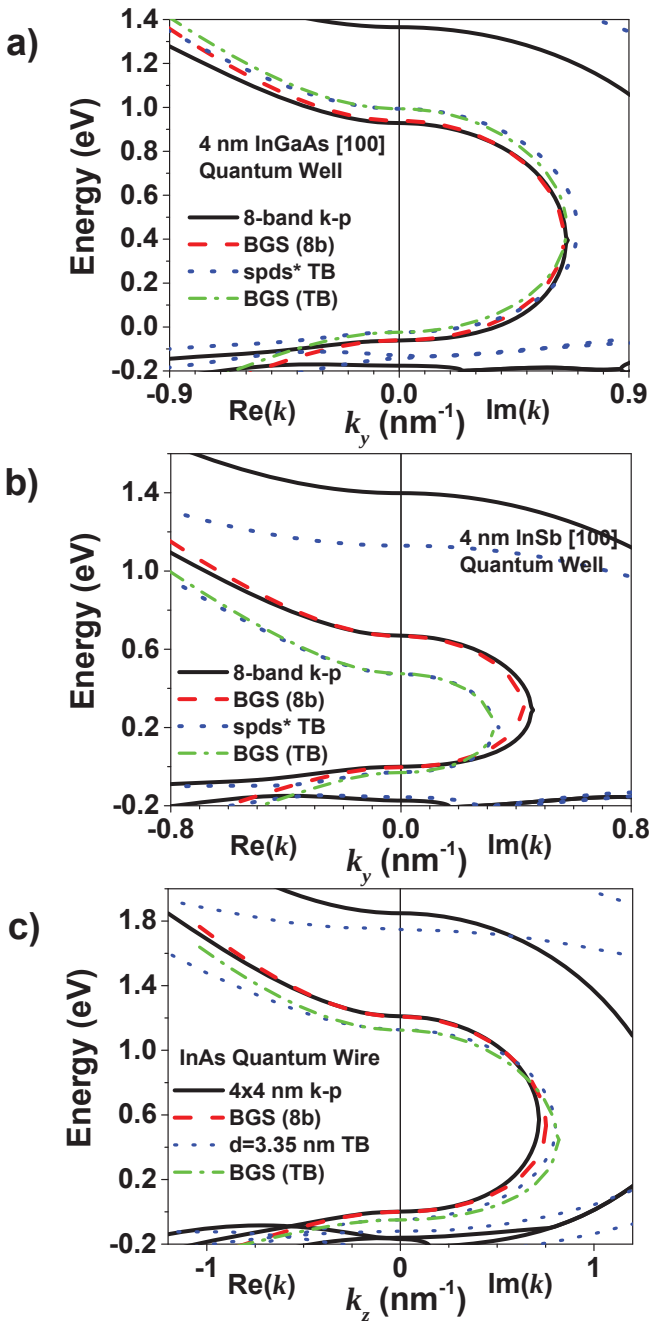

Figure 3.12: Real and imaginary dispersions for 4 nm (a) InGaAs quantum well (b) and InSb quantum well calculated using the eight-band $\mathbf{k}\cdot\mathbf{p}$ and *spds** methods as well as two-band predictions using BGS for the eight-band (8b) and TB gaps. (c) Dispersions for InAs quantum wires with 4x4 nm rectangular cross section (using eight-band $\mathbf{k}\cdot\mathbf{p}$) and 3.35 nm diameter cylindrical cross section (using TB); the latter is taken from Ref. [4].

we see that the BGS two-band model is more accurate at describing the imaginary than the real dispersions, and hence is suitable for approximating interband tunneling in QC devices. Since the $\mathbf{k}\cdot\mathbf{p}$ and TB calculations give different $E_{g,QC}$ under strong confinement, we perform BGS to the extracted gaps for both methods and find good agreement in every case, demonstrating that our argument is not restricted to eight-band $\mathbf{k}\cdot\mathbf{p}$ and hence not substantially affected by the presence of other bands. Nonetheless, BGS is still an approximation; even more exact fits to the full-band complex dispersions could be obtained by freely adjusting m and P_2, for instance[118, 119]. If detailed knowledge of the complex band structure is available through calculations like the ones shown here, such fitting should give the most accurate results. However, a huge variety of nanostructures can be realized in practice by modifying physical geometry, materials and alloying, strain, etc.; while theoretical and experimental studies of such structures frequently quote the associated energy gaps and sometimes the subband masses, comparatively few report the imaginary dispersion (which cannot be easily probed experimentally). Therefore, when the latter information is not available, BGS provides an alternative starting point for studying tunneling or other phenomena dependent on the complex dispersion.

3.5 Interband Tunneling in Confined Materials

3.5.1 Applying Semiclassical Tunneling Models using BGS

The Kane formula[96] gives the interband tunneling current under constant electric field F

$$J = A(F)\int_0^{\Delta E} dE\ \exp\left(-\frac{B}{F}\right)\left[f_L(E) - f_R(E)\right] \tag{3.20}$$

with $f_{L,R}$ being the Fermi distribution functions on either side of the tunnel junction and ΔE equal to the energy interval over which tunneling is possible. For bulk materials the

parameters A and B are

$$A_{\text{bulk}} = \frac{q^2 B_{bulk} F}{9\pi^2 \hbar E_g^2} \tag{3.21a}$$

$$B_{\text{bulk}} = \frac{\pi m_r^{1/2} E_g^{3/2}}{2q\hbar}, \tag{3.21b}$$

where the reduced mass m_r is defined as

$$m_r = \left(m_{CB}^{-1} + m_{LH}^{-1}\right)^{-1} = \frac{\hbar^2 E_g}{4P_2^2}, \tag{3.22}$$

the last expression following from Eq. 3.19. Based on our complex band structure results, we use BGS to modify the Kane model for QC structures by replacing the bulk E_g with $E_{g,QC}$, assuming that only the lowest conduction and valence subbands are important for tunneling. Since P in Eq. 3.22 remains constant, the reduced mass increases from its bulk value via

$$m_{r,BGS} = m_r \frac{E_{g,QC}}{E_g} = \frac{1}{m_{CB}^{-1} + m_{LH}^{-1}} \frac{E_{g,QC}}{E_g}. \tag{3.23}$$

Substituting $E_{g,QC}$ and $m_{r,BGS}$ into Eq. 3.21b leads to

$$B_{BGS} = B_{bulk} \left(\frac{E_{g,QC}}{E_g}\right)^2, \tag{3.24}$$

whereas the $E_{g,QC}$ dependence cancels for A and it retains its bulk form. Eq. 3.24 is equivalent to computing the WKB action integral for the BGS two-band model. This provides a satisfyingly simple way to adapt the Kane formula for QC using only knowledge of the bulk band structure and $E_{g,QC}$.

In Section 3.3, we demonstrate that the accuracy of the Kane formula can be improved for bulk materials using corrections for the split-off (SO) band and transverse nonparabolicity. Band mixing in QC devices means that a distinct "SO subband" cannot be defined and there is no simple way to develop an analytical three- or four-band model, so we will stick to the two-band B_{BGS}. However, the Kane formula should be adjusted for the different joint density of states (DOS) in lower dimensions.

3.5.2 Lower-Dimensional Tunneling Coefficients

The parameter A in the Kane tunneling formula for dimension d can be found using

$$A = \frac{q\pi}{9\hbar} \int_0^\infty \frac{dk_\perp^{d-1}}{(2\pi)^{d-1}} \exp\left(-\frac{2E_\perp}{\bar{E}_\perp}\right), \tag{3.25}$$

where $k_\perp$ is the transverse momentum, $\bar{E}_\perp = 2q\hbar F/(\pi m_r^{1/2} E_g^{1/2})$, and $E_\perp$ for nonparabolic bands is

$$E_\perp = E_g \left(\sqrt{1 + \frac{\hbar^2 k_\perp^2}{m_r E_g}} - 1\right), \tag{3.26}$$

which reduces to the parabolic result $E_\perp = \hbar^2 k_\perp^2 / 2m_r$ for small $k_\perp$. In quantum wells $d = 2$, so substituting Eq. 3.26 in Eq. 3.25 and changing variables we obtain

$$A_{\text{2-D}} = \frac{q\sqrt{m_r E_g}}{18\hbar^2} \int_0^\infty dx \exp\left(-\beta\sqrt{1+x^2} - 1\right), \tag{3.27}$$

such that $\beta = 2E_g/\bar{E}_\perp = 2B/qF$. No exact solution exists for this integral; however, using the lowest order Taylor expansion of the square root in the exponential (the parabolic or effective mass approximation) gives precisely $\sqrt{\pi/2\beta}$. We therefore take

$$\int_0^\infty dx \exp\left(-\beta\sqrt{1+x^2} - 1\right) \approx \sqrt{\frac{\pi}{2\beta}} + \frac{a}{\beta^c}, \tag{3.28}$$

where a and c are fitted to the numerical integration results. We find that $a = 0.425$ and $c = 1.2$ give less than 4% error for all $\beta > 0.1$, which is nearly always the case for realistically achievable electric fields in devices (at such high fields, real device performance might in any case be limited by series resistance or other mechanisms like impact ionization). Using this expression leads to Eq. 3.29. In 1-D, the integration in Eq. 3.25 disappears entirely, leaving Eq. 3.30.

In general, the prefactor A is found from integration over transverse modes; for 2-D quantum wells this approximately leads to

$$A_{\text{2-D}} = \frac{q\sqrt{m_r E_{g,QC}}}{18\hbar^2} \left[\sqrt{\frac{q\pi F}{4B_{BGS}}} + 0.185\left(\frac{qF}{B_{BGS}}\right)^{1.2}\right]. \tag{3.29}$$

In 1-D, there are no transverse modes and

$$A_{\text{1-D}} = \frac{2q\pi^2}{9h}, \tag{3.30}$$

which is equal to the quantum conductance times a factor of $\pi^2/9$ from the Kane transmission coefficient[96]. Similar expressions have been derived assuming parabolic bands for graphene and 2-D semiconductors[120, 121]. Aside from constant prefactors, the key change to the bulk Kane formula in lower dimensions is the reduction of the power of F in A. Eqs. 3.29-3.30 are more appropriate for lower-dimensional devices, though many existing compact and TCAD models are based on the 3-D Eq. 3.21a.

Since tunneling depends sensitively on potential, we also need to consider the effects of QC on the electrostatics. For TFETs, the quasi-Fermi levels E_f change in confined structures due to DOS, altering the built-in voltage V_{bi} between source and drain (which also increases due to a larger $E_{g,QC}$). In quantum wells, for instance, E_f can be evaluated assuming nonparabolic subbands[122]. In TCAD simulations, the electron effective DOS can then be adjusted to give the same Fermi level. For holes, 3-D formulas for E_f give relatively minor errors because heavy holes dominate the DOS. The elevation of the lowest conduction subband above the bulk CB edge also reduces the electron affinity, which leads to a "threshold voltage" shift when matching semiclassical and quantum I-V curves but does not otherwise affect the results.

To summarize, when applying the Kane model to quantum confined devices like TFETs, the bulk value of B in the Kane formula should be replaced by Eq. 3.24. When possible the dimensionally appropriate A should also be used. Despite its neglect of other quantum effects, BGS matches NEGF simulations very well as shown below, demonstrating its practical value. The BGS model focuses on tunneling along the unconfined direction(s) in nanostructures. In structures which transport occurs along the confined direction, i.e., "line tunneling" perpendicular to the gate in TFETs, electric field-dependent confinement may work to increase the effective band gap. In this case, using BGS alone will provide an "upper limit" on the tunneling since it neglects the field-induced gap increase, and additional field-dependent corrections will be needed to model these effects[48, 123].

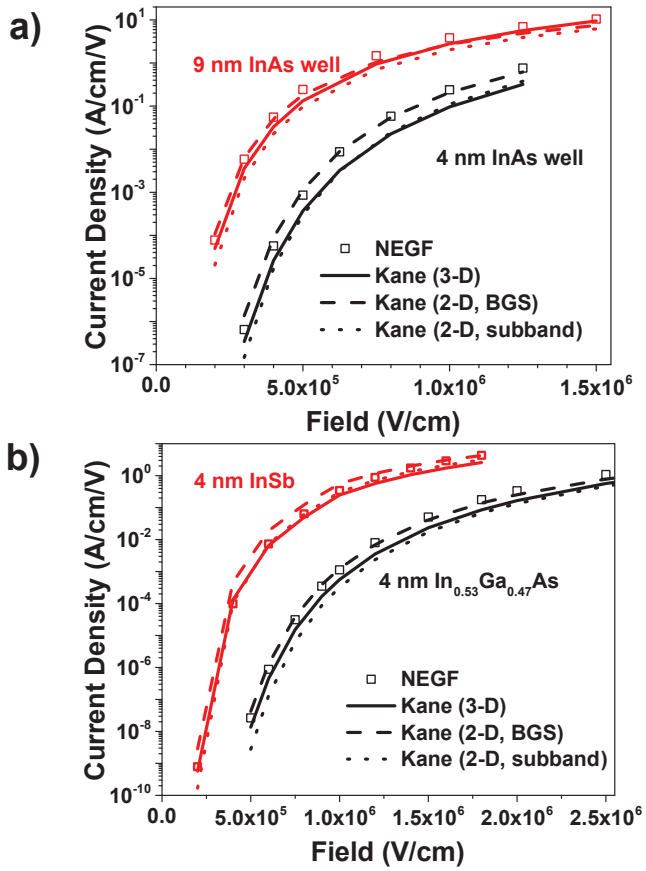

Figure 3.13: Normalized current densities $J/\Delta E$ versus applied field for (a) 4 and 9 nm thick InAs quantum wells and (b) 4 nm InGaAs and InSb quantum wells using eight-band NEGF and the 3-D and 2-D Kane formulas.

3.5.3 Tunneling in Constant Fields

To validate these analytical semiclassical models, we compare them with eight-band $\mathbf{k}\cdot\mathbf{p}$ NEGF calculations. Since the semiclassical formulas are derived assuming constant electric field, we first simulate quantum wells and wires with linear voltage drops. Hard wall boundary conditions are imposed, i.e., the wave functions vanish at the boundaries of the confined directions.

The NEGF and analytical tunneling currents (normalized by the tunneling energy window ΔE) for two different InAs quantum wells are shown in Fig. 3.13a. The 3-D Kane curves correspond to the use of A_{bulk} with B_{BGS} in Eq. 3.20, while the 2-D Kane

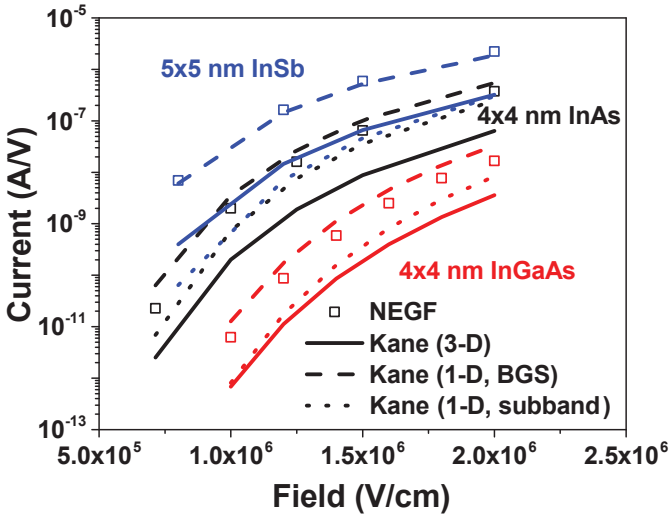

Figure 3.14: Normalized current $J/\Delta E$ versus applied field for InGaAs, InAs, and InSb quantum wires using eight-band NEGF and the 3-D and 1-D Kane formula calculations.

curves utilize the DOS-corrected $A_{2\text{-}D}$; for the latter we compare B_{BGS} with the subband scaling approximation, where the subband effective masses are used to evaluate m_r. In both cases we find that bulk Kane model tends to underestimate the current density with the discrepancy increasing for stronger confinement, whereas the 2-D BGS model gives good quantitative agreement. Using the subband effective masses also leads to underestimation of current, as expected and explained from the complex band structure in Section 3.4. The same trends hold for the $In_{0.53}Ga_{0.47}As$ and InSb quantum wells we simulated, as shown in Fig. 3.13b.

In Fig. 3.14 we calculate several quantum wires with NEGF and compare the 3-D and 1-D Kane models using BGS and subband scaling. Again, the 1-D BGS model gives the best agreement with the NEGF simulations, though it slightly overestimates the current for InAs and InGaAs. Quantum confinement and band mixing effects are very strong in these wires, so that the tunneling current may be particularly sensitive to details of the band structure in these cases and the BGS approximation leads to larger errors.

On the whole, these results show that the Kane formula, properly adjusted for dimensionality and QC through BGS, is still qualitatively and often quantitatively useful for modeling tunneling in lower-dimensional structures. However, the bulk Kane model,

which is sometimes the only available option for modeling interband tunneling in TCAD device simulators, tends to underestimate current even after applying BGS.

3.6 Conclusion

The primary physical limitations of the Kane formula, namely the neglect of multiband corrections to the transverse gap and the tunneling probability, can be mostly overcome using simple parameter-free adjustments of the standard tunneling equations. These corrections can be easily implemented in analytical device models and simulation software. Of course, NEGF or equivalent quantum transport calculations are always preferable if exact results are required or new physical effects need to be considered. However, in many practical applications they may be infeasible or unnecessary; in these cases our results indicate that semiclassical models, properly used, provide good qualitative and even quantitative accuracy for bulk tunneling.

We find from the complex band structure evaluated using $\mathbf{k}\cdot\mathbf{p}$ and TB calculations that the main correction to the two-band Hamiltonian in lower-dimensional nanostructures is the scaling of the QC band gap. This leads to a BGS adjustment of the Kane formula that allows the tunneling current to be approximated using only the bulk properties and the QC gap, which is especially useful for rapid device design studies or when limited information is available about the band structure. Through comparisons with constant field simulations for different materials and dimensions, we demonstrate that the BGS scheme leads to good agreement with NEGF simulations and therefore enables semiclassical modeling of direct tunneling.

CHAPTER 4

Electrostatic and Current Models for TFETs

The world is charged with the grandeur of God.

G. M. Hopkins, "God's Grandeur"

Because TFETs operate according to fundamentally different principles compared to traditional MOSFETs, their device characteristics are qualitatively as well as quantitatively different. Most of the research on TFETs has been on the device level, attempting to realize the full potential of the tunneling process. However, the unique advantages and disadvantages of a new device like the TFET mean that it cannot be simply substituted for MOSFETs everywhere but rather deployed intelligently to derive the maximum benefit. To evaluate viability of TFET adoption in future applications, circuit-scale modeling and design will be crucial. Despite the extensive experimental and simulation work done on TFETs, there remains a need for simple, predictive models that provide design insight and facilitate circuit and system-level modeling.

There is increasing interest in TFET analytical modeling to provide physical insight and facilitate compact modeling for circuit-level studies. As with MOSFETs, both electrostatics and transport need to be considered for proper modeling. Because interband tunneling is highly sensitive to the band bending, the full 2-D channel potential might be expected to play an important role in understanding TFET characteristics. However, in an arbitrary device structure the 2-D Poisson equation may be analytically intractable, particularly for lateral TFET structures where the current flow is roughly parallel to the gate. Some researchers use series expansions[124, 125, 126] to compute the channel potential in ideal lateral TFETs and integrate the tunneling current over the device vol-

ume. These provide rigorous and exact solutions for the cases under study, but require numerical evaluation and are not easily generalizable to more complicated conditions. Attempts have also been made to model the 2-D effects by assuming geometric tunneling paths[127, 128], which do not however necessarily correspond to the actual electric field lines (or paths of least action) in realistic TFET structures. Still others use analytical approximations[129, 130, 131], i.e. a pseudo-2-D approach, to model the potential at the channel surface or center, effectively assuming tunneling is dominated by 1-D effects. Many of these works only analyze certain idealized structures and do not account for important device variables like source and drain doping. Additionally, analytical solutions of the Poisson equation preclude inclusion of the nonlinear mobile charge, though a few works use approximations to capture some key effects of channel carriers [131, 132].

In this chapter we present our work in this area. We use the pseudo-2-D approach[133] as the basis for our analytical models, striving to unify the analysis of different device structures, including double gate (DG), nanowire (NW), and SOI TFETs, and investigate nonidealities. While approximations and limitations are inevitable in any analytical approach, it is of great practical interest to explore how far this framework can be stretched. We focus on the device electrostatics, using standard approximations to describe tunneling; our detailed study of the potential yields new general device insights and clarifies common modeling assumptions. Our approach is semiclassical, neglecting quantum effects like channel subband formation, which we expect will introduce quantitative changes but preserve the relevant device trends we identify. We demonstrate the accuracy of the model through comparison with numerical simulations and experiments and use it to explain why 2-D tunneling may be neglected in well-scaled devices. We carefully study the impact of degeneracy on TFETs, providing a simple analytical model to explain its influence and showing how it alters both the device transport and electrostatics. By examining the modeling and simulation results, we discover that Thomas-Fermi electrostatic screening has a major impact on the tunneling current in TFETs made with low density-of-states materials. Much of the work in this chapter has been published in various forms in Refs. [7, 8, 9].

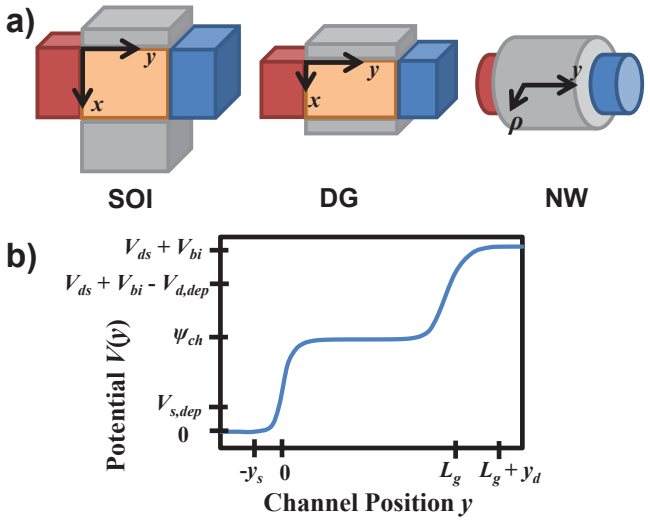

Figure 4.1: a) Structures and coordinate systems under consideration. b) Potential and variables along the channel at arbitrary x or ρ in device.

4.1 TFET Model Development

We use the pseudo-2-D approach to calculate the electrostatic potential of lateral TFETs. By assuming a polynomial form for the potential perpendicular to the gate, the 2-D Poisson equation reduces to an analytically solvable 1-D equation. This procedure, also known as the polynomial potential model, has been used to analyze MOSFET short-channel effects (SCE) in subthreshold[134, 135, 136]. We will see that the method allows easy generalization to different structures and that it is valid over the design space for future multigate TFETs.

4.1.1 2-D Electrostatic Potential Model

The basic TFET structure we study is a gated lateral p-i-n diode. We consider three device configurations: the silicon-on-insulator (SOI) thin body, double-gate (DG), and gate-all-around nanowire (NW), as illustrated in Fig. 4.1(a). These multigate devices, besides being straightforward to model, are also the basic structures most likely to be used in future technology nodes. For simplicity we first review the DG derivation. Our sign

conventions are for n-type TFETs, but the model is equally applicable to p-type devices by reversing the polarities of the voltages. We consider the operating regime when the mobile charge in the channel can be neglected (corresponding to $\psi_{ch} < V_{ds} + V_{bi}$ in Fig. 4.1(b)) and hence the Poisson equation in the channel is

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) V(x,y) = \frac{qN_c}{\epsilon_{ch}}, \tag{4.1}$$

where N_c is the channel doping concentration and ϵ_{ch} is the channel dielectric constant. In most TFETs the channel is lightly doped and N_c can be neglected since it does not significantly affect the potential; it is included here for generality. We make three standard approximations in our derivation: 1) the potential in the x-direction (perpendicular to the gate) is parabolic with the form $V(x,y) = a(y) + b(y)x + c(y)x^2$, 2) the electric field lines in the oxide region are perpendicular, and 3) the potential in the source and drain is described by the depletion approximation. For the DG, $b(y) = -c(y)t_{ch}$ from the symmetry requirement that the vertical electric field at the center must be zero. Approximation 2 corresponds to neglect of lateral fringing fields in the gate oxide and gives

$$-\left.\frac{\partial V}{\partial x}\right|_{x=0} = \frac{\epsilon_{ox}}{\epsilon_{ch} t_{ox}}(V'_{gs} - a(y)) = -b(y). \tag{4.2}$$

Here $V'_{gs} = V_{gs} - q\Phi_{ms}$ is the gate bias relative to the source Fermi level, ϵ_{ox} is the oxide permittivity, t_{ox} is the oxide thickness, and Φ_{ms} is the difference between the gate work function and the source Fermi level. Substituting the parabolic potential into the Poisson equation to remove the partial derivative in x, and re-arranging terms, we find

$$\frac{d^2V(x,y)}{dy^2} - \frac{V(x,y) - \psi_{ch}(x)}{\lambda_{DG}^2(x)} = 0, \tag{4.3}$$

where

$$\lambda_{DG}(x) = \sqrt{\frac{\epsilon_{ch} t_{ox} t_{ch}}{2\epsilon_{ox}}\left(1 + \frac{\epsilon_{ox}(t_{ch}x - x^2)}{\epsilon_{ch} t_{ch} t_{ox}}\right)} \tag{4.4}$$

and

$$\psi_{ch}(x) = V'_{gs} + \frac{qN_c\lambda^2(x)}{\epsilon_{ch}}. \tag{4.5}$$

It is easy to verify that Eq. 4.4 reduces to the well-known forms for the surface and center potential in a DG structure[135]

$$\lambda_{DG,surf} = \sqrt{\frac{\epsilon_{ch} t_{ch} t_{ox}}{2\epsilon_{ox}}} \tag{4.6}$$

$$\lambda_{DG,cent} = \sqrt{\frac{\epsilon_{ch} t_{ch} t_{ox}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox} t_{ch}}{4\epsilon_{ch} t_{ox}}\right)}. \tag{4.7}$$

All information about the structure (i.e. SOI, DG, NW) is contained in λ and ψ_{ch} and hence the subsequent expressions hold equally true for any geometry, provided the parabolic potential is a good approximation. The boundary conditions for Eq. 4.3 are continuity of potential and electric field in the depletion region of the source, assuming constant doping N_s:

$$V(x,0) = V_{s,dep}(x), \tag{4.8a}$$

$$\frac{\partial V(x,0)}{\partial y} = \frac{qN_s y_{s,dep}(x)}{\epsilon_{ch}} = \sqrt{\frac{2qN_s V_{s,dep}(x)}{\epsilon_{ch}}}, \tag{4.8b}$$

where $V_{s,dep}$ and $y_{s,dep}$ are the source depletion potential and width, respectively. Similar conditions hold for the drain side depletion potential $V_{d,dep}$ and depletion width $y_{d,dep}$ at the boundary $y = L_g$, where L_g is the gate length. These can be solved to yield the depletion potentials (suppressing x dependences for clarity):

$$V_{s,dep} = \psi_{ch} + V_{s0} - \sqrt{2\psi_{ch} V_{s0} + V_{s0}^2}, \tag{4.9}$$

$$V_{d,dep} = V_{ds} + V_{bi} - \psi_{ch} + V_{d0} - \sqrt{2(V_{ds} + V_{bi} - \psi_{ch}) V_{s0} + V_{d0}^2}, \tag{4.10}$$

$$V_{s0} = \frac{qN_s \lambda^2}{\epsilon_{ch} \coth^2(\frac{L_g}{\lambda})}, \tag{4.11}$$

$$V_{d0} = \frac{qN_d \lambda^2}{\epsilon_{ch} \coth^2(\frac{L_g}{\lambda})}. \tag{4.12}$$

Here V_{ds} is the applied drain bias, V_{bi} is the built-in voltage between the source and drain, and N_d is the drain doping. The relevant potentials are illustrated in Fig. 4.1(b). The

general 2-D solution for the channel potential is then

$$V(y) = \begin{cases} \dfrac{qN_s}{2\epsilon_{ch}}(y + y_s(x))^2 & y_s(x) \le y < 0 \\ \psi_{ch}(x) + \dfrac{(V_{s,dep} - \psi_{ch}(x))}{\sinh(\frac{L_g}{\lambda(x)})} \sinh\left(\dfrac{L_g - y}{\lambda(x)}\right) & \\ + \dfrac{(V_{bi} + V_{ds} - \psi_{ch}(x) - V_{d,dep})}{\sinh(\frac{L_g}{\lambda(x)})} \sinh\left(\dfrac{y}{\lambda(x)}\right) & 0 \le y < L_g \\ V_{bi} + V_{ds} - \dfrac{qN_d}{2\epsilon_{ch}}(y - L_g - y_d(x))^2 & L_g \le y \le L_g + y_d(x) \end{cases} \tag{4.13}$$

which gives the potential at arbitrary channel depth x, in contrast to previous works which only derived the surface ($x = 0$) or center ($x = \dfrac{L_g}{2}$) potential. The source and drain depletion widths are given by

$$y_s(x) = \sqrt{\frac{2\epsilon_{ch} V_{s,dep}(x)}{qN_s}} \tag{4.14}$$

$$y_d(x) = \sqrt{\frac{2\epsilon_{ch} V_{d,dep}(x)}{qN_d}}. \tag{4.15}$$

These results can be easily generalized to heterojunctions by considering the band offsets and continuity of displacement (rather than electric field) at the source/channel interface. The model shares some features with Ref. [137], but by placing all the x-dependence in the characteristic length λ and effective channel potential ψ_{ch}, the results are more easily manipulated and can be trivially generalized to the SOI or NW by using the appropriate λ and ψ_{ch}, as we will now show.

The SOI structure can be analyzed similarly to the DG device. The primary difference is the presence of different front and back gate biases V'_{gf} and V'_{gb}, as well as oxide permittivities and thicknesses $\epsilon_{ox,f}$, $\epsilon_{ox,b}$, $t_{ox,f}$, and $t_{ox,b}$. Hence the boundary condition Eq. 4.2 is altered for the front and back gates. The corresponding channel potential and

characteristic length as a function of x are

$$\psi_{ch,SOI} = (1 + k_f x)\frac{k_f V'_{gf} + k_b V'_{gb} + k_f k_b t_{ch} V'_{gf}}{k_f + k_b + k_f k_b t_{ch}}$$
$$- k_f x V'_{gf}, \tag{4.16}$$

$$\lambda_{SOI} = \sqrt{\frac{t_{ch}(2 + k_b t_{ch})(1 + k_f x - \frac{(k_f + k_b + k_f k_b t_{ch})x^2}{t_{ch}(2+k_b t_{ch})})}{2(k_f + k_b + k_f k_b t_{ch})}}, \tag{4.17}$$

where $k_f = \frac{\epsilon_{ox,f}}{\epsilon_{ch} t_{ox,f}}$ and $k_b = \frac{\epsilon_{ox,b}}{\epsilon_{ch} t_{ox,b}}$. Setting the front and back voltages and oxides equal reduces the results to $\psi_{ch,DG}$ and λ_{DG} as expected.

For NWs, the Poisson equation is solved for the cylindrical system, so the radial coordinate ρ is substituted for x (see Fig. 4.1) and Eq. 4.1 for the DG is replaced by

$$\frac{1}{\rho}\left[\frac{\partial}{\partial \rho}\left(\rho \frac{\partial V}{\partial \rho}\right) + \rho \frac{\partial^2 V}{\partial y^2}\right] = 0. \tag{4.18}$$

Similarly, the oxide field (previously Eq. 4.2) in cylindrical coordinates is equal to

$$-\left.\frac{\partial V}{\partial \rho}\right|_{\rho = r_s} = -\frac{\epsilon_{ox}}{\epsilon_{ch}} \frac{V'_{g,eff} - V(r_s)}{r_s \ln\left(1 + \frac{t_{ox}}{r_s}\right)}, \tag{4.19}$$

where r_s is the radius of the nanowire. The channel potential $\psi_{ch,NW} = \psi_{ch,DG}$ is unchanged from the DG case due to symmetry, but the characteristic length for the channel potential is

$$\lambda_{NW} = \sqrt{\frac{2\epsilon_{ch} r_s^2 \ln\left(1 + \frac{t_{ox}}{r_s}\right) + \epsilon_{ox} r_s^2}{4\epsilon_{ox}} - \frac{\rho^2}{4}}. \tag{4.20}$$

It is easily shown by substitution that this generalized λ reduces to the corresponding SOI and NW scaling lengths previously given in the literature[134, 136].

4.1.1.1 Inclusion of Mobile Channel Charge

Next we examine the case when the mobile charge is large and Eq. 4.1 is not valid. When the channel potential approaches that of the drain, it is well known that an inversion layer is formed near the drain, leading to two effects: ψ_{ch} is almost pinned near the drain potential $V_{ds} + V_{bi}$ and only changes slowly with gate bias, while the channel charge pinches the

potential near the source and reduces the tunneling distance[132]. These effects account for the nonlinear behavior of $I_d - V_{ds}$ at low V_{ds}. Exact solutions require numerical calculation, but variational treatments of the nonlinear Poisson equation show these effects may be approximated with a variable λ. Here we adopt the depth-independent form

$$\frac{1}{\lambda^2} = \frac{1}{\lambda_0^2} - \frac{\alpha N_{inv}}{\epsilon_{ch} t_{ch} \psi_{ch}}, \tag{4.21}$$

where λ_0 is the previously given characteristic length for the structure and α is a free parameter, which can be set to 8 according to [132]. Pinning occurs near $V_{th} = V_{ds} + V_{bi}$, so we set ψ_{ch} to V_g' when it is less than V_{th} and equal to V_{th} otherwise. The inversion charge is approximated by $N_{inv} = 2C_{ox}(V_{gs}' = V_{th})$. This is reasonable far above threshold, but the results for $V_{gs}' \approx V_{th}$ are improved by using a smoothing function, such as $\alpha = \alpha_0 \left[1 - c \exp\left(-\left(\frac{V_{gs}' - V_{th} - \eta}{\sigma}\right)^2\right)\right]$, which will be seen when we compare the model results with simulations for low V_{ds}.

4.1.1.2 Extension to Nonabrupt Doping Profiles

The pseudo-2-D approach can also help us to analytically study nonidealities in the device structure. For instance, junction abruptness is an important factor in determining TFET performance. A lateral doping profile near the source or drain can be straightforwardly incorporated in our model. For specificity we consider a Gaussian junction overlap underneath the gate, as illustrated in Fig. 4.2(a). Then in the channel ($0 < y < L_g$), Eq. 4.3 is extended by including a Gaussian doping profile with decay length σ:

$$\frac{d^2V(x,y)}{dy^2} - \frac{V(x,y) - \psi_{ch}(x)}{\lambda_{DG}(x)} = \frac{qN_s}{\epsilon_{ch}} \exp(-\frac{y^2}{2\sigma^2}). \tag{4.22}$$

In contrast to the ideal p-i-n structure, we find that the presence of significant mobile charge near the overlap edge ($y = 0$) strongly screens the source depletion, so that the source potential may be well approximated as a constant value (zero for convenience) and

electric field continuity at $y = 0$ may be neglected. The solution of Eq. 4.22 is then

$$V_1(y) = A\exp\left(\frac{y}{\lambda}\right) + B\exp\left(-\frac{y}{\lambda}\right) + \frac{\sqrt{\pi}\alpha\lambda\sigma}{4}\exp\left(\frac{\sigma^2}{4\lambda^2}\right)$$
$$\left(\exp\left(\frac{y}{\lambda}\right)\mathrm{erf}\left(\frac{y}{\sigma}+\frac{\sigma}{2\lambda}\right) - \exp\left(\frac{-y}{\lambda}\right)\mathrm{erf}\left(\frac{y}{\sigma}-\frac{\sigma}{2\lambda}\right)\right), \tag{4.23}$$

where

$$A = \frac{1}{2\sinh\left(\frac{L_g}{\lambda}\right)}\left(\psi_{ch}\exp\left(-\frac{L_g}{\lambda}\right) - \frac{\sqrt{\pi}\alpha\lambda\sigma}{4}\exp\left(\frac{\sigma^2}{4\lambda^2}\right)\right.$$
$$\left.\times\left\{\exp\left(\frac{L_g}{\lambda}\right)\mathrm{erf}\left(\frac{L_g}{\sigma}+\frac{\sigma}{2\lambda}\right) - \exp\left(-\frac{L_g}{\lambda}\right)\times\left[\mathrm{erf}\left(\frac{L_g}{\sigma}-\frac{\sigma}{2\lambda}\right) + 2\mathrm{erf}\left(\frac{\sigma}{2\lambda}\right)\right]\right\}\right), \tag{4.24a}$$

$$B = \frac{1}{2\sinh\left(\frac{L_g}{\lambda}\right)}\left(-\psi_{ch}\exp\left(\frac{L_g}{\lambda}\right) + \frac{\sqrt{\pi}\alpha\lambda\sigma}{4}\exp\left(\frac{\sigma^2}{4\lambda^2}\right)\right.$$
$$\left.\times\left\{\exp\left(\frac{L_g}{\lambda}\right)\left[\mathrm{erf}\left(\frac{L_g}{\sigma}+\frac{\sigma}{2\lambda}\right) - 2\mathrm{erf}\left(\frac{\sigma}{2\lambda}\right)\right] - \exp\left(-\frac{L_g}{\lambda}\right)\mathrm{erf}\left(\frac{L_g}{\sigma}-\frac{\sigma}{2\lambda}\right)\right\}\right), \tag{4.24b}$$

and $\alpha = \dfrac{qN_s}{\epsilon_{ch}}$. Eq. 4.23 can be substituted for the second term in Eq. 4.13 as the source-side contribution to the channel potential. Expressions for other doping profiles, including constant or linearly graded profiles in the overlap region, can be similarly derived. Underlap effects can also be incorporated using the appropriate depletion approximation in the source and/or drain.

The primary limitation of this model is its neglect of carriers in the overlap region which reduce the band bending. However, we can find the model's region of validity by noting that under normal operating conditions, the overlap potential should always be higher than the source, whereas Eq. 4.23 allows an unphysical "hump" to occur when the overlap is too long or heavily doped. Fig. 4.2(b) contrasts a well-behaved solution (Gaussian 1) with one for large σ (Gaussian 2) where the neglect of mobile charge creates an accumulation region near the source. Hence if Eq. 4.23 has a minimum in the channel ($y > 0$), we expect the model to be inaccurate. Finding the minimum via differentiation,

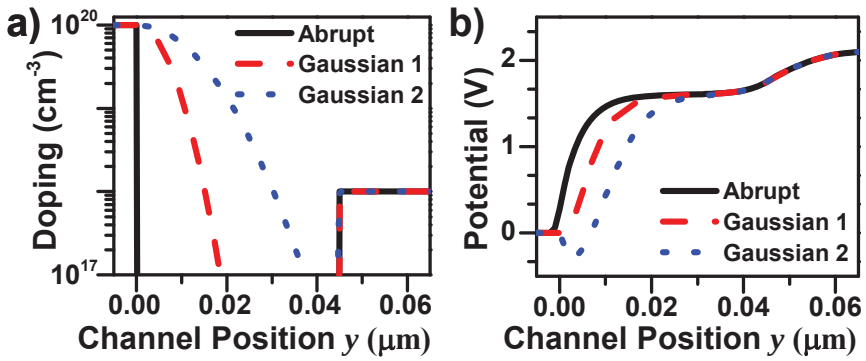

Figure 4.2: a) Schematic doping and b) analytical potential along channel for abrupt and Gaussian doping profiles in the channel. Gaussian 2 represents an analytical solution (large σ) beyond the region of validity that leads to an unphysical accumulation "hump" in the potential.

we obtain the validity condition

$$\psi_{ch} > \frac{\sqrt{\pi}\alpha\lambda\sigma}{2} \exp\left(\frac{\sigma^2}{4\lambda^2}\right)\left(1 - \text{erf}\left(\frac{\sigma}{2\lambda}\right)\right). \tag{4.25}$$

In the absence of direct source-to-drain leakage the threshold for tunneling occurs when the channel potential crosses the bandgap E_g, so $\psi_{ch} = E_g/q$ can be used in Eq. 4.25 to determine the operating regime of the model.

4.2 Developing TFET $I - V$ Models

Using the electrostatic model we have developed, we can now find the tunneling distances by inverting the potential. Equation 4.13 gives the potential explicitly as a function of y. We can invert these equations to obtain y as a function of potential. Whereas inversion of $V(y)$ in the source and drain is straightforward, the presence of multiple sinh functions in the channel presents a greater challenge. However, by rewriting the hyperbolic sine functions in terms of exponentials, we obtain

$$V(y) = \psi_{ch} + A\exp\left(\frac{y}{\lambda}\right) + B\exp\left(-\frac{y}{\lambda}\right) \tag{4.26}$$

where

$$A = \frac{V_{ds} + V_{bi} - V_{d,dep} - \psi_{ch}\left(1 - \exp - \left(\frac{L_g}{\lambda}\right)\right) - V_{s,dep}\exp\left(-\frac{L_g}{\lambda}\right)}{2\sinh\left(\frac{L_g}{\lambda}\right)} \tag{4.27}$$

$$B = \frac{V_{d,dep} - V_{ds} - V_{bi} + \psi_{ch}\left(1 - \exp\left(\frac{L_g}{\lambda}\right)\right) + V_{s,dep}\exp\left(\frac{L_g}{\lambda}\right)}{2\sinh\left(\frac{L_g}{\lambda}\right)}. \tag{4.28}$$

This equation can be inverted to find

$$y = \lambda \ln\left[\frac{V - \psi_{ch} + \sqrt{(V-\psi_{ch})^2 - 4AB}}{2A}\right]. \tag{4.29}$$

The tunneling distance from the source to channel occurs at

$$y_t(V) = \lambda \ln\left[\frac{V + (E_g/q) - \psi_{ch} + \sqrt{(V + (E_g/q) - \psi_{ch})^2 - 4AB}}{2A}\right] + y_{s,dep} - \sqrt{\frac{2\epsilon_{ch}V}{qN_s}} \tag{4.30}$$

while from the channel to the drain it is

$$y_t(V) = L_g + y_{d,dep} - \sqrt{\frac{2\epsilon_{ch}}{qN_d}(V_{bi} + V_{ds} - V - (E_g/q))} - \lambda \ln\left[\frac{V - \psi_{ch} + \sqrt{(V-\psi_{ch})^2 - 4AB}}{2A}\right]. \tag{4.31}$$

The largest contribution to current comes at the energy at which the tunneling distance is minimum. By straightforward manipulation, this energy for source-side tunneling can be found

$$V_{min} = \psi_{ch} + V_{s,eff} - \sqrt{(\psi_{ch} + V_{s,eff})^2 + 4AB - (\psi_{ch} - (E_g/q))^2}. \tag{4.32}$$

On the drain side,

$$V_{min} = \psi_{ch} - V_{d,eff} + \sqrt{(\psi_{ch} - V_{d,eff})^2 + 4AB - \psi_{ch}^2 + 2V_{d,eff}(V_{bi} + V_{ds} - (E_g/q))}. \tag{4.33}$$

Here $V_{s,eff} = \dfrac{qN_s\lambda^2}{\epsilon_{ch}}$ and $V_{d,eff} = \dfrac{qN_d\lambda^2}{\epsilon_{ch}}$.

We note the importance of the source and drain doping in deriving the tunneling path: because the source depletion causes the electric field to reach its maximum at $y = 0$, the minimum tunneling path will occur somewhere between $y = -y_s$ and $y = 0$ ($V = 0$ and

$V = V_{s,dep}$). This is in contrast to the usual models[124, 127, 129] which do not account for the source doping and assume tunneling paths beginning at the source edge ($y = 0$).

Once the energy-dependent and minimum tunneling distances have been determined from the appropriate electrostatic model, the tunneling current can be computed using the analytical tunneling models discussed in Chapter 3. We adapt the Kane formula for nonuniform fields by defining the effective field $F = E_g/qy_{tunn}$ at a given energy. For greatest accuracy, the current should be calculated by integrating the contributions of all energies and along the channel thickness x via

$$I = A_{tunn} \int dE \int dx \left(\frac{E_g}{qy_{tunn}(x,E)} \right)^n \exp\left(-\frac{qB_{tunn} y_{tunn}(x,E)}{E_g} \right) \left[f_s(E) - f_d(E) \right]. \tag{4.34}$$

where $f_{s,d}(E)$ are the Fermi-Dirac equations for the source and drain respectively. For a fully analytic model, we can approximate that the tunneling is dominated by contributions at x in the channel surface or center and at the minimum tunneling distance $y_{tunn}(E = qV_{min,s})$. The current then reduces to

$$I = A_{tunn} t_{ch} \left(\frac{E_g}{qy_{tunn,minn}} \right)^n \min(q\Psi_{ch} - E_g, qV_{ds}) \exp\left(\frac{qB_{tunn} y_{tunn,min}}{E_g} \right). \tag{4.35}$$

4.3 Results and Discussion

4.3.1 Validation of Electrostatic Model

To test the usefulness of our model, we compare it with potential profiles from numerical TCAD simulations[114]. We will work at the level of the Poisson equation and WKB approximation in both our model and simulations, which at least ensures consistency in our comparisons. Comparisons with NEGF simulations including BGS will be presented at the end of this chapter. For brevity, we will focus on results from silicon DG and $In_{0.53}Ga_{0.47}As$ NW p-i-n TFETs simulated using Fermi-Dirac statistics and the nonlocal tunneling model[114]; the TCAD model parameters are calibrated using experimental tunnel diode data. Although we sweep a wide range of parameters, our baseline devices have $L_g = 45$ nm, t_{ch} or $2r_s = 8$ nm, 1 nm t_{ox} SiO_2 oxide, and abrupt source/drain junctions

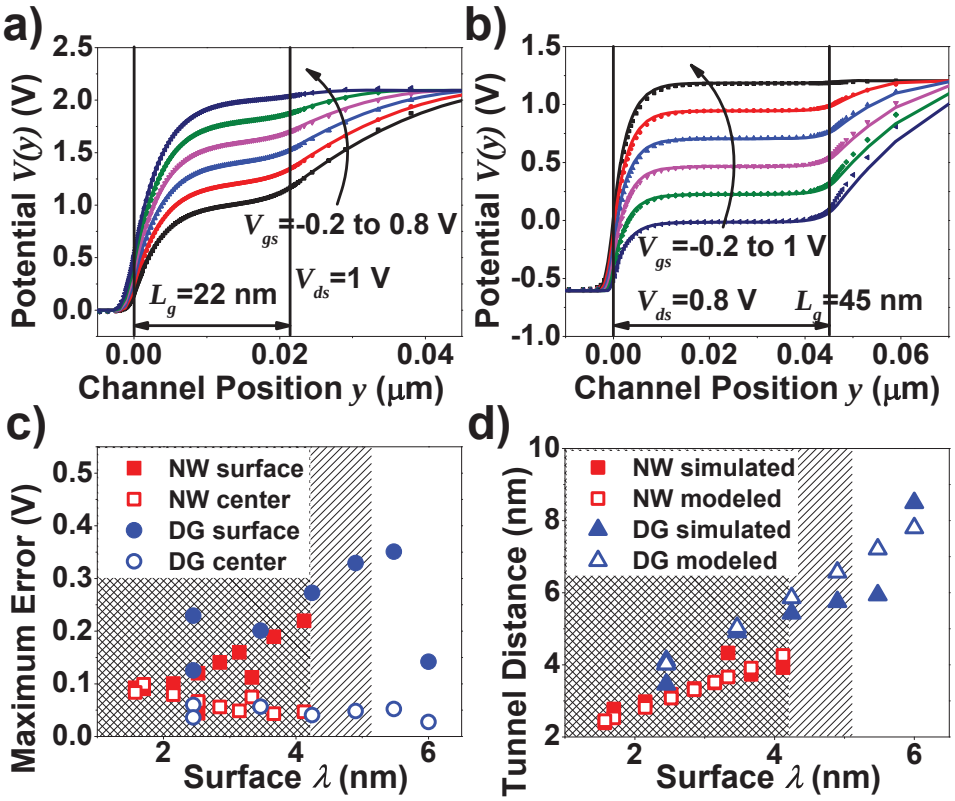

Figure 4.3: Comparison of simulations (symbols) and model (lines) for a) Si DG center potential ($x = \frac{t_{ch}}{2}$), b) InGaAs NW surface potential ($\rho = r_s$), c)-d) largest extracted error in surface and center potentials, and comparison of shortest tunneling distance at channel surface, respectively, of 17 devices at $V_{gs} = 0.4$ V and $V_{ds} = 1$ V for Si DG and $V_{gs} = 0.52$ V and $V_{ds} = 0.8$ V for InGaAs NW TFETs. Light- and dark-hatched regions indicate λ less than ITRS multigate projections for 21 and 15 nm technology nodes, respectively.

with $N_s = 10^{20}$ cm^{-3} and $N_d = 10^{18}$ cm^{-3}. Variations from these values for specific cases are indicated in the figures. In Fig. 4.3(a)-(b) we show the channel potential for sample structures, illustrating the excellent agreement between the model and simulations. No fitting parameters are used; to evaluate the Fermi levels and built-in voltages for heavily degenerate regions, we use Nilsson's analytical approximation[138].

To quantify the model's degree of accuracy, we simulate TFETs with different geometric and doping parameters and find the largest difference between the analytical and simulated potentials across the entire body of the device. In Fig. 4.3(c), we aggregate results from a large number of DG and NW simulations to show the maximum potential error along the channel surface and center. The largest error is generally found at the source/channel interface, which is expected given our neglect of fringing fields and use of the polynomial approximation; ignoring source depletion would greatly increase the error. The error is strongly localized at the interface and is further reduced when we consider the average electric field or tunneling distance, which is the main quantity of interest for TFETs and is shown in Fig. 4.3(d), where we plot the extracted shortest tunneling distance at the channel surface for simulated and modeled TFETs. ITRS projections[139] for multigate devices yield surface $\lambda_{DG} = 5.2$ nm and 4.2 nm for 21 nm and 15 nm node technologies, respectively, as indicated by the shaded regions in Fig. 4.3(c)-(d). We see the modeled tunneling distance is reasonably accurate within the dimensions of interest. As λ decreases with scaling, the improved electrostatics more closely coincide with the parabolic ansatz and errors further reduce.

We calculate current by adapting the Kane tunneling formula[96] and numerically integrating over the analytical 1-D tunneling lengths using Eq. 4.34. Again, no fitting parameters are employed in this semiclassical approach as the matrix elements used in the simulator and model are identical. We note that most results obtained in this way agree up to a constant prefactor with those calculated using the analytical expression Eq. 4.35. In Fig. 4.4 we present comparisons for InGaAs devices with varying nanowire radius and oxide thickness. We observe excellent agreement, further validating the model's usefulness.

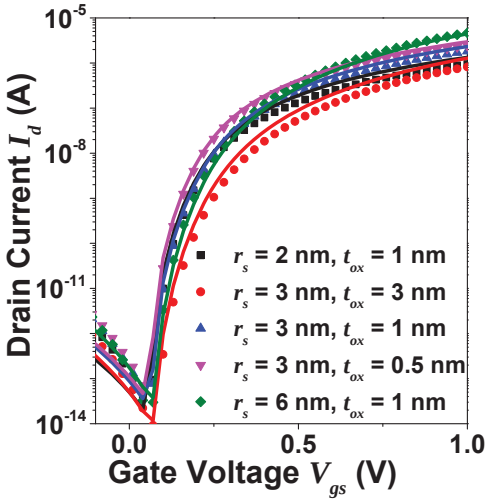

Figure 4.4: Simulated (symbols) and modeled (lines) $I_d - V_{gs}$ ($V_{ds} = 0.8$) V for InGaAs NW n-TFETs with different radii and oxide thicknesses.

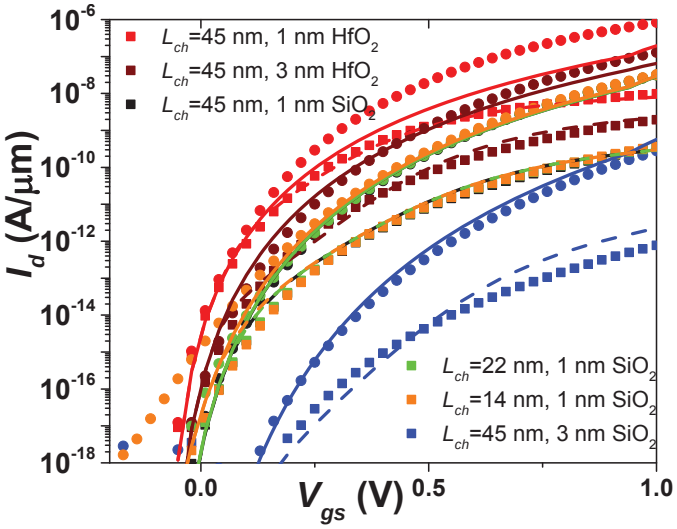

Figure 4.5: $I_d - V_{gs}$ for Si DG TFETs (V_{ds}=1 V) as function of doping and thickness. Symbols are simulated, lines are modeled. For all devices, 1 nm SiO2 gate oxide is used with an undoped channel, 10^{18} cm^{-3} drain doping, 45 nm gate length, and 4.05 eV gate work function. Source doping is 10^20 cm^{-3} unless otherwise indicated.

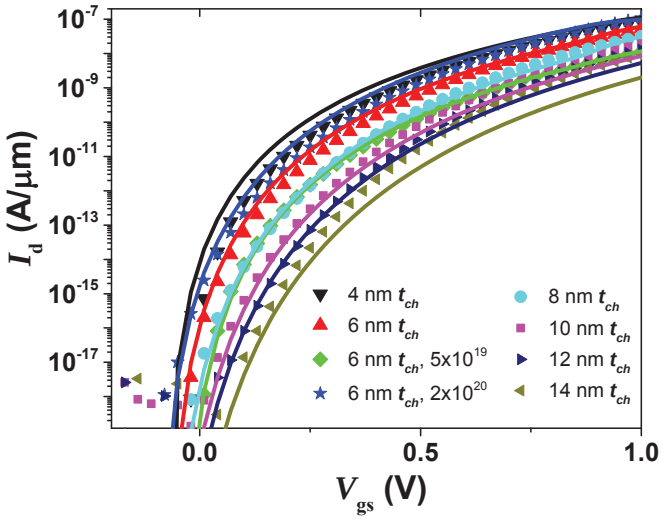

Figure 4.6: I_d-V_{gs} for 8 nm tch TFET with varying oxides and channel length. Other device parameters are the same as Fig. 4.5. Symbols (squares for V_{ds}=0.1 V, circles for V_{ds}=1 V) are simulated, lines (dotted for V_{ds}=0.1 V, straight for V_{ds}=1 V) are modeled. For SiO2, $\alpha_0 = 8$, $c = 0.4$, $\eta = 0.2$, and $\sigma = 0.4$ for the smoothing function; for HfO2, $\alpha_0 = 12$, c= 0.8, $\eta = 0.1$, and $\sigma = 0.6$.

We also simulate silicon DG TFETs and compare with fully analytical calculations with Eq. 4.35, using the center potential as the reference point. As seen in Figs. 4.5-4.6, good agreement is observed for a variety of material thicknesses and doping levels.

To examine the impact of graded junctions, we also simulate DG TFETs with different Gaussian overlap decay lengths σ. When comparing model and simulated $I - V$ curves, the presence of the error functions in Eq. 4.23 prevents analytical inversion of the channel potential. Therefore the tunneling distance must be calculated numerically. In Fig. 4.7, we see that reasonable agreement is achieved up to $\sigma = 6$ nm, beyond which the model error increases due to neglect of the mobile charge as mentioned earlier. For the tested device, the condition Eq. 4.25 yields $\sigma = 5$ nm as the upper limit of validity in agreement with the plotted results. The model makes clear that the reduced performance with increasing overlap is due to the lower electric field near the source, as the dopant ions terminate field lines within the channel.

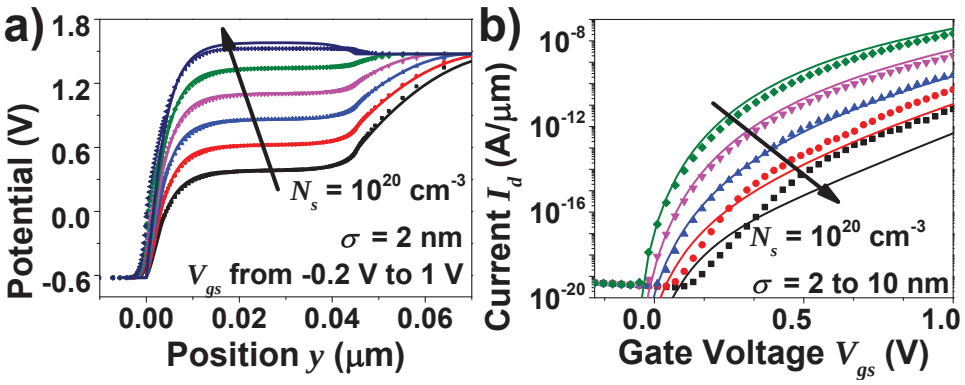

Figure 4.7: a) Channel potential for simulated (symbols) and modeled (lines) Si DG TFET with a Gaussian source overlap in the channel. b) Transfer characteristics as a function of σ. $V_{ds} = 1$ V for all devices.

4.3.2 2-D Effects on Tunneling

With an accurate electrostatic model in hand, we can examine the effects of the 2-D potential on transport. Within the WKB approximation, the tunneling probability is dominated by the least action path; however, the rigorous analytical evaluation of this quantity is difficult[128, 140]. Here we make a simple electrostatic argument, making the common assumption that tunneling is determined by the shortest distance between available conduction and valence band states. Our model allows us to evaluate this length and compare it with the more analytically convenient 1-D tunneling distance. As illustrated in Fig. 4.8(a), for abrupt source/channel junctions the shortest tunneling length must take place near the largest electric fields, hence one endpoint must lie on the surface at $(0, y_2)$ and the other must be located in the source depletion region at (x_1, y_1) since the lateral electric field is maximized at the source/channel junction. We can analytically define the endpoints and the total tunneling distance t_d as

$$y_2 = L_g - \lambda_s \sinh^{-1}\left(\frac{\psi_{ch} - V - (E_g/q)}{\psi_{ch} - V_{s,dep}} \sinh\left(\frac{L_g}{\lambda_s}\right)\right), \tag{4.36}$$

$$y_1 = -y_{s,dep}(x) + \sqrt{\frac{2\epsilon_{ch} V}{q N_s}}, \tag{4.37}$$

$$t_d = \sqrt{x_1^2 + (y_2 - y_1)^2}. \tag{4.38}$$

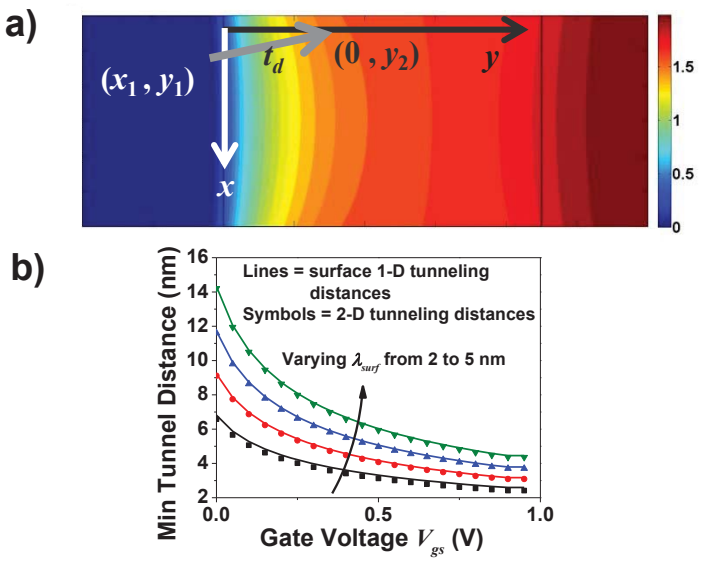

Figure 4.8: a) 2-D channel potential and endpoints of the shortest tunneling distance t_d. b) Shortest 2-D and 1-D tunneling distances as a function of gate voltage and λ for DG structures.

Here we simplify slightly by assuming that L_g is long enough such that the drain side potential does not significantly affect the tunneling distance. Our goal is to find the energy or potential V at which t_d is minimized. Setting the derivative of t_d to zero, we obtain an implicit equation for x_1 to minimize t_d

$$x_{1,min} = \frac{Ct_{ch}(y_2 - y_1(x_{1,min}))}{2C(y_2 - y_1(x_{1,min})) - 1}, \tag{4.39}$$

where

$$C = \sqrt{\frac{qN_s}{8\epsilon_{ch}V_{s,dep}\coth^4(L_g/\lambda)}}\left(1 - \frac{L_g\coth(L_g/\lambda)}{\lambda\sinh^2(L_g/\lambda)}\right)\left[1 - \frac{\psi_{ch} + V_{s,0}\lambda^2}{\sqrt{2\psi_{ch}V_{s,0} + V_{s,0}^2}}\right]\Bigg|_{x=x_{1,m}}. \tag{4.40}$$

For a given structure, we can solve this equation numerically and compare it with the shortest 1-D tunneling distance, which also takes place along the surface ($x = 0$). Fig. 4.8(b) plots these distances as a function of gate voltage and surface λ; we observe that over the range of validity of our model, the correction introduced by the 2-D tunneling distance is negligible. This can be understood by observing that in the polynomial model, the potential varies exponentially in the lateral (y) direction near the source while it only

changes quadratically in the vertical (x) direction. This result makes the 1-D assumption in semiclassical tunneling models of TFETs more plausible and is further supported by the excellent agreement already presented between simulated $I - V$ curves and currents calculated using 1-D tunneling distances. For larger structures, as the parabolic potential approximation breaks down and the vertical field increases, the 2-D effects may become more pronounced. We reiterate that this argument rests on the assumption that the shortest geometric length is a good proxy for the tunneling probability; in real situations, channel subband formation and corrections to the semiclassical tunneling model may impact the results.

4.3.3 Contact Doping and Tunneling Modeling

Source and drain doping have a major effect on TFET operation, making the use of the depletion approximation of great importance. Our electrostatics-based analysis neglects quantum effects like bandgap narrowing and impurity scattering which may play a role at heavy doping; a rigorous analysis of these effects is in progress but beyond the scope of this work, though we note that TCAD simulations including bandgap narrowing show the same trends discussed below. Fig. 4.9(a) illustrates how our model accurately predicts the ambipolar tunneling leakage which appears in TFETs with heavily doped drains, a calculation which would be impossible if the drain doping were neglected. As expected, lowering the drain dopant concentration widens the depletion region and reduces electric field, dramatically reducing leakage. Fig. 4.9(b) shows the effect of the source doping level on TFET characteristics. Interestingly we observe that raising the doping concentration initially boosts device performance, but past a certain level (about 2×10^{20} cm^{-3} in Si) further increases degrade the subthreshold swing and saturate the on-current. Similar effects have been discussed in Refs. [19, 141, 142].

To explain this effect, we argue that the tunneling current for strongly degenerate sources has two kinds of contributions, as depicted in Fig. 4.9(c): occupied states at the energy $E_{s,min}$ where the tunneling distance $y_{t,min}$ is shortest ("gate-controlled tunneling"

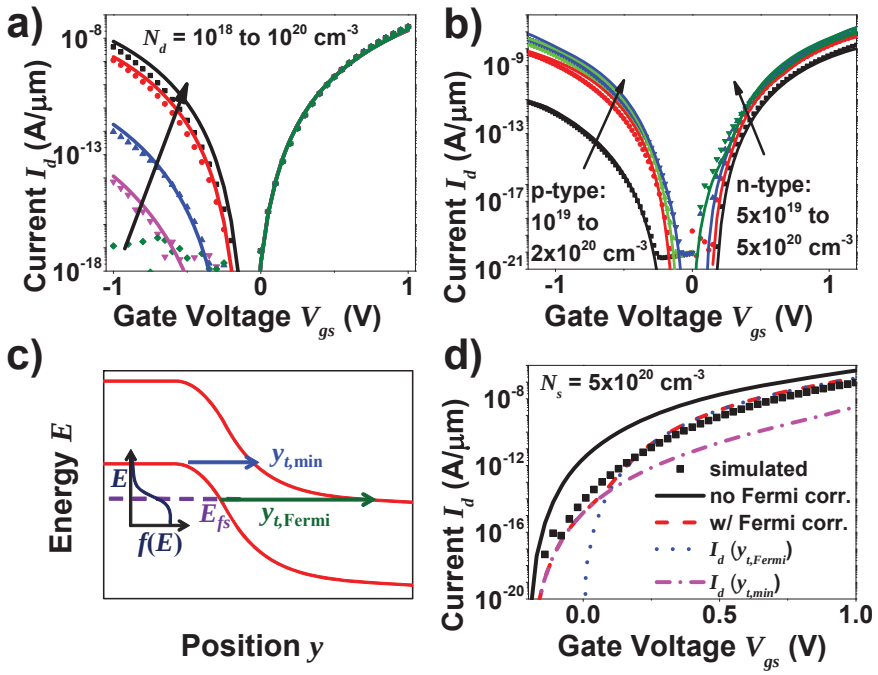

Figure 4.9: a) Ambipolar $I-V$ of Si DG n-TFETs with different drain doping. b) Source doping dependence of $I-V$ for Si DG n- and p-TFETs, where the model (lines) currents are calculated using Eq. 4.41. c) Schematic band diagram of tunneling at the minimum tunneling distance and Fermi energies. d) Contributions of minimum tunneling distance and Fermi energy to current in degenerate situations. ($V_{ds} = 1$ V and abrupt source/drain junctions are used for all simulations.)

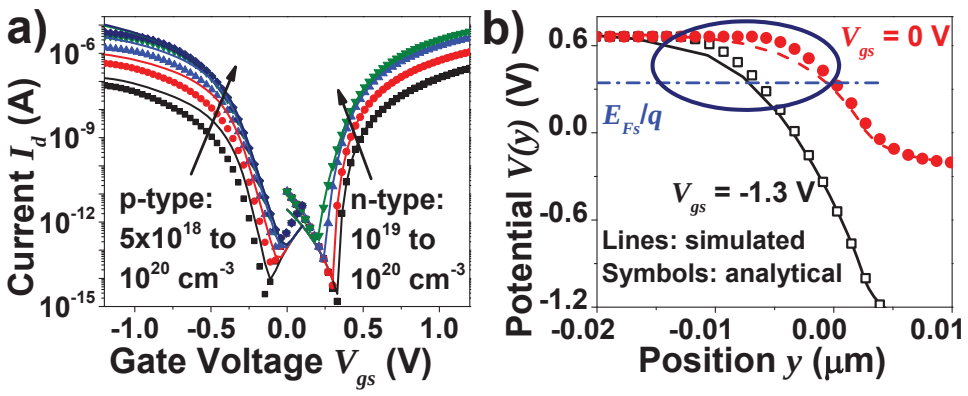

Figure 4.10: a) Transfer characteristics ($|V_{ds}| = 0.8$ V) for InGaAs NW n- and p-TFETs, b) surface potential near the source/channel interface for p-TFET with $N_s = 10^{19}$ cm^{-3} doping. The discrepancy at the depletion region edge is circled.

in the terminology of Ref. [142]), and states near the source Fermi energy E_{Fs} ("cold carriers") with tunneling distance $y_{t,Fermi}$. Gate-controlled states have a large tunneling probability, but if $E_{s,min}$, which occurs near the source band edge, lies far above E_{Fs}, then these states are sparsely occupied. In contrast, cold carriers face long $y_{t,Fermi}$ but have near unity occupation. As gate bias increases, the gate-controlled tunneling states become available first but $y_{t,min}$ does not change strongly with bias, so the subthreshold swing becomes limited by the Fermi factor. The cold carrier states dominate the on-current due to their high occupancy, but their relatively long tunneling distances limit its magnitude. These contributions can be modeled using the equation

$$I_{tunn} = A_t t_{ch}(q\psi_{ch} - E_g)\left[\frac{(\frac{E_g}{qy_{t,min}})^n \exp(-\frac{qBy_{t,min}}{E_g})}{1+\exp(\frac{E_{s,min}-E_{Fs}}{kT})} + \frac{(\frac{E_g}{qy_{t,Fermi}})^n \exp(-\frac{qBy_{t,Fermi}}{E_g})}{2}\right], \tag{4.41}$$

where A_t, n, and B_t are tunneling coefficients[96]. The first term in the brackets of Eq. 4.41 approximates gate-controlled tunneling and the second cold carrier tunneling. Fig. 4.9(b) and (d) demonstrate that this analytical approximation qualitatively and even semi-quantitatively explains the simulations, whereas neglect of degeneracy leads to overestimation of the device characteristics.

For low density of states (DOS) materials like $In_{0.53}Ga_{0.47}As$, we expect the effects of

degeneracy to become visible at lower doping levels. In Fig. 4.10(a), we plot simulated and modeled $I-V$ curves for both n- and p-type InGaAs TFETs. We use Fermi statistics with a parabolic band 3-D DOS in the simulations; since band nonparabolicity increases the DOS at higher energies and reduces degeneracy, we also resimulated the structures using nonparabolic statistics and confirmed that though the built-in voltages are altered, the channel potential is simply shifted and current nearly unchanged by a nonparabolic DOS. For InGaAs TFETs, we find that saturation of on-current takes place around $N_s = 5 \times 10^{19}$ cm^{-3}, lower than in silicon as expected. The degradation in InGaAs n-TFETs can be attributed to the effects of "cold carrier" tunneling and is accurately predicted by our model.

4.3.4 Degenerate Electrostatic Screening in TFETs

The most interesting and unexpected feature in Fig. 4.10(a), however, is the model's overestimation of current for p-TFETs, especially for N_s around 5×10^{18} to 2×10^{19} cm^{-3}, in contrast with the good agreement for n-type and silicon TFETs. This discrepancy must arise from a different physical effect from that discussed above, since the impact of DOS on *tunneling* has already been accounted for through the choice of tunneling prefactor and integration over the Fermi distribution. Therefore, this result must be understood as a new DOS-related *electrostatic* effect as follows. In p-TFETs the source is n-doped and the conduction band density of states for $In_{0.53}Ga_{0.47}As$ is very low (2.5×10^{17} cm^{-3}). The resulting heavy degeneracy, besides affecting the occupancy factors in Eq. 4.41, also modifies the electrostatics directly through screening. Our model uses the depletion approximation, ignoring mobile carriers that perturb the potential on the scale of the screening length (usually given by the Debye length $\lambda_{Debye} = \sqrt{\frac{\epsilon kT}{q^2 N}}$). For heavily doped semiconductors ($N > 10^{20}$ cm^{-3}), the screening length is very small ($\sim$Å) and has no visible impact. However, in low DOS materials, screening is reduced: while classical carriers respond equally to perturbations, in a degenerate system only electrons near the Fermi level screen effectively because of the Pauli exclusion principle[143]. Hence the

quantum mechanical (Thomas-Fermi) screening length becomes longer than the classical value:

$$\frac{1}{\lambda_{TF}^2} = \frac{q^2}{\epsilon}\int -\frac{\partial f(E)}{\partial E} g(E) dE = \frac{q^2 N}{\epsilon kT}\left(\frac{\mathcal{F}_{-1/2}(\frac{E_F}{kT})}{\mathcal{F}_{1/2}(\frac{E_F}{kT})}\right). \tag{4.42}$$

Here $f(E)$ and $g(E)$ are the Fermi distribution function and the DOS, respectively, and $\mathcal{F}_i$ is the i^{th}-order normalized Fermi-Dirac integral. The term in parentheses in Eq. 4.42 is less than one for degenerate materials, increasing the screening length. (For lower-dimensional DOS in quantum wells or wires, the form is altered but the physical result of reduced screening still occurs.) For instance, for $N = 10^{19}$ cm^{-3} in $In_{0.53}Ga_{0.47}As$, the Debye length $\lambda_{Debye} = 1.4$ nm while the Thomas-Fermi length $\lambda_{TF} = 4.2$ nm. Reduced screening widens the depletion region and lowers the source electric field, decreasing the tunneling current relative to n-TFETs. This quantum effect is captured in the simulator through the use of Fermi-Dirac statistics and the material-dependent effective DOS when solving the Poisson equation. We verify this by comparing the simulated and calculated potential in the source of a p-TFET shown in Fig. 4.10(b). We observe a reduced electric field and extended depletion region compared to the prediction of the depletion approximation; this error is not due to the neglect of fringing fields, since no similar discrepancy is found in Si or n-type InGaAs TFETs (because of their higher DOS). At $V_{gs} = -1.3$ V, screening increases the p-TFET tunneling distance at the Fermi level by about 0.7 nm compared to the model prediction, compared to discrepancies smaller than 0.1 nm for the analogous n-TFET. Such a large change substantially reduces the tunneling probability and explains the results of Fig. 8. It is interesting to note that this effect occurs in any low DOS material p-n junction, but the extreme sensitivity of tunneling to band bending makes it noticeably important for TFETs.

We can estimate the relative importance of the electrostatic and "gate-controlled tunneling" and/or "cold carrier" effects arising from the low source DOS by comparing the simulated and modeled characteristics for the InGaAs TFETs in Fig. 4.10. We extract the subthreshold swing and on-current as a function of source doping. For a consistent comparison, we define on-current at $|V_{gs} - V_{th}| = |V_{ds}|$, where V_{th} is the gate voltage at which the source tunneling current exceeds the leakage due to diode current and drain-

Table 4.1: Minimum subthreshold swing (mV/dec) for simulated and modeled InGaAs n- and p-TFETs from Fig. 4.10(a)

Doping (cm^{-3})	**n-TFET (sim)**	**n-TFET (mod)**	**p-TFET (sim)**	**p-TFET (mod)**
10^{19}	26.92	26.51	46.70	48.93
2×10^{19}	25.18	25.81	54.26	56.68
5×10^{19}	22.96	23.80	58.84	59.92
10^{20}	37.04	39.81	59.52	59.55

Table 4.2: On-current (in μA) for simulated and modeled InGaAs NW n- and p-TFETs from Fig. 4.10(a) at $|V_{gs} - V_{th}| = |V_{ds}| = 0.8$ V

Doping (cm^{-3})	**n-TFET (sim)**	**n-TFET (mod)**	**p-TFET (sim)**	**p-TFET (mod)**
10^{19}	0.236	0.277	0.181	0.392
2×10^{19}	0.918	1.05	0.771	1.49
5×10^{19}	2.97	3.35	2.26	3.12
10^{20}	4.91	5.43	2.91	3.21

side tunneling in the simulated device (defined as the point when total current is at a minimum in Fig. 4.10(a)). The results for subthreshold swing are shown in Table 4.3.4; we observe that both simulations and the model capture the degradation in swing for InGaAs p-TFETs due to reduced DOS. We have confirmed that the relative change between model and simulation results in Tables 4.3.4 and 4.3.4 are basically independent of the choice of bias or threshold voltage, as would be expected by visual inspection of Fig. 4.10.

For the on-current displayed in Table 4.3.4, we observe the n-TFET model and simulations agree on the magnitude to about 10%. However, the discrepancy between the model (which accurately includes the effects of DOS on tunneling via "cold carriers") and

simulations (which also include degenerate screening) is much greater for p-TFETs (>50% at 10^{19} cm^{-3} doping). This indicates that degenerate screening (the only new effect not captured by the model) is in fact a major effect at lower source doping. Finally and most surprisingly, we find that the model predicts larger on-current for p-TFETs compared to n-TFETs at low source doping. This counterintuitive result is due to the normalization of the gate overdrive through V_{th}. Because the source is extremely degenerate for p-TFETs, the built-in voltage between source and drain is significantly larger (V_{bi} = 0.89 V versus 1.23 V for n- and p-TFETs, respectively, at 2×10^{19} source and 10^{18} drain doping). If we ignore all degeneracy effects, we would expect tunneling to occur when the channel conduction band overlaps the source valence band edge (at $q\psi_{ch} = E_g$). However, when the source is heavily doped, the gate voltage must pull the channel down to near the source Fermi energy before interband tunneling from the source can overcome the device leakage current. Therefore, for the same $|V_{gs} - V_{th}|$, the p-TFET has a considerably larger voltage drop across the source and hence higher tunneling probability. At lower source doping, this would lead to a larger overall current at the same gate overdrive *if* degenerate screening effects were absent. (If instead, the n- and p-TFET operate with the same voltage drop across the source, i.e. same $|\psi_{ch}|$, then the current for the p-TFET would be lower as expected.) We note in passing that our model predicts that the p-TFET source tunneling current before ψ_{ch} approaches the source Fermi energy, when $V_{gs} < V_{th}$, has a Boltzmann-limited subthreshold swing of 59.5 mV/decade in agreement with theory[19, 144].

Our analytical treatment allows us to separate the impact of degeneracy on transport and electrostatics, clarifying the device physics. The penalties on both fronts arising from low DOS make the design of p-TFETs potentially problematic, since low bandgap materials generally have low electron DOS. Furthermore, the low solid solubilities of donors in many bulk and nanoscale III-V materials[145] may limit the dopant levels to the range where these problems are most pronounced ($N_s \sim 10^{19}$ cm^{-3} for $In_{0.53}Ga_{0.47}As$) and have an impact on complementary TFET architectures.

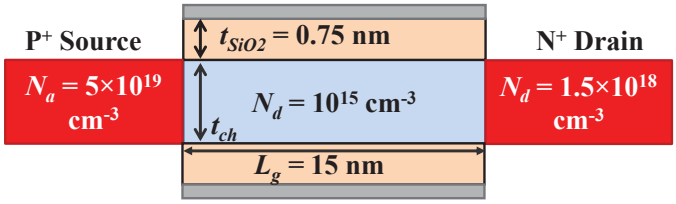

Figure 4.11: Layout of DG TFET structure in NEGF simulations. The channel thickness t_{ch} and material are varied in our simulations.

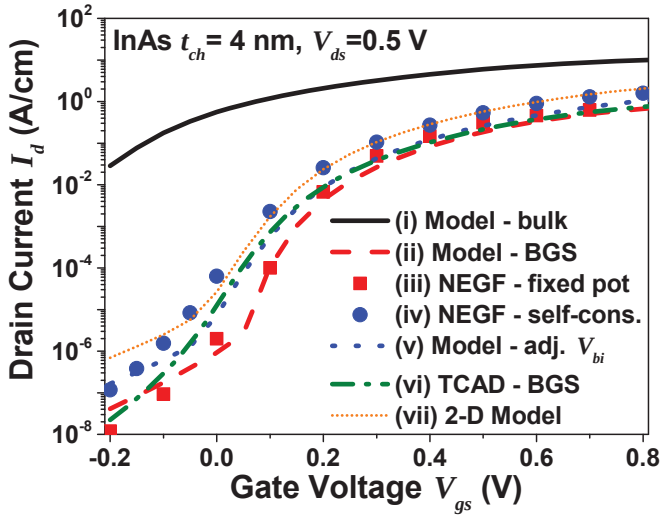

Figure 4.12: Analytical model, TCAD, and NEGF simulations of 4 nm InAs TFET under varying assumptions discussed in the text ($V_{ds} = 0.5$ V).

4.3.5 Comparison with NEGF Simulations and Experiment

We next test the utility of BGS-adjusted tunneling parameters for reproducing quantum TFET simulations. For these scenarios, we perform NEGF simulations of 2-D homojunction double-gate (DG) TFETs where the real-space discretized eight-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian is solved self-consistently with the Poisson equation, except where otherwise noted. Scattering is neglected since only direct tunneling is considered. The schematic TFET structure is illustrated in Fig. 4.11. Since such devices are frequently studied using semiclassical device simulators, we also perform TCAD simulations using the nonlocal tunneling model[114] with parameters adjusted according to the 3-D Kane BGS scheme. No additional quantum models like MLDA (modified local density approximation) or

density gradient are used[114].

In Fig. 4.12, we compare the I-V curves for a 4 nm-thick InAs device using the analytical TFET model with (i) bulk and (ii) BGS parameters, versus NEGF simulations using either (iii) the same electrostatic potential as the analytical model or (iv) a self-consistently evaluated potential. Using all bulk values (i) gives unrealistic results due to the small bulk gap. This illustrates that calibrating the Kane model to bulk material data is totally inadequate for QC devices. However, for identical electrostatic potentials, the BGS Kane model (ii) provides a much better match with the NEGF results (iii). Compared to (i), the current decreases and threshold voltage shifts due to the increased $E_{g,QC}$. The self-consistent NEGF simulation (iv) is similar to the fixed potential result at large bias but has higher off-current; this is because the self-consistent potential has a larger V_{bi} than the 3-D value assumed by the fixed potential. This leads to higher electric fields and hence increased tunneling in the off-state, illustrating the importance of QC electrostatics. If we also change V_{bi} to its quantized value in the analytical TFET model, shown by case (v) in Fig. 4.12, the model yields similar results to self-consistent NEGF throughout the whole bias range. Case (vi) is a BGS TCAD simulation which agrees quite well with the analytical results, as expected from our semiclassical based study, though both still somewhat underestimate the NEGF current. Finally, (vii) using the 2-D Kane model with the analytical TCAD potential gives good quantitative agreement with NEGF.

Fig. 4.13a shows the congruence of BGS and NEGF simulations still holds when the channel thickness (and hence $E_{g,QC}$) is varied, demonstrating the key role of band gap scaling. From these results we see a tradeoff in TFET scaling between the on-current, which is reduced by increasing $E_{g,QC}$, and the on-off ratio and subthreshold swing, which are improved by larger $E_{g,QC}$ as well as better electrostatic control. In Fig. 4.13b, we see that BGS remains valid for different channel materials like InGaAs and InSb. In Figs. 4.12 and 4.13, using the bulk Kane tunneling formula with BGS leads to reduced currents; however, we found that an lateral voltage shift (corresponding to a 20-50 meV difference in gate work function or threshold voltage) of the corresponding simulated and analytical

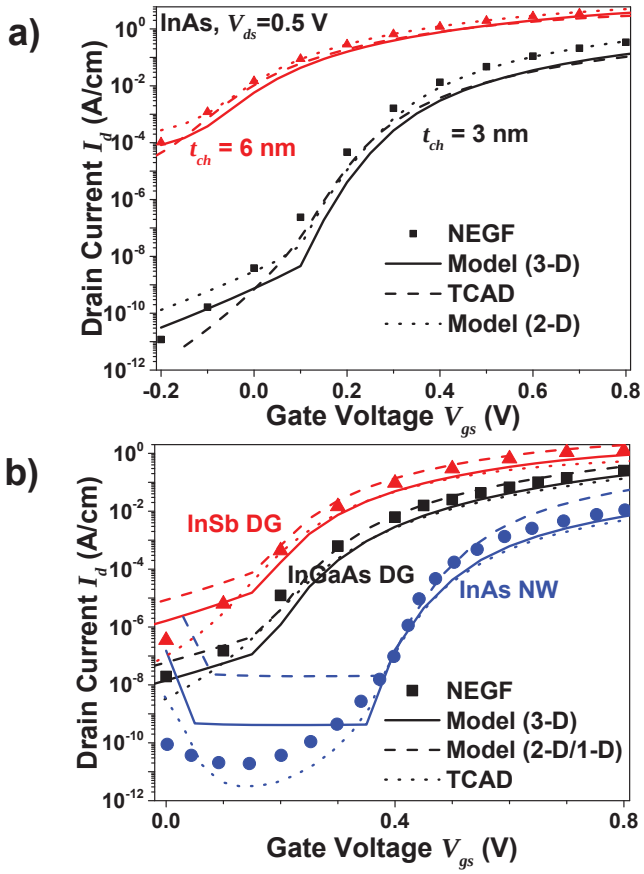

Figure 4.13: (a) Analytical model, TCAD, and NEGF simulations of 3 nm and 6 nm thick InAs TFETs (V_{ds} = 0.5 V). (b) Analytical, TCAD, and NEGF I-V for 4 nm thick InGaAs (V_{ds} = 0.8 V) DG, InSb (V_{ds} = 0.4 V) DG, and a 3.35 nm diameter ($E_{g,QC}$ = 1.175 eV) InAs NW TFET (V_{ds} = 0.5 V). The NW NEGF simulations are from Ref. [4] and have been horizontally shifted by 0.4 V for clarity.

I-V curves brings them into close agreement to the NEGF results. This is because at small and moderate fields the magnitude of the current is determined by the exponential dependence of transmission on B_{BGS}/F, which is the same for all dimensions. The power law dependence of the prefactor A only dominates at very high fields, where the voltage dependence is weaker. The net effect is that while the 3-D Kane formula underestimates current, its percentage error changes slowly as a function of field (as seen in Figs. 3.13 and 3.14) and can thus be roughly compensated by a voltage shift.

Though we do not perform TB or 1-D TFET quantum transport calculations due to

computational constraints, we do compare BGS predictions with TB NEGF simulations for an InAs nanowire from the literature[4] as shown in Fig. 4.13b. The gate work function was not reported for this device, so we shifted our model and TCAD I-V curves to fit the extracted data points. We observe qualitative agreement although the 3-D BGS underestimates the current and 1-D BGS overestimates it, similar to our constant field NW simulations. The disagreement between the analytical and NEGF curves in the off-state (when tunneling occurs directly between source and drain) is due to the TFET model's use of the depletion approximation; TCAD gives somewhat more representative results here due to its self-consistent potential.

We emphasize that no arbitrary fitting parameters or extraneous simulation models are used to achieve the results in Figs. 4.12-4.13, except for the gate work function shifting of the NW TFET; scaling is performed directly using the known bulk masses and band gap and the calculated $E_{g,QC}$. Using BGS and the dimensionally appropriate Kane model gives the best match to quantum transport calculations without further adjustment. The 3-D Kane formula tends to underestimate the current, though heuristically this error can be partly masked by a small threshold voltage shift. This implies that semiclassical TCAD simulations, properly adjusted using BGS, can still be valuable for qualitative studies, though incorporation of dimensionality should improve their accuracy. Our results also incidentally demonstrate that analytical TFET electrostatic models are capable of describing quantum devices, though modeling of drain depletion needs to be improved to quantitatively reproduce the off-state current.

While full quantum calculations remain irreplaceable for maximum accuracy or exploration of novel nanostructure device physics, their complexity makes them inaccessible for many practical engineering studies, as is also the case for many bulk devices. There is clear value in a simple method like the BGS scheme that allows widely used semiclassical models and simulators to be applied with good qualitative and often quantitative accuracy. Alternatively, if calibrated parameters are available for a particular device[146], BGS can be performed to extrapolate them to different dimensions.

Finally, we compare our analytical models with measured data for high performing

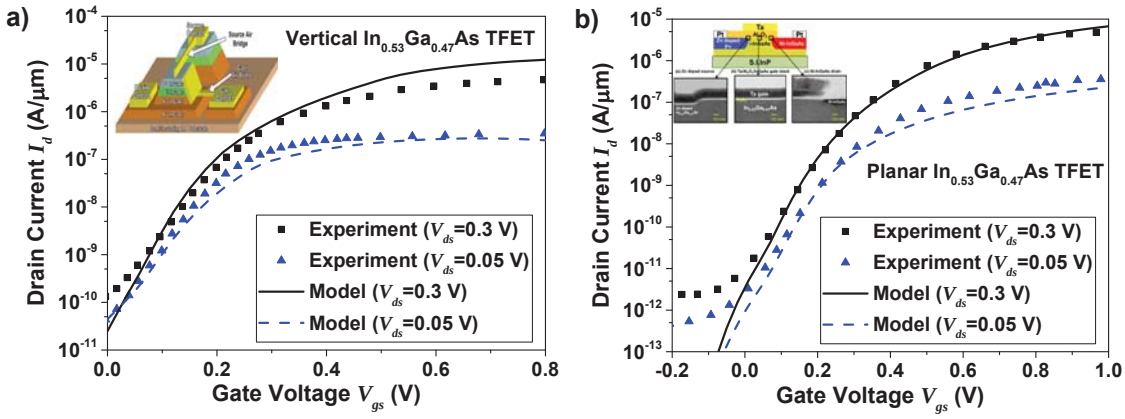

Figure 4.14: (a) Comparison of analytical model fit with experimental data for planar InGaAs device reported in Ref. [5]. (b) Comparison of analytical model fit with experimental data for vertical InGaAs device reported in Ref. [6].

III-V TFETs[6, 5]. Because of the large sizes and complexity of the structures realized experimentally, a fitting-free comparison is no longer possible. To reduce the arbitrariness of the fitting procedure, we assume that the bulk tunneling coefficients for different device materials derived in Chapter 3 are fixed; this is reasonable since the large dimensions of the experimental devices should make quantum confinement effects negligible. We use λ as an adjustable parameter to approximate the electrostatics of the complicated device geometries. In Fig. 4.14, we show the results of our comparison. We see good qualitative and even quantitative agreement in the subthreshold and above threshold regions, though the model does not incorporate the leakage currents observed at small V_{gs}. We note that the vertical TFET reported in [6] and fitted in Fig. 4.14(b) has a small heterojunction near the source/channel interface which complicates the tunneling process; for this exercise we neglect the difference, which may lead to quantitative corrections in the current.

4.4 Conclusion

We present general equations to describe the 2-D potential in lateral TFETs. The model can be seamlessly converted to describe different structures, including nonabrupt doping

profiles. We verify via simulations that the model is applicable for well-scaled multigate devices and demonstrate that 2-D tunneling effects are minor. We use our framework to present new analytical treatments of Gaussian source/channel junctions, ambipolar leakage, and the tunneling current under degenerate doping. Comparison of the model with simulation reveals the major, formerly unappreciated role of degenerate screening on TFET electrostatics and transport characteristics. Validation with the BGS correction via comparison with NEGF TFET simulations show that effects of lower-dimensional DOS generally lead to higher currents, though they can sometimes be approximated by a threshold voltage shift. Because of its flexibility and the device insight it offers, our model provides an attractive foundation for TFET analysis and compact modeling.

CHAPTER 5

Designing Doping-Independent Tunneling Transistors: the GISTFET

They put arsenic in his meat
And stared aghast to watch him eat

A. E. Housman, "Terence, this is stupid stuff"

Tunneling field-effect transistors (TFETs) are alluring because in theory their use of gate modulated interband tunneling allows for extremely low leakage currents and steep subthreshold swings (SS), enabling new kinds of ultralow power electronics[147]. In practice, however, major engineering challenges still impede the progress of such devices, including low on-currents, n- and p-device asymmetry, large-scale reproducibility and variability challenges, and parasitic leakage. Most of these problems are due to the material- and doping profile-related difficulties in realizing high quality tunneling junctions at the source-channel interface.

For example, TFETs using small band gap III-V materials are highly desirable for increasing drive current and reducing supply voltage. However, p-type TFETs, which require n-doped sources, face two significant obstacles: 1) low active donor concentrations (on the order of 10^{19} cm^{-3} for many III-V bulk materials[148, 149] and even lower for nanostructures[145]) due to solid solubility, incomplete ionization, or defect compensation limits, and 2) low conduction band (CB) DOS. Low doping lengthens the tunneling length and strongly reduces drive current, while strong carrier degeneracy due to low DOS degrades the SS by increasing the contribution of "thermal tail" states in the source distribution function as shown in Chapter 4 as well as other works[19, 150]. As a result,

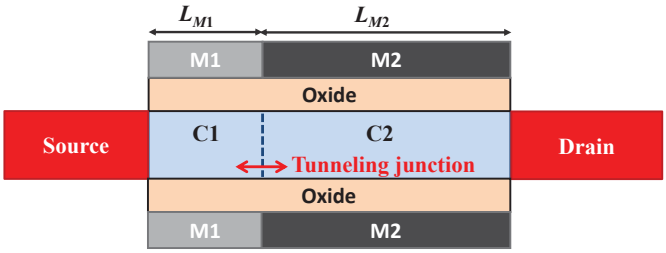

Figure 5.1: GISTFET in a double gate design. M1 and M2 are shorted and biased together by the gate voltage, but their WF difference sets up a tunneling junction between the lightly doped channel sections C1 and C2.

virtually all experimental III-V TFETs in the literature are n-type[147]. Since p-type devices are essential for complementary circuits, their unavailability would doom TFETs for conventional logic applications. Furthermore, in both n- and p-type TFETs the current is highly sensitive to the position and abruptness of the doping profile[151, 152], which are difficult to control precisely. This leads to poor nominal performance and increased variability due to random dopant fluctuations (RDF)[153, 154]. Disorder induced by heavy doping also creates band tails[155] which may significantly worsen the SS[156, 157].

To circumvent these difficulties, we present a new device structure, the gate-induced source tunneling FET (GISTFET), which relies on gated electrostatic doping to decouple the tunneling process from the chemical dopant junction. We introduce this concept and discuss its merits compared to other novel TFET proposals in Section II, demonstrate its primary features and advantages using quantum simulations in Section III, and summarize our conclusions in Section IV. The work in this chapter has been published in Ref. [12].

5.1 Defining Tunneling Junctions Electrostatically

5.1.1 Operating Mechanism of the GISTFET

The proposed GISTFET is shown in Fig. 5.1. We will use double gate (DG) p-type TFETs to illustrate the concept throughout this paper without losing generality, since

the same principle clearly applies to n-type transistors and other device architectures. The GISTFET resembles the usual lateral TFET structure, but with a gate electrode comprised of two metals M1 and M2 with work functions (WFs) ϕ_1 and ϕ_2 such that

$$q\Delta\phi = q\phi_2 - q\phi_1 > E_{g,QC} + qV_{dd} \tag{5.1}$$

where $E_{g,QC}$ is the quantum confined band gap of the channel material, i.e., the gap between the lowest subbands of the conduction band (CB) and valence band (VB), and V_{dd} is the maximum operating voltage. M1 and M2 are electrically shorted together and share the same external gate bias. We will refer to the channel sections "under" M1 and M2 as C1 and C2, respectively; the (bias-dependent) potential difference between them defines the relevant tunneling junction and is denoted by $q\Delta\psi_c$. For each individual channel C1 or C2, the gate bias can strongly modulate the potential when the channel is in depletion (i.e., the electron and hole quasi-Fermi levels lie deeply in the band gap), but weakly if the channel is in accumulation or strong inversion (such that one quasi-Fermi level is degenerate, leading to "electrostatic doping"). Modulation of $\Delta\psi_c$ arises because the different metal WFs offset the gate voltage thresholds for which C1 and C2 pass between accumulation, depletion, and inversion.

The resulting bias stages of device operation are schematically indicated by the band diagrams in Fig. 5.2. We assume in the following discussion that scattering is strong and quantization weak enough such that local equilibrium holds and semiclassical carrier distributions can be described by local quasi-Fermi levels throughout the device. (In the subsequent section we will see that ballistic and quantum effects can lead to quantitative changes, though the basic operating principle remains the same.) In the off-state a), C1 is pulled by M1 into electron accumulation, partially pinning the channel potential, and C2 is in depletion but $q\Delta\psi_c < E_g$, suppressing tunneling. As V_g becomes more negative b), the C2 energy bands are pulled up and interband tunneling occurs once $q\Delta\psi_c$ exceeds E_g. With further gate bias c), C1 passes from accumulation to depletion and $q\Delta\psi_c$ reaches a maximum value close to $q\Delta\phi$. Eventually d), the valence band edge of C2 crosses the drain Fermi level and undergoes hole accumulation, causing $\Delta\psi_c$ to decrease; this leads

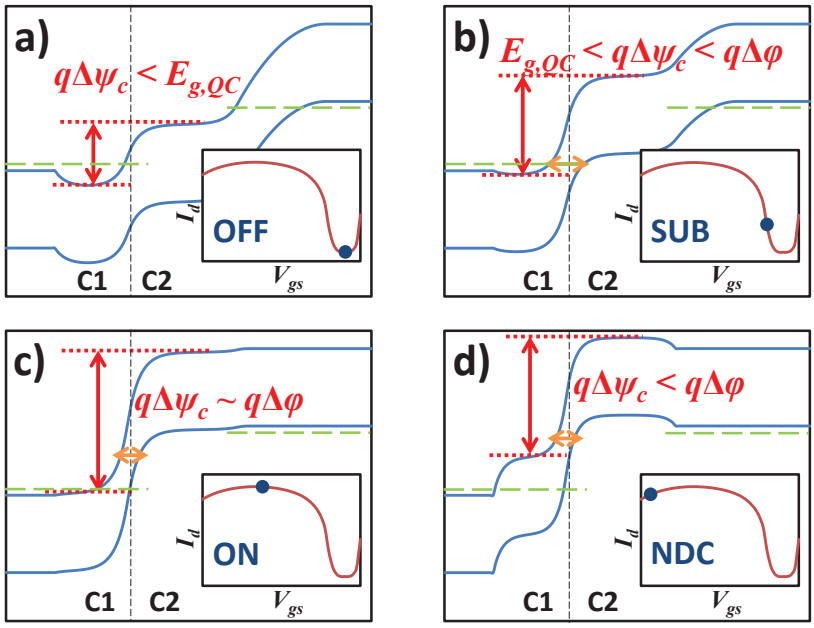

Figure 5.2: Band diagrams along the p-type GISTFET channel as V_g decreases assuming local equilibrium. The energy interval indicated by the red vertical arrows and dashed lines is $q\Delta\psi_c$; the horizontal orange solid arrows indicate the tunneling lengths. The dashed green arrows indicate the source and drain quasi-Fermi levels, respectively. The insets indicate the corresponding bias points on the device $I_d - V_{gs}$ curve.

to a drop in current and hence negative differential conductance (NDC). However, if Eq. 5.1 is fulfilled this occurs at biases well outside the operating range of the system and the NDC will have no practical impact.

Because the junction is electrostatically induced in the lightly doped channel region and controlled by $q\Delta\phi$, the tunneling length is decoupled from the placement and magnitude of the source doping. This eliminates the previously described challenges in creating heavily doped abrupt junctions. Furthermore, because the source doping is no longer critical and can be kept relatively low, the adverse effects of degeneracy and low DOS on the SS may be alleviated. While lower source doping may increase series resistance, the latter should not be a limiting factor for the low power applications in which tunneling devices are likely to be used. Therefore the use of electrostatic doping, as implemented in the GISTFET, may be particularly suitable for realizing high performance III-V p-TFETs.

Qualitatively, we can explain the advantage of the GISTFET over the conventional TFET by comparing their electrostatic properties. The potential drop near the junction of a gated TFET channel occurs over a distance set by the characteristic length λ, which is determined by structural parameters of the device and defines its electrostatic integrity. For instance, for the surface potential in DG devices,

$$\lambda = \sqrt{\frac{\epsilon_{ch} t_{ch} t_{ox}}{2\epsilon_{ox}}} \tag{5.2}$$

where ϵ_{ch} and ϵ_{ox} are the channel and oxide permittivities and t_{ch} and t_{ox} are the channel and oxide thicknesses, respectively. In a conventional TFET, the potential drop across the source-side tunneling junction is divided between the gated channel and the depletion region in the source, so it extends over a distance on the order of $\lambda + w_s$, where w_s is the source depletion width set by the doping profile. By contrast, both sides of the tunnel junction in the GISTFET are gated, so the potential drop $q\Delta\psi_c$ occurs over a distance equal to about twice the characteristic electrostatic length 2λ. Roughly speaking, when $\lambda < w_s$, the potential barrier will become narrower in the GISTFET and its tunneling current can then exceed that of a TFET. Referring to Eq. 5.2, this implies that the GISTFET becomes comparatively more attractive as the channel thickness or gate oxide

thickness and permittivity are scaled. λ can also be reduced by using tri-gate or nanowire structures with stronger gate electrostatic control, as illustrated in Chapter 4. These are of course also the general directions in which most transistor geometries are moving. By contrast, electrically active doping concentrations are not easily scalable and may even be reduced from their bulk values in nanostructures[145], making doping-centric concepts for improving tunneling performance more difficult to implement[158]. We note that in isolation, body thickness scaling in TFETs and GISTFETs may be counterproductive past a certain point because size quantization effects will increase the band gap $E_{g,QCD}$ and reduce tunneling[159].

5.1.2 Comparison with Alternative Schemes

The GISTFET operating principle and requirements are distinct from the multiple WF designs previously explored for TFETs[160, 161], as the latter primarily employ different metals to reduce drain leakage and the device operation still relies on a heavily doped source. The doping-less device proposed in [162] bears greater resemblance to the GISTFET but requires narrowly spaced and separately gated accumulation and inversion layers, which increase the tunneling length and raise the possibility of unwanted contact shorting. In the GISTFET, the tunnel junction width is ultimately modulated by the (potentially atomically) abrupt M1/M2 interface rather than the lithographically defined gate and source separation.

Another design with similar objectives is the electron-hole bilayer (EHB) TFET, which utilizes "vertical" tunneling between accumulation and inversion layers on opposite sides of the undoped channel body[163]. However, since EHB tunneling occurs perpendicular to the gate, field-induced quantum confinement (FIQC) effects are large[123] and the lowest FIQC valence subband is heavy hole-like[164], reducing the gate efficiency and current. By contrast, transport in the GISTFET occurs along the unconfined direction parallel to the gate where FIQC is negligible and the lowest valence states are light hole-like, increasing the tunneling probability as demonstrated in Chapter 3. We emphasize that

a major conceptual difference between the GISFET and the doping-less or EHB TFETs arises because the former relies on the difference between the M1/M2 WFs (instead of asymmetric applied voltages) to directly define the tunnel junction, not just to shift the threshold voltage.

By minimizing the role of chemical dopants, the GISTFET greatly simplifies the associated design and processing considerations. However, the oxide quality and abruptness of the M1-M2 interface will in turn become key factors for GISTFET performance; in particular, metal intermixing and effects of WF pinning or variability must be minimized, which heavily depend on details of material system and processing conditions. CMOS-compatible metal combinations are also needed with work function differences fulfilling Eq. 5.1, which will be on the order of 1 eV in practice[159]; combinations of Ti or Al (with work functions around 4 eV), with Pt, Ni, or W (with work functions greater than 5 eV) may be promising in this regard[165, 166, 167]. It is also encouraging to note that theoretical[168] and experimental[167] studies of such as-deposited bilayer metal stacks indicate that the change in WF occurs within just a few atomic monolayers (<1 nm) across the heterointerface.

5.2 Simulation and Validation of the GISTFET Concept

To demonstrate the proposed device concept, we perform ballistic non-equilibrium Green's function (NEGF) simulations of InAs TFETs and GISTFETs with varying levels of source doping. We use the program we developed which is explicated in detail in Chapter 2. For computational efficiency we use a four-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonian to describe InAs, neglecting spin-orbit effects; this may underestimate the current by giving a larger effective band gap compared to full band tight-binding predictions, as indicated by the band structure analysis in Chapter 3, but suffices for qualitative comparisons[169, 170]. The simulated devices are DG structures like the one shown in Fig. 5.1.

We simulate GISTFETs using $q\phi_1 = 4$ eV and $q\phi_2 = 5.4$ eV, which can be achieved experimentally using Al and Pt, for instance[166]. The results are shown in Fig. 5.3

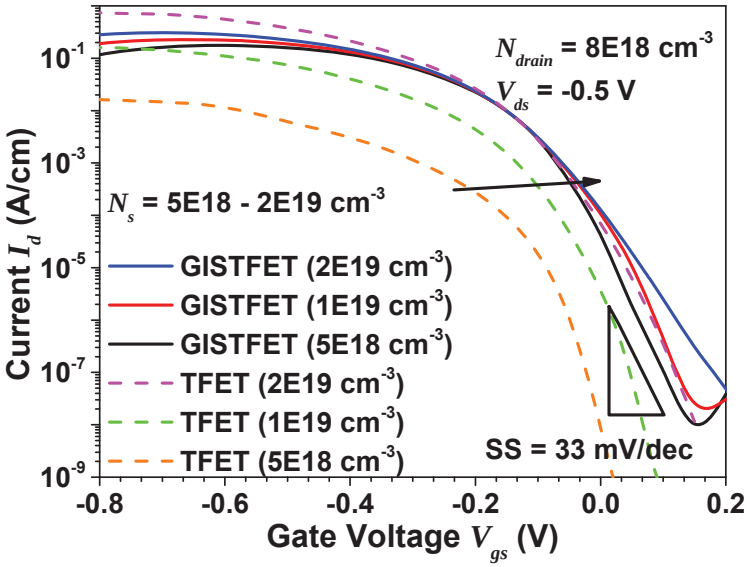

Figure 5.3: Simulated $I_d - V_{gs}$ for InAs DG GISTFETs (solid) and TFETs (dashed) with different abrupt source doping concentrations N_s. The gate oxide is 3 nm thick HfO_2 and channel thickness is 4 nm for all devices. $L_{M1} = 5$ nm and $L_{M2} = 25$ nm for GISTFETs while TFETs have gate length of 30 nm and metal WF of 5.4 eV. The increasing current at positive V_{gs} is due to drain-side tunneling as no drain underlap is employed.

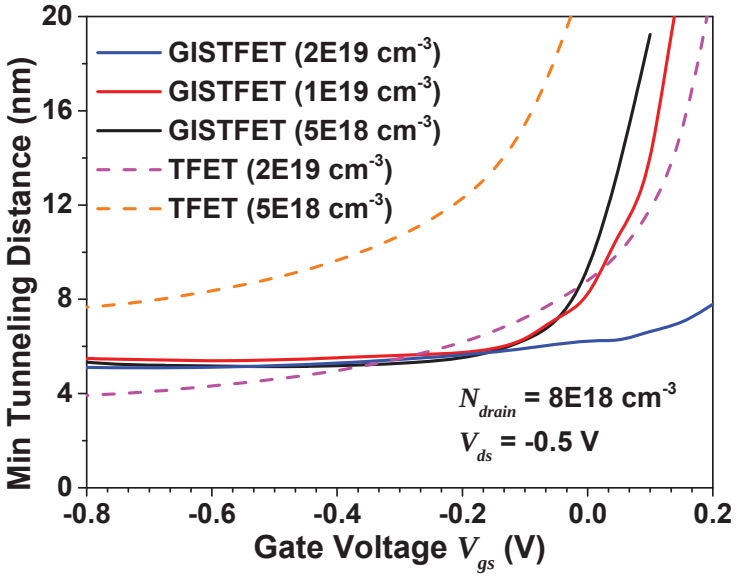

Figure 5.4: Minimum source-side tunneling lengths as a function of gate bias for GISTFET and TFET devices with different source doping.

and demonstrate several important characteristics. First, while the TFET I_{on} varies by almost two orders of magnitude with a 4X change in source doping, the GISTFET I_{on} is independent of source doping and about equal to that of the $N_s = 2 \times 10^{19}$ cm^{-3} TFET (which we note is a doping level that may already exceed the solid solubility limit of donors in InAs[149]). Secondly, both the TFET and GISTFET SS tend to degrade somewhat with increased N_s because of the low DOS of InAs[19]. Therefore, low n-type source doping levels are ideal for p-GISTFETs to combine steep SS with high current. As an example, the $N_s = 5 \times 10^{18}$ cm^{-3} GISTFET offers an on-off ratio of 10^7, minimum SS = 33 mV/dec, and $I_{on} = 13$ μA/μm over a 0.5 V bias range (between 0.1 to -0.4 V), comparable in performance to that of the 2×10^{19} cm^{-3} TFET. Finally, the sensitivity of threshold voltage to N_s is significantly less for GISTFETs compared to TFETs due to the greater influence of the source doping-dependent built-in voltage on the latter; this implies RDF-induced variability will be smaller for GISTFETs.

To understand these characteristics in more depth, in Fig. 5.4 we extract the doping- and bias-dependent minimum tunneling lengths for our simulated TFETs and GISTFETs, defined as the shortest distance between CB and VB subband edges for energies above

the lowest conduction subband energy in the source. We observe that as the devices turn on (V_{gs} <-0.1 V), the tunneling lengths are virtually identical for the different GISTFETs since the C1/C2 junction electrostatics are basically independent of doping, whereas the longer depletion regions at lower source doping lengthen the TFET tunneling distances. At low gate bias, however, the tunneling length is significantly shorter for highly doped GISTFETs. This is because the higher degeneracy of heavily doped GISTFETs increases the source Fermi level and leads to an effective threshold voltage shift, such that overlap of the source CB and channel VB edges still occurs up to $V_{gs} = 0.2$ V. We can see this most clearly by examining the band diagrams and spectral currents within the lowest and highest doped GISTFETs in the off-state in Fig. 5.5. Whereas all the current in the $N_s = 5 \times 10^{18}$ cm^{-3} device flows via tunneling at the C2/drain interface (and can thus be suppressed using lower drain doping, underlaps, etc), a separate current path also occurs near the source in the 2×10^{19} cm^{-3} GISTFET due to energetic overlap of band states in C2 and the source.

One may object via inspection of the band diagrams in Fig. 5.5 that source-side tunneling should still occur in the low doped GISTFET at the selected voltage, since we observe that the C2 VB in the off-state is below the CB edge of the source, but not that of C1. The reason why no tunneling current flows at these energies is due to quantization effects and the assumption of ballistic transport. This can be best understood by examining the local density of states (LDOS) for the GISTFET in the off- and on-states, shown in Fig. 5.6. In the off-state a), the narrow electrostatically induced potential well causes spatially localized states appear at energies below the source CB edge as indicated by the narrow lines in the LDOS; the effects of such states can only be properly analyzed using intrinsically quantum mechanical simulations like NEGF. (In actuality the well is an effective 1-D electrostatically gated quantum "wire" because of the spatial confinement of the DG structure.) Ordinarily we would expect tunneling to occur between energetically overlapping states within the C1 well and the C2 VB. However, the C1 states are spatially localized and hence cannot support carrier flow unless they can couple to continuum, current-carrying states on either side. Because these states lie in the band

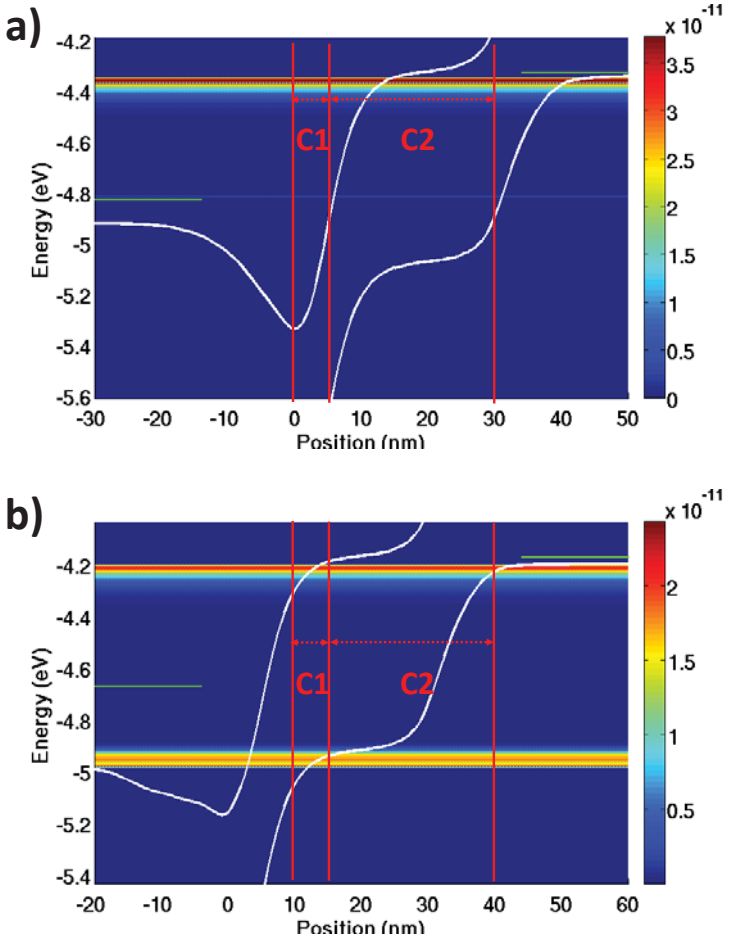

Figure 5.5: GISTFET band diagrams and spectral currents densities in the off-state at $V_{gs} = 0.2$ V and $V_{ds} = -0.5$ V for a) $N_s = 5 \times 10^{18}$ cm^{-3} and b) $N_s = 2 \times 10^{19}$ cm^{-3}. Green lines indicate the source and drain Fermi energies.

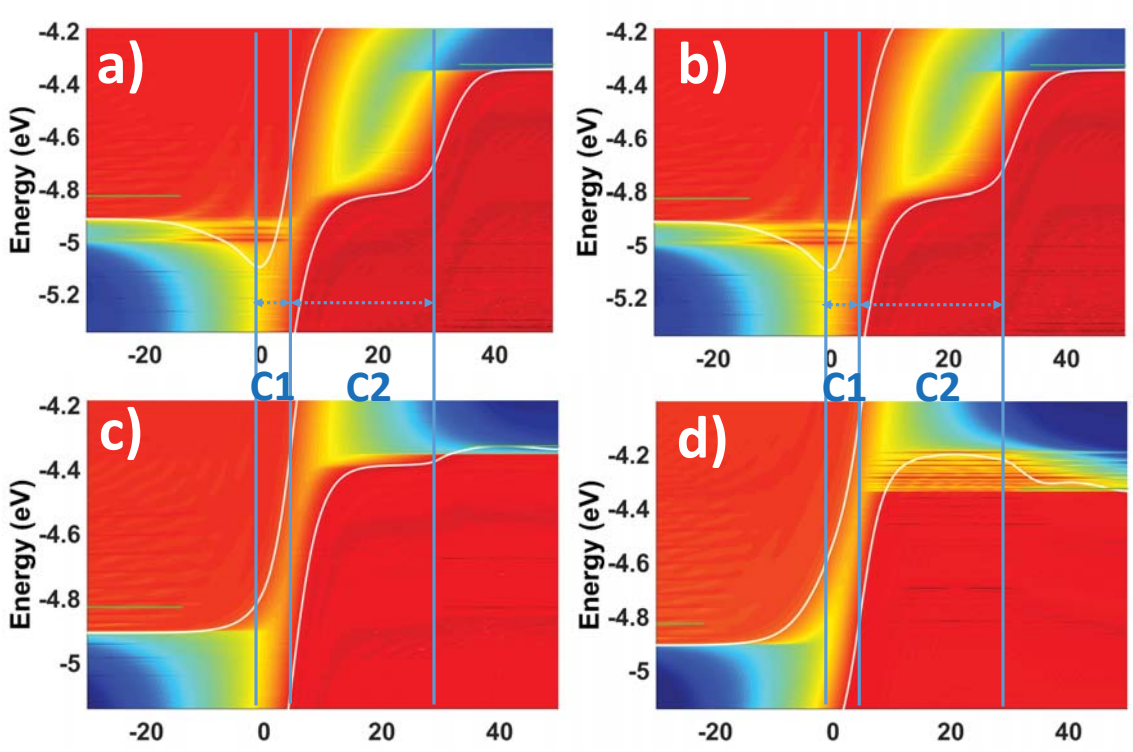

Figure 5.6: Band diagram and LDOS along the channel of a GISTFET with $N_s = 5 \times 10^{18}$ cm^{-3}, $L_{M1} = 5$ nm, $L_{M2} = 25$ nm, and $V_{ds} = -0.5$ V for a) $V_{gs} = 0.2$ V b) $V_{gs} = 0$ V and $V_{ds} = -0.5$ V, c) $V_{gs} = -0.5$ V, and d) $V_{gs} = -0.8$ V, corresponding to the semiclassical operating regions in Fig. 5.2. The green lines indicate the source and drain Fermi energies, and white lines correspond to the lowest conduction and valence subbands. The separation between C1 and C2 is also indicated by the vertical lines. LDOS is shown on a log scale.

gap of the source, no continuous elastic transport process connects them to the source electrode, and therefore no current will flow. Therefore, in the ballistic limit, once the VB edge of C2 falls below the source CB edge, source-side tunneling ceases. A similar phenomenon has been observed in quantum simulations of accumulation layers in pocket-doped TFETs[158].

If inelastic scattering processes occur, they can couple the C1 localized states to the continuum CB and provide a continuous current path. Inclusion of inelastic scattering may therefore increase the off-state leakage current by allowing parasitic tunneling through the localized states in C1. We do not considered these processes here owing

to the substantially greater computational requirements of multiband NEGF simulations with phonon scattering; further work is needed to quantify and assess these effects, which will be relevant for GISTFETs as well as other types of TFETs where localized accumulation regions appear[158]. Qualitatively, we expect these effects to be less important in III-V GISTFETs with narrow C1 channel lengths and high mobilities because of weaker electron-phonon coupling. In principle, provided proper metal work functions can be found, the threshold of the device can always be shifted such that no C1/C2 overlap occurs at all in the off-state.

The behavior of the off-state in the nanoscale GISTFET is therefore slightly complicated by the presence of ballistic and quantum effects. In subthreshold and the on-state, the band bending and device operation are as expected from semiclassical arguments, as shown in Fig. 5.6b) and c), while the NDC region behaves as the obverse of the off state as is clear from Fig. 5.6d). We note our ballistic simulations underestimate one potential benefit of the GISTFET because they do not account for the self-energy effects of impurity scattering[155], which if fully incorporated would create a DOS tail in the heavily doped source and further degrade the SS of conventional TFETs. This effect should be negligible in the GISTFET since the doped regions are separated from the tunneling junction, giving it another comparative advantage. Electron-electron effects are also neglected here as they primarily narrow the band gap without producing substantial DOS distortion and hence should not impact SS.

As an illustration of how device structure affects performance, we also simulate GISTFETs and TFETs with different channel and oxide thicknesses as shown in Fig. 5.7. Interestingly, all else being equal, performance degrades for both device types if channel thickness is reduced because the band gap increases rapidly for very thin InAs devices, suppressing tunneling current. However, the GISTFET undergoes a significantly larger performance boost than the TFET when oxide thickness is scaled down. This shows the potential scaling benefits of the GISTFET. Overall, the use of gate-induced electrostatic coupling offers greater benefits from better gate control, small E_g materials, and low V_{dd} operation, which fortunately are also the primary development goals for TFETs. Device

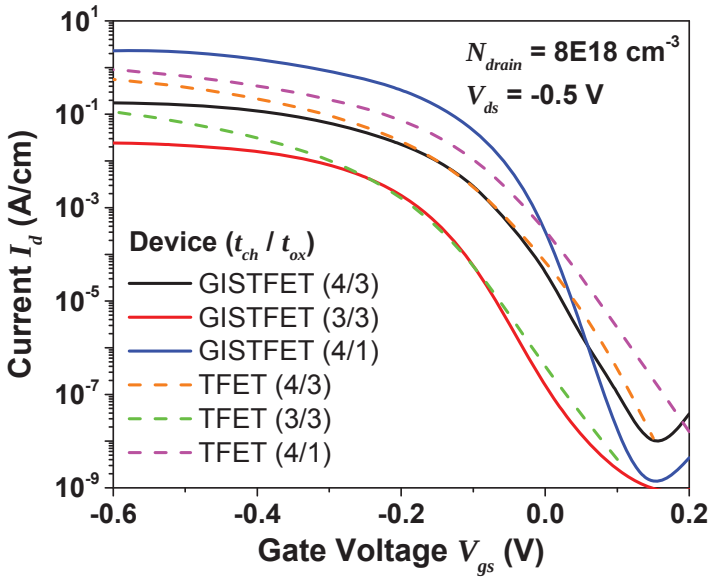

Figure 5.7: Simulated $I_d - V_{gs}$ for InAs DG GISTFETs (solid) and TFETs (dashed) with different channel and oxide thicknesses t_{ch} and t_{ox}. The source doping N_s is 5×10^{18} cm^{-3} for the GISTFETs and 2×10^{19} cm^{-3} for the TFETs, respectively, with other device characteristics are as in Fig. 5.3.

performance can be further improved if a heterojunction can be aligned in the channel alongside the M1-M2 interface, which might be achievable using a vertical device layout for instance[6, 34]. Finally, 2-D semiconductors such as graphene nanoribbons or transition metal dichalcogenides like MoS_2 are also promising GISTFET channel candidates due to their very small λ[106].

5.3 Conclusion

The GISTFET offers the possibility of high performance complementary tunneling transistors by utilizing gate work function offsets rather than doping to induce tunneling. Quantum transport simulations confirm the operating principle as well as subtleties related to quantization effects near the C1/C2 junction, while crucially exhibiting substantially improved p-type GISTFET characteristics over conventional TFETs.

CHAPTER 6

Modeling Intraband Tunneling Leakage in Ultrascaled MOSFETs

> If you can mock a leek, you can eat a leek.
>
> Shakespeare, *Henry V*

III-V semiconductors are promising alternatives to silicon in nanoscale FET channels because of their high mobilities, but their small effective masses (EM) increase source-drain tunneling (SDT) current. This is a crucial concern since off-state leakage is a major and increasing cause of power consumption in state-of-the-art CMOS. SDT therefore must be considered in device design and projection; indeed, its dominance in sub-10 nm silicon[171, 172, 173] and III-V FETs[174, 175, 176, 177, 178, 179] has been established by quantum mechanical studies using NEGF. However, the computational requirements of quantum simulations limit their use and simple models are desirable for more efficient device studies. Analytical treatments of intraband tunneling presently only offer qualitative guidelines[119] or require iterative calculations and externally supplied band structures[180]. The density gradient model attempts to mimic tunneling[181] in technology computer-aided-design (TCAD) simulations but requires calibration and is generally implemented in EM form, which is questionable for nanoscale devices. In this paper, we use band structure and NEGF calculations to quantify conduction band (CB) nonparabolicity corrections in nanoscale III-V FETs, derive parameter-free analytical formulas for SDT in bulk and quantum confined devices, and numerically validate these models using our simulations. Our results enable accurate and efficient calculations of intraband

tunneling in FET models and TCAD simulations. The work in this chapter has been published in Ref. [11].

6.1 NEGF Simulations of Intraband Tunneling

Accurate band structure models are critical when studying quantum transport, particularly for nanoscale devices where quantum confinement (QC) occurs. Ideally the electronic structure should be self-consistently calculated using large basis set approaches like *ab initio* density functional theory (DFT) or semi-empirical tight-binding (TB), but such methods are time-consuming. A common compromise is to use EM but to adjust the mass(es) to the energy dispersion or density of states (DOS) computed using more detailed models. Alternatively, since the primary correction to the CB EM in direct gap III-V materials is nonparabolicity arising from mixing of the Γ valley valence band (VB) states, a 4-band (neglecting spin-orbit coupling) or 8-band $\mathbf{k}\cdot\mathbf{p}$ Kane Hamiltonian[77] can be used to include CB-VB coupling within a comparatively small basis. However, the Kane model does not include more remote band extrema which may be important at high energies or under strong confinement.

We compare these methods in Fig. 6.1 by computing the complex subband structure of 8.5 nm and 3.7 nm thick $In_{0.53}Ga_{0.47}As$ quantum wells using *spds** TB, 4-band $\mathbf{k}\cdot\mathbf{p}$, and EM adjusted to fit the TB results. Parameters for the multiband Hamiltonians are taken from Table 2.1 in Chapter 2 and spin-orbit coupling is neglected because its effects are minor for low energy CB states. For the lowest conduction subbands of interest, the TB and $\mathbf{k}\cdot\mathbf{p}$ calculations agree quite well, indicating that VB-induced nonparabolicity dominates at these dimensions. The adjusted EM fits the lowest subband at real k but is qualitatively inaccurate for imaginary k because it lacks CB-VB coupling.

To observe how these differences impact devices, we compare ballistic NEGF simulations of InGaAs FETs using the adjusted EM and 4-band $\mathbf{k}\cdot\mathbf{p}$ Hamiltonians. We study double gate (DG) structures conforming to ITRS projections for the 15 and 6 nm nodes, as summarized in Table 6.1[139]. The resulting $I-V$ and $C-V$ curves are shown in Fig.

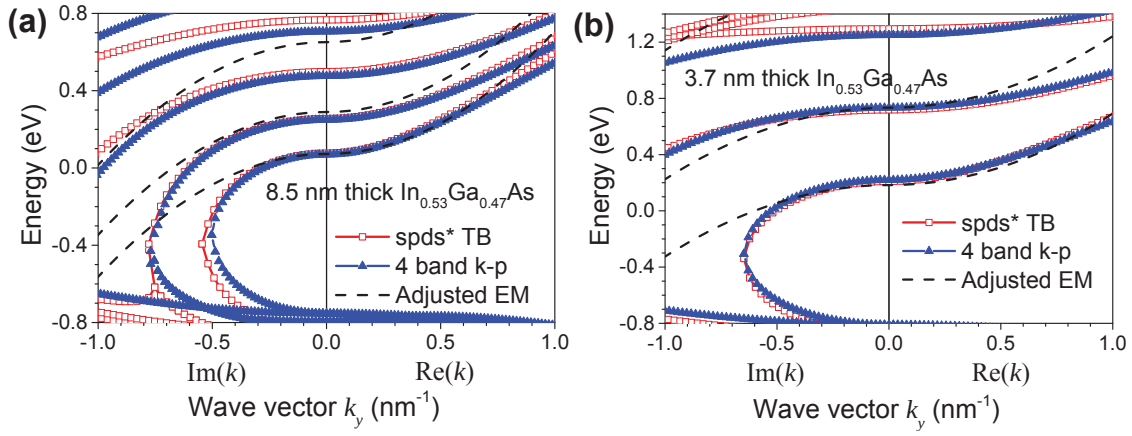

Figure 6.1: Complex subband structure of (a) 8.5 nm and (b) 3.7 nm thick InGaAs ideal quantum wells from TB, k-p, and EM assuming infinite potential boundary conditions. The fitted EM m are (a) 0.072 and 0.06 and (b) 0.15 and 0.075 in the confined and unconfined directions, respectively.

Table 6.1: Device geometrical parameters for simulated DG FETs.

Node	15 nm	6 nm
L_g (nm)	12.8	5.9
t_{ox} (nm)	0.68	0.45
t_{ch} (nm)	8.5	3.7
V_{ds} (V)	0.73	0.57

6.2. In the on-state, the currents computed using adjusted EM and $\mathbf{k}\cdot\mathbf{p}$ are in fair agreement, whereas EM noticeably underestimates the capacitance for the 15 nm device. This is because nonparabolicity and higher subbands both increase DOS and decrease carrier velocity; these trends partly compensate each other in the current, but lead to increased quantum capacitance since only DOS impacts the latter. The effect is less pronounced at 6 nm because the lower operating voltage and increased QC render the higher subbands and nonparabolic region less significant.

By contrast, EM simulations dramatically underestimate subthreshold leakage, especially at 6 nm when SDT dominates the spectral current as seen in Fig. 6.3. This is because the parabolic imaginary dispersion of the EM model, shown in Fig. 6.1, artificially suppresses the tunneling probability. Hence even intraband tunneling calculations must include CB-VB coupling as well as QC effects. These effects have been accurately described for interband tunneling using the band gap scaled (BGS) two-band model developed in Chapter 3, which we will therefore adapt to SDT.

6.2 Modeling Subthreshold Electrostatic Barriers

Before modeling the tunneling current, the potential barrier must be known from the device electrostatics. Pseudo-2-D approximations are widely used to analytically describe MOSFET subthreshold electrostatics[133], but most formulations do not account for the source and drain depletion regions, which are significant at short gate lengths. We therefore present a pseudo-2-D channel formula which includes these depletion regions, adapting similar models previously derived for interband tunneling FETs (TFETs) in Chapter 4. Assuming abrupt symmetric source and drain doping N_D and using the notation in Fig. 6.4, the resulting electrostatic potential energy along the channel direction y is given

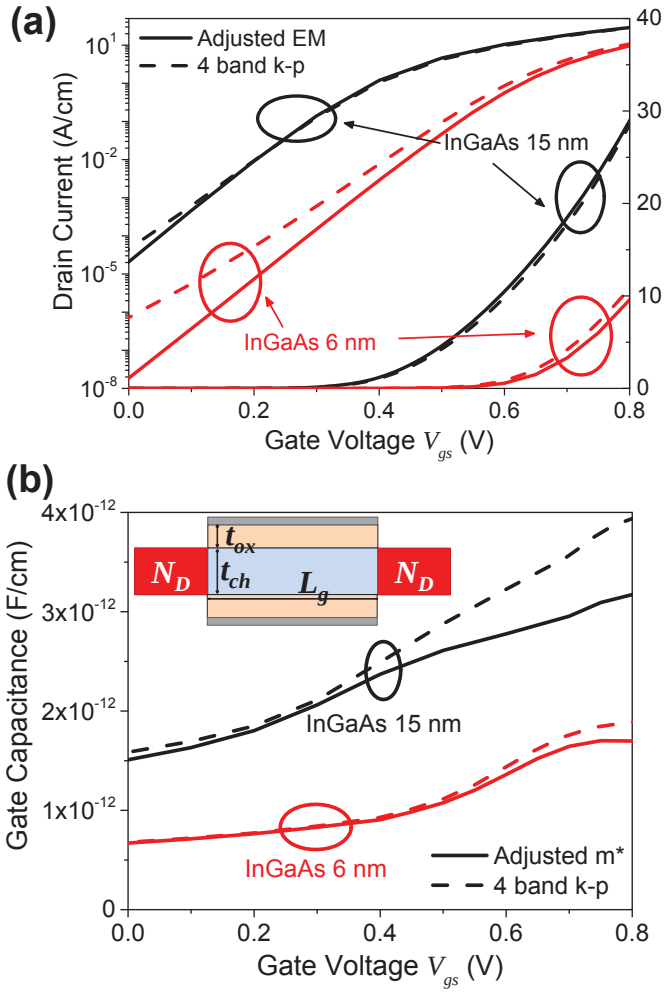

Figure 6.2: (a) NEGF $I - V$ simulations of InGaAs DG FETs with source and drain doping $N_D = 10^{19}$ cm^{-3} on log (left axis) and linear (right axis) scales. (b) DC $C - V$ simulations of same devices. Inset: simulated DG structure.

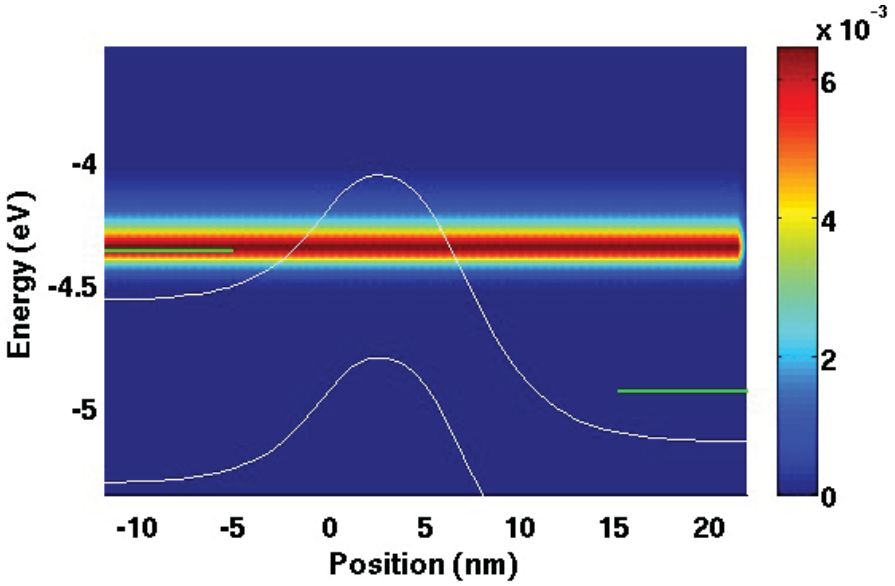

Figure 6.3: Energy-resolved current density and band diagram for 6 nm InGaAs device ($V_{gs} = 0$, $V_{ds} = 0.73$ V).

by

$$V(y) = \begin{cases} \frac{q^2 N_D}{2\epsilon_{ch}}(y + y_s)^2 & y_s \leq y < 0 \\ \psi_{ch} - A\exp(\frac{y}{\lambda}) - B\exp(-\frac{y}{\lambda}) & 0 \leq y < L_g \\ -V_{ds} + \frac{q^2 N_D}{2\epsilon_{ch}}(y_d - y)^2 & L_g \leq y \leq y_d \end{cases} \tag{6.1}$$

where ϵ_{ch} is the channel permittivity, L_g is the gate length, λ is the characteristic length, and ψ_{ch} is related to the gate voltage via $\psi_{ch} = \Delta V - V_{gs}$, where ΔV is the flatband shift. The coefficients A and B are

$$A = \frac{\psi_{ch} + V_{ds} - V_{d,dep} + (V_{s,dep} - \psi_{ch})\exp(-\frac{L_g}{\lambda})}{2\sinh(\frac{L_g}{\lambda})} \tag{6.2}$$

$$B = \frac{V_{d,dep} - V_{ds} - \psi_{ch} + (\psi_{ch} - V_{s,dep})\exp(\frac{L_g}{\lambda})}{2\sinh(\frac{L_g}{\lambda})} \tag{6.3}$$

and the source and drain depletion widths y_s and y_d equal

$$y_s = -\sqrt{\frac{2\epsilon_{ch} V_{s,dep}}{qN_D}} \tag{6.4}$$

$$y_d = L_g + \sqrt{\frac{2\epsilon_{ch} V_{d,dep}}{qN_D}} \tag{6.5}$$

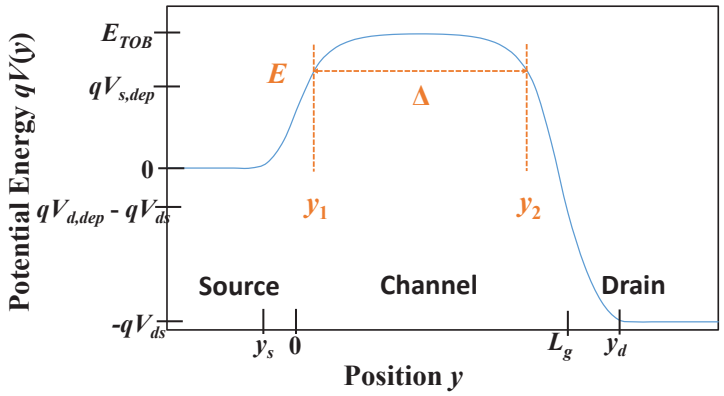

Figure 6.4: Conventions for pseudo-2-D electrostatic model of FETs in subthreshold. At a given energy E, the classical turning points y_1, y_2 define the tunneling width Δ.

where

$$V_{s,dep} = \psi_{ch} + V_{s0} - \sqrt{2\psi_{ch}V_{s0} + V_{s0}^2}, \tag{6.6}$$

$$V_{d,dep} = V_{ds} + \psi_{ch} + V_{s0} - \sqrt{2(V_{ds} + \psi_{ch})V_{s0} + V_{s0}^2}, \tag{6.7}$$

$$V_{s0} = \frac{q^2 N_D \lambda^2}{\epsilon_{ch} \coth^2(\frac{L_g}{\lambda})}. \tag{6.8}$$

From these equations the top of the potential barrier is found to be $E_{TOB} = q(\psi_{ch} - 2\sqrt{AB})$, and the location y for a given V between $-qV_{ds}$ and E_{TOB} is given by

$$y(V) = \begin{cases} -y_s + \sqrt{\dfrac{2\epsilon_{ch}V}{q^2 N_D}} & 0 < V < V_{s,dep} \\ y_d - \sqrt{\dfrac{2\epsilon_{ch}}{q^2 N_D}(V + V_{ds})} & -V_{ds} < V < V_{d,dep} - V_{ds} \end{cases} \tag{6.9}$$

if V lies in the source or drain depletion regions and

$$y = \lambda \ln \left[\frac{\psi_{ch} - V}{2A} \mp \frac{\sqrt{(V - \psi_{ch})^2 - 4AB}}{2A} \right] \tag{6.10}$$

if it lies within the channel. These equations are valid in depletion (the absence of strong inversion or accumulation) for a variety of device structures, including thin-film material-on-insulator, DG, and nanowire geometries, differing only in the appropriate choice of

characteristic length λ. For the DG devices simulated here, we use the λ value derived by Suzuki *et al.*[135]

$$\lambda = \sqrt{\frac{\epsilon_{ch} t_{ch} t_{ox}}{2\epsilon_{ox}} \left(1 + \frac{\epsilon_{ox} t_{ch}}{4\epsilon_{ch} t_{ox}}\right)} \tag{6.11}$$

where ϵ_{ox} and ϵ_{ch} are the permittivities of the oxide and channel material and t_{ox} and t_{ch} are the oxide and channel thicknesses.

6.3 Intraband Tunneling Modeling

Next we model the tunneling current. The elastic SDT current density is given by

$$J_{intra} = \frac{2q}{h} \int \frac{d\vec{k}_\perp}{(2\pi)^{d-1}} \int_0^{E_{TOB}} dE\, T(E(\vec{k}_\perp))[f_S(E) - f_D(E)] \tag{6.12}$$

and we will use the Wentzel-Kramers-Brillouin (WKB) approximation for the tunneling probability

$$T(E(\vec{k}_\perp)) = \exp\left(-2 \int_{y_1}^{y_2} \kappa(E(\vec{k}_\perp), y) dy\right). \tag{6.13}$$

Here $\vec{k}_\perp$ is the transverse momentum, d is the device dimensionality, $f_S(E)$ and $f_D(E)$ are the Fermi-Dirac distribution functions for the source and drain respectively, $\kappa(E(\vec{k}_\perp), y)$ is the imaginary part of the energy and position-dependent momentum, and y_1 and y_2 are the classical turning points defining the width of the tunneling barrier at a given $\vec{k}_\perp$ and energy E. Once the form of κ is determined, the action integral inside the exponential can be integrated to obtain the tunneling probability. Eq. 6.12 is easily generalized to the case of multiple subbands though we will focus on the lowest subband, which usually dominates leakage.

The EM approximation gives $\kappa(E, y) = \sqrt{2m(qV(y) - E + E_\perp)}/\hbar$ where $E_\perp$ is the energy associated with the transverse momentum. To include nonparabolic CB-VB coupling, we instead use a BGS 2-band model to obtain

$$\kappa(E, y) = \frac{1}{P} \sqrt{(E_{g,QC} + E - qV(y))(V(y) - E)} \tag{6.14}$$

where $E_{g,QC}$ is the QC band gap obtained from band structure calculations or measurements and the momentum matrix element $P = \hbar \sqrt{\frac{E_g}{4}\left[m_{CB}^{-1} + m_{LH}^{-1}\right]}$ in terms of the

bulk band gap E_g and CB and light hole (LH) masses m_{CB} and m_{LH}. In a bulk device, $E_{g,QC} = E_g$. Detailed justification of the BGS model and evidence of its agreement with more sophisticated band structure and interband tunneling calculations has already been given in Chapter 3.

The computation of current using Eq. 6.12 is onerous because it requires integrations over the tunneling path, energy, and transverse momentum. In 1-D devices like nanowires, the transverse term disappears, while in 2-D and 3-D devices like DG or bulk MOSFETs, we can approximate the transverse integration by Taylor expanding Eq. 6.14 around $E_\perp = 0$. However, the tunneling probability Eq. 6.13 still requires an integration over the potential barrier, which is not analytically solvable in general. To obtain a tractable result for the transverse integration, we use the EM approximation for the transverse energy $E_\perp = \hbar^2 k_\perp^2/2m$ and treat the potential as a $E_\perp = 0$ square barrier of height $E_{TOB} - E$ and width $\Delta = y_2(E) - y_1(E)$. This immediately leads to

$$J_{intra} = \frac{q}{\pi\hbar}\int_0^{E_{TOB}} dE T(E)[f_L(E) - f_R(E)] \tag{6.15}$$

with the transverse-integrated transmission coefficients for different dimensions given by

$$T_{1\text{-D}}(E) = \exp\left(-\frac{2}{P}\int_{y_1}^{y_2}\sqrt{(E_{g,QC} + E - qV(y))(qV(y) - E)}dy\right) \tag{6.16}$$

$$T_{2\text{-D}}(E) = \sqrt{\frac{m}{2\pi\hbar^2\alpha(E)}}\text{erf}\left(\sqrt{E\alpha(E)}\right)T_{1\text{-D}}(E) \tag{6.17}$$

$$T_{3\text{-D}}(E) = \frac{m}{2\pi\hbar^2\alpha(E)}\left[1 - \exp\left(\sqrt{E\alpha(E)}\right)\right]T_{1\text{-D}}(E) \tag{6.18}$$

where

$$\alpha(E) = \frac{\Delta\left(\sqrt{E_{g,QC} - E_{TOB} + E} + \sqrt{E_{TOB} - E}\right)}{P\sqrt{(E_{g,QC} - E_{TOB} + E)(E_{TOB} - E)}}. \tag{6.19}$$

Eqs. 6.15-6.19 are the central results of this chapter and can be used to calculate the intraband tunneling current given a channel potential $V(y)$ from numerical simulations or an analytical model like Eqs. 6.9-6.10. We find that energies around the source Fermi level E_{fs} usually dominate the tunneling current, such that Eq. 6.15 can often be approximated by computing only $T(E_{fs})$ and replacing the integral over E with $\min(2kT, qV_{ds})$. The limiting case of parabolic bands, relevant for indirect gap semiconductors like silicon, is

obtained via the substitution of $\sqrt{2m_{CB}}/\hbar$ for $\sqrt{E_{g,QC}+E-qV(x)}/P$ in Eq. 6.16 and $\alpha_{\mathrm{EM}}=\sqrt{2m^*/(E_{\mathrm{TOB}}-E)}\Delta/\hbar$ for Eq. 6.19.

The only inputs to the electrostatic and tunneling models we have derived are the device geometry, source-drain doping and Fermi energies, and the QC band gap $E_{g,QC}$ and mass m. No adjustable parameters are needed. In Fig. 6.5(a) we compare the 2-D tunneling current using Eq. 6.15 with our NEGF DG simulations and find excellent agreement in the subthreshold regime. Because our electrostatic model Eq. 6.1 only applies for subthreshold, we limit $q\psi_{ch} \geq E_{Fs}$ in our calculations. The high V_{gs} characteristics are not captured by our model since they are dominated by current flow over the barrier rather than tunneling; if we add the former via a ballistic virtual source model[182], for instance, the total current is in good agreement over the entire bias range, as shown by the dashed lines in Fig. 6.5(a).

To further demonstrate the model's general applicability, we simulate InGaAs devices with different source/drain doping and an InAs FET in Fig. 6.5(b) and again find good agreement for our BGS model in subthreshold. The same model also applies to silicon devices, where nonparabolicity effects are smaller and the EM approximation may hold. For example, in Fig. 6.6 we observe that the tunneling current and source doping dependence of 6 nm silicon DG FinFETs calculated via EMA NEGF dominate the leakage current and are indeed correctly predicted by our analytical model. Overall, regardless of material we observe an interesting trend where higher doping leads to improved on-currents by mitigating source starvation issues[183] but also increases leakage by orders of magnitude due to narrower tunneling barriers[178, 179]. In practical devices, the importance of high source/drain doping for reducing series resistance (not considered in these scattering-free ballistic simulations) will further complicate the picture. This suggests that careful contact doping optimization will be necessary in sub-10 nm FETs to balance leakage power and performance, making a SDT model like the one presented here particularly crucial for evaluation and design purposes.

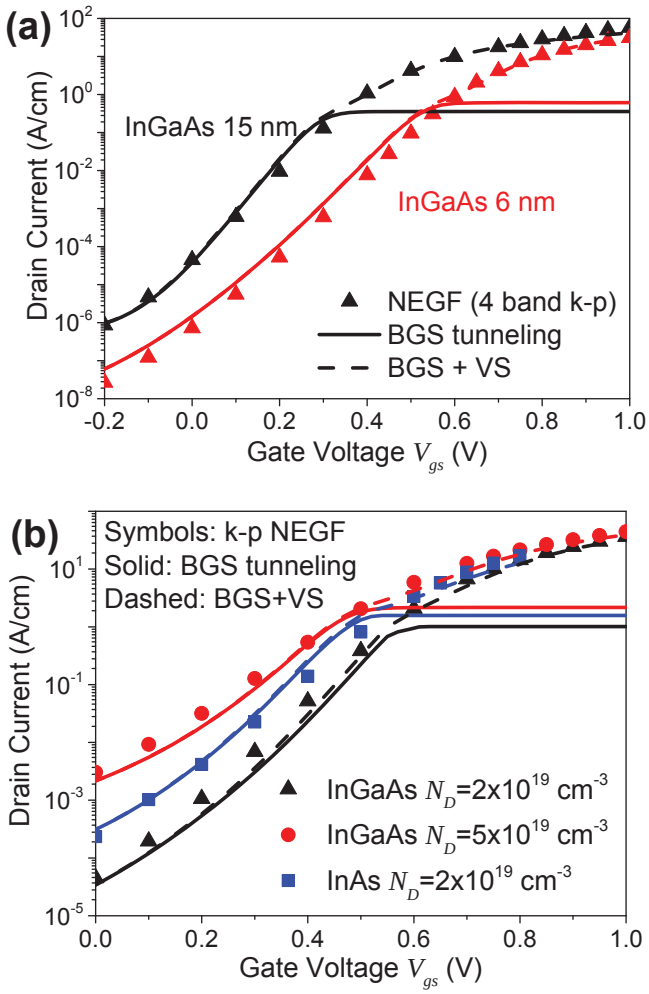

Figure 6.5: (a) Analytical versus NEGF $I-V$ for InGaAs DG FETs of Fig. 6.2. Solid lines are for the BGS tunneling model only and dashed lines are the sum of BGS with a virtual source (VS) model. (b) Simulated and modeled $I-V$ for the 6 nm device with different doping and channel material.

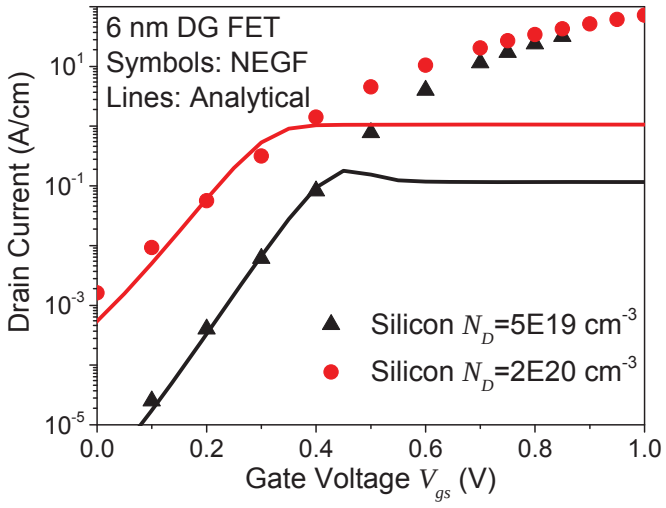

Figure 6.6: (a) Analytical versus NEGF $I-V$ for 6 nm silicon DG FETs with different source/drain doping.

6.4 Conclusion

Our results demonstrate that intraband tunneling can be modeled using simple, parameter-free formulas which are easily applied to different device geometries and dimensions for both III-V materials and silicon. The models presented here are also generalizable for scaled transistors using novel materials like direct-gap 2-D semiconductors (including transition metal dichalcogenides), or carbon nanotubes (CNTs). These models will be crucial in the development of sub-10 nm technologies where SDT is expected to dominate FET subthreshold characteristics.

CHAPTER 7

Summary and Future Directions

In my end is my beginning.

T. S. Eliot, "East Coker"

7.1 New Results of this Work

In this dissertation, we have investigated the physics and design of tunneling in MOSFETs and TFETs using a combination of quantum mechanical and semiclassical simulations and analytical methods. We have designed and implemented a complete semiconductor device simulator using the NEGF formalism with which we can study basic transport physics as well as calculate device characteristics of realistic structures. Whenever possible we have attempted to bridge the gap between the different levels of device analysis, using quantum calculations to obtain rigorous results and providing physically motivated models to accurately approximate tunneling effects within semiclassical numerical and analytical methods. We have developed an improved analytical theory of interband tunneling in bulk semiconductors and validated its accuracy against NEGF calculations. We introduced the band gap scaling (BGS) approximation to extend tunneling calculations to quantum confined structures and shown the physical reasons for its success. Building on our basic transport modeling, we derived the first physically well-defined, quantitatively useful models of tunneling currents in TFETs and MOSFETs. Using the device engineering insights gained in our theoretical study, we proposed a new device, the GISTFET, which offers the potential for ultralow voltage, high performance complementary tunneling III-V transistors for future low power applications. The work presented here thus

spans the gamut of theoretical semiconductor device analysis, from fundamental transport physics through device-level modeling to the development of new device concepts.

7.2 Future Directions

The theory of tunneling and tunneling devices we have developed in this dissertation can be further extended to include other phenomena and emerging technologies.

7.2.1 Modeling Tunneling in Heterojunctions

The band offset in semiconductor heterostructures can lead to smaller tunneling barriers and hence higher drive currents. For this reason, heterojunction TFETs with a type-II or type-III material interface between the source and channel are becoming increasingly popular in theoretical and experimental studies. Analytical models of such devices generally rely on WKB-based calculations without careful justification or comparison with experiment. Generalization of the BGS model to heterojunctions can provide a physically justifiable and more accurate approach to calculate tunneling currents in these structures.

7.2.2 Theory of Scattering-Assisted Tunneling

We have focused throughout this dissertation on coherent transport in semiconductors, where tunneling occurs through band coupling introduced by the electric field. However, incoherent tunneling is an important process in many realistic scenarios. For instance, in indirect gap semiconductors, interband tunneling is accompanied by scattering off of phonons, impurities, or other excitations to conserve momentum[96]. Alternatively, defect-assisted tunneling (which may involve multiphonon cascade processes[184]) often limits the leakage of realistic devices and may obscure the steep SS behavior of experimental TFETs. The fundamental theory of these mechanisms remains unsatisfactory in many cases, and its agreement with experimental data unclear. Detailed study of tunneling in the presence of scattering in NEGF may provide a more rigorous theory and give insight

into the mechanisms limiting performance in experimental tunneling devices.

7.2.3 Modeling and Understanding Limits of Experimental TFETs

In general, the exact causes of the performance limitations in experimental TFETs, particularly high leakage currents and SS, are not truly understood, though a plethora have been explanations have been offered, including band tailing due to disorder, tunneling through interface states, leakage via bulk traps, and poorly optimized electrostatics. A device theory that can quantitatively explain experiments with minimal, physically well-defined adjustable parameters will be critical in elucidating the performance bottlenecks and designing devices which overcome these problems. Crucial to this effort will be the development of a quantitatively accurate modeling of defect-assisted tunneling, as mentioned above.

A detailed understanding of leakage and other limitations in TFETs will also be important in developing more realistic projections and assessments of TFET-based electronics. Many circuit-level comparisons of tunneling technologies rely on models which are likely to be too optimistic for practical TFETs; a satisfactory model should be used with an equally realistic device-circuit methodology like PROCEED[3] to draw useful conclusions about the place of TFETs in future electronic systems.

7.2.4 GISTFET Development

In this work, we introduced the GISTFET concept and provided preliminary evidence using NEGF simulations that it can offer high performance complementary device operation. However, many fundamental and practical questions remain. Theoretically, it will be important to assess how scattering effects, particularly via optical phonons, may smear out the electron distribution in the induced source and affect performance. It will also be important to assess the impact of nonidealities like gate tunneling leakage and gradients in work function between M1 and M2. Since improved device electrostatics is especially critical for good GISTFET operation, the use of 2-D materials like transition

dichalcogenides or 3-D structures like nanowires or FinFETs should also be explored.

It is clearly desirable to demonstrate the device concept experimentally. Assuming such a demonstration, further experimental study of the choice of gate metals, structural design, and leakage will be critical to develop the GISTFET into a commercially viable device.

References

[1] W. Haensch, E. J. Nowak, R. H. Dennard, P. M. Solomon, A. Bryant, O. H. Dokumaci, A. Kumar, X. Wang, J. B. Johnson, and M. V. Fischetti, "Silicon CMOS devices beyond scaling," *IBM J. Res. & Dev.*, vol. 50, pp. 339–361, July 2006.

[2] L. Chang, D. Frank, R. Montoye, S. Koester, B. Ji, P. Coteus, R. Dennard, and W. Haensch, "Practical Strategies for Power-Efficient Computing Technologies," *Proc. IEEE*, vol. 98, pp. 215–236, Feb. 2010.

[3] S. Wang, A. Pan, C. O. Chui, and P. Gupta, "PROCEED: A pareto optimization-based circuit-level evaluator for emerging devices," in *Design Automation Conference (ASP-DAC), 2014 19th Asia and South Pacific*, pp. 818–824, IEEE, Jan. 2014.

[4] M. Luisier and G. Klimeck, "Simulation of nanowire tunneling transistors: From the Wentzel–Kramers–Brillouin approximation to full-band phonon-assisted tunneling," *J. Appl. Phys.*, vol. 107, no. 8, p. 084507, 2010.

[5] M. Noguchi, S. Kim, M. Yokoyama, S. Ji, O. Ichikawa, T. Osada, M. Hata, M. Takenaka, and S. Takagi, "High Ion/Ioff and low subthreshold slope planar-type InGaAs tunnel FETs with Zn-diffused source junctions," in *Proc. IEDM 2013*, pp. 28.1.1–28.1.4, IEEE, Dec. 2013.

[6] G. Dewey, B. Chu-Kung, J. Boardman, J. M. Fastenau, J. Kavalieros, R. Kotlyar, W. K. Liu, D. Lubyshev, M. Metz, N. Mukherjee, P. Oakey, R. Pillarisetty, M. Radosavljevic, H. W. Then, and R. Chau, "Fabrication, characterization, and physics of III-V heterojunction tunneling Field Effect Transistors (H-TFET) for steep subthreshold swing," in *Electron Devices Meeting (IEDM), 2011 IEEE International*, pp. 33.6.1–33.6.4, IEEE, Dec. 2011.

[7] A. Pan and C. O. Chui, "A Quasi-Analytical Model for Double-Gate Tunneling Field-Effect Transistors," *IEEE Electron Device Lett.*, vol. 33, pp. 1468–1470, Oct. 2012.

[8] A. Pan, S. Chen, and C. O. Chui, "Electrostatic Modeling and Insights Regarding Multigate Lateral Tunneling Transistors," *IEEE Trans. Electron Devices*, vol. 60, pp. 2712–2720, Sept. 2013.

[9] A. Pan and C. O. Chui, "Modeling direct interband tunneling. II. Lower-dimensional structures," *J. Appl. Phys.*, vol. 116, p. 054509, Aug. 2014.

[10] A. Pan and C. O. Chui, "Modeling direct interband tunneling. I. Bulk semiconductors," *J. Appl. Phys.*, vol. 116, p. 054508, Aug. 2014.

[11] A. Pan and C. O. Chui, "Modeling source-drain tunneling in ultimately scaled III–V transistors," *Appl. Phys. Lett.*, vol. 106, p. 243505, June 2015.

[12] A. Pan and C. O. Chui, "Gate-Induced Source Tunneling FET (GISTFET)," *IEEE Trans. Electron Devices*, vol. 62, pp. 2390–2395, Aug. 2015.

[13] R. Dennard, F. Gaensslen, V. Rideout, E. Bassous, and A. LeBlanc, "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits*, vol. 9, pp. 256–268, Oct. 1974.

[14] Y. Taur and T. H. Ning, *Fundamentals of modern VLSI devices*. Cambrige, UK ; New York: Cambridge University Press, 1998.

[15] K. Gopalakrishnan, P. Griffin, and J. Plummer, "Impact Ionization MOS (I-MOS)–Part I: Device and Circuit Simulations," *IEEE Trans. Electron Devices*, vol. 52, pp. 69–76, Jan. 2005.

[16] V. Pott, H. Kam, R. Nathanael, J. Jeon, E. Alon, and T.-J. King Liu, "Mechanical Computing Redux: Relays for Integrated Circuit Applications," *Proc. IEEE*, vol. 98, pp. 2076–2094, Dec. 2010.

[17] S. Salahuddin and S. Datta, "Use of Negative Capacitance to Provide Voltage Amplification for Low Power Nanoscale Devices," *Nano Lett.*, vol. 8, pp. 405–410, Feb. 2008.

[18] A. M. Ionescu and H. Riel, "Tunnel field-effect transistors as energy-efficient electronic switches," *Nature*, vol. 479, pp. 329–337, Nov. 2011.

[19] J. Knoch, S. Mantl, and J. Appenzeller, "Impact of the dimensionality on the performance of tunneling FETs: Bulk versus one-dimensional devices," *Solid-State Electron.*, vol. 51, pp. 572–578, Apr. 2007.

[20] S. Hofstein and G. Warfield, "The insulated gate tunnel junction triode," *IEEE Trans. Electron Devices*, vol. 12, pp. 66–76, Feb. 1965.

[21] S. Banerjee, W. Richardson, J. Coleman, and A. Chatterjee, "A new three-terminal tunnel device," *IEEE Electron Device Lett.*, vol. 8, pp. 347–349, Aug. 1987.

[22] T. Baba, "Proposal for Surface Tunnel Transistors," *Jpn. J. Appl. Phys.*, vol. 31, pp. L455–L457, Apr. 1992.

[23] W. M. Reddick and G. A. J. Amaratunga, "Silicon surface tunnel transistor," *Appl. Phys. Lett.*, vol. 67, no. 4, p. 494, 1995.

[24] J. Koga and A. Toriumi, "Negative differential conductance in three-terminal silicon tunneling device," *Appl. Phys. Lett.*, vol. 69, no. 10, p. 1435, 1996.

[25] J. Appenzeller, Y.-M. Lin, J. Knoch, and P. Avouris, "Band-to-Band Tunneling in Carbon Nanotube Field-Effect Transistors," *Phys. Rev. Lett.*, vol. 93, Nov. 2004.

[26] K. K. Bhuwalka, J. Schulze, and I. Eisele, "Performance Enhancement of Vertical Tunnel Field-Effect Transistor with SiGe in the δp $^{+}$ Layer," *Jpn. J. Appl. Phys.*, vol. 43, pp. 4073–4078, July 2004.

[27] W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. K. Liu, "Tunneling Field-Effect Transistors (TFETs) With Subthreshold Swing (SS) Less Than 60 mV/dec," *IEEE Electron Device Lett.*, vol. 28, pp. 743–745, Aug. 2007.

[28] F. Mayer, C. Le Royer, J.-F. Damlencourt, K. Romanjek, F. Andrieu, C. Tabone, B. Previtali, and S. Deleonibus, "Impact of SOI, Si1-xGexOI and GeOI substrates on CMOS compatible Tunnel FET performance," in *Proc. IEDM 2008*, pp. 1–5, IEEE, Dec. 2008.

[29] T. Krishnamohan, D. Kim, S. Raghunathan, and K. Saraswat, "Double-Gate Strained-Ge Heterostructure Tunneling FET (TFET) With record high drive currents and <60mv/dec subthreshold slope," in *Proc. IEDM 2008*, pp. 1–3, IEEE, Dec. 2008.

[30] S. H. Kim, H. Kam, C. Hu, and T.-J. K. Liu, "Germanium-source tunnel field effect transistors with record high ION/IOFF," in *Proc. VLSIT 2009*, pp. 178–179, IEEE, June 2009.

[31] D. Leonelli, A. Vandooren, R. Rooyackers, A. S. Verhulst, S. D. Gendt, M. M. Heyns, and G. Groeseneken, "Performance Enhancement in Multi Gate Tunneling Field Effect Transistors by Scaling the Fin-Width," *Jpn. J. Appl. Phys.*, vol. 49, p. 04DC10, Apr. 2010.

[32] K. Jeon, W.-Y. Loh, P. Patel, C. Y. Kang, J. Oh, A. Bowonder, C. Park, C. S. Park, C. Smith, P. Majhi, H.-H. Tseng, R. Jammy, T.-J. K. Liu, and C. Hu, "Si tunnel transistors with a novel silicided source and 46mv/dec swing," in *Proc. VLSIT 2010*, pp. 121–122, IEEE, June 2010.

[33] R. Gandhi, Z. Chen, N. Singh, K. Banerjee, and S. Lee, "Vertical Si-Nanowire n-Type Tunneling FETs With Low Subthreshold Swing ($\leq \hbox{50}\ \hbox{mV/decade}$) at Room Temperature," *IEEE Electron Device Lett.*, vol. 32, pp. 437–439, Apr. 2011.

[34] H. Zhao, Y. Chen, Y. Wang, F. Zhou, F. Xue, and J. Lee, "InGaAs Tunneling Field-Effect-Transistors With Atomic-Layer-Deposited Gate Oxides," *IEEE Trans. Electron Devices*, vol. 58, pp. 2990–2995, Sept. 2011.

[35] G. Zhou, R. Li, T. Vasen, M. Qi, S. Chae, Y. Lu, Q. Zhang, H. Zhu, J.-M. Kuo, T. Kosel, M. Wistey, P. Fay, A. Seabaugh, and Huili Xing, "Novel gate-recessed vertical InAs/GaSb TFETs with record high ION of 180 uA/um at VDS =0.5 V," in *Proc. IEDM 2012*, pp. 32.6.1–32.6.4, IEEE, Dec. 2012.

[36] R. Bijesh, H. Liu, H. Madan, D. Mohata, W. Li, N. V. Nguyen, D. Gundlach, C. A. Richter, J. Maier, K. Wang, T. Clarke, J. M. Fastenau, D. Loubychev, W. K. Liu, V. Narayanan, and S. Datta, "Demonstration of In0.9ga0.1as/GaAs0.18sb0.82 near broken-gap tunnel FET with ION = 740 uA/um, GM = 70 uS/um and gigahertz switching performance at VDD=0.5v," in *Proc. IEDM 2013*, pp. 28.2.1–28.2.4, IEEE, Dec. 2013.

[37] B. Rajamohanan, R. Pandey, V. Chobpattana, C. Vaz, D. Gundlach, K. P. Cheung, J. Suehle, S. Stemmer, and S. Datta, "0.5 V Supply Voltage Operation of In0.65ga0.35as/GaAs0.4sb0.6 Tunnel FET," *IEEE Electron Device Lett.*, vol. 36, pp. 20–22, Jan. 2015.

[38] K. Tomioka, M. Yoshimura, and T. Fukui, "Steep-slope tunnel field-effect transistors using III-V nanowire/Si heterojunction," in *VLSI Technology, 2012 Symposium on*, pp. 47–48, IEEE, June 2012.

[39] K. Tomioka and T. Fukui, "Current increment of tunnel field-effect transistor using InGaAs nanowire/Si heterojunction by scaling of channel length," *Appl. Phys. Lett.*, vol. 104, p. 073507, Feb. 2014.

[40] B. Ganjipour, J. Wallentin, M. T. Borgström, L. Samuelson, and C. Thelander, "Tunnel Field-Effect Transistors Based on InP-GaAs Heterostructure Nanowires," *ACS Nano*, vol. 6, pp. 3109–3113, Apr. 2012.

[41] L. Knoll, Q. T. Zhao, A. Nichau, S. Richter, G. V. Luong, S. Trellenkamp, A. Schafer, L. Selmi, K. K. Bourdelle, and S. Mantl, "Demonstration of improved transient response of inverters with steep slope strained Si NW TFETs by reduction of TAT with pulsed I-V and NW scaling," in *Electron Devices Meeting (IEDM), 2013 IEEE International*, pp. 4.4.1–4.4.4, IEEE, Dec. 2013.

[42] A. Vandooren, A. M. Walke, A. S. Verhulst, R. Rooyackers, N. Collaert, and A. V. Y. Thean, "Investigation of the Subthreshold Swing in Vertical Tunnel-FETs Using ${\rm H}_{2}$ and ${\rm D}_{2}$ Anneals," *IEEE Trans. Electron Devices*, vol. 61, pp. 359–364, Feb. 2014.

[43] E. F. Schubert, *Doping in III-V semiconductors*. No. 1 in Cambridge studies in semiconductor physics and microelectronic engineering, Cambridge [England] ; New York, NY, USA: Cambridge University Press, 1993.

[44] K.-Y. Shen, *A Metal-Oxide-Semiconductor Tunneling Effect Transistor*. PhD thesis, University of California, Los Angeles, 2005.

[45] A. Bowonder, P. Patel, Kanghoon Jeon, Jungwoo Oh, Prashant Majhi, Hsing-Huang Tseng, and Chenming Hu, "Low-voltage green transistor using ultra shallow junction and hetero-tunneling," in *Proc. IWJT 2008*, pp. 93–96, IEEE, May 2008.

[46] Y. Morita, T. Mori, S. Migita, W. Mizubayashi, A. Tanabe, K. Fukuda, T. Matsukawa, K. Endo, S. O'uchi, Y. X. Liu, M. Masahara, and H. Ota, "Performance Enhancement of Tunnel Field-Effect Transistors by Synthetic Electric Field Effect," *IEEE Electron Device Lett.*, vol. 35, pp. 792–794, July 2014.

[47] A. M. Walke, A. Vandooren, R. Rooyackers, D. Leonelli, A. Hikavyy, R. Loo, A. S. Verhulst, K.-H. Kao, C. Huyghebaert, G. Groeseneken, V. R. Rao, K. K. Bhuwalka, M. M. Heyns, N. Collaert, and A. V.-Y. Thean, "Fabrication and Analysis of a ${\rm Si}/{\rm Si}_{0.55}{\rm Ge}_{0.45}$ Heterojunction Line Tunnel FET," *IEEE Trans. Electron Devices*, vol. 61, pp. 707–715, Mar. 2014.

[48] W. G. Vandenberghe, B. Sorée, W. Magnus, G. Groeseneken, and M. V. Fischetti, "Impact of field-induced quantum confinement in tunneling field-effect devices," *Appl. Phys. Lett.*, vol. 98, no. 14, p. 143503, 2011.

[49] D. E. Nikonov and I. A. Young, "Overview of Beyond-CMOS Devices and a Uniform Methodology for Their Benchmarking," *Proc. IEEE*, vol. 101, pp. 2498–2533, Dec. 2013.

[50] S. Selberherr, *Analysis and simulation of semiconductor devices*. Wien, Austria ; New York: Springer-Verlag, 1984.

[51] P. Y. Yu and M. Cardona, *Fundamentals of Semiconductors*. Graduate Texts in Physics, Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[52] W. Frensley, "Boundary conditions for open quantum systems driven far from equilibrium," *Rev. Mod. Phys.*, vol. 62, pp. 745–791, July 1990.

[53] L. P. Kadanoff and G. Baym, *Quantum statistical mechanics: Green's function methods in equilibrium and nonequilibrium problems*. Advanced book classics, Cambridge, Mass: Perseus Books, 1962.

[54] L. V. Keldysh, "Diagram technique for nonequilibrium processes," *Sov. Phys. JETP*, vol. 20, pp. 1018–1026, Apr. 1965.

[55] J. Schwinger, "Brownian Motion of a Quantum Oscillator," *J. Math. Phys.*, vol. 2, no. 3, p. 407, 1961.

[56] C. Caroli, R. Combescot, P. Nozieres, and D. Saint-James, "Direct calculation of the tunneling current," *J. Phys. C.*, vol. 4, pp. 916–929, June 1971.

[57] C. Caroli, R. Combescot, P. Nozieres, and D. Saint-James, "A direct calculation of the tunnelling current: IV. Electron-phonon interaction effects," *Journal of Physics C: Solid State Physics*, vol. 5, pp. 21–42, Jan. 1972.

[58] A.-P. Jauho, "Nonequilibrium green function techniques applied to hot electron quantum transport," *Solid-State Electron.*, vol. 32, pp. 1265–1271, Dec. 1989.

[59] S. Datta, *Electronic Transport in Mesoscopic Systems*. Cambridge: Cambridge University Press, 1995.

[60] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, "Single and multiband modeling of quantum electron transport through layered semiconductor devices," *J. Appl. Phys.*, vol. 81, no. 12, pp. 7845–7869, 1997.

[61] S. Datta, *Quantum Transport: From Atom to Transistor*. Cambridge, UK: Cambridge University Press, 2005.

[62] T. Kubis and P. Vogl, "Assessment of approximations in nonequilibrium Green's function theory," *Phys. Rev. B*, vol. 83, p. 195304, May 2011.

[63] A. A. Abrikosov, L. P. Gor'kov, and I. E. Dzyaloshinski, *Methods of quantum field theory in statistical physics*. New York: Dover Publications, 1960.

[64] A. L. Fetter and J. D. Walecka, *Quantum theory of many-particle systems*. Mineola, N.Y: Dover Publications, 1971.

[65] G. D. Mahan, *Many-particle physics*. New York: Kluwer Academic/Plenum Publishers, 2000.

[66] A. Altland, *Condensed matter field theory*. Cambridge, UK ; New York: Cambridge University Press, 2006.

[67] S. Datta, "A simple kinetic equation for steady-state quantum transport," *J. Phys.: Condens. Matter*, vol. 2, pp. 8023–8052, Oct. 1990.

[68] M. Anantram, M. Lundstrom, and D. Nikonov, "Modeling of Nanoscale Devices," *Proc. IEEE*, vol. 96, pp. 1511–1550, Sept. 2008.

[69] M. P. L. Sancho, J. M. L. Sancho, J. M. L. Sancho, and J. Rubio, "Highly convergent schemes for the calculation of bulk and surface Green functions," *J. Phys. F.*, vol. 15, pp. 851–858, Apr. 1985.

[70] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, "Two-dimensional quantum mechanical modeling of nanotransistors," *J. Appl. Phys.*, vol. 91, no. 4, p. 2343, 2002.

[71] H. Haug and A.-P. Jauho, *Quantum kinetics in transport and optics of semiconductors*. Berlin; New York: Springer, 2008.

[72] R. Venugopal, Zhibin Ren, and M. Lundstrom, "Simulating quantum transport in nanoscale mosfets: ballistic hole transport, subband engineering and boundary conditions," *IEEE Trans. Nanotechnol.*, vol. 2, pp. 135–143, Sept. 2003.

[73] S. E. Laux, A. Kumar, and M. V. Fischetti, "Analysis of quantum ballistic electron transport in ultrasmall silicon devices including space-charge and geometric effects," *J. Appl. Phys.*, vol. 95, no. 10, p. 5545, 2004.

[74] M. G. Burt, "The justification for applying the effective-mass approximation to microstructures," *J. Phys.: Condens. Matter*, vol. 4, pp. 6651–6690, Aug. 1992.

[75] J. Luttinger and W. Kohn, "Motion of Electrons and Holes in Perturbed Periodic Fields," *Phys. Rev.*, vol. 97, pp. 869–883, Feb. 1955.

[76] E. O. Kane, "Band structure of indium antimonide," *J. Phys. Chem. Solids*, vol. 1, pp. 249–261, Jan. 1957.

[77] E. O. Kane, "The k •p Method," in *Semiconductors and Semimetals*, vol. 1, pp. 75–100, Elsevier, 1966.

[78] B. Foreman, "Elimination of spurious solutions from eight-band k.p theory," *Phys. Rev. B*, vol. 56, pp. R12748–R12751, Nov. 1997.

[79] R. Veprek, S. Steiger, and B. Witzigmann, "Ellipticity and the spurious solution problem of k.p envelope equations," *Phys. Rev. B*, vol. 76, p. 165320, Oct. 2007.

[80] I. Vurgaftman, J. R. Meyer, and L. R. Ram-Mohan, "Band parameters for III–V compound semiconductors and their alloys," *J. Appl. Phys.*, vol. 89, no. 11, p. 5815, 2001.

[81] T. Boykin, G. Klimeck, R. Bowen, and F. Oyafuso, "Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory," *Phys. Rev. B*, vol. 66, p. 125207, Sept. 2002.

[82] X. Cartoixà, D. Z.-Y. Ting, and T. C. McGill, "Numerical spurious solutions in the effective mass approximation," *J. Appl. Phys.*, vol. 93, no. 7, p. 3974, 2003.

[83] J. C. Slater and G. F. Koster, "Simplified LCAO Method for the Periodic Potential Problem," *Phys. Rev.*, vol. 94, pp. 1498–1524, June 1954.

[84] J.-M. Jancu, R. Scholz, F. Beltram, and F. Bassani, "Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters," *Phys. Rev. B*, vol. 57, pp. 6493–6507, Mar. 1998.

[85] S. Lee, H. Chung, K. Nahm, and C. Kim, "Band structure of ternary-compound semiconductors using a modified tight-binding method," *Phys. Rev. B*, vol. 42, pp. 1452–1454, July 1990.

[86] S. Lee, F. Oyafuso, P. von Allmen, and G. Klimeck, "Boundary conditions for the electronic structure of finite-extent embedded semiconductor nanostructures," *Phys. Rev. B*, vol. 69, p. 045316, Jan. 2004.

[87] V. Eyert, "A Comparative Study on Methods for Convergence Acceleration of Iterative Vector Sequences," *J. Comput. Phys.*, vol. 124, pp. 271–285, Mar. 1996.

[88] M. Anantram and A. Svizhenko, "Multidimensional Modeling of Nanotransistors," *IEEE Trans. Electron Devices*, vol. 54, pp. 2100–2115, Sept. 2007.

[89] L. Esaki, "Long journey into tunneling," *Rev. Mod. Phys.*, vol. 46, pp. 237–244, Apr. 1974.

[90] A. Pethe, T. Krishnamohan, D. Kim, S. Oh, H.-S. P. Wong, Y. Nishi, and K. C. Saraswat, "Investigation of the performance limits of III-V double-gate n-MOSFETs," in *Proc. IEDM 2005*, (Washington, DC), pp. 605–608, IEEE, 2005.

[91] W. Anderson, "Tunnel contribution to Hg1-xCdxTe and Pb1-xSnxTe p-n junction diode characteristics," *Infrared Phys.*, vol. 20, pp. 353–361, Nov. 1980.

[92] M. Hermle, G. Létay, S. P. Philipps, and A. W. Bett, "Numerical simulation of tunnel diodes for multi-junction solar cells," *Prog. Photovoltaics*, vol. 16, pp. 409–418, Aug. 2008.

[93] C. B. Duke, *Tunneling in Solids*. No. 10 in Solid State Physics Supplement, New York: Academic Press, 1969.

[94] E. Burstein and S. Lundqvist, eds., *Tunneling Phenomena in Solids*. New York: Plenum Press, 1969.

[95] E. O. Kane, "Zener tunneling in semiconductors," *J. Phys. Chem. Solids*, vol. 12, pp. 181–188, Jan. 1960.

[96] E. O. Kane, "Theory of Tunneling," *J. Appl. Phys.*, vol. 32, no. 1, p. 83, 1961.

[97] A. Di Carlo, P. Vogl, and W. Potz, "Theory of Zener tunneling and Wannier-Stark states in semiconductors," *Phys. Rev. B*, vol. 50, pp. 8358–8377, Sept. 1994.

[98] O. Morandi, "Multiband Wigner-function formalism applied to the Zener band transition in a semiconductor," *Phys. Rev. B*, vol. 80, p. 024301, July 2009.

[99] K. Majumdar, "Band to band tunneling in III-V semiconductors: Implications of complex band structure, strain, orientation, and off-zone center contribution," *J. Appl. Phys.*, vol. 115, p. 174503, May 2014.

[100] C. Benz, M. Claassen, and D. Liebig, "Tunneling and impact ionization at high electric fields in abrupt GaAs p-i-n structures," *J. Appl. Phys.*, vol. 81, no. 7, p. 3181, 1997.

[101] Q. Smets, D. Verreck, A. S. Verhulst, R. Rooyackers, C. Merckling, M. Van De Put, E. Simoen, W. Vandervorst, N. Collaert, V. Y. Thean, B. Sorée, G. Groeseneken, and M. M. Heyns, "InGaAs tunnel diodes for the calibration of semi-classical and quantum mechanical band-to-band tunneling models," *J. Appl. Phys.*, vol. 115, p. 184503, May 2014.

[102] W. Vandenberghe, B. Sorée, W. Magnus, and G. Groeseneken, "Zener tunneling in semiconductors under nonuniform electric fields," *J. Appl. Phys.*, vol. 107, no. 5, p. 054520, 2010.

[103] A. Schenk, R. Rhyner, M. Luisier, and C. Bessire, "Analysis of Si, InAs, and Si-InAs tunnel diodes and tunnel FETs using different transport models," in *SISPAD 2011*, pp. 263–266, IEEE, Sept. 2011.

[104] K. Ganapathi and S. Salahuddin, "Zener tunneling: Congruence between semi-classical and quantum ballistic formalisms," *J. Appl. Phys.*, vol. 111, no. 12, p. 124506, 2012.

[105] T. Krishnamohan, Z. Krivokapic, K. Uchida, Y. Nishi, and K. Saraswat, "High-mobility ultrathin strained Ge MOSFETs on bulk and SOI with low band-to-band tunneling leakage: experiments," *IEEE Trans. Electron Devices*, vol. 53, pp. 990–999, May 2006.

[106] D. Jena, "Tunneling Transistors Based on Graphene and 2-D Crystals," *Proc. IEEE*, vol. 101, pp. 1585–1602, June 2013.

[107] D. Sarkar, M. Krall, and K. Banerjee, "Electron-hole duality during band-to-band tunneling process in graphene-nanoribbon tunnel-field-effect-transistors," *Appl. Phys. Lett.*, vol. 97, no. 26, p. 263109, 2010.

[108] D. Fredkin and G. Wannier, "Theory of Electron Tunneling in Semiconductor Junctions," *Phys. Rev.*, vol. 128, pp. 2054–2061, Dec. 1962.

[109] L. Kleinman, "Theory of Phonon-Assisted Tunneling in Semiconductors," *Phys. Rev.*, vol. 140, pp. A637–A648, Oct. 1965.

[110] R. Shuey, "Theory of Tunneling Across Semiconductor Junctions," *Phys. Rev.*, vol. 137, pp. A1268–A1277, Feb. 1965.

[111] J. B. Krieger, "Theory of electron tunneling in semiconductors with degenerate band structure," *Ann. Phys.*, vol. 36, pp. 1–60, Jan. 1966.

[112] J. Conley and G. Mahan, "Tunneling Spectroscopy in GaAs," *Phys. Rev.*, vol. 161, pp. 681–695, Sept. 1967.

[113] S. Agarwal and E. Yablonovitch, "Using dimensionality to achieve a sharp tunneling FET (TFET) turn-on," in *Proc. 69th IEEE Device Res. Conf.*, pp. 199–200, IEEE, June 2011.

[114] *Sentaurus Device User Guide*. Synopsys, Inc., 2012.

[115] G. Hurkx, "On the modelling of tunnelling currents in reverse-biased p-n junctions," *Solid-State Electron.*, vol. 32, pp. 665–668, Aug. 1989.

[116] C.-H. Shih and N. Dang Chien, "Physical properties and analytical models of band-to-band tunneling in low-bandgap semiconductors," *J. Appl. Phys.*, vol. 115, p. 044501, Jan. 2014.

[117] A. Zunger, "On the Farsightedness (hyperopia) of the Standard k - p Model," *phys. stat. sol. (a)*, vol. 190, pp. 467–475, Apr. 2002.

[118] X. Guan, D. Kim, K. C. Saraswat, and H.-S. P. Wong, "Complex Band Structures: From Parabolic to Elliptic Approximation," *IEEE Electron Device Lett.*, vol. 32, pp. 1296–1298, Sept. 2011.

[119] S. S. Sylvia, H.-H. Park, M. A. Khayer, K. Alam, G. Klimeck, and R. K. Lake, "Material Selection for Minimizing Direct Tunneling in Nanowire Transistors," *IEEE Trans. Electron Devices*, vol. 59, pp. 2064–2069, Aug. 2012.

[120] D. Jena, T. Fang, Q. Zhang, and H. Xing, "Zener tunneling in semiconducting nanotube and graphene nanoribbon p-n junctions," *Appl. Phys. Lett.*, vol. 93, no. 11, p. 112106, 2008.

[121] N. Ma and D. Jena, "Interband tunneling in two-dimensional crystal semiconductors," *Appl. Phys. Lett.*, vol. 102, no. 13, p. 132102, 2013.

[122] V. A. Altschul, A. Fraenkel, and E. Finkman, "Effects of band nonparabolicity on two-dimensional electron gas," *J. Appl. Phys.*, vol. 71, no. 9, p. 4382, 1992.

[123] S. Agarwal, J. T. Teherani, J. L. Hoyt, D. A. Antoniadis, and E. Yablonovitch, "Engineering the Electron-Hole Bilayer Tunneling Field-Effect Transistor," *IEEE Trans. Electron Devices*, vol. 61, pp. 1599–1606, May 2014.

[124] Y. Lu, A. Seabaugh, P. Fay, S. J. Koester, S. E. Laux, W. Haensch, and S. O. Koswatta, "Geometry dependent tunnel FET performance - dilemma of electrostatics vs. quantum confinement," in *Proc. 68th IEEE Device Res. Conf.*, pp. 17–18, June 2010.

[125] M. J. Lee and W. Y. Choi, "Analytical model of single-gate silicon-on-insulator (SOI) tunneling field-effect transistors (TFETs)," *Solid-State Electron.*, vol. 63, pp. 110–114, Sept. 2011.

[126] L. Liu, D. Mohata, and S. Datta, "Scaling Length Theory of Double-Gate Interband Tunnel Field-Effect Transistors," *IEEE Trans. Electron Devices*, vol. 59, pp. 902–908, Apr. 2012.

[127] A. S. Verhulst, B. Sorée, D. Leonelli, W. G. Vandenberghe, and G. Groeseneken, "Modeling the single-gate, double-gate, and gate-all-around tunnel field-effect transistor," *J. Appl. Phys.*, vol. 107, no. 2, p. 024518, 2010.

[128] K.-H. Kao, A. S. Verhulst, W. G. Vandenberghe, B. Sorée, G. Groeseneken, and K. D. Meyer, "Modeling the impact of junction angles in tunnel field-effect transistors," *Solid-State Electron.*, vol. 69, pp. 31–37, Mar. 2012.

[129] P. M. Solomon, D. J. Frank, and S. Koswatta, "Compact model and performance estimation for tunneling nanowire FET," in *Proc. 69th IEEE Device Res. Conf.*, pp. 197–198, June 2011.

[130] J. Wan, C. Le Royer, A. Zaslavsky, and S. Cristoloveanu, "A tunneling field effect transistor model combining interband tunneling with channel transport," *J. Appl. Phys.*, vol. 110, no. 10, p. 104503, 2011.

[131] L. Zhang, X. Lin, J. He, and M. Chan, "An Analytical Charge Model for Double-Gate Tunnel FETs," *IEEE Trans. Electron Devices*, vol. 59, pp. 3217–3223, Dec. 2012.

[132] C. Shen, S.-L. Ong, C.-H. Heng, G. Samudra, and Y.-C. Yeo, "A Variational Approach to the Two-Dimensional Nonlinear Poisson's Equation for the Modeling of Tunneling Transistors," *IEEE Electron Device Lett.*, vol. 29, pp. 1252–1255, Nov. 2008.

[133] Q. Xie, J. Xu, and Y. Taur, "Review and Critique of Analytic Models of MOSFET Short-Channel Effects in Subthreshold," *IEEE Trans. Electron Devices*, vol. 59, pp. 1569–1579, June 2012.

[134] K. Young, "Short-channel effect in fully depleted SOI MOSFETs," *IEEE Trans. Electron Devices*, vol. 36, pp. 399–402, Feb. 1989.

[135] K. Suzuki, T. Tanaka, Y. Tosaka, H. Horie, and Y. Arimoto, "Scaling theory for double-gate SOI MOSFET's," *IEEE Trans. Electron Devices*, vol. 40, pp. 2326–2329, Dec. 1993.

[136] C. Auth and J. Plummer, "Scaling theory for cylindrical, fully-depleted, surrounding-gate MOSFET's," *IEEE Electron Device Lett.*, vol. 18, pp. 74–76, Feb. 1997.

[137] M. G. Bardon, H. P. Neves, R. Puers, and C. Van Hoof, "Pseudo-Two-Dimensional Model for Double-Gate Tunnel FETs Considering the Junctions Depletion Regions," *IEEE Trans. Electron Devices*, vol. 57, pp. 827–834, Apr. 2010.

[138] J. Blakemore, "Approximations for Fermi-Dirac integrals, especially the function used to describe electron density in a semiconductor," *Solid-State Electron.*, vol. 25, pp. 1067–1076, Nov. 1982.

[139] ITRS, "ITRS Roadmap," 2012.

[140] R. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals*. New York: Dover Publications, 1965.

[141] W. G. Vandenberghe, A. S. Verhulst, K.-H. Kao, K. D. Meyer, B. Sorée, W. Magnus, and G. Groeseneken, "A model determining optimal doping concentration and material's band gap of tunnel field-effect transistors," *Appl. Phys. Lett.*, vol. 100, no. 19, p. 193509, 2012.

[142] S. S. Sylvia, M. A. Khayer, K. Alam, and R. K. Lake, "Doping, Tunnel Barriers, and Cold Carriers in InAs and InSb Nanowire Tunnel Transistors," *IEEE Trans. Electron Devices*, vol. 59, pp. 2996–3001, Nov. 2012.

[143] J. M. Ziman, *Principles of the Theory of Solids*. Cambridge: Cambridge University Press, 2nd ed., 1972.

[144] P. M. Solomon, "Inability of Single Carrier Tunneling Barriers to Give Subthermal Subthreshold Swings in MOSFETs," *IEEE Electron Device Lett.*, vol. 31, pp. 618–620, June 2010.

[145] D. R. Khanal, J. W. L. Yim, W. Walukiewicz, and J. Wu, "Effects of Quantum Confinement on the Doping Limit of Semiconductor Nanowires," *Nano Lett.*, vol. 7, pp. 1186–1190, May 2007.

[146] A. Biswas, S. S. Dan, C. L. Royer, W. Grabinski, and A. M. Ionescu, "TCAD simulation of SOI TFETs and calibration of non-local band-to-band tunneling model," *Microelec. Eng.*, vol. 98, pp. 334–337, Oct. 2012.

[147] H. Lu and A. Seabaugh, "Tunnel Field-Effect Transistors: State-of-the-Art," *IEEE J. Electron Devices Soc.*, vol. 2, pp. 44–49, July 2014.

[148] K. S. Jones, A. G. Lind, C. Hatem, S. Moffatt, and M. C. Ridgeway, "A Brief Review of Doping Issues in III-V Semiconductors," *ECS Trans.*, vol. 53, pp. 97–105, May 2013.

[149] J. C. Ho, A. C. Ford, Y.-L. Chueh, P. W. Leu, O. Ergen, K. Takei, G. Smith, P. Majhi, J. Bennett, and A. Javey, "Nanoscale doping of InAs via sulfur monolayers," *Appl. Phys. Lett.*, vol. 95, no. 7, p. 072108, 2009.

[150] A. S. Verhulst, D. Verreck, M. A. Pourghaderi, M. Van de Put, B. Sorée, G. Groeseneken, N. Collaert, and A. V.-Y. Thean, "Can p-channel tunnel field-effect transistors perform as good as n-channel?," *Appl. Phys. Lett.*, vol. 105, p. 043103, July 2014.

[151] H.-Y. Chang, S. Chopra, B. Adams, J. Li, S. Sharma, Y. Kim, S. Moffatt, and J. C. Woo, "Improved subthreshold characteristics in tunnel field-effect transistors using shallow junction technologies," *Solid-State Electron.*, vol. 80, pp. 59–62, Feb. 2013.

[152] A. Villalon, C. Le Royer, M. Casse, D. Cooper, J.-M. Hartmann, F. Allain, C. Tabone, F. Andrieu, and S. Cristoloveanu, "Experimental Investigation of the Tunneling Injection Boosters for Enhanced I_{ON} ETSOI Tunnel FET," *IEEE Trans. Electron Devices*, vol. 60, pp. 4079–4084, Dec. 2013.

[153] G. Leung and C. O. Chui, "Stochastic Variability in Silicon Double-Gate Lateral Tunnel Field-Effect Transistors," *IEEE Trans. Electron Devices*, vol. 60, pp. 84–91, Jan. 2013.

[154] N. Damrongplasit, S. H. Kim, and T.-J. K. Liu, "Study of Random Dopant Fluctuation Induced Variability in the Raised-Ge-Source TFET," *IEEE Electron Device Lett.*, vol. 34, pp. 184–186, Feb. 2013.

[155] E. O. Kane, "Band tails in semiconductors," *Solid-State Electron.*, vol. 28, pp. 3–10, Jan. 1985.

[156] M. A. Khayer and R. K. Lake, "Effects of band-tails on the subthreshold characteristics of nanowire band-to-band tunneling transistors," *J. Appl. Phys.*, vol. 110, no. 7, p. 074508, 2011.

[157] S. Agarwal and E. Yablonovitch, "Band-Edge Steepness Obtained From Esaki/Backward Diode Current-Voltage Characteristics," *IEEE Trans. Electron Devices*, vol. 61, pp. 1488–1493, May 2014.

[158] D. Verreck, A. S. Verhulst, B. Sorée, N. Collaert, A. Mocuta, A. Thean, and G. Groeseneken, "Improved source design for p-type tunnel field-effect transistors: Towards truly complementary logic," *Appl. Phys. Lett.*, vol. 105, p. 243506, Dec. 2014.

[159] Q. Zhang, Y. Lu, C. A. Richter, D. Jena, and A. Seabaugh, "Optimum Bandgap and Supply Voltage in Tunnel FETs," *IEEE Trans. Electron Devices*, vol. 61, pp. 2719–2724, Aug. 2014.

[160] S. Saurabh and M. J. Kumar, "Novel Attributes of a Dual Material Gate Nanoscale Tunnel Field-Effect Transistor," *IEEE Trans. Electron Devices*, vol. 58, pp. 404–410, Feb. 2011.

[161] H. Wang, S. Chang, Y. Hu, H. He, J. He, Q. Huang, F. He, and G. Wang, "A Novel Barrier Controlled Tunnel FET," *IEEE Electron Device Lett.*, vol. 35, pp. 798–800, July 2014.

[162] M. J. Kumar and S. Janardhanan, "Doping-Less Tunnel Field Effect Transistor: Design and Investigation," *IEEE Trans. Electron Devices*, vol. 60, pp. 3285–3290, Oct. 2013.

[163] L. Lattanzio, L. De Michielis, and A. M. Ionescu, "Complementary Germanium Electron-Hole Bilayer Tunnel FET for Sub-0.5-V Operation," *IEEE Electron Device Lett.*, vol. 33, pp. 167–169, Feb. 2012.

[164] E. Yablonovitch and E. O. Kane, "Correction to "Reduction of lasing threshold current density by the lowering of valence band effective mass"," *J. Lightw. Technol.*, vol. 4, p. 961, July 1986.

[165] Y.-C. Yeo, T.-J. King, and C. Hu, "Metal-dielectric band alignment and its implications for metal gate complementary metal-oxide-semiconductor technology," *J. Appl. Phys.*, vol. 92, no. 12, p. 7266, 2002.

[166] H. Kim, P. C. McIntyre, C. O. Chui, K. C. Saraswat, and S. Stemmer, "Engineering chemically abrupt high-k metal oxide/silicon interfaces using an oxygen-gettering metal overlayer," *J. Appl. Phys.*, vol. 96, no. 6, p. 3467, 2004.

[167] C.-H. Lu, G. Wong, M. Deal, W. Tsai, P. Majhi, C. O. Chui, M. Visokay, J. Chambers, L. Colombo, B. Clemens, and Y. Nishi, "Characteristics and mechanism of tunable work function gate electrodes using a bilayer metal structure on SiO2 and HfO2," *IEEE Electron Device Lett.*, vol. 26, pp. 445–447, July 2005.

[168] S. Park, L. Colombo, Y. Nishi, and K. Cho, "Ab initio study of metal gate electrode work function," *Appl. Phys. Lett.*, vol. 86, no. 7, p. 073118, 2005.

[169] K. Ganapathi, Y. Yoon, and S. Salahuddin, "Analysis of InAs vertical and lateral band-to-band tunneling transistors: Leveraging vertical tunneling for improved performance," *Appl. Phys. Lett.*, vol. 97, no. 3, p. 033504, 2010.

[170] E. Baravelli, E. Gnani, R. Grassi, A. Gnudi, S. Reggiani, and G. Baccarani, "Optimization of n- and p-type TFETs Integrated on the Same InAs/Al{x}Ga_{1-x}Sb Technology Platform," *IEEE Trans. Electron Devices*, vol. 61, pp. 178–185, Jan. 2014.

[171] G. Fiori and G. Iannaccone, "Three-Dimensional Simulation of One-Dimensional Transport in Silicon Nanowire Transistors," *IEEE Trans. Nanotechnol.*, vol. 6, pp. 524–529, Sept. 2007.

[172] N. Seoane and A. Martinez, "A detailed coupled-mode-space non-equilibrium Green's function simulation study of source-to-drain tunnelling in gate-all-around Si nanowire metal oxide semiconductor field effect transistors," *J. Appl. Phys.*, vol. 114, no. 10, p. 104307, 2013.

[173] M. Salmani-Jelodar, S. Kim, K. Ng, and G. Klimeck, "Transistor roadmap projection using predictive full-band atomistic modeling," *Appl. Phys. Lett.*, vol. 105, p. 083508, Aug. 2014.

[174] K. D. Cantley, Y. Liu, H. S. Pal, T. Low, S. S. Ahmed, and M. S. Lundstrom, "Performance Analysis of III-V Materials in a Double-Gate nano-MOSFET," in *Proc. IEDM 2007*, pp. 113–116, IEEE, 2007.

[175] M. Luisier, M. Lundstrom, D. A. Antoniadis, and J. Bokor, "Ultimate device scaling: Intrinsic performance comparisons of carbon-based, InGaAs, and Si field-effect transistors for 5 nm gate length," in *Proc. IEDM 2011*, pp. 11.2.1–11.2.4, IEEE, Dec. 2011.

[176] S. Koba, M. Ohmori, Y. Maegawa, H. Tsuchiya, Y. Kamakura, N. Mori, and M. Ogawa, "Channel length scaling limits of III–V channel MOSFETs governed by source–drain direct tunneling," *Jpn. J. Appl. Phys.*, vol. 53, p. 04EC10, Apr. 2014.

[177] D. Basu, R. Kotlyar, C. E. Weber, and M. A. Stettler, "Ballistic Band-to-Band Tunneling in the OFF State in InGaAs MOSFETs," *IEEE Trans. Electron Devices*, vol. 61, pp. 3417–3422, Oct. 2014.

[178] R. Kim, U. E. Avci, and I. A. Young, "Source/Drain Doping Effects and Performance Analysis of Ballistic III-V n-MOSFETs," *IEEE J. Electron Devices Soc.*, vol. 3, pp. 37–43, Jan. 2015.

[179] Z. Jiang, B. Behin-Aein, Z. Krivokapic, M. Povolotskyi, and G. Klimeck, "Tunneling and Short Channel Effects in Ultrascaled InGaAs Double Gate MOSFETs," *IEEE Trans. Electron Devices*, vol. 62, pp. 525–531, Feb. 2015.

[180] A. Szabo and M. Luisier, "Under-the-Barrier Model: An Extension of the Top-of-the-Barrier Model to Efficiently and Accurately Simulate Ultrascaled Nanowire Transistors," *IEEE Trans. Electron Devices*, vol. 60, pp. 2353–2360, July 2013.

[181] M. G. Ancona, "Density-gradient theory: a macroscopic approach to quantum confinement and tunneling in semiconductor devices," *J. Comput. Electron.*, vol. 10, pp. 65–97, June 2011.

[182] A. Khakifirooz, O. M. Nayfeh, and D. Antoniadis, "A Simple Semiempirical Short-Channel MOSFET Current-Voltage Model Continuous Across All Regions of Operation and Employing Only Physical Parameters," *IEEE Trans. Electron Devices*, vol. 56, pp. 1674–1680, Aug. 2009.

[183] M. V. Fischetti, S. Jin, T.-W. Tang, P. Asbeck, Y. Taur, S. E. Laux, M. Rodwell, and N. Sano, "Scaling MOSFETs to 10 nm: Coulomb effects, source starvation, and virtual source model," *J. Comput. Electron.*, vol. 8, pp. 60–77, June 2009.

[184] P. Landsberg, *Recombination in Semiconductors*. Cambridge: Cambridge University Press, 1991.