Tight-Binding Models, Their Applications to Device Modeling, and Deployment to a Global Community

Gerhard Klimeck
Network for Computational Nanotechnology, School of Electrical and Copmputer Engineering,
Purdue University, West Lafayette, IN 47906

Timothy B. Boykin
Department of Electrical and Computer Engineering,
The University of Alabama in Huntsville, Huntsville, AL 35899

Abstract—Tight-binding has become the state-of-the-art for nano-scale device modeling. It has been adopted by many device modeling research groups, industry and commercial vendors. This impact on the field is due to the virtually ideal combination of inclusion of critical material and device description, numerical efficiency, and simple integration into quantum transport approaches. We start by setting forth the modeling requirements for realistically extended devices which include a full quantum mechanical treatment, atomistic interface treatments, atomistic representations of crystal symmetries, polarization, strain, and bond directions, and embedding into macroscopic fields such electromagnetic potentials and long-range strain. We describe the essential definitions followed by numerical issues in terms of transfer matrices, Green's functions, and parallel scaling. The second half of this text is dedicated to applications in realistically large devices. We highlight million-atom quantum dot simulations and focus on carrier transport though silicon nanowires. We demonstrate how effective masses and bandgaps become design parameters at the nanoscale, how heavy masses are desirable for end-of-roadmap transistors, and how coherent transport assumptions break down. The nanowire simulations can be duplicated by everyone on nanoHUB.org

1. INTRODUCTION

Tight-binding has become the state-of-the-art for nano-scale device modeling. It has been implemented by multiple advanced device modeling research groups in conjunction with multiple quantum transport methodologies. Industry has begun to adopt some of the advanced codes and is beginning to implement company internal versions. Commercial software vendors are adopting the methodologies and are beginning to deploy the approaches into their commercial TCAD products. Intense software development combining tight-binding with the non-equilibrium Green Function (NEGF) approach and Quantum Transmitting Boundary Conditions (QTBM) for quantum transport began in 1994 at Texas Instruments. Acceptance of NEGF and tight-binding began with wider adoption in about 2004. The past 25 years have resulted in many model advancements, numerical technology development, and physics exploration that is by far too large to be covered here comprehensively. We start by setting forth the requirements for realistic modeling of extended nano-scale device (Sections 1.1-1.2) which include a full quantum mechanical treatment, atomistic interface treatments, atomistic representations of crystal symmetries, polarization, strain, and bond directions, and embedding into macroscopic fields such electromagnetic potentials and long-range strain. We then proceed to describe the essential definitions and features for empirical tight-binding (Section 1.3). Sections 1.4 to 1.6 address numerical issues of tight-binding in terms of transfer matrices, Green's functions, and parallel scaling. Section 1.7 is dedicated to several applications around Quantum Dots and nanowires. We highlight million-atom quantum dot simulations and focus on carrier transport though silicon nanowires. We demonstrate how effective masses and bandgaps become design parameters at the nanoscale, how heavy masses are desirable for end-of-roadmap transistors, and how coherent transport assumptions break down. The nanowire simulations can be duplicated by everyone on nanoHUB.org. Section 1.8 highlights the widespread use of tight-binding within nanoHUB applications and Sec. 1.9 concludes.

1.1. NANODEVICE CHARACTERISTICS

Nanodevices by definition have extremely small active regions, so small that the number of atoms is as a practical matter countable. The small size has consequences for nanodevice physics, so that these devices operate far differently from traditional semiconductor devices. The most important aspects of nanodevice physics which follow from their small size are: (1) length scales smaller than the electron wavelength; (2) a much higher surface-area-to-volume ratio, necessitating detailed surface modeling; (3) the importance of crystal orientation, rendering continuum models as invalid; (4) the interaction of long-range Coulomb potentials and strain fields with short-range quantum confinement. These physical consequences of the exceptionally small nanodevice size determine the requirements for modeling nanodevices.

Prototypical nanodevices, listed roughly in order of historical technical developments, are the resonant tunneling diode (RTD), quantum dots, and modern nanoFETs. The RTD consists of 1D vertically stacked, nano-scaled layers and can be considered the first room temperature quantum device geared for electronic circuits. The RTD has, however, never found widespread technical adoption due to a variety of technical issues. While no longer really technically relevant, RTDs offer a wonderful utility to study and understand the intricacies of quantum transport in simple 1D geometries that become even more complicated in 3D. Quantum dots confine electrons in 3D to nano-scale dimensions, are sometimes called artificial atoms, and they have found utility in optical applications. Because of the three-dimensional confinement, they offer an extraordinarily clear view of the interplay of geometry and bandstructure effects on optical interactions in these artificial atoms. Nano-scaled field effect transistors (nanoFETs) are ubiquitous and literally in everyone's pocket inside modern smart phones, and certainly in our laptops and notebooks. Understanding and optimizing carrier transport in such highly sophisticated 3D geometries is complex and still undergoing extensive research and development. We next briefly summarize the dimensions of these three prototypical devices.

The central part of a typical RTD heterostructure consists of two 3-5nm thick barriers surrounding a 3-5nm thick quantum well. The material system is typically GaAs/AlGaAs or InGaAs/InAlAs for the well/barrier. The central RTD is typically undoped and surrounded by 10-30nm of undoped spacer layers to reduce ionized dopant scattering near the central quantum device. This undoped region is surrounded by heavily, degenerately doped regions, so that under forward (reverse) bias electron reservoirs form adjacent to the emitter (collector) barriers. The long, heavily-doped regions which eventually terminate in quasi-ohmic metal contacts. The overall semiconductor device that needs to be modeled for quantitative agreement with experiment is typically 100-200nm long in the growth direction. This length represents a surprisingly small atomic count: 8 atoms in one dimension occupy a roughly 1nm length. Thus a 128 nm long device corresponds to a one-dimensional string of around 1000 atoms. For an RTD, the most critical consequences of small size manifest as:

1) Electron confinement to a 5nm quantum well is clearly significantly smaller than the phase coherence length of about 50 nm for typical GaAs or InGaAs at room temperature [1].
2) The cross section of the RTD is typically defined by mesas that are of the order of 1-100 μm and therefore effectively infinite. The mesa surface is designed to play no role. The heterostructure, however needs to be treated in atomistic detail and the surfaces as such define the device.
3) Typical RTDs have been grown in III-V material systems in the [100] direction. Substrate orientation has not played a very critical role in these typical electron devices, but does play a role in understanding hole transport, due to anisotropy of the valence bands.

The long-range effects of the electrostatic potentials in the semiconductor do, however, play a critical role, and a multi-scale approach to the modeling of the central 15nm long RTD connected to 100nm long semiconductor regions is critical.

In self-assembled Quantum Dots (QDs), the dot materials provide the confinement. These QDs are typically of the size of 10-30nm in the lateral dimension and 5-10nm height. These dots are typically embedded in a buffer material which confines the electrons to the quantum dots and

passivates any open surfaces. A typical QD/Buffer/Substrate material system is InGaAs/InAlAs/GaAs. The typically used Stranski-Krastanov growth mode [2] ensures that the overall structure is virtually defect free, even though complex 3D geometries in the form of (capped) pyramids or domes are formed. QDs can also be realized via electrostatic confinement. This class of QD typically consists of a 1D heterostructure that confines electrons in 1 dimension and electrostatic 2D gates that confine electrons in the lateral dimensions. This 2D lateral confinement is limited in its down-scaling by lithography and processing technology and is typically larger than the self-assembled quantum dots. Typical material systems are the III-V and column IV semiconductor families. Yet another structurally different quantum dot is one formed by an individual impurity such as P embedded in Si.

Typical self-assembled QDs have 50,000-300,000 atoms. For example, typical dome shaped QDs have dimensions of $5\times30\times30nm^3$ corresponding to roughly 200,000. The electron and hole wavefunction and the eigenstates are not completely confined to just the central QD but leak out, or "feel" the buffer layers. The typical electronic structure domain size is around 1-4 million atoms and the strain domain must yet be much larger as discussed further below. Electron states in a single P impurity have a very shallow binding energy below the conduction band and the wavefunction spreads out to a space of $30x30x30nm^3$ corresponding to about 1.7 million atoms. These total atom numbers start to be large, but are still countable. Typically, these geometries have one dimension that is significantly smaller than the other two – generally the growth direction and electron are confined to less than 5nm, making the geometry look similar to an RTD in that aspect. The physical consequences of the small QD size are:

1) Electron confinement of 5-30nm is even at room temperature smaller than typical decoherence lengths. Modeling therefore requires a quantum mechanical treatment.
2) QDs are strongly influenced by the details of the heterostructures that confine them, and have high surface-area-to-volume ratios. The treatment of the interfaces is thus critical.
3) Crystal symmetry plays a critical role in any of the three described quantum dots in terms of the intricate details of optical polarization effects [3] as well as valley splitting in indirect-gap bandstructure materials such as Si [4].
4) Self-assembled quantum dots form by differences in the lattice constants of the constituent materials and are subject to significant long-range strain. The strain can reach 20-30nm down into the substrate and is must be managed in the design [5] of overlayer growth. Gated quantum dots and also gated impurity dots are embedded in long range fields that extend tens to hundreds of nm. These long-range, classical fields must be coupled to the detailed central device models.

Field effect transistors (FETs) have evolved from planar to full 3D geometries. The core active material is some form of strained Si connected to SiGe contacts with sophisticated 3D gate stacks. A modern finFET looks a little bit like an old RTD (oriented vertically), except that the current flows in the plane/fin. The Si/Ge transistors are strained and subject to large electric fields. Many of the principles of operation of RTDs and QDs translate to modern finFETs. Today's commercially wide-spread "old" nano-FETs dating back to 2011 are 3D geometries that already have channel widths as small as 8 nm or 59 atoms and channel heights that are about 3 or 4 times larger[6-8]. The active device cross section has only around 1,000 atoms. Channel lengths have been decreasing from nominally 22nm in 2012[6,7] to around 7nm in 2016[8]. The centrally active device might be well-described by 100,000 to 1 million atoms. The physical consequences of ultra-scaling finFETs are:

1) Quantum mechanical treatment is required because the confinement within a finFET is clearly smaller than the electron coherence length, as with an RTD.
2) Device performance is determined by interface details of the finFET or nanowire FET geometry in terms of shape and interface roughness; these effects must be included in the device model.
3) Crystal orientation has a profound effect on device performance. It has been a critical transistor device design parameter since the 90nm transistor node [9], where Texas Instruments was able to avoid the very costly SiGe-based stressors employed by Intel[10]. A simple 45 degree rotation of the wafer / device provided the needed performance boost for PMOS transistors. This invention was motivated by observations derived from orientation effects in NEMO-modeled Quantum Dots. The rotated substrate approach is standard technology today, indicating the continued critical importance of crystal orientation.
4) Alternative materials beyond the traditional Si have been introduced into 3D device geometries. Intel introduced SiGe in source & drain for strain-engineering the Si channel in 2002 [10] in their 90nm node. The 2016 7nm channel transistor introduced by IBM, GLOBAL FOUNDRIES, and Samsung features SiGe channels [8]. These new materials must be rapidly included in the modeling suite.

nanoFETs are strained by design [6-8,10] to improve performance and are subject to electrostatic fields to control carrier flow. This classical long-range physics must be coupled to local, relatively small central devices.

1.2. NANODEVICE MODELING APPROACHES

Thus, nanoscale feature size and its consequences are common across a broad range of devices: RTDs, QDs, and nanoFETs. Models for these devices must accurately take into account the quantum mechanical nature of the electrons, surface and substrate effects, and long range Coulomb potentials and strains. *Taken together, these characteristics point to a basis that is fundamentally atomistic, localized, and transferrable.*

This conclusion to use an atomistic basis was, however, not obvious to most researchers involved with quantum device modeling over the past 30 years. The principal focus of the modeling community that studies quantum devices or nanodevices as we call them today had been on the quantum

mechanical aspects of the modeling. This focus was natural, since nanostructure dimensions in the range of 5-30nm are of the length scale of the electron wavelength and significantly smaller than the coherence length. It was clear even then that only quantum mechanical models would work, but quantum mechanical models differ widely in the basis sets used for the electronic states.

At the most basic level, single-(quasi) electron models fall into one of two broad categories: 1) those that explicitly include the potential due to the atoms and 2) continuum models. Continuum models, called envelope-function or effective-mass approaches, deal with electrons (holes) in a single conduction (valence) band. Materials differences are incorporated only in terms of band gaps and effective masses (inverse band curvatures). In these models, carriers in the conduction and valence bands are different quasi-particles. These models treat the crystalline environment as a continuum; the potential due to the atoms does not appear beyond the effective masses and energy gaps. The basis states are plane waves or decaying exponentials. In addition, these models are valid for energies only near the band edges. In a periodic crystal these models do not account for non-parabolic bands (unless they are empirically expanded *ad hoc*) and are useful only for wavevectors near the band extrema. The k.p method is a multi-band continuum approach, and has similar drawbacks (and is traditionally plagued by numerical stability issues).

Beyond their deficiencies for periodic crystals, continuum approaches are completely unsuited for nanodevice modeling in other ways that relate to realistic device representations. They cannot model surface effects and cannot account for differences in substrate orientation, both of which are critically important. In addition, multi-band effects, such as valley splitting in Si, and the Rashba and Dresselhaus effects, can only be included crudely via *ad hoc* specifically empirically fitted parameters. There is no predictive value in models that must use such *ad hoc* parameters. Effects critical to realistic device models such as atomistic interfaces and their roughness, 3D geometry effects that impose non-trivial strain fields, alloy disorder, alloy-grading, and full-band representations can in general not be included in continuum models, unless they are highly calibrated through ad-hoc methods.

While the traditional device modeling community pursues efforts scale their large device models down towards small nano-devices, there is another community, originating in material science, that pushes their models up from a few atoms towards devices. This materials science (MS) approach has some commonality with electrical engineering (EE). At the nanometer scale, the distinction between material and device vanishes and the two approaches meet. MS brings the knowledge of the placement of atoms and the electronic ground state to the common problem. EE brings approaches to quantum transport under highly non-equilibrium conditions with the need for atomistic material resolution.

2.

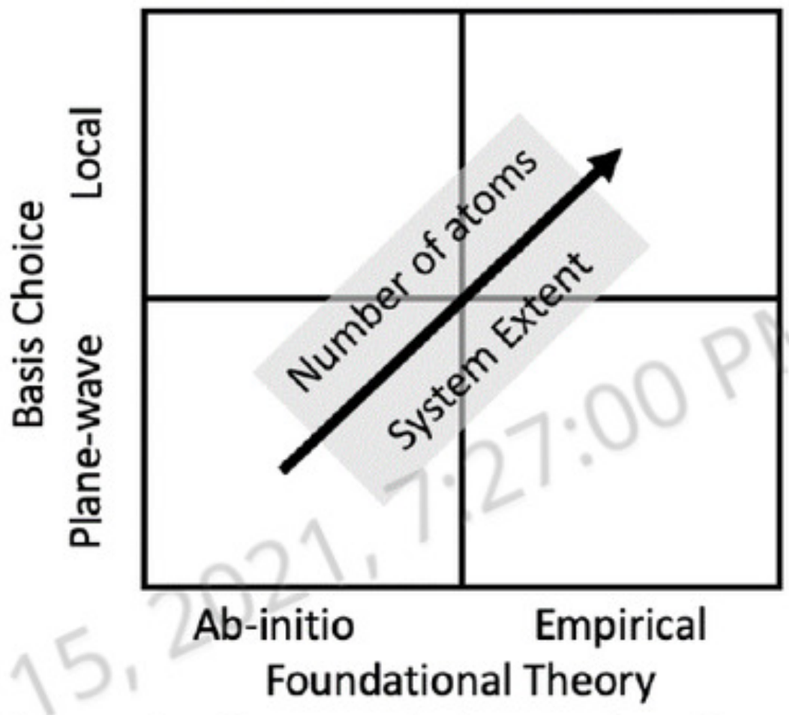

Figure 1: Conceptual layout for the choice of foundational electronic structure theories and choice of basis sets. Empirical theories with localized basis sets can scale to larger and realistically extended system sizes.

Myriad atomistic structure methods can be found in the area of material science [11]. The methods can be roughly distinguished into a 2×2 table, with one axis representing basis sets (plane waves vs. local orbitals) [12,13] and the other axis fundamental theory (*ab-initio* vs. empirical) [14,15,] (Figure 1). Plane waves are appealing since they offer a quantifiable path to improve accuracy by adding more waves and are very efficient for modeling perfectly periodic crystals. However, plane waves are not well suited for modeling typical nanodevices, which are finite, with electrical contacts, and not periodic. Therefore, local orbitals are preferable, and computationally less expensive, for nanodevice modeling. Fundamental many-electron correlated methods based on perturbation theory [16], quantum Monte Carlo method [17], or GW approach [18] offer intellectual appeal, but can only predict electronic structure for very small systems (around 100 atoms). Empirical methods lose some fundamental claim to correctness but do much better in terms of incorporating correct bulk bandgaps and masses via their Hamiltonian parameters and can be scaled up to larger system sizes. Almost all ab-initio and quantum chemistry codes treat closed systems close to or at equilibrium; that excludes far-from-equilibrium irreversible electron transport needed to solve our problem. Beyond these considerations, the standards for sufficient accuracy in bulk bandstructure calculations are very different in MS and EE.

Accuracy of bandgaps and masses must be significantly greater in EE than in MS. In MS ground state properties are usually the focus, and gaps off by 200 meV and masses within a factor of 2 are sufficient. In contrast, the interest in EE is on the upper valence bands and the conduction bands; gaps and band offsets need to be within about 10 meV and masses within 5%.

Forming and breaking bonds vs. stable systems: MS is primarily concerned with the formation and establishment of the material systems and their equilibrium conditions. Realistic semiconductor devices must have stable bonds and transport characteristics are determined by the valence electrons only. Therefore, full ab-initio methods are not necessary to simulate current/spin flow. Approximate methods that resolve the physics of the valence electrons with

stable bonds are appropriate (and computationally feasible) for large systems.

Empirical Orthogonal Tight-binding (ETB) (sp^3s^* and $sp^3d^5s^*$) is one of the oldest and conceptually simplest atomistic material descriptions available. The 'empirical' refers to fitting the Hamiltonian matrix elements to bulk bandstructures. ETB has proven to be scalable to very large system sizes and valid for standard semiconductors electronic structure and transport simulations under high bias current conditions [19-32]. ETB lacks an explicit basis and projection methods are needed to compute coulomb matrix elements [33]. ETB treats the atomic positions as inputs. NEMO3D bases the positioning of the atoms on a Valence Force Field (VFF) [24,25,34].

1.3. Empirical Tight-Binding Method

1.3.1. Bases for the Hamiltonian

Tight-binding methods (or more specifically semi-empirical tight-binding methods) are ideally suited for nanodevice simulations. With a relatively small basis of usually no more than 20 orbitals per atom (when spin-orbit interaction is included) the bands of the device materials in the relevant energy range can be accurately reproduced throughout the Brillouin zone. Surface and interface parameters can be fit to special structures. The limited range of interactions, especially for nearest-neighbor models, allows for highly efficient device electronic structure and transport calculations. These characteristics are all highly desirable for nanodevice modeling.

The modern use of tight-binding for crystalline electronic structure calculations begins with the work of Slater and Koster[35]. The tight-binding method traditionally uses an orthonormal basis of atomic-like orbitals centered on the atoms of the crystal with the orbitals having a limited range of interaction. Most applications are semi-empirical tight-binding, in which the orbitals do not have explicit spatial representations and are instead defined through their Hamiltonian parameters[35], although explicit, non-orthogonal, basis sets can be used, such as with the closely related extended Hückel theory[36]. In orthogonal tight-binding, there is no unique method for generating orthonormal atomic-like orbitals from true atomic orbitals, but the Löwdin orthogonalization[37] is one of the most useful. A very significant advantage of this method is that the resulting orthonormal orbitals, while no longer possessing full atomic symmetry, nevertheless transform as would their atomic antecedents under the symmetry operations of the crystal and remain generally localized. For semi-empirical tight-binding models, the mere existence of such a basis is sufficient for most device modeling applications, since explicit wavefunctions are not needed most of the time.

Tight-binding bases are specified by their orbital compositions and the range of interaction. Nearest-neighbor models are best suited for nanostructures and nanodevices, which generally have a high proportion of surfaces or interfaces. In nanodevices it is often preferable to use a nearest-neighbor model with a larger basis versus a smaller basis of more extended orbitals. A limited range of interactions also results in a more computationally efficient calculation. The basis set chosen must accurately reproduce the energy gaps and effective masses of all bands in the relevant energy range for the bulk materials. The earliest models for semiconductors employed a sp^3 basis (one s- and three p-orbitals per atom, doubled when the spin-orbit interaction is included), but proved unable to accurately reproduce the conduction bands. Vogl, Hjalmarson, and Dow[38] significantly improved conduction-band reproduction along the [100] directions by adding an excited s-like (s^*) orbital to the basis (sp^3s^*). Still, problems in the conduction band remained. Most problematic was the inability of the nearest-neignbor sp^3s^* model to accurately reproduce the X-valley transverse effective mass. Two remedies for this deficiency were developed: (1) within the sp^3s^* basis, extending the range of interaction to second-nearest neignbors[39]; or (2) retaining nearest-neighbor interactions with a basis including d-orbitals, the $sp3d^5s^*$ basis proposed by Jancu, et. al.[40]. The nearest-neighbor $sp3d^5s^*$ basis is usually preferred for nanodevices due to its minimal range of interaction.

The parameters of tight-binding Hamiltonians are referred to as Slater-Koster[35] parameters and are characterized as two- and three-center integrals. The two-center integrals include only potentials centered the two atoms of the orbitals involved in the matrix element; three-center integrals include the effects of potentials on other atoms as well. Most nanoelectronic device tight-binding models restrict interactions to nearest-neighbors, so that the two-center approximation is justified. An example of one such matrix element in silicon or germanium is that between an s-orbital at the origin and a p_x-orbital at **d**:

$$\langle s;\mathbf{0}|\hat{H}|x;\mathbf{d}\rangle = lV_{sp\sigma} \quad (1)$$

In eq. (1) l denotes the x-direction cosine for the vector **d**. $V_{sp\sigma}$ is the Slater-Koster[35] parameter for this type of matrix element. Cyclic permutations of coordinates and direction cosines give the remaining two matrix elements for p_y- and p_z-orbitals. In compound semiconductors such as GaAs there are separate parameters for the s-orbital on the anion and p_x-orbital on the cation ($V_{sapc\sigma}$) and vice-versa ($V_{scpa\sigma}$). The atomic locations are fixed by the crystal structure and lattice constant; the Slater-Koster parameters are determined via a fitting process.

Tight-binding basis states take different forms depending on the number of confined dimensions; in the non-confined dimensions they obey the appropriate form of Bloch's Theorem. In bulk three-dimensional materials the basis states for the perfect crystal Hamiltonian are:

$$|\omega;l;\mathbf{k}\rangle = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\exp\left[i\mathbf{k}_j\bullet\left(\mathbf{R}_j+\mathbf{d}_l\right)\right]|\omega;\mathbf{R}_j+\mathbf{d}_l\rangle \quad (2)$$

In eq. (2) denotes the orbital type (s-, p_x-, etc., including spin where appropriate), $\boldsymbol{R}_j$ the location of the j-th primitive cell, $\boldsymbol{d}_l$ the location relative to the cell origin of the l-th atom in the primitive cell (for polyatomic primitive cells), $\boldsymbol{k}$ the wavevector (crystal momentum), and N the number of primitive cells. These states obey Bloch's Theorem in three dimensions. Superlattice basis states are formally the same as expressed in Eq. (2), except that $\boldsymbol{R}_j$ denotes the supercell

location, N the number of supercells, and $\boldsymbol{d}_l$ the location relative to the supercell origin of the l-th atom. (For a supercell of N_c primitive cells with N_a atoms per primitive cell, there are $N_c N_a$ different $\boldsymbol{d}_l$.) Basis states for a two-dimensional material such as graphene are constructed analogously, except that the wavevector and primitive cell vectors are two-dimensional. Basis states for nanowires are constructed as are three-dimensional superlattice states, except that there is translational symmetry in only one direction and thus only a one-dimensional wavevector.

Three-dimensional materials with confinement or broken translational symmetry in one or more dimensions require different basis states. In layered structures with only planar translational symmetry, the basis states are Bloch sums in the plane:

$$\left|\omega;l;L;\mathbf{k}_{\|}\right\rangle = \frac{1}{\sqrt{N_{\|}}}\sum_{j=1}^{N_{\|}} \exp\left[i\mathbf{k}_{\|}\bullet\left(\mathbf{R}_{j\|}(L)+\mathbf{d}_{l\|}\right)\right]\left|\omega;L;\mathbf{R}_{j\|}(L)+\mathbf{d}_{l\|}\right\rangle \quad (3)$$

In eq. (3) the subscript '||' denotes a vector in the plane, and L the layer index. A layer is usually defined as a grouping of one or more atomic planes so that there are only nearest-neighbor couplings between layers. Note that the bulk primitive cell origin may be offset depending on the layer index.

1.3.2. Parameter Fitting

The fitting process for three-dimensional materials begins with the perfect crystal Hamiltonian in the tight-binding basis of Eq. (2). Perfect two-dimensional Hamiltonians are constructed analogously. The eigenvalues and eigenvectors of the Hamiltonian depend on the parameters and thus are used to fit them so that the relevant effective masses and energy gaps reproduce those of the bands in the relevant energy range. Until very recently gaps and masses were the only targets used in the fitting process; the values for these targets are taken from *ab initio* calculations or experimental results. At high symmetry points there may be analytic formulas for some gaps and masses in terms of the parameters[39,41-43]. These expressions are most useful for the band fitting process when the dependencies on the parameters is clear, meaning that in practice the Hamiltonian matrix $\underline{\boldsymbol{H}}(\boldsymbol{k})$ should be analytically block-diagonizable, with some of the blocks being of dimension 1 or 2. (Analytic solutions are always possible for up to quartic equations and thus the eigenvalues and eigenvectors of matrices up to 4×4, but the complicated solutions to quartic and cubic equations generally obscure parameter depenedences.) Expressions for inverse effective masses (band curvatures) are possible once the Hamiltonian eigenvalues and eigenvectors are found at the **k**-point of interest. The expression for the inverse effective mass is different from the standard formula found in solid-state textbooks due to the incompleteness of the basis set[44,45]. The issue of incompleteness and its consequences are discussed in Secs. 1.3.4-1.3.7 below.

Although analytic gap and mass formulas help in the fitting process, in practice iterative computer algorithms are almost always used. Computational optimization is needed when fitting mass and gap targets not at high-symmetry points. Klimeck, et. al.[46] introduced Genetic algorithms for global optimization in a very high dimensional space to the tight-binding parameter fitting problem. This genetic algorithm employs user-assigned weights for targets and from them computes a fitness function in each step. Then random mutations of the current parameter set are generated and the one with the best overall fitness is preserved as the parent for the next mutation step. The algorithm terminates when a solution with acceptable fitness is found. While fitting algorithms such as this are essential for parameter determination, there remains the possibility that the best solution might not be found. Sometimes very different parameter sets can reproduce the targeted gaps and masses to nearly the same accuracy. However, they may well differ in their wavefunction quality. The analytical expressions for bandgaps and effective masses discussed above are extremely helpful as they serve as additional constraints to the problem solution and guide the high-dimensional fitting problem to convergence.

To address the fundamental wavefunction variations for different parameter sets a new fitting method has been developed which includes matches to *ab initio* wavefunctions in addition to gaps and masses in the fitting process. Tan, et. al.[47-49] use DFT wavefunctions to construct real-space tight-binding wavefunctions. A low-rank approximate DFT Hamiltonian is then transformed into the tight-binding basis using the transformation between the DFT and tight-binding wavefunctions. The tight-binding Hamiltonian is iteratively improved by comparing gaps, masses, and wavefunctions at high-symmetry points to DFT results. Including wavefunction fitting removes much of the uncertainty about whether the true best fit has been found.

1.3.3. Strain

Strains displace the atoms of a crystal from their ideal positions and can be either regular or irregular. Regular strains affect all primitive cells identically, and occur in strained layer superlattices (GaAs/InAs, Si/Ge, etc.) or when external strains are applied. Irregular strains affect differing primitive cells differently and occur in random alloys or the presence of defects, for example, as well as in geometry variations in two- or three-dimensions. Both types affect the electronic structure of the material by altering the Hamiltonian in terms of both its potentials and basis states. Changes appear in both the neighboring-atom and onsite Hamiltonian matrix elements. The most obvious change is in the relative positions (distances and orientations) of the potentials centered on the two atoms involved in a two-center integral. The second, less obvious change, is to the orbitals themselves. Recall that orthonormal basis sets are generally constructed from non-orthogonal bases using Löwdin's procedure[37]. In the Löwdin method an orthonormal set $\{|\varphi_j\rangle; j=1,\ldots,N\}$ is constructed from a non-orthogonal set (e.g., atomic orbitals) $\{|\eta_j\rangle; j=1,\ldots,N\}$ using the square root of the overlap matrix:

$$\left|\varphi_j\right\rangle = \sum_{m=1}^{N}\left[\underline{\mathbf{S}}^{-1/2}\right]_{m,j}\left|\eta_m\right\rangle, \quad \left[\underline{\mathbf{S}}\right]_{m,n} = \left\langle\eta_m\middle|\eta_n\right\rangle. \quad (4)$$

Displacing the atoms, and thus the non-orthogonal orbitals centered on them, changes the overlap matrix elements and thus the orthonormal basis of the Hamiltonian.

Both types of changes are incorporated into the neighboring-atom matrix elements (generally two-center integrals) via a power-law scaling relation. This scaling is referred to as the generalized Harrison's Law. Harrison[42] fit the bands of a simple cubic crystal to the free-electron bands at $k = 0$ and $k = \pi/a$, which required the *s-s* and *s-p* matrix elements to vary as d^{-2}, where d is the nearest-neighbor distance. The generalization allows for the power to vary with the bond type ($ss\sigma, sp\sigma, pd\pi$, etc.):

$$V_{ab\gamma}(d) = V_{ab\gamma}^{(0)}\left(\frac{d_0}{d}\right)^{\eta_{ab\gamma}} \tag{5}$$

In eq. (5) $V_{ab\gamma}^{(0)}$ is the ideal crystal two-center integral between orbitals *a*, and *b* (*s*, *p*, *d*, etc.) with bond type (σ, π, δ, etc.); d_0 is the ideal distance between the atoms and d the actual distance; and $\eta_{ab\gamma}$ the scaling exponent (nominally 2). The exponents are fit in a process similar to the two-center integrals in order to reproduce the gaps and masses as functions of strain, using as targets either *ab inito* calculations or experimental results.

The issue of scaling the onsite matrix elements under strain is somewhat more subtle. While it is tempting to think of the onsite parameters as being atomic orbital energies, eq. (4) and consideration of the crystalline environment make it clear that this cannot be the case. From eq. (4) it is clear that the onsite terms are affected by the Löwdin orthogonalization[37]. These terms are likewise affected by the neighboring atom potentials. In a free atom there is no p_x-p_y Hamiltonian matrix element due to the spherical symmetry of the atomic potential. In diamond or zincblende this term vanishes as well due to symmetry: Nearest neighbors are located at (a/4)(1,1,1), (a/4)(-1,-1,1), (a/4)(1,-1,-1), (a/4)(-1,1,-1). Under uniaxial [110] strain, however, the first two ($z > 0$) neighbors are at one common distance from the central atom while the second two ($z < 0$) are at a different common distance, leading to an incomplete cancellation of their effects. As a result, onsite matrix elements absent from the ideal crystal Hamiltonian can appear under some strain conditions.

Within the context of the $sp3d^5s^*$ basis[40] four different approaches to the strain treatment of the onsite matrix elements have been proposed. Jancu, et. al.[40] scale only the onsite *d* parameters for regular strains, the only new parameter being the *d*-state shear paramenter. In order to treat randomly-varying strains, such as one finds in a random alloy, adjustments computed separately for each neighboring atom, and summed over neighbors to find the complete shift, have been proposed[49-52]. The implementation of each of these methods is sufficiently involved that the reader is referred to the original references for the details. Briefly, however, the differences between them are that Niquet, et. al.[50] include both neighboring atom potential and Löwdin renormalization effects into a pair of parameters for each type of onsite matrix element. These adjustments are relatively simple to calculate, but because both effects are bundled together into each matrix element, establishing physically meaningful limits on the strain parameters is difficult. Boykin, et. al. first included Löwdin renormalization[37] shifts only[51], then later added nearest-neighbor potential effects (for *p*- and *d*-orbitals only)[52]. Although this approach is more complicated to implement, it does have the advantage of allowing physical limits on the parameters arising from the two separate effects to be determined[51,52]. The most recent approach is the DFT-mapped method of Tan, et. al.[49]; the DFT mapping reduces uncertainty in the parameter values. This method takes into account changes in neighboring atom potentials, but has the largest number of strain parameters of the methods discussed here.

1.3.4. Consequences of Discreteness and Incompleteness

Perhaps the least well understood and least appreciated properties of tight-binding models are their discreteness and incompleteness. Actually, any finite model is incomplete, and any model implemented on a computer is discrete. Even large-basis models, such as the Linearized Augmented Plane Wave (LAPW) method, are incomplete and the incompleteness must be properly taken into account[53]. The consequence of incompleteness is that many familiar relations from continuous quantum mechanics must be modified, because steps in their derivation assumed a complete basis. Not surprisingly, incompleteness is most readily seen in connection with the matrix elements of the momentum operator, which has a real-space representation: $\hat{p}^{(x)} = -i\hbar\partial/\partial x$, etc. Faithful representation of the derivative operator in a real-space basis implies an infinitesimal spatial resolution, which is clearly impossible in any finite basis. The momentum operator can only be exactly represented (for the included states) in a purely plane-wave basis, because plane waves are momentum eigenfunctions. Beyond this difficulty, commutators of the momentum operator are problematic in any finite basis. Graf and Vogl[44] point out that in any finite basis the position and momentum matrices no longer satisfy the commutator: $[\underline{r}^{(\alpha)}, \underline{p}^{(\beta)}] \neq i\hbar\underline{\delta}_{\alpha,\beta}$. The reason is that for any finite matrices $\underline{\mathbf{M}}, \underline{\mathbf{N}}$, $\mathrm{trace}(\underline{\mathbf{M}}\,\underline{\mathbf{N}}) = \mathrm{trace}(\underline{\mathbf{N}}\,\underline{\mathbf{M}})$ so that the trace of the commutator is zero. Because these operators and commutators appear in many different formulas, the consequences of incompleteness are far-reaching.

1.3.5. Effective-Mass Formula

The effective mass formula found in nearly all solid-state physics texts is incorrect and must be modified in an incomplete basis, because the leading term comes from the position-momentum commutator. Graf and Vogl[44] and Boykin[45] derive the correct formula (Ref. [45] also includes modifications for degenerate bands), which agrees exactly with the effective mass as calculated from the band curvature $\partial^2 E_n(\mathbf{k})/\partial k^{(\alpha)}\partial k^{(\beta)}$. The inverse effective mass tensor for a non-degenerate band state $|n, \mathbf{k}\rangle$ is[44,45]:

$$\left[\frac{1}{m^*}(\mathbf{k})\right]_{\alpha,\beta} = \frac{-i}{m_0\hbar}\langle n,\mathbf{k}|\left[\hat{r}^{(\alpha)}, \hat{p}^{(\beta)}\right]|n,\mathbf{k}\rangle + \frac{1}{m_0^2}\sum_{j\neq n}\frac{\langle n,\mathbf{k}|\hat{p}^{(\alpha)}|j,\mathbf{k}\rangle\langle j,\mathbf{k}|\hat{p}^{(\beta)}|n,\mathbf{k}\rangle + \langle n,\mathbf{k}|\hat{p}^{(\beta)}|j,\mathbf{k}\rangle\langle j,\mathbf{k}|\hat{p}^{(\alpha)}|n,\mathbf{k}\rangle}{E_n(\mathbf{k}) - E_j(\mathbf{k})} \tag{6}$$

In eq. (6) the sum runs over all states *j* in the finite basis. In terms of the Hamiltonian matrix, these matrices and commutators are[45]:

$$\frac{\hbar}{m_0}\underline{\mathbf{p}}^{(\alpha)}(\mathbf{k})=\frac{\partial\underline{\mathbf{H}}(\mathbf{k})}{\partial k^{(\alpha)}},\quad \underline{\mathbf{p}}^{(\alpha)}(\mathbf{k})=\frac{m_0}{i\hbar}\left[\underline{\mathbf{r}}^{(\alpha)},\underline{\mathbf{H}}\right](\mathbf{k}) \tag{7}$$

$$\frac{-i\hbar}{m_0}\left[\underline{\mathbf{r}}^{(\alpha)},\underline{\mathbf{p}}^{(\beta)}\right](\mathbf{k})=\frac{\partial^2\underline{\mathbf{H}}(\mathbf{k})}{\partial k^{(\alpha)}\partial k^{(\beta)}} \tag{8}$$

Note in particular eq. (8) which is not (up to constants) $\underline{\boldsymbol{\delta}}_{\alpha,\beta}$. We remark that so long as the spin-orbit Hamiltonian $\underline{\mathbf{H}}_{so}$ has only same-atom couplings, as in Chadi's prescription[A54,], and the position operator is diagonal (see 1.3.7 below), $\left[\underline{\mathbf{r}}^{(\alpha)},\underline{\mathbf{H}}_{so}\right]=\underline{\mathbf{0}}$. As a result the momentum and velocity matrices are proportional.

1.3.6. Localized-Orbital Representations

One cannot simply calculate localized-orbital position and momentum matrix elements as one does in atomic physics. The existence of discrete translational symmetry in an ideal crystal imposes a discrete real-space grid which puts constraints on the position and momentum matrices. First, these matrices must be consistent with the band velocities and effective masses of the tight-binding model itself, eqs. (6)-(8) above. Calculating momentum matrix elements using the continuous-space momentum operator fails the consistency test[55]. Second, the formalism must be gauge invariant, which in an arbitrary localized-orbital basis requires commuting position matrices[55]. In an arbitrary basis the only guaranteed gauge-invariant definition of the position matrix is diagonal:

$$\left\langle\gamma;\mathbf{R}_m+\mathbf{d}_n\left|\hat{r}^{(\alpha)}\right|\omega;\mathbf{R}_j+\mathbf{d}_l\right\rangle=\delta_{\gamma,\omega}\delta_{m,j}\delta_{n,l}\left(\mathbf{R}_j^{(\alpha)}+\mathbf{d}_l^{(\alpha)}\right) \tag{9}$$

These position matrices commute[44,55,56], and, from the states (2), also satisfy the consistency tests, eqs. (6)-(8).

The momentum matrices must therefore be calculated using the position operator:

$$\hat{p}^{(\alpha)}=\frac{m_0}{i\hbar}\left[\hat{r}^{(\alpha)},\hat{H}\right] \tag{10}$$

$$\left\langle\gamma;\mathbf{R}_m+\mathbf{d}_n\left|\hat{p}^{(\alpha)}\right|\omega;\mathbf{R}_j+\mathbf{d}_l\right\rangle=\frac{-im_0}{\hbar}\left(\mathbf{R}_m^{(\alpha)}\right.$$

$$\left.+\mathbf{d}_n^{(\alpha)}-\mathbf{R}_j^{(\alpha)}-\mathbf{d}_l^{(\alpha)}\right)\left\langle\gamma;\mathbf{R}_m+\mathbf{d}_n\left|\hat{H}\right|\omega;\mathbf{R}_j+\mathbf{d}_l\right\rangle \tag{11}$$

Using the diagonal form of the position operator therefore results in no same-atom position or momentum matrix elements (in an orthonormal basis). Same-atom, different-orbital position matrix elements generally result in non-commuting position matrices, thus violating gauge invariance: The only exception is for unusual, restricted bases. Foreman[57] has shown that gauge invariance with intra-atomic position matrix elements requires unconventional basis choices, for example in cubic crystals: s^1, $s^1p^3d^2$, $s^1p^3d^3f^1$, or $s^1p^3d^5f^3$. Conventional bases, such as sp^3, sp^3s^*, and $s^1p^3d^5s^*$ are ruled out. Therefore, in an arbitrary basis, gauge invariance and consistency are only guaranteed with diagonal position matrices.

1.3.7. Electromagnetic Coupling Hamiltonian

In continuous quantum mechanics the vector potential couples into the Hamiltonian via the momentum operator. Because the discrete, finite-basis expression should become the continuous-space one in the limit of a complete basis, the vector potential should enter with the momentum operator in tight-binding as well. A further requirement is that the expression be gauge invariant. The matrix elements of the tight-binding Hamiltonian with a vector potential are given by the Peierls substitution[58], which weights the Hamiltonian matrix elements with a phase (called the Peierls phase)[44,55,56]:

$$\left\langle\gamma;\mathbf{R}_{m,n}\left|\hat{H}^{(\mathbf{A})}\right|\omega;\mathbf{R}_{j,l}\right\rangle=$$

$$\exp\left[-\frac{ie}{\hbar}\int_{\mathbf{R}_{j,l}}^{\mathbf{R}_{m,n}}\mathbf{A}(\mathbf{s},t)\bullet d\mathbf{s}\right]\left\langle\gamma;\mathbf{R}_{m,n}\left|\hat{H}^{(0)}\right|\omega;\mathbf{R}_{j,l}\right\rangle \tag{12}$$

In eq. (12) $\hat{H}^{(\mathbf{A})}(\hat{H}^{(0)})$ denotes the Hamiltonian with (without) vector potential $\mathbf{A}$; and $\mathbf{R}_{a,b}=\mathbf{R}_a+\mathbf{d}_b$ the position of atom b (for polyatomic unit cells) in unit cell a. Eq. (12) is manifestly gauge invariant, but requires further justification on two points: (1) Its relationship to the minimal coupling Hamiltonian of continuous quantum mechanics and; (2) The path to be taken in the Peierls integral. Concerning the latter point, Graf and Vogl[44] propose the straight-line path connecting the two atoms. This choice is clearly the simplest and most natural, especially for nearest-neighbor models. It is also reasonable in view of the discrete nature of the model. If additionally, one uses the trapezoidal rule for the integral, which is in fact exact for linear $\mathbf{A}$, (e.g., a uniform $\mathbf{B}$-field as in $\mathbf{A}=(\mathbf{B}_0\times\mathbf{r})/2$), the result bears a striking resemblance to minimal-coupling Hamiltonian:

$$\left\langle\gamma;\mathbf{R}_{m,n}\left|\hat{H}^{(\mathbf{A})}\right|\omega;\mathbf{R}_{j,l}\right\rangle=\exp\left[-\frac{ie}{2\hbar}\left(\mathbf{R}_{m,n}-\mathbf{R}_{j,l}\right)\right.$$

$$\left.\left[\mathbf{A}\left(\mathbf{R}_{m,n},t\right)+\mathbf{A}\left(\mathbf{R}_{j,l},t\right)\right]\right]\left\langle\gamma;\mathbf{R}_{m,n}\left|\hat{H}^{(0)}\right|\omega;\mathbf{R}_{j,l}\right\rangle \tag{13}$$

In particular, note that an expansion of the exponential to first order in $\mathbf{A}$ gives for the first-order term the average of the momentum matrix element, eq. (11), times $\mathbf{A}$ evaluated on the two atomic sites. The full power series for the exponential gives matrix elements of multiple commutators of the position and Hamiltonian. Thus there are solid physical reasons for favoring the straight-line path.

More formally, Boykin[59] has shown that in the continuous $\{|\mathbf{r}\rangle\}$ basis, the matrix elements

$$\left\langle\mathbf{r}\left|\hat{H}^{(\mathbf{A})}\right|\mathbf{r}'\right\rangle=\exp\left[-\frac{ie}{\hbar}\int_{\mathbf{r}'}^{\mathbf{r}}\mathbf{A}(\mathbf{s},t)\bullet d\mathbf{s}\right]\left\langle\mathbf{r}\left|\hat{H}^{(0)}\right|\mathbf{r}'\right\rangle \tag{14}$$

result in the usual minimal-coupling Schrödinger equation:

$$\left\langle\mathbf{r}\left|\hat{H}^{(\mathbf{A})}\right|\psi(t)\right\rangle=\frac{1}{2m}\left[\hat{\mathbf{p}}+e\mathbf{A}(\mathbf{r},t)\right]^2\psi(\mathbf{r},t) \tag{15}$$

Note the analogy of eq. (14) to eq. (12). In eq. (14) it is critical to realize that the continuous-space basis matrix elements are generalized functions. In the continuous $\{|\mathbf{r}\rangle\}$ basis the path is immaterial, as one might expect. In addition, Boykin, Bowen, and Klimeck[56] derive eq. (12) in a different way. Using a multiple-commutator expansion and diagonal position operators (9), they obtain an infinite series in terms of the momentum operator and the vector potential. In a complete basis the series terminates at second order in the vector potential, resulting in the usual minimal coupling Hamiltonian. Furthermore, they show that when the straight line path between the two atoms is used in the Peierls integral

in eq. (12), the multiple commutator series and the Peierls substitution give the same matrix elements. Thus, when one uses diagonal (i.e., commuting) position matrices and takes into account the incomplete basis, the vector potential is accommodated by weighting the Hamiltonian matrix elements by the Peierls phase.

External, scalar potentials are handled consistently with the diagonal form of the position matrices, eq. (9). The applied scalar potential, $U(\mathbf{r})$, can be expanded in a Taylor series in the position operators, so that its matrix elements are:

$$\left\langle \gamma;\mathbf{R}_{m,n}\left|\hat{U}\left(\mathbf{r}\right)\right|\omega;\mathbf{R}_{j,l}\right\rangle = \delta_{\gamma,\omega}\delta_{m,j}\delta_{n,l}U\left(\mathbf{R}_{j,l}\right) \quad (16)$$

This prescription for the scalar potential also satisfies gauge invariance[44,55,56].

1.4. Interfaces and Transport

1.4.1. Complex Bands

The complex bands are the surface modes of a crystal[60-66]. These surface states are used in constructing the open boundary conditions of planar devices, such as nanowire transistors, or resonant tunneling diodes (RTDs). They also determine the confinement, as with barriers in RTDs or the cladding of a quantum dot. In RTDs, for example, the confinement determines the sharpness of the resonances and thus greatly affects the current-voltage characteristics. In a quantum dot it determines the wavefunction leakage out of the dot. The complex bands are found by imposing periodic boundary conditions only in the plane of the surface, using basis states of Eq. (3); one or more atomic planes are grouped together into layers, so that there are only nearest-layer couplings. The resulting Hamiltonian sub-matrices are defined as:

$$\left[\underline{\mathbf{H}}_{(L',L)}\right]_{(\omega'l',\omega l)} = \left\langle \omega';l';L';\mathbf{k}_{\parallel}\left|\hat{H}\right|\omega;l;L;\mathbf{k}_{\parallel}\right\rangle \quad (17)$$

The total state is:

$$\left|\psi\left(\mathbf{k}_{\parallel}\right)\right\rangle = \sum_{\omega',l',L'} C_{L'}^{\omega',l'}\left|\omega';l';L';\mathbf{k}_{\parallel}\right\rangle \quad (18)$$

The Schrödinger equation therefore has rows:

$$\underline{\mathbf{H}}_{(L-1,L)}^{\dagger}\mathbf{C}_{L-1} + \left[\underline{\mathbf{H}}_{(L,L)} - \underline{\mathbf{1}}E\right]\mathbf{C}_{L} + \underline{\mathbf{H}}_{(L,L+1)}\mathbf{C}_{L+1} = \underline{\mathbf{0}} \quad (19)$$

In eq. (19) $\mathbf{C}_L$ is the vector of all coefficients for layer L; in bulk, $\underline{\mathbf{H}}_{(\mathrm{L-1,L})} = \underline{\mathbf{H}}_{(\mathrm{L,L+1})}$.

The propagation equation is found by introducing $\mathbf{C}_L = \mathbf{C}_L$ along with eq. (19):

$$\begin{bmatrix} \underline{\mathbf{H}}_{(L,L+1)} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\mathbf{1}} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{L+1} \\ \mathbf{C}_{L} \end{bmatrix} = \begin{bmatrix} -\left(\underline{\mathbf{H}}_{(L,L)} - \underline{\mathbf{1}}E\right) & -\underline{\mathbf{H}}_{(L,L+1)}^{\dagger} \\ \underline{\mathbf{1}} & \underline{\mathbf{0}} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{L} \\ \mathbf{C}_{L-1} \end{bmatrix} \quad (20)$$

The traditional way to solve eq. (20) is to multiply by the inverse of the left-hand matrix to generate a transfer matrix:

$$\underline{\mathbf{M}}_{(L+1,L)} = \begin{bmatrix} \underline{\mathbf{H}}_{(L,L+1)} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\mathbf{1}} \end{bmatrix}^{-1}\begin{bmatrix} -\left(\underline{\mathbf{H}}_{(L,L)} - \underline{\mathbf{1}}E\right) & -\underline{\mathbf{H}}_{(L,L+1)}^{\dagger} \\ \underline{\mathbf{1}} & \underline{\mathbf{0}} \end{bmatrix};$$

$$\begin{bmatrix} \mathbf{C}_{L+1} \\ \mathbf{C}_{L} \end{bmatrix} = \underline{\mathbf{M}}_{(L+1,L)}\begin{bmatrix} \mathbf{C}_{L} \\ \mathbf{C}_{L-1} \end{bmatrix} \quad (21)$$

Setting $\mathbf{C}_{L+1} = \lambda_{+}\mathbf{C}_L$ gives an ordinary eigenvalue equation for the transfer matrix. Its eigenstates are the surface states for the crystal and they can be propagating or growing/decaying.

While this equation is intuitively satisfying, it unfortunately has a severe problem: It might not exist. For certain parameter sets or simply values of $\mathbf{k}_{\boxdot}$ the left- and/or right-hand matrices in eq. (20) become singular and the transfer matrix does not exist [64,65]. The alternative, and most numerically stable way to find the surface states is to solve either the forward $\mathbf{C}_{L+1} = \lambda_{+}\mathbf{C}_L$ or reverse $\mathbf{C}_{L-1} = \lambda_{-}\mathbf{C}_L$ propagation problem as a generalized eigenproblem[64]; an ordinary eigenproblem with a shift strategy can also be used[66]. The forward generalized eigenproblem is:

$$\lambda_{+}\begin{bmatrix} \underline{\mathbf{H}}_{(L,L+1)} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\mathbf{1}} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{L} \\ \mathbf{C}_{L-1} \end{bmatrix} = \begin{bmatrix} -\left(\underline{\mathbf{H}}_{(L,L)} - \underline{\mathbf{1}}E\right) & -\underline{\mathbf{H}}_{(L,L+1)}^{\dagger} \\ \underline{\mathbf{1}} & \underline{\mathbf{0}} \end{bmatrix}\begin{bmatrix} \mathbf{C}_{L} \\ \mathbf{C}_{L-1} \end{bmatrix} \quad (22)$$

The eigenvalues come in pairs $(\lambda_{\pm}, \lambda_{\pm}^{*})$ and eigenvalues $\lambda_{\pm} \neq 0$ come in pairs $(\lambda_{\pm}, 1/\lambda_{\pm}^{*})$; for every m-fold $\lambda_{+} = 0$ there is likewise an m-fold $\lambda_{-} = 0$ [64]. Thus for every immediately decaying state there is an immediately growing one; for every forward propagating state there is a reverse-propagating one. The $\lambda_{\pm} \neq 0$ are conveniently written as $\exp\left[ik_{\perp}a_{\perp}\right]$, where $k_{\perp}$ is the wavevector normal to the surface and is in general complex (real for Bloch states) and $a_{\perp}$ is the layer thickness. The forward or reverse generalized eigenproblem therefore yields wavevectors $k_{\perp}(E, \mathbf{k}_{\boxdot})$.

An example complex bandstructure is plotted in Fig. 2, part of the AlAs complex bands[64] for the [001] surface, calculated using the sp^3s^* parameter set given in [62]. In Fig. 2 a is the face-centered cubic conventional unit cube edge. The real bands are plotted with black solid lines; purely imaginary bands with blue dashed lines; and the real and imaginary parts of complex bands are plotted with red dotted lines. Real bands and real parts of complex bands are plotted for $k_z > 0$ only; imaginary bands and imaginary parts of complex bands are plotted for $k_z < 0$ only. The imaginary bands connect the valence band maxima (degenerate heavy/light hole and split-off hole) with conduction band minima. The lowest X-valley minimum occurs at $k_z \approx 1.5\pi/a$, so that at the Brillouin zone face the lowest conduction band has a shallow local maximum. This maximum is connected to the minimum of the next lowest band by a small complex band between around 2.0 - 2.2eV with $\mathrm{Re}\{k_z\} \approx 2\pi/a$ (i.e., at the zone face). Another complex band begins off the X-valley minimum.

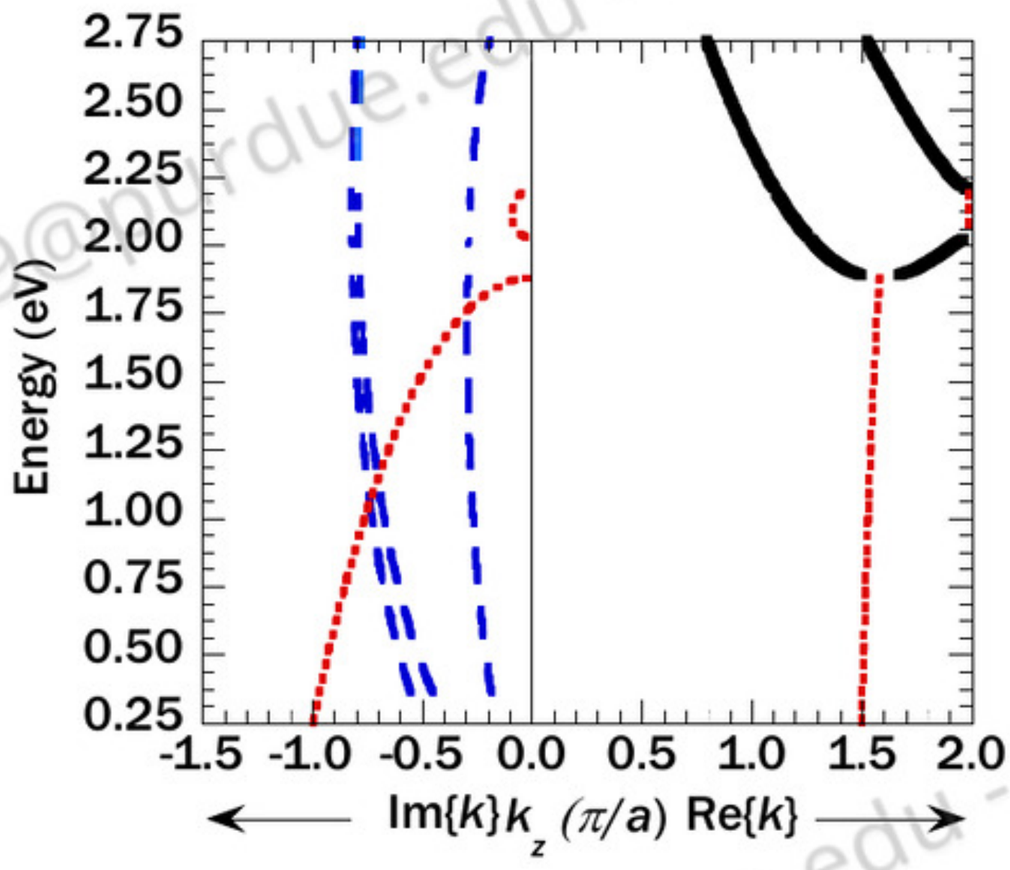

Figure 2: The complex bands[64] of AlAs calculated using the parameter set of [62], All three types of bands are present: Real (solid black lines), Imaginary (dashed blue lines), and Complex (dotted red lines), which have wavevectors with both real and imaginary parts.

1.4.2. Transmission Calculations with Transfer Matrices: Numerical Stability

Wavefunction-based transport calculations involve injecting propagating state(s) into a device and calculating the reflection and transmission coefficients (probabilities), and are thus open systems. The energy is continuous, not quantized, and the boundary conditions on the terminal wavefunction coefficients come from the complex bandstructure calculations discussed above. Mathematically, in a direct transmission calculation, this results in an $\underline{\mathbf{A}}\mathbf{x} = \mathbf{b}$ problem at each energy, with the inhomogeneous vector representing the open-system boundary conditions.

The transfer matrix equation (21) is a conceptually simple way to calculate the wavefunction in a layered device or structure such as an RTD, superlattice, or nanowire. Its structure is similar to that of the transfer matrix used to calculate the transmission through an RTD in the envelope function (i.e., effective mass) model[67], or to propagate electromagnetic waves through layered dielectrics. Because of its success in envelope function modeling of RTDs, the transfer matrix was the method first applied in early tight-binding models of these devices in the late 1970s-early 1980s. For a structure with boundaries at layers 0 and (N+1), the wavefunction coefficients were calculated as:

$$\begin{bmatrix} \mathbf{C}_{N+1} \\ \mathbf{C}_{N} \end{bmatrix} = \prod_{L=0}^{N-1} \underline{\mathbf{M}}_{(L+1,L)} \begin{bmatrix} \mathbf{C}_{1} \\ \mathbf{C}_{0} \end{bmatrix} \tag{23}$$

The boundary conditions applied on the layer 0 side include only a forward-propagating state and reflected reverse-propagating and reverse-decaying states; those on the layer (N+1) side include only transmitted forward-propagating and forward-decaying states.

This procedure often worked for very thin structures, but once the number of transfer matrices exceeded around 20 or so, the calculation rapidly became severely inaccurate[68]. The sum of reflection and transmission probabilities, which *must* be unity, would instead turn out to be much higher, often in the range $10^2 - 10^6$. As a result tight-binding models were not used to model realistic devices, which have long space-charge regions.

The cause of the instability is easily seen in the AlAs complex band plot of Fig. 2. Notice that the lowest conduction-band states (in the *X*-valleys) appear at around 1.90 eV (lowest black solid line). In addition to these purely propagating states, at the same energy, there are also three imaginary bands (blue dashed lines) connecting valence-band maxima to conduction-band minima at zone center; these states grow and decay. Generally then, there are evanescent (growing/decaying) states at the same energy as propagating states, and that is the source of the numerical problem.

The boundary conditions result from accurate, but still finite-precision, numerical calculations of the eigenvectors eq. (22), whether via an ordinary (transfer-matrix) or generalized eigenproblem. Thus, the boundary condition vectors will be corrupted with tiny amounts of other, disallowed eigenvectors, namely those which grow in forward propagation; at a given energy, these evanescent states are generally present with propagating states. At each subsequent transfer-matrix multiplication, these initially small, erroneous eigenvectors grow in magnitude, and the growth is exponential. It does not take many transfer-matrix multiplications before the corrupt growing eigenvectors overwhelm the true eigenvectors, leading to a catastrophic failure of the calculation[68,69]. In principle, this sort of failure can occur in an envelope-function calculation, but it requires extraordinarily thick barrier regions because each material supports only a single pair of states, propagating or evanescent, at a given energy.

1.4.3. Direct Transmission Methods

The numerical instabilities discussed above prevented multi-band tight-binding models from being used to model realistic aperiodic structures and devices for around a decade. Then, in the late 1980s-early 1990s a series of solutions to the problem were developed[68,70,71]. Although the three methods differ in their details, they all recast the transmission problem as a large, sparse, matrix equation. This reformulation into a large, sparse linear system is the key to preventing corruption of the true eigenvectors by erroneous growing eigenvectors. The renormalization method[71] was initially used for superlattice electronic structure calculations (thus with periodic, as opposed to open-system boundary conditions). The first numerically-stable resonant-tunneling diode calculations were those of Boykin, et. al.[68,72], which included long, realistically-sized space-charge layers. They divided the device into segments of a few layers over which wavefunction propagation using transfer-matrices was still accurate. The result is a large, sparse, linear system consisting of many small transfer-matrix problems, though the dimension is considerably smaller than the Hamiltonian with open-system boundary conditions. The next year, Ting, et. al.[70] published a similar method, in which they solve the $\underline{\mathbf{A}}\mathbf{x} = \mathbf{b}$ problem resulting from the Hamiltonian with open-system boundary conditions directly, the large, sparse, linear system is larger than of Boykin, et. al.[68] but avoids transfer matrices and their products.

The methods of Boykin, et. al.[68] and Ting, et. al.[70] work well for structures having periodicity in the plane normal to the transport or quantization axis, such as RTDs and quantum wells (QWs). However, they can be costly for nanowires (NWs) which have finite cross-sections, because the resulting layer-basis now consists of perhaps dozens of atoms (versus as few as 2), each with many (often 20) orbitals. The layer matrices in eq. (19) are therefore very large, which for a general basis and NW orientation can significantly slow the computation. However, for a nearest-neighbor basis and for [100]- and [111]-oriented NWs in zincblende/diamond materials, these matrices are very sparse because each atomic plane contains atoms belonging to only one of the two face-centered-cubic (FCC) sublattices, so there are no same-plane inter-atomic couplings. In 2008 Boykin, et. al. [73,74] published an optimized renormalization method[71] to exploit this sparsity.

The optimized method[73,74] is easily illustrated for the case of a [100]-oriented NW in diamond or zincblende. In these NWs, four atomic planes make up a layer (i.e., a primitive cell for a perfect NW). In terms of layer matrices and vectors, the Schrödinger equation for a wire of L layers is written as[73]:

$$\begin{bmatrix} \underline{\underline{\mathbf{H}}}_{1,1}+\underline{\underline{\boldsymbol{\Sigma}}}_1-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{H}}}_{1,2} & \underline{\underline{\mathbf{0}}} & \cdots & \underline{\underline{\mathbf{0}}} \\ \underline{\underline{\mathbf{H}}}_{1,2}^{\dagger} & \underline{\underline{\mathbf{H}}}_{2,2}-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{H}}}_{2,3} & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{0}}} \\ \underline{\underline{\mathbf{0}}} & \ddots & \ddots & \ddots & \underline{\underline{\mathbf{0}}} \\ \vdots & \ddots & \ddots & \ddots & \underline{\underline{\mathbf{H}}}_{L-1,L} \\ \underline{\underline{\mathbf{0}}} & \cdots & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{H}}}_{L-1,L}^{\dagger} & \underline{\underline{\mathbf{H}}}_{L,L}+\underline{\underline{\boldsymbol{\Sigma}}}_L-\underline{\underline{\mathbf{1}}}E \end{bmatrix} \begin{bmatrix} \vec{\mathbf{C}}_1 \\ \vec{\mathbf{C}}_2 \\ \vdots \\ \vdots \\ \vec{\mathbf{C}}_L \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{V}}_l \\ \vec{\mathbf{0}} \\ \vdots \\ \vdots \\ \vec{\mathbf{0}} \end{bmatrix} \tag{24}$$

where the self-energies $\underline{\underline{\boldsymbol{\Sigma}}}_1, \underline{\underline{\boldsymbol{\Sigma}}}_L$ couple the emitter (1) and collector (L) layers to the semi-infinite contact regions (unbiased NW), the vectors $\vec{\mathbf{C}}_j$ are the j-th layer orbital coefficients, and the injection into the emitter is the vector $\vec{\mathbf{V}}_l$.

As mentioned above, the layer matrices in eq. (24) are themselves sparse: layers l and (l+1) are only coupled by the last atomic plane of l and the first atomic plane of (l+1). This means that the NW can be very efficiently treated one atomic plane at a time; the renormalization method[71] is ideally suited to the task. Renormalization[71] decouples one plane from its neighbors and then re-couples the surrounding planes; no other planes are altered. Decoupling such a plane is accomplished thusly[73]:

$$H = M_{L,p}^{-1}\tilde{H}M_{R,p}^{-1} \tag{25}$$

$$H = \begin{bmatrix} \ddots & \ddots & & \\ \ddots & \underline{\underline{\mathbf{H}}}_{p-1,p-1}-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{H}}}_{p-1,p} & \underline{\underline{\mathbf{0}}} \\ \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{H}}}_{p-1,p}^{\dagger} & \underline{\underline{\mathbf{H}}}_{p,p}-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{H}}}_{p,p+1} \\ & & \underline{\underline{\mathbf{H}}}_{p,p+1}^{\dagger} & \underline{\underline{\mathbf{H}}}_{p+1,p+1}-\underline{\underline{\mathbf{1}}}E \\ & & \ddots & \ddots \end{bmatrix} \tag{26}$$

$$\tilde{H} = \begin{bmatrix} \ddots & \ddots & & \\ \ddots & \hat{\underline{\underline{\mathbf{H}}}}_{p-1,p-1}-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{0}}} & \hat{\underline{\underline{\mathbf{H}}}}_{p-1,p+1} \\ \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{H}}}_{p,p}-\underline{\underline{\mathbf{1}}}E & \underline{\underline{\mathbf{0}}} \\ & \hat{\underline{\underline{\mathbf{H}}}}_{p-1,p+1}^{\dagger} & \underline{\underline{\mathbf{0}}} & \hat{\underline{\underline{\mathbf{H}}}}_{p+1,p+1}-\underline{\underline{\mathbf{1}}}E \\ & & & \ddots \end{bmatrix} \tag{27}$$

$$M_{L,p}^{-1} = \begin{bmatrix} \ddots & \underline{\underline{\mathbf{0}}} & \ddots & \cdots & \cdots \\ \ddots & \underline{\underline{\mathbf{1}}} & \underline{\underline{\mathbf{X}}}_p^{\dagger} & \underline{\underline{\mathbf{0}}} & \ddots \\ \ddots & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{1}}} & \underline{\underline{\mathbf{0}}} & \ddots \\ \vdots & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{Y}}}_p^{\dagger} & \underline{\underline{\mathbf{1}}} & \ddots \\ \vdots & \ddots & \ddots & \underline{\underline{\mathbf{0}}} & \ddots \end{bmatrix}, \quad M_{R,p}^{-1} = \left(M_{L,p}^{-1}\right)^{\dagger} \tag{28}$$

$$\underline{\underline{\mathbf{X}}}_p = \left(\underline{\underline{\mathbf{H}}}_{p,p}-\underline{\underline{\mathbf{1}}}E\right)^{-1}\underline{\underline{\mathbf{H}}}_{p-1,p}^{\dagger}, \quad \underline{\underline{\mathbf{Y}}}_p = \left(\underline{\underline{\mathbf{H}}}_{p,p}-\underline{\underline{\mathbf{1}}}E\right)^{-1}\underline{\underline{\mathbf{H}}}_{p,p+1} \tag{29}$$

$$\begin{aligned} &\hat{\underline{\underline{\mathbf{H}}}}_{p-1,p+1} = -\underline{\underline{\mathbf{H}}}_{p-1,p}\underline{\underline{\mathbf{Y}}}_p, \ \hat{\underline{\underline{\mathbf{H}}}}_{p-1,p-1} = \underline{\underline{\mathbf{H}}}_{p-1,p-1} \\ &-\underline{\underline{\mathbf{H}}}_{p-1,p}\underline{\underline{\mathbf{X}}}_p, \ \hat{\underline{\underline{\mathbf{H}}}}_{p+1,p+1} = \underline{\underline{\mathbf{H}}}_{p+1,p+1}-\underline{\underline{\mathbf{H}}}_{p,p+1}^{\dagger}\underline{\underline{\mathbf{Y}}}_p \end{aligned} \tag{30}$$

Clearly, each plane is decoupled only once. Also, no complicated inversions are required for the decoupling matrices M:

$$M_{L,p} = \begin{bmatrix} \ddots & \underline{\underline{\mathbf{0}}} & \ddots & \cdots & \cdots \\ \ddots & \underline{\underline{\mathbf{1}}} & -\underline{\underline{\mathbf{X}}}_p^{\dagger} & \underline{\underline{\mathbf{0}}} & \ddots \\ \ddots & \underline{\underline{\mathbf{0}}} & \underline{\underline{\mathbf{1}}} & \underline{\underline{\mathbf{0}}} & \ddots \\ \vdots & \underline{\underline{\mathbf{0}}} & -\underline{\underline{\mathbf{Y}}}_p^{\dagger} & \underline{\underline{\mathbf{1}}} & \ddots \\ \vdots & \ddots & \ddots & \underline{\underline{\mathbf{0}}} & \ddots \end{bmatrix}, \quad M_{R,p} = M_{L,p}^{\dagger} \tag{31}$$

The matrices $\underline{\underline{\mathbf{X}}}_p, \underline{\underline{\mathbf{Y}}}_p$ and the inverses $\hat{\underline{\underline{\mathbf{H}}}}_{p,p}^{-1}$ for later wavefunction reconstruction. The emitter and collector layers require a somewhat different, but not overly complicated, treatment because the open-system boundary conditions make them non-Hermitian[73].

The method becomes highly efficient when the decoupling proceeds by *alternating* interior atomic planes. In the initial step, inverting the diagonal blocks is trivial in the no-spin-orbit case and nearly trivial in the spin-orbit case. The inter-plane coupling matrices are themselves very sparse, so that in the initial step one-half of the NW is decoupled with trivial inversions and sparse-matrix multiplies. Subsequent steps use the alternating-plane method as well, and savings accrue in the second step (decoupling a further one-fourth of the NW), which requires a mixture of sparse- and full-matrix operations. Only the last fourth of the NW requires full-matrix operations. The computational efficiency improvement can be substantial [73,74].

1.5. Transport with Green Functions

The numerical instability of transfer matrices can also be totally avoided by using a completely different Ansatz to compute transmission and transport: The non-equilibrium Green Function (NEGF) approach. The approach is based on the formulation of an explicit, real-space representation of the whole device in a (sparse) Hamiltonian. The discussion of the NEGF approach is beyond the scope of this text. The foundational theory for NEGF was developed in the 1960s by

Kadanoff and Baym [75] and Keldysh [76], but it remained confined to the physics literature. Ferry[77] and Datta[78] took NEGF from an obscure physics approach to a useful tool in electrical engineering by placing quantum transport simulations on a solid theoretical footing in the early 1990s. Datta refined the essential description of the methodology into very good introductions to NEGF for engineers in the early 2000s [79-81] and published a series of well received text books [82-84]. The latest two books [85,86] accompany two self-paced study courses in nanoHUB-U [87,88].

The NEMO code developed at Texas Instruments to model resonant tunneling diodes provided the first tight-binding implementation within NEGF [19,89]. The concept of an impulse response at every site inside of the device eliminates altogether the transfer matrix issue. However, the computation of the open boundary conditions of the infinite contact layer is still battling similar issues depending on the detailed approach. Bowen provided a solution to the wavefunction based or QTBM-based boundary conditions [90,66] in case of singular coupling matrices [91]. Completely stable iterative surface green function treatments have been developed [92] and refined to a point of geometrically fast convergence [93].

The perception in the field of quantum transport is that NEGF's strength is the fundamentally solid capability to include incoherent scattering. Yet most researchers in the field shy away from actually treating incoherent scattering due to the high computational burden and use NEGF in a purely coherent, ballistic approach. Bowen[94] demonstrated the coherent wavefunction approach based on Quantum Transmitting Boundary Conditions (QTBM) [90,66] is equivalent to the coherent NEGF approach. Luisier showed that QTBM is computationally less intensive than the NEGF approach in 2D and 3D systems [95].

It is the opinion of these authors that the real strength of the NEGF approach is the treatment of the open boundary conditions and their influence to broaden the interior quantum states which provide finite lifetimes. This concept can now be conveyed to undergraduate students in Electrical Engineering by Datta's books[85,86] and courses[87,88]. Such general boundary conditions enable the treatment of complex, irregular contacts which inject carriers from Quasi-Bound States [96] formed in quantum wells, geometrically L-shaped contacts, superlattice energy filters [97], generally shaped [98] and disordered [98,99] contacts. Such contact treatment includes an empirical broadening that compiles together scattering mechanisms of any sort. Such broadening is specifically critical in geometries such as heterostructure tunnel FETs (TFETs) which typically contain quasi-bound states confined close to the tunneling region due to heterostructure design [100] or traps [101]. With the assumption of a local quasi-Fermi level in the contact regions one can easily link the NEGF approach with drift diffusion in sophisticated, quantum state dominated contact regions in 1D [102] and 2D/3D [103] to enable truly multi-scale transport modeling that ranges from ab-initio-based atomistic representations to realistic device domains that extend hundreds of nanometers [96,102,103]. The use of tight-binding enables the multi-scale modeling of long-range, spatially dependent strain or piezo-electric effects [3,5,24,25,26,27] and electron-nucleus interactions[28]. Tight-binding approaches are used in references [100-102] and the extensive references therein.

1.6. Large-Scale Numerical Aspects

1.6.1. Hamiltonian Matrix Structures and Scaling for Closed Systems

One critical application-oriented advantage of the nearest neighbor orthogonal tight-binding model is its very sparse matrix representation. The valence electrons of each atom are represented with a number of orbitals where typical choices are s, sp^3, sp^3s^*, or $sp^3d^5s^*$ with 1, 4, 5, and 10 orbitals respectively which ultimately determine the per-atom matrix size. If spin is explicitly included, which is typically necessary for most applications, then the number of orbitals simply doubles. Each atom forms a well-structured block matrix that is coupled with the same block size to any neighboring atom (typically 4 in Si/Ge and III-V semiconductors). The on-site block matrix elements are diagonal only with a few off-diagonal terms coupling up and down spin through spin-orbit interaction. The block matrices coupling neighboring atoms are typically close to full, but couple only the same spin. Spin-up and spin-down coupling blocks are typically identical, unless modified by an external magnetic field. Rather general Hamiltonian matrix construction algorithms were implemented early in the NEMO1D work [19]. The NEMO5 [104,105] implementation was further generalized to include any number of orbitals such as f which resulted in a rather rapid incorporation of piezoelectric materials such as SmSe and SmTe [106].

The specific block matrix Hamiltonian structure offers significant opportunities for computational optimization with regard to: (1) optimized Hamiltonian construction (pre-compiled Slater-Koster tables); (2) memory storage requirement reductions (number of stored elements and real/complex custom storage); (3) optimized matrix-vector multiplications; and (4) short-integer based pointers to address individual atoms. With these optimizations it was possible in NEMO3D to compute million-atom electronic structures in the year 1999 on the first Beowulf cluster computers built at Caltech/NASA/JPL [30,31,107]. Further optimized matrix-vector optimizations based on Intel-specific SSE instruction sets enabled scaling to 50 million atoms, which roughly corresponds to a chunk of semiconductor of size $100\times100\times100\text{nm}^3$.

Figure 3 shows the excellent scaling on a moderate size cluster of 2008 [3]. The compute time represents the end-to-end computing time including set-up and output for 4 conduction band and 4 valence band states of a InAs quantum dot embedded in GaAs. The fixed-size InAs QD is in the middle of the simulation domain and the GaAs buffer size is increased in this numerical experiment to demonstrate several key aspects of NEMO3D and its algorithms: a) that the eigenenergies and eigenstates should be independent of the GaAs buffer size as long as the buffer is large enough; b) that the calculation of interior eigenvalues remains stable regardless of the system size; c) that memory requirements remain under control even in such large system sizes (special data structures and addressing schemes had to be developed to minimize storage and avoid duplicate storage on the

distributed MPI-based algorithms); d) that parallel scaling performance remains linear with system size. MPI communication requirements increase with the number of cores and with the number of atoms at each MPI parallel overlap region. The required times increase from 28.3 hours per million atoms on 8 cores to 83 hours per million atoms on 128 cores. As the overall system size increases more cores must be used to fit the overall computation into memory.

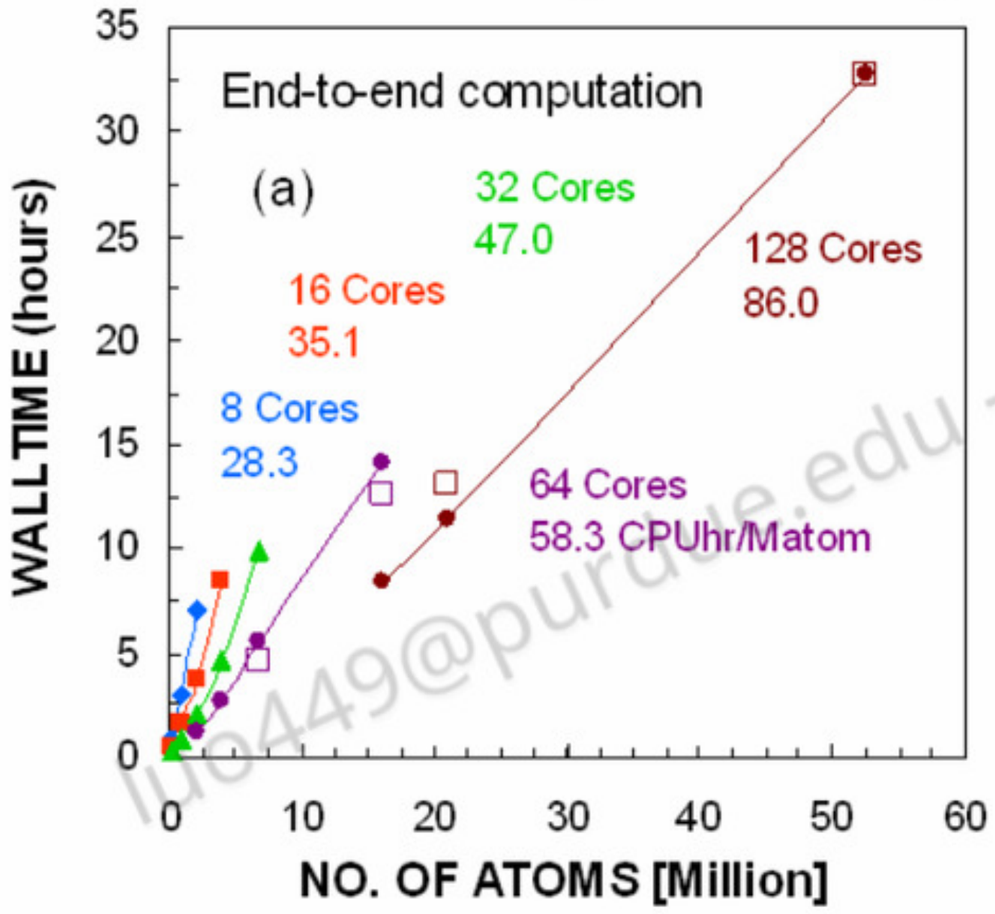

Figure 3: from Ref. [3] Fig 10a - Wall clock time vs. number of atoms for end-to-end computations of the electronic structure of a quantum dot, for various numbers of cores on the PU/Woodcrest cluster. Listed next to the number of cores are the CPU hours/Million atoms needed in the simulation.

1.6.2. Scaling Issues with Open Boundary Conditions

Scaling electronic structure calculations to multiple millions of atoms has been critical for the study of quantum dots and single shallow impurities such as As or P in Si as discussed below. In these problems the system is considered to be closed and the eigenvalues are real (even though the Hamiltonian is in general complex). The introduction of open boundary conditions turns the system Hamiltonian into a non-Hermitian system. Any atom inside a unit cell that is connected to a semi-infinite lead is part of a full boundary condition sub-matrix that is completely full. Typical Unit cells of a 2.1nm Si nanowire cross section with transport in the [100], [110], and [111] crystal directions are depicted in Figure 4 below. Numerical inversion stability in double precision limits such matrix sizes to typically dimension 2,000 to 10,000. With a typical number 20 orbitals only 100 to 500 atoms can be in open system boundary condition. Ignoring spin or ignoring *d* orbitals can result in a 2× or 10× increase of allowed atoms in the open surface. Either way the open cross sections in nanowires can barely reach 5-10nm diameters.

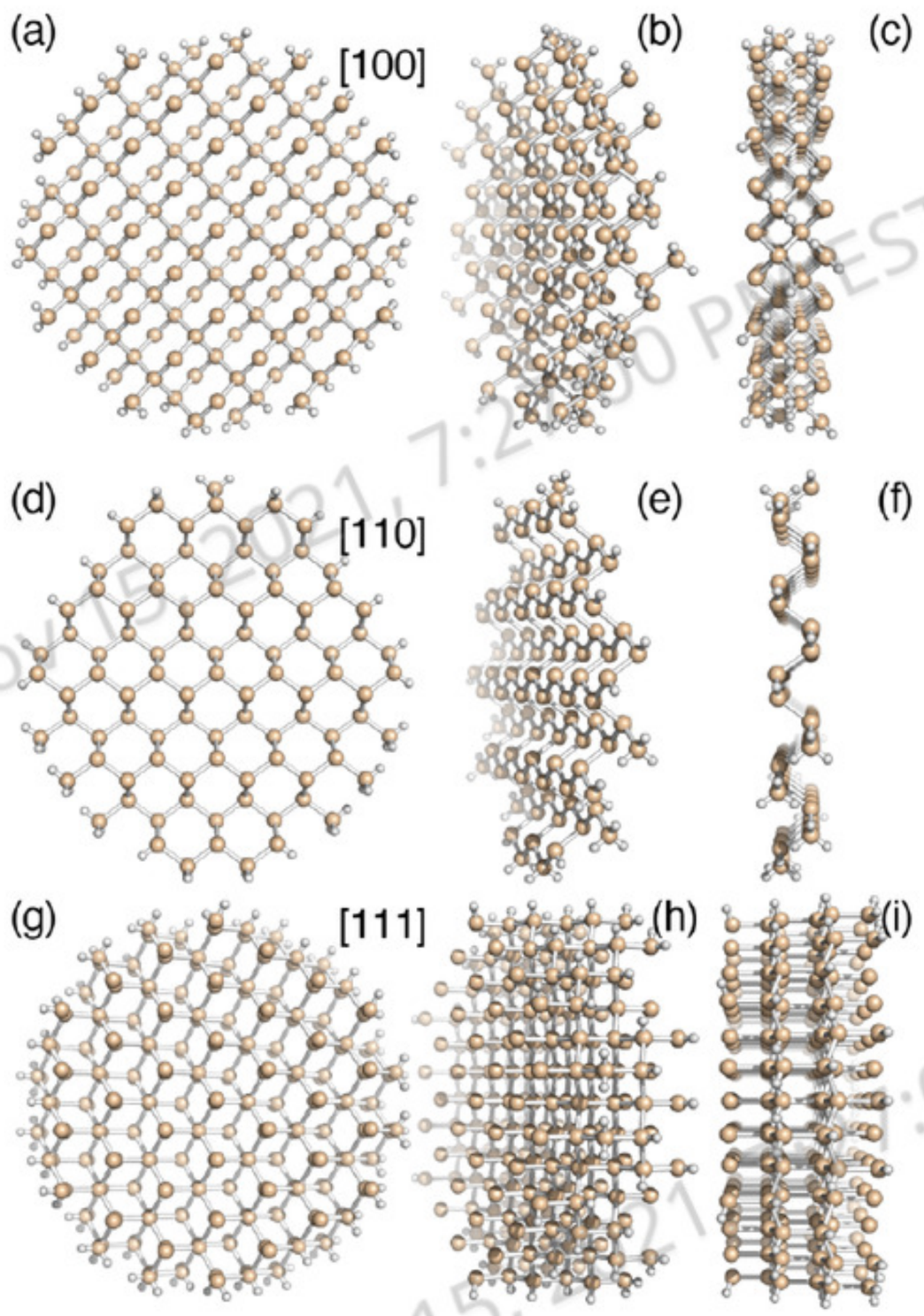

Figure 4: Unit cells of three 2.5 nm thick nanowires with transport in 3 different crystal directions [100], [110] and [111]. The full cross sections are shown in the left column, a tilted side view in the middle column, and a side view of the repeated cell in the right column. The images created with *Bandstructure Lab* on nanoHUB [108].

The non-Equilibrium Green Function (NEGF) method has now been established as the state-of-the-art in quantum transport modeling. The retarded Green function is the inverse of the open system non-Hermitian Hamiltonian and is for realistic systems computationally prohibitively expensive and in almost all cases is really not needed. The first and last column of the off-diagonal matrix elements are needed to compute the charge injection from the left and the right end of the device. The diagonals represent the local density of states in the system and are needed if some sort of physics-based decoherence or scattering is included in the model [89,96]. The Hamiltonian of the open system remains sparse as described above, except for the block of atoms at the contacts that create a full-matrix sub-block. This sub-block is determined by the number of atoms in the contact regions such as source and drain, that are modeled as infinite leads.

The size of the cross section scales the computational load in the number of atoms in the open surface in $O[(NB)^3]$ where N is the number of atoms in the repeated boundary cell and B is the number of orbitals in the basis (per atom). The length of the nanowire in terms of number of layers (L) enters the scaling of the computation basically linearly $O[L(NB)^3]$ in the typical recursive Greens function algorithm (RGF) [89]. RGF remains the fastest general-purpose algorithm to compute the diagonals and selected off-diagonals of the

system's Green's function, which is effectively the inverse of the open-system Hamiltonian. Algorithm enhancements to reduce the cubic scaling dependence $O[L(NB)^3]$) to lower order, reduce the layer coupling, decompose the blocks, and to enable efficient parallelization have been developed over the years [71,73,74,109-114].

The computational load can also be reduced by compressing the real-space basis in the cross section into modes [115]. This approach works very well for effective mass models and has been applied extensively [116,117]. This basis transformation enables reduced computation times and increased cross section sizes. The tool *nanowire* [118] on nanoHUB.org allows anyone to visualize the modes and explore larger cross section devices interactively. The mode space method, however, has serious issues in multi-band (**k.p**) and atomistic basis sets, where the quantization in the cross section couples different orbitals. Careful construction of the mode space can be performed for **k.p** models [119-122]. The use of mode space representations in real space tight-binding models has been limited to carbon nanotubes and graphene, where simple p_z-orbitals where used for each atom[123-125]. (A one-dimensional -coupled chain is mathematically identical to a one dimensional *s*-orbital chain.) The mode approach has been shown to be extremely challenging if not impossible to implement systematically for atomistic highly coupled tight-binding approaches [126]. A related reduced-order method not requiring explicit eigenstates has been demonstrated for effective-mass models [127]. Relatively recently Mil'nikov et al. developed a methodology [128] that eliminates spurious eigenvalues in a specific energy window. This method has been used for nanowires MOSFETs [128] and Tunnel FETs [129]. The method was further used in non-orthogonal DFT methods [130] and 2D Ultra-Thin Body (UTB) devices [131]. The Mil'nikov method has been further automated and stabilized such that larger cross sections can be considered and implemented in NEMO5 [132].

If only coherent transport is considered and no quasi-bound, lead-decoupled states exist in the device, then fully atomistic implementations of wave-function-based approaches such as the Quantum Transmitting Boundary Method (QTBM) [90,66] and optimized renormalization [73,74] can scale significantly better than $O[L(NB)^3]$ due to the usage of highly optimized linear solvers. For ultra-thin-body transistors this method has been shown to be at least 4× faster than NEGF in the highly optimized and parallelized OMEN code [95].

1.6.3. Quantum Transport: Parallel Computing Scaling

With QTBM or NEGF the computational cost for quantum transport under high bias is so intense that large supercomputers can be utilized efficiently, if the transport algorithms are implemented efficiently. The art of parallel computing is to maximize the number of computing cores, to keep them busy computing, and to reduce the communication between nodes as much as possible. Communication can be minimized by the design and grouping of independent computing tasks. In an ideal world one would write a scientific program and a compiler would understand the levels of parallelism and distribute the computational load automatically. This has been a dream for a long time and still has not reached practical utility beyond shared memory use in a single CPU with OpenMP for real application programs. Many mathematical libraries use the shared memory OpenMP or threaded paradigm for parallelization within a single chip to about 24 or 64 cores. The Message Passing Interface (MPI) paradigm requires the explicit design and programming of messages between cores or groups of cores. Such programming requires careful design, intense testing, and fine tuning. OMEN was designed to have 4 or even 5 levels of parallelism. At the highest level are independent groups of bias points that do not need to communicate at all until assembly of the final *I-V* curve. The next two levels parallelize the internal transport kernels of momentum (k) and energy (E) integrals. The (outer) k-integral requires the evaluation of (inner) E-integrals for fixed k. At the lowest level a transport kernel needs to be solved as a function of (E,k). That kernel consists of the computation of the open boundary conditions in source and drain and then the subsequent QTBM or NEGF transport equation solution. In QTBM this requires an eigenvalue solution for the boundary conditions and a solution of a linear system of equations. Each of these can utilize standard libraries that have been parallelized in MPI or even OpenMP or a combination of both. The dual-level MPI+OpenMP parallelization has not really delivered better computational performance within OMEN than the pure MPI parallelization. Within the 4-level parallelization OMEN scaled almost perfectly to over 220,000 computational cores [74,95]. In an ideal world of so-called strong scaling, one would divide the compute time by two if the number of computational cores is doubled. This strong scaling results in a straight line of reduced compute time on a logarithmic scale. OMEN delivers almost perfect strong scaling as depicted in Figure 5. Running for 1 hour on 221,400 cores corresponds to about 25.3 years compute time on a single CPU. The OMEN implementation was recognized with the Gordon Bell Prize Honorable Mention as the first Engineering Code running at the Peta-scale in 2011.

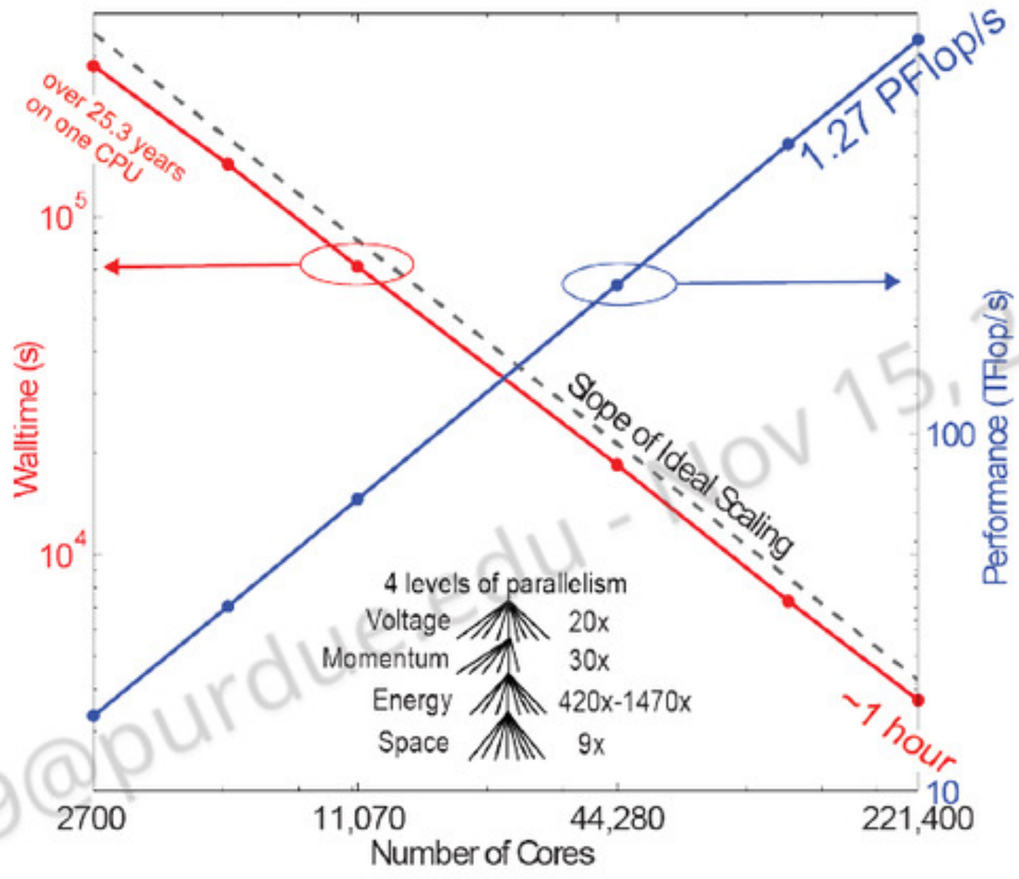

Figure 5: OMEN parallel compute times for an ultra-thin body transistor as a function of number of computational cores on a logarithmic scale. The dashed line indicates the slope of perfect strong scaling. Also indicated on the right axis is the computational throughput in Tera Floating Point

Operations per second. OMEN was the first engineering code to reach the peta-scale [95].

1.6.4. Surface Passivation

The strength of empirical tight-binding methods are the valence electron bonding electronic structure representation within an atomistic representation of the physical geometry. This representation lends itself very well to the numerical description of nanoscale structures. There is, however, one critical pitfall that needs special attention due to finite size structures: dangling bonds.

This issue of dangling bonds first emerged in the NEMO3D work, where InAs QDs embedded in a GaAs matrix were to be simulated. A finite amount of GaAs buffer layer needs to be included to properly model the wavefunction penetration into the buffer layer and to include the effects of spatial distortion to built-in strain. A periodic boundary condition would create a 3D array of QDs which would result in bandstructure and not in discrete states. Furthermore, such a periodic boundary condition requires that the boundary surfaces are perfectly flat and periodic which is typically not the case in the strained and bulged systems. The simulation domain needs to be finite. If the simulation domain is simply terminated then something really interesting can be numerically "observed" in the computed eigen spectrum of this finite system: eigen states localized at the surfaces and edges emerge in the simulation domain with eigenenergies distributed across the whole energy spectrum including the nominal band edge. These surface states are in this case of course an artifact due to the numerical requirement of a finite simulation domain. Naïve attempts to shift the onsite matrix element energies up or down to push the surface states out of the bandgap failed to achieve any generally reliable solution. Lee *et al.* [133] developed a numerical methodology for the typical sp^3s^* and $sp3d^5s^*$ models through a basis transformation into the direction of the dangling bonds and adding an adjustable energy shift to that bond energy. This method results in a very robust and stable method to eliminate the effects of the dangling bonds and has been highly adopted and cited in the literature.

Other, non-embedded, nanostructures actually do have exposed surfaces that require a numerical passivation. The passivation method in reference [133] has found very wide use in the field manage the surface states that naturally emerge in the tight-binding calculations. Twelve years later the NEMO5 team generalized this passivation method [134] and linked it to chemical bonding. For the very thinnest active regions, however, explicit hydrogen passivation is often the most accurate method. Explicit passivation parameters have been determined by mapping ab initio wavefunctions and bands to tight-binding Hamiltonians[48].

1.7. Applications

1.7.1. Quantum Dots: Closed Systems

Quantum dots (QDs) enable designers to custom-create optical transitions at specific wavelengths or confine electros at specific spatial locations with specific energies. Therefore, QDs are sometimes also called "artificial atoms". These systems are studied very widely in the literature. To highlight the enabling capabilities of tight-binding we just highlight a few quantum dot applications here[135].

Figure 6 depicts graphically appealing an InAs QD on an InAs wetting layer on a GasAs substrate, capped by an InGaAs capping layer. With such carefully engineered designs one can achieve optical emission at 1.5μm for optical communication. A detailed description of the required modeling and the understanding of the design space is given in [5].

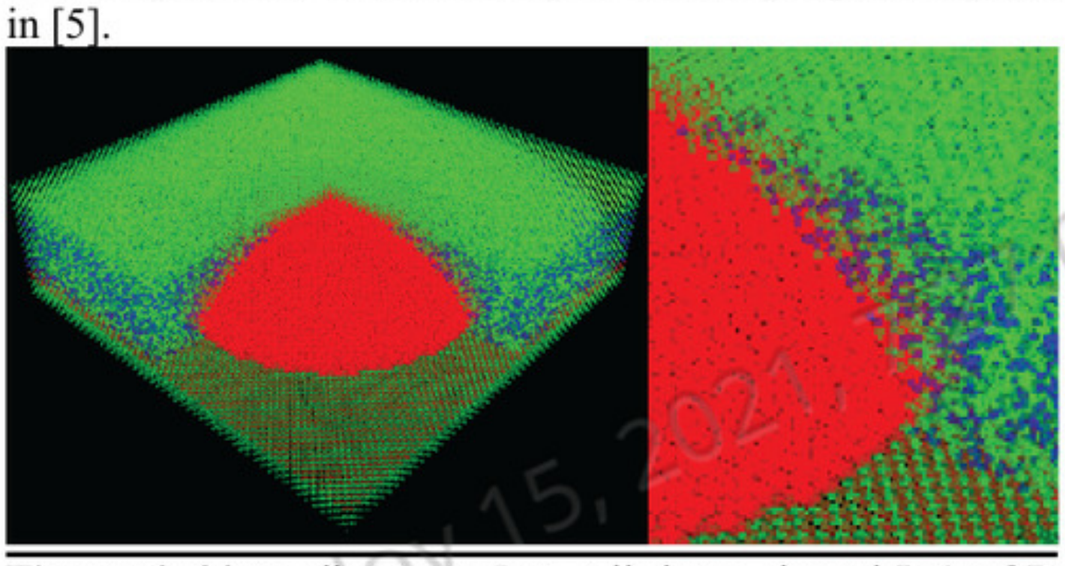

Figure 6: 20nm diameter, 5nm tall dome shaped InAs QD embedded in an AlGaAs buffer layer grown on top of a GaAs substrate[5]. Image is accessible also on nanoHUB [136].

A typical dome shaped QD of 20nm diameter and 5nm height might nominally only contain around 100,000 atoms. However, such QDs are formed by significant crystal lattice mismatches in the GaAs and InAs system via the Stranski-Krastanov growth method [2]. As such these systems have a significant built-in strain that reaches far beneath the QD into the substrate. Cutting off the finite, closed system simulation domain too small results in significantly false eigenvalues with variations in conduction band states of over 150meV and valence band states over 40meV underestimating the gap by over 20% [107]. Figure 7 taken from Reference [3] shows that eigenvalue convergence and proper state polarization significantly depend on the size of the simulation domain. In Figure 7(b) the substrate is extended down to 50nm underneath the QD. Eigenstate tuning can be controlled by the capping layer as indicated in Figure 7(b).

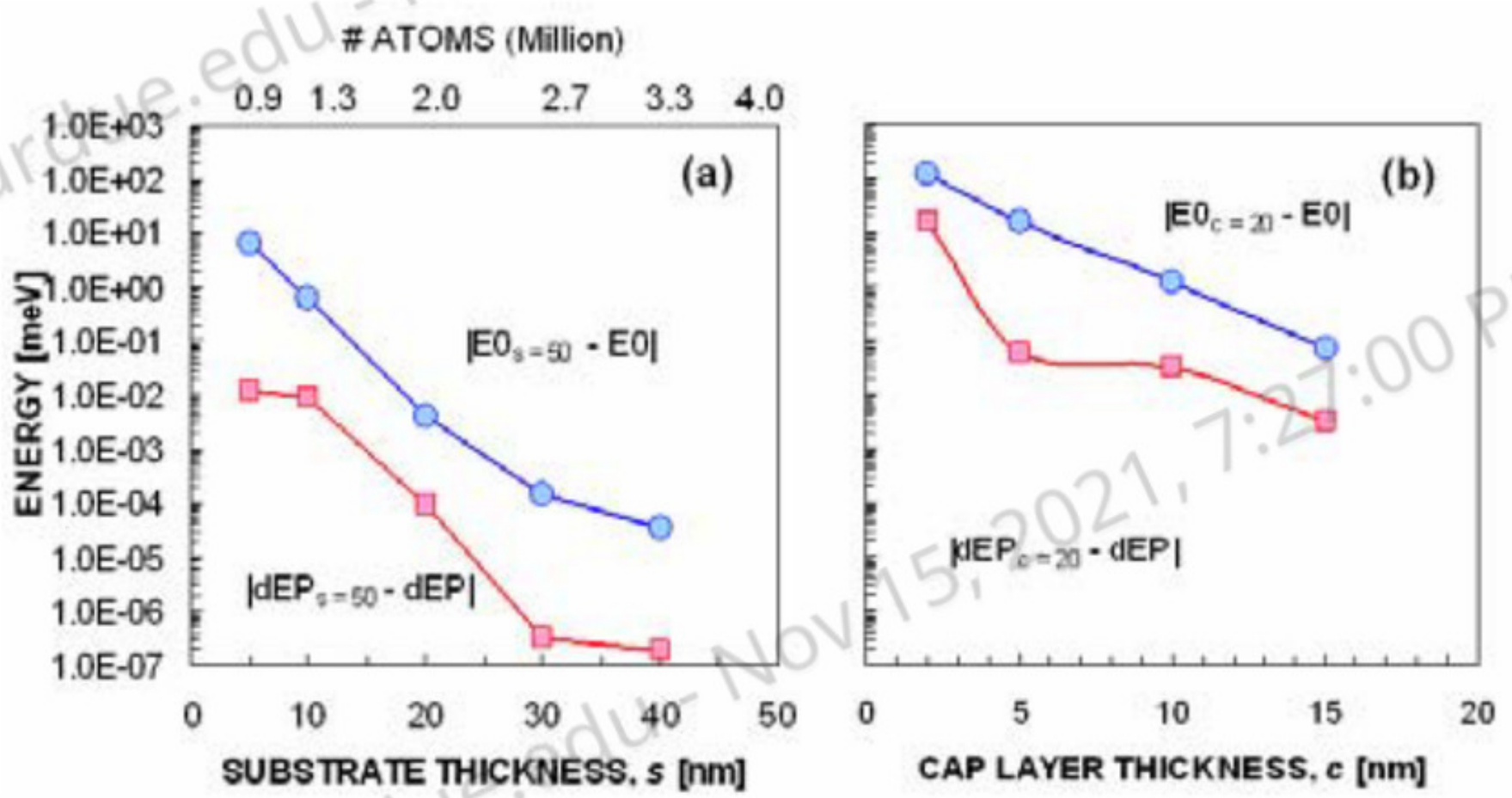

Figure 7: From [3] (a) Substrate layer thickness dependence of the conduction band minimum and the P level splitting. Other structural parameters remain constant (h = 5.65 nm, d = 11.3 nm, c = 10 nm, and D = 31.3 nm). (b) The impact of cap layer thickness (with substrate, s = 30 nm and other structural parameters remaining the same). Lanczos convergence tolerance = 1×10^{-7}. The changes in both these quantities are calculated with respect to the largest s (50 nm) and c (20 nm) respectively.

1.7.2. Nanowire Electronic Structure: Quasi-Periodic in 1D and Closed Systems in 2D

Nanowires as depicted in Figure 8 are of significant technical relevance for transistor scaling at the end of Moore's Law. The last generations of CMOS-based transistors in the so-called 5nm node might be vertical stacks of such nanowires with wrap-around gate control.

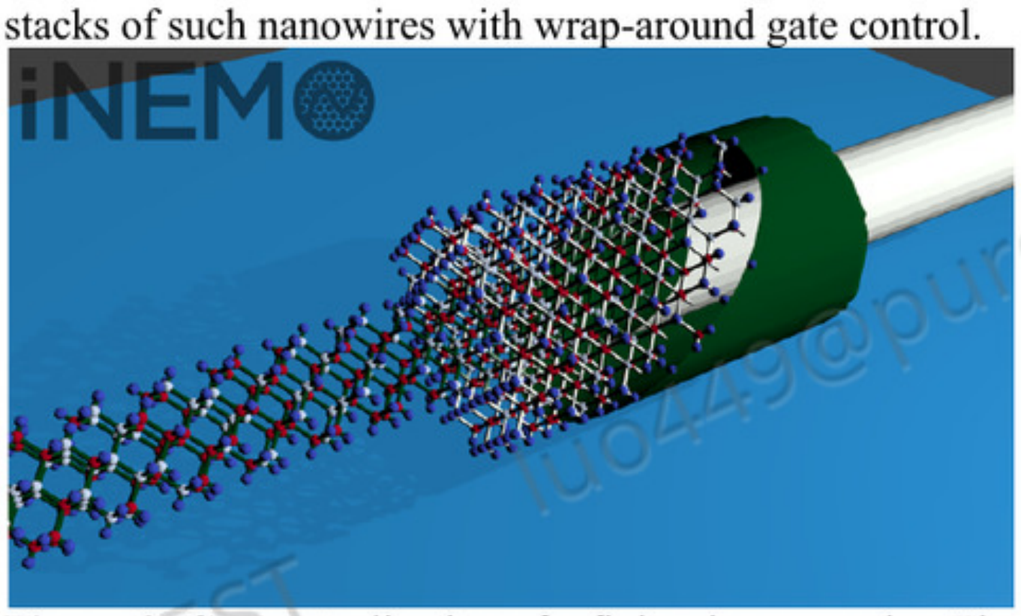

Figure 8: Conceptualization of a finite size nanowire. The difference between a typical continuum representation (right end) and an atomistic representation (left end) is eye opening. Figure published online [137] created by Dr. Daniel Mejia in the NEMO5 development team.

Figure 4 above depicts unit cells of two different 2.5nm thick nanowires in the [100], [110], and [111] transport directions. If periodic boundary conditions are applied in the transport direction, then a dispersion in the transport direction can be computed with confinement in two dimensions. Band folding and band-to-band interactions result in complex new bandstructures due to the confinement as discussed extensively in the literature. For early tight-binding examples see references [138-141].

Visual inspection shows that the dispersions shown in Figure 9 for [100], [110], and [111] Si nanowires vary dramatically. The [100] wire has four almost degenerate minimum band-edges at the Γ point with an effective mass of 0.31. The [110] wire has two almost degenerate bands of very light masses 0.14 and 0.18 at the Γ point. The [111] wire has three minimum band-edges off the Gamma point with an effective mass, two with masses m*= 0.48, and one with m*= 0.68. At the Γ point the [111] oriented wire has a negative effective mass of -0.26, effectively repelling carriers back into the source. Clearly one can expect the [100] and [110] nanowires to conduct current much better than the [111] oriented wire.

The four lowest [100] wire bands are almost degenerate with a mass of 0.31. In contrast, the two lowest bands of the [110] wire have effective masses 0.14 and 0.18. Determining which one of the two nanowires has "better" electron transport requires more quantitative transport models.

At the macroscopic scale bandgap and effective mass are determined solely by bulk material properties. At the nano-scale geometry, surfaces, and crystal orientation are equally important. Consequently, electron transport in nanowires varies significantly as a function of geometry. This discussion demonstrates conclusively that *effective masses and bandgaps can be designed.*

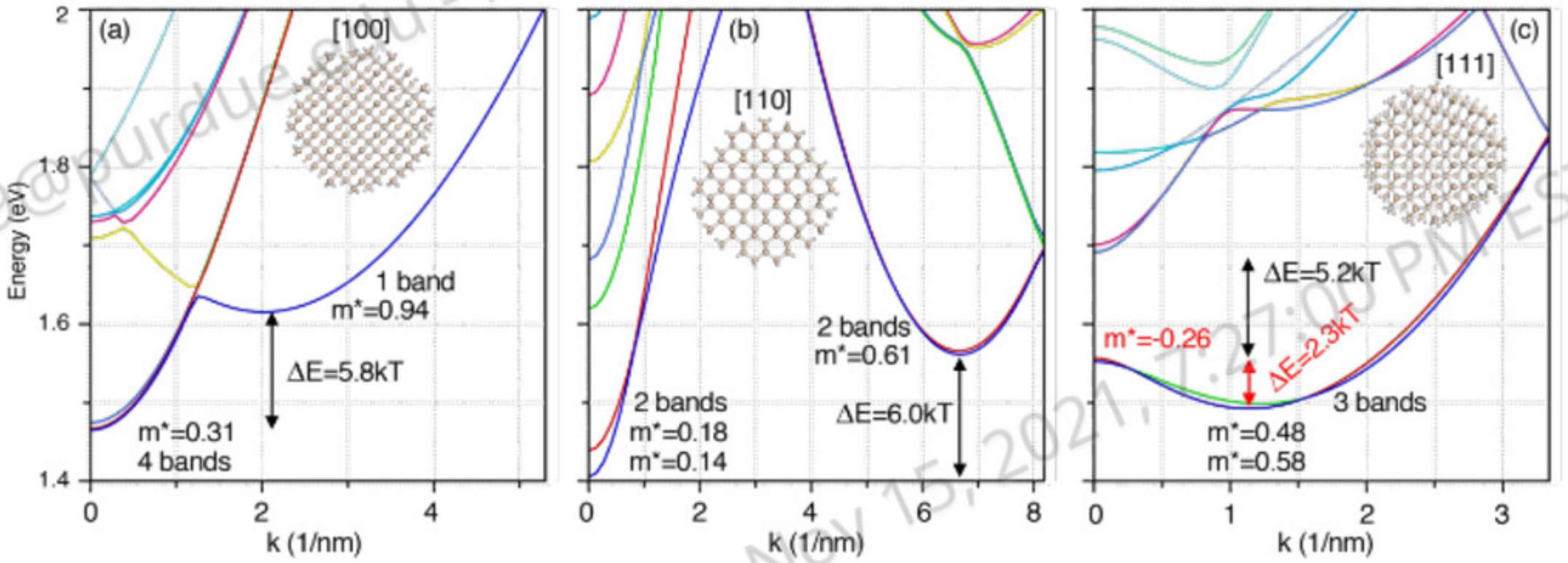

Figure 9: Dispersion in the transport direction of infinitely long, ungated, and hydrogen passivated 2.1nm Si nanowires in the (a) [100], (b) [110], (c) [111] transport directions. (a) Four bands with m*= 0.31 (units are the free-electron mass, m_0) at Γ are separated by about an energy of $5.8k_BT$ (at room temperature) from the next band with a heavier mass of m*= 0.94. The heavy band with m*= 0.94 is not occupied at the source injection of a transistor. (b) Two very light effective mass bands of m*= 0.14 and m*= 0.18 at Γ, are separated by $6k_BT$ in energy from the next conduction band. (c) Three bands with the minimal conduction band off Γ, two with masses m*= 0.48 and one with m*= 0.68. These bands are close in energy of $2.3k_BT$ of hole-like masses of m*= -0.26. Transport in the [111] direction experiences a significantly heavier mass and even carrier rejection due to the negative mass.

1.7.3. BALLISTIC TRANSPORT FROM NANOWIRE DISPERSIONS WITH THE TOP-OF-THE-BARRIER MODEL

The beautifully simple top-of-the-barrier (TOB) model developed by Lundstrom [142] can provide intuitive insight into the effects of material, geometry, and strain on the transport in nano-scaled transistors. Significant work has been published using the top of the barrier model to examine and illustrate: atomistic vs effective mass models in Si nanowires [143]; injection velocity in ultra-thin-body Si [144] and III-V MOSFETs [145]; geometry effects on electrons [146] and holes [147,148] in Si nanowires (as discussed above). The early application of the top-of-the-barrier model considered just the dispersion diagram without consideration of the effects of electrostatics. Later implementations as now available in NEMO5 or *Bandstructure Lab* on nanoHUB [108] consider also the electrostatic effects in the device which self-consistently relate bandstructure and local potential. Analytical extensions even help with estimating the effects of tunneling through the gated region with an "under-the-barrier model" [149].

To estimate the transistor performance of the three nanowires considered here we utilize the TOB model implemented in the nanoHUB tool *FETtoy* [150]. *FETtoy* is configured with a small set of model parameters that a reminiscent of a compact model. The material parameters are entered as effective masses, band degeneracies, and dielectric constants. Multiple non-degenerate bands cannot be entered into the published version of *FETtoy*, but can, of course, be considered in the top-of-the-barrier model. For simplicity we choose the lowest lying bands to be degenerate and average out the masses. For the three considered wires we used: [100]: m*= 0.31 4-fold degenerate, [110]: m*= 0.16 2-fold degenerate, and [111]: m*= 0.51 3-fold degenerate. Shape and size of the nanowire can are also entered. A 1nm thick gate dielectric is chosen, with dielectric constant 3.9. The TOB model can estimate effects of drain-induced barrier lowering (DIBL) and the degradation of the sub-threshold swing with a gate control parameter α_G and a drain control parameterα_D. For clarity of the arguments we set $\alpha_G = 1$ and $\alpha_D = 0$ ignoring these gate issues and focus on the ideal, long wire performances.

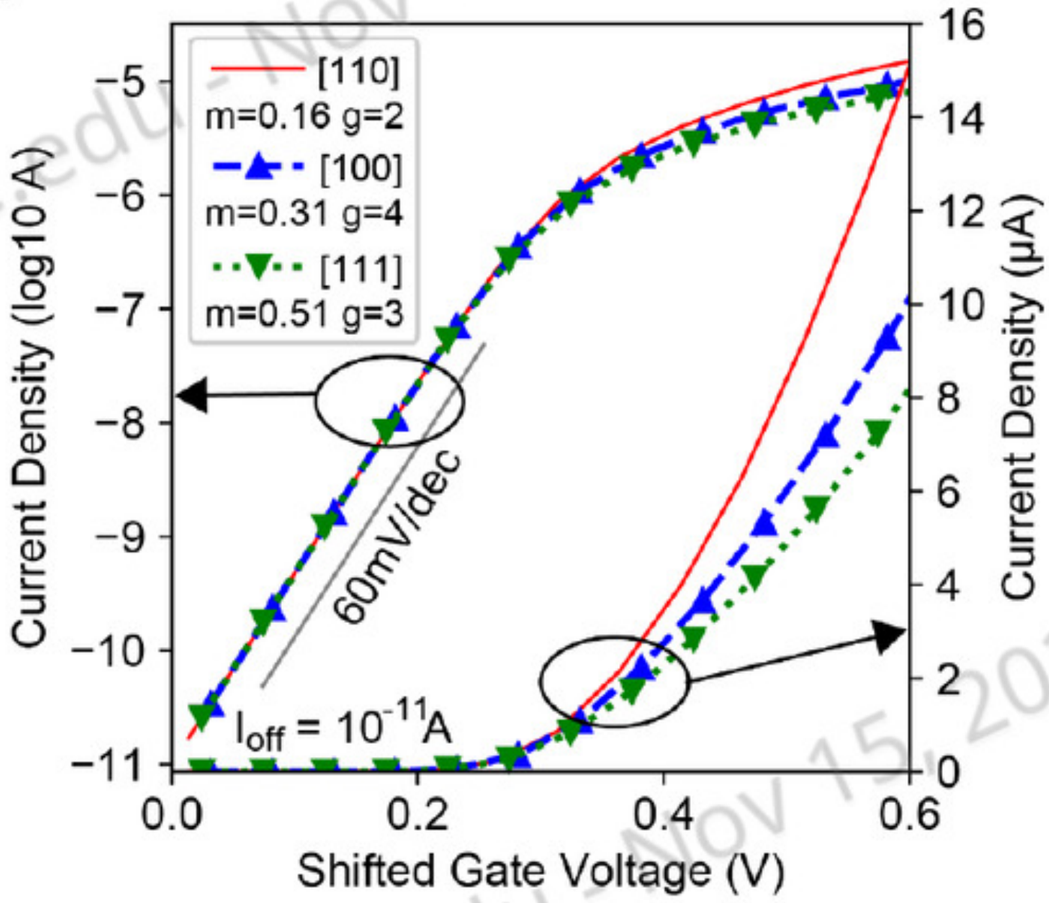

Figure 10: Drain current versus gate voltage computed with the top-of-the barrier model implemented in the nanoHUB tool *FETtoy* [150]. Effective masses and band degeneracies taken from atomistic dispersions. The [110] wire delivers the highest ON current and the [111] wire delivers the lowest ON current.

Figure 10 depicts the drain current versus gate voltage computed within the top-of-the barrier model (TOB) as implemented in the nanoHUB tool *FETtoy* [150]. The left vertical axis is on a logarithmic scale and the right vertical axis on a linear scale for the drain current. The tool allows the user to set a threshold voltage which effectively adjusts

the OFF current. We have taken the three current voltage characteristics for the three wires and shifted them to the same OFF current $I_{OFF} = 10^{-11}$A. Since we consider ideal wires and neglect electrostatic effects along the channel all three transistors behave perfectly in the OFF region and deliver a subthreshold swing of 60mV/dec. The three wires behave differently in their delivered ON current. As expected, the [111] wire delivers the lowest ON current. The low effective mass [110] with just 2 degenerate bands provides a higher ON current than the 4-fold degenerate [100] nanowire with a mass that is roughly doubled compared to the [110] wire.

With these performance estimates for the three wires we emphasize here again that at the nano-scale geometry and crystal direction modify parameters such as bandgap and effective masses that in the past we considered material constants. *Bandgap and effective mass have become design choices.*

From the dispersion alone and the TOB significant insights can be gleaned about the transport properties of the nanowires. These calculations are computationally relatively cheap as they principally involve the computation of a set of eigenvalues of a relatively small unit cell as depicted in Figure 4. Full quantum transport simulations are significantly more expensive and are discussed in the next section.

1.7.4. Full Quantum Transport in "Long" 15nm Nanowires: 3D representation

The top-of-the-barrier model is an excellent choice to screen for the fundamental effects of a variety of crystal directions and confinement effects. The TOB model has the ability to include some electrostatic effects to estimate DIBL and a subthreshold degradation. Analytical extensions might even help with estimating the effects of tunneling through the gated region with an "under-the-barrier Model" [149]. A full TCAD based, 3D simulation of a wire, however, requires a full 3D representation of the carrier charges, gates, tunneling processes, and relaxation in the reservoirs. One of the first full 3D implementations of atomistic quantum transport was performed by Luisier in the OMEN code [66,74,151]. Luisier continues to develop his OMEN code at ETH and Klimeck et *al.* continue to develop NEMO5 [104,105] at Purdue. Within the scope of these modeling capabilities different cross section shapes, materials and stress conditions have been explored in a variety of materials systems such as Si, graphene, SiGe, III-V, and 2D materials.

OMENwire is a tool published in nanoHUB [151] powered by the OMEN simulation engine. *OMENwire* can easily be configured through a graphical user interface for different wire and ultra-thin-body transistors. *OMENwire* runs on parallel machines typically for an hour or so to deliver the results in a fully graphical and interactive form. Different simulations for user choices such as geometries, crystal orientation, and material properties can be run and compared interactively. The results presented in this section have been generated with *OMENwire* code [151] and anyone can duplicate these results on nanoHUB.

We consider the same nanowire geometries as depicted in Figure 4 above: 2.5nm diameter Si in [100]. [110], and [111] crystal direction. 10nm source and drain at 2×10^{20} doping. 15nm gate length with 1nm EOT oxide. The drain voltage is set to 0.6V and the gate voltage is ramped from 0V to 0.6V. We use the default parameters of the dielectric constant of the Si channel (11.9), the channel material affinity (4.05eV), dielectric function of the oxide (3.9), and the gate contact workfunction (4.1eV). Figure 11 shows the drain current as a function of gate voltage on a logarithmic scale (left vertical axis) and a linear scale (right vertical axis).

Figure 11(a) depicts the direct numerical results as delivered by *OMENwire* given the specific geometries. Figure 11(b) shifts the three performance curves to the same OFF current $I_{OFF} = 10^{-11}$A. We note that the [111] wire simulations do not converge to meaningful results for gate voltages above 0.4V as further discussed below.

The threshold-shifted performance curves in Figure 11(b) show that for the long gate length of 15nm the three nanowires perform perfectly identical at 60mV/dec in the subthreshold regime. With full 3D electrostatics the three wires also perform virtually the same in the ON region in this long channel device with ballistic transport. We note here that this result of virtually identical performance is different from the analytical TOB result depicted in Figure 10.

Figure 11(a) shows the specific performance curves without adjustment of the threshold voltage. The [110] wire delivers the highest current and the [111] wire delivers the lowest current. Understanding the different performances requires a deeper understanding of the bandstructure, density of states and charge densities in the different wires.

Figure 12 shows a composite picture of localized band edge (black line), bandstructure in the source (green lines), bandstructure in the drain (orange lines), normalized current density (blue line), and transmission coefficient (brown line) for the three considered nanowires with a source-drain voltage of 0.6V and a gate voltage of 0.3V. Each of these plots is available in *OMENwire* [151]. The rather different bandstructures discussed in Figure 9 above are evident resulting in rather dramatic transmission coefficients. The normalized current density under high bias is the product of the transmission and the Fermi function in the source. This composite picture Figure 12 illustrates the relative energy scales and spatial positions in the nanowires. The different effective masses in the three wires combined with the same doping level results in three very different Fermi levels in the source and the drain, as indicated by the horizontal red lines. For the very heavy mass [111] wire the Fermi level is just under the conduction band by 5.4meV while the [100] and [110] Fermi levels are above the conduction band by 26.1meV and 98.6meV, respectively. From a classical transport perspective, we can now look at the maximum barrier height in the channel measured against the Fermi level in the source. For the [111] wire that energy distance is 141meV or about 5.5k_BT. For the [100] and [110] wires this energy difference is much smaller at 85meV and 69meV, respectively, corresponding to about 3.3k_BT and 2.7k_BT. With about 3k_BT higher barrier in the [111] wire one can expect a factor of e^3, so about 20x lower current in the [111] wire, regardless of transmission coefficient details. The

simulated I_{OFF} difference in Figure 11(a) is about a factor of 10x indicating that the Fermi level difference estimate is off by a factor of 2. The I_{OFF} difference in Figure 11(b) between [110] and [100] is a factor of 8.8/2.5 ≈ 3.5×. The Fermi level energy difference to the top of the barrier is just 0.6k_BT corresponding to a Fermi tail difference of $e^{0.6}$ or a factor of 1.8×. Also this Fermi level estimate difference is off by a factor of 2. The difference in current between the [100] and [110] requires some more detailed understanding.

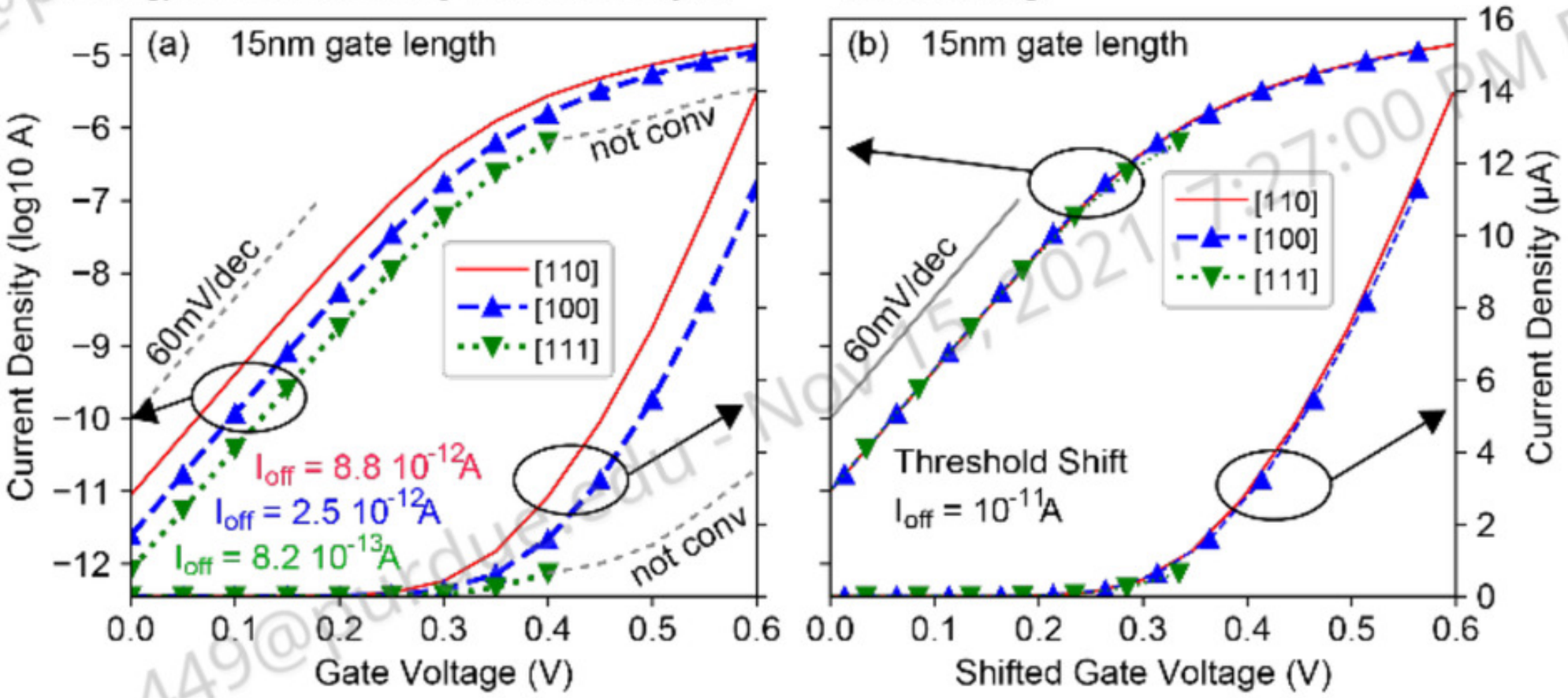

Figure 11: (a) $I_d - V_g$ performance curves of the discussed 2.5nm diameter [100], [110], and [111] nanowire transistors. Source-Drain bias is 0.6V. The gate length is long at 15nm. (a) Simulation results with specific gate work function value of 4.05eV. The [110] wire delivers the highest current and the [111] wire delivers the smallest current. (b) Performance curves of (a) shifted to a common OFF current 10^{-11}A. All transistor curves are virtually identical.

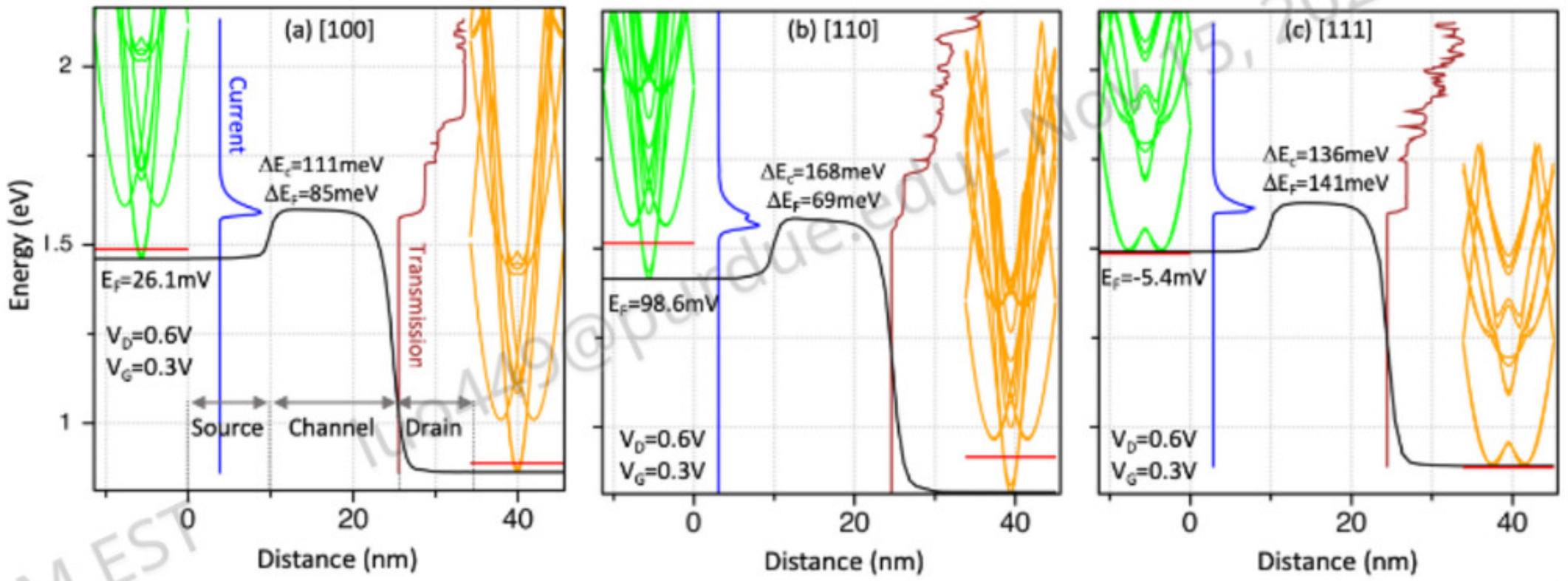

Figure 12: Localized band edge (black line), bandstructure in the source (green lines), bandstructure in the drain (orange lines), normalized current density (blue line), and transmission coefficient (brown line) for the three considered nanowires with a source-drain voltage of 0.6V and a gate voltage of 0.3V. (a) [100] wire, (b) [110] wire, and (c) [111] wire. The Fermi levels in the source are indicated as E_F, the maximum barrier height in the channel measured against the conduction band and Fermi level in the source as ΔE_C, ΔE_F, respectively.

The normalized current density and transmission coefficients shown in Figure 12 are depicted for more clarity and ease of comparison in Figure 13 again (taken directly from the *OMENwire* code) [151]. The figures are augmented by the bandstructure information in Figure 9, which measures the number and effective mass of the lowest bands. For each available injection band one can see a close to step like feature to a discrete value of transmission as expected. As discussed above in the context of effective masses of the three different wires one can expect the [111] wire to conduct the current less well than the other two wires due to its much larger effective masses. As discussed above the [100] wire has a nearly 4-fold degenerate band set with effective mass of 0.31 while the [110] wire has two (2) bands close to degenerate with masses 0.14 and 0.18.

The transmission coefficient of the [111] wire with its 3 lowest bands with effective mass 0.48 shows a rather steep rise at the band edge to a value of 3. The [100] wire has with its four lowest bands of effective mass 0.31 has a more gradual slope to a maximum transmission of 4 above the band-onset. The [110] wire has the broadest onset of the transmission and the two bands that are not quite degenerate result in a transmission step to values of 1 and 2. These different slopes of the transmission coefficients

have a significant effect on the three different normalized current densities. Each of the three plots in Figure 13 show a dashed vertical line labeled with E_{TOB} corresponding to the top-of-the barrier conduction band edge in the channel under 0.3V gate bias. Even for this large gate length geometry of 15nm *there is a significant tunneling current flow under the conduction band edge*. Just by visual inspection it appears that the [110] wire has the largest energy range in its tunneling current contribution. The tunneling current alone does not, however, yet explain the differences in the current between the [100] and [110] wire.

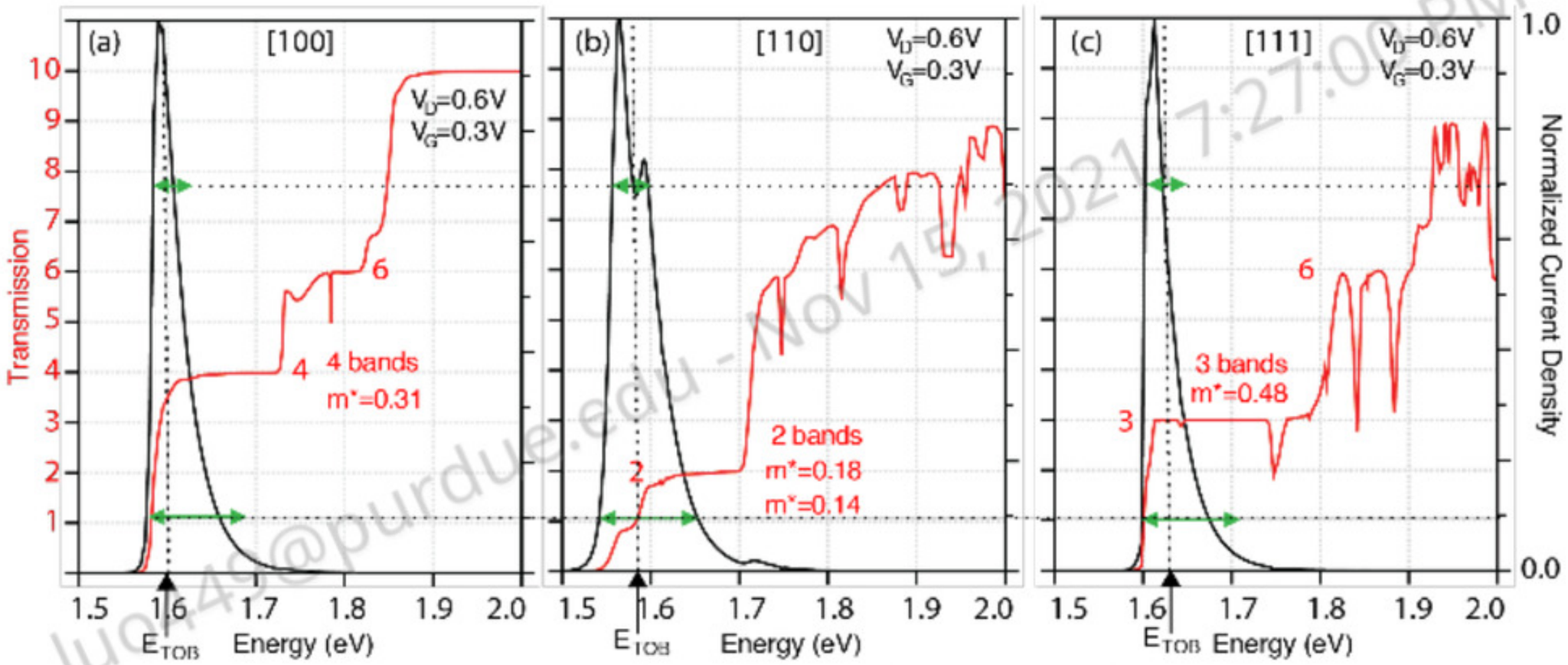

Figure 13: Transmission (red lines) and normalized current density (black lines) for the three considered nanowires with a source-drain voltage of 0.6V and a gate voltage of 0.3V. The effective masses of the lowest bands correspond to the values in Figure 9. The vertical dashed line labeled with E_{TOB} indicates the energy of the top of the barrier inside the channel. The green double arrows in (b) indicate the energetic width of the current flow at a normalized value of 10% and 70%. These two green vectors are also inserted in (a) and (c) indicating that the energetic width of current flow in the [100] and [111] wires is much narrower than the [110] wire.

Two horizontal arrows in Figure 13(b) "measure" the energetic width of the current density at a level of 10% and 70% of the peak value of the normalized current. These two arrows are transposed in the figures for the [100] and the [111] wire showing that the [110] wire transmits current over a significantly wider energy range.

Figure 14 examines in more detail why the [110] wire has a much wider energetic range in its current density. Figure 14(a) basically repeats Figure 12(b) for a [110] wire, where, however, the gate bias is reduced to 0.15V, further into the OFF region. Clearly there is a lot of current flowing under the conduction band edge in the center of the device. Figure 14(b) superposes in addition the local density of states (DOS) in the simulation domain. It is rather evident that there is a significant DOS under the nominal conduction band edge. In the source and drain the DOS aligns with the bandstructure of an ideal wire. However, in the central device region the bandstructure really does not align perfectly anymore with the local DOS. There is a significant DOS in the bandgap and also the state spacing is (slightly) different to the additional OFF-state electrical confinement in the middle of the channel.

The conclusive reason why the [110] wire delivers a higher current than the [100] wire is the interplay of electrostatic potentials and tunneling through the gate-controlled potential.

1.7.5. Full Quantum Transport vs. An Analytical Model in "Long" 15nm Nanowires

The full quantum transport simulations presented above provide significant insights into the device physics of the 2.5 nm thin cross section wires. Tunneling and modeling of the detailed potential are critical in the understanding of the device performance. It is, however, particularly interesting to see that the overall device characteristics for the long 15nm gate length the wires in [100], [110], and [111] directions perform virtually the same when they are shifted to the same OFF current (see Figure 11(b)). Not only are the subthreshold performance results the same at 60mV/dec, but also the ON currents are virtually the same. Figure 15 compares the full 3D quantum transport simulation results threshold-shifted from Figure 11(b) and the analytical top-of-the-barrier (TOB) results depicted in Figure 10 for the [100] and [110] wires. The TOB model is configured in the *FETtoy* simulation on nanoHUB to ignore any assumptions of short channel effects. These simulations are agnostic of a channel length and the results are shown in solid lines. The full 3D *OMENwire* simulations are labeled in dashed lines with symbols. The subthreshold behavior of the considered cases is identical performing at 60mV/dec. Only the linear scale shows that the TOB model predicts slightly different ON currents for the two wires, while the full 3D model predicts virtually identical ON currents. If we were to consider the full 3D model as ground truth, then the TOB model underpredicts the [100] wire ON current by about 30%. There are plenty of uncertainties in the full 3D quantum transport model and one

is compelled to ask: "are the hours of CPU time consumed in full quantum transport simulations worth it?" The next section will show the strength of a full 3D quantum transport model in a regime, where the analytical TOB model really breaks down – going from "long" 15nm channel lengths to 5nm.

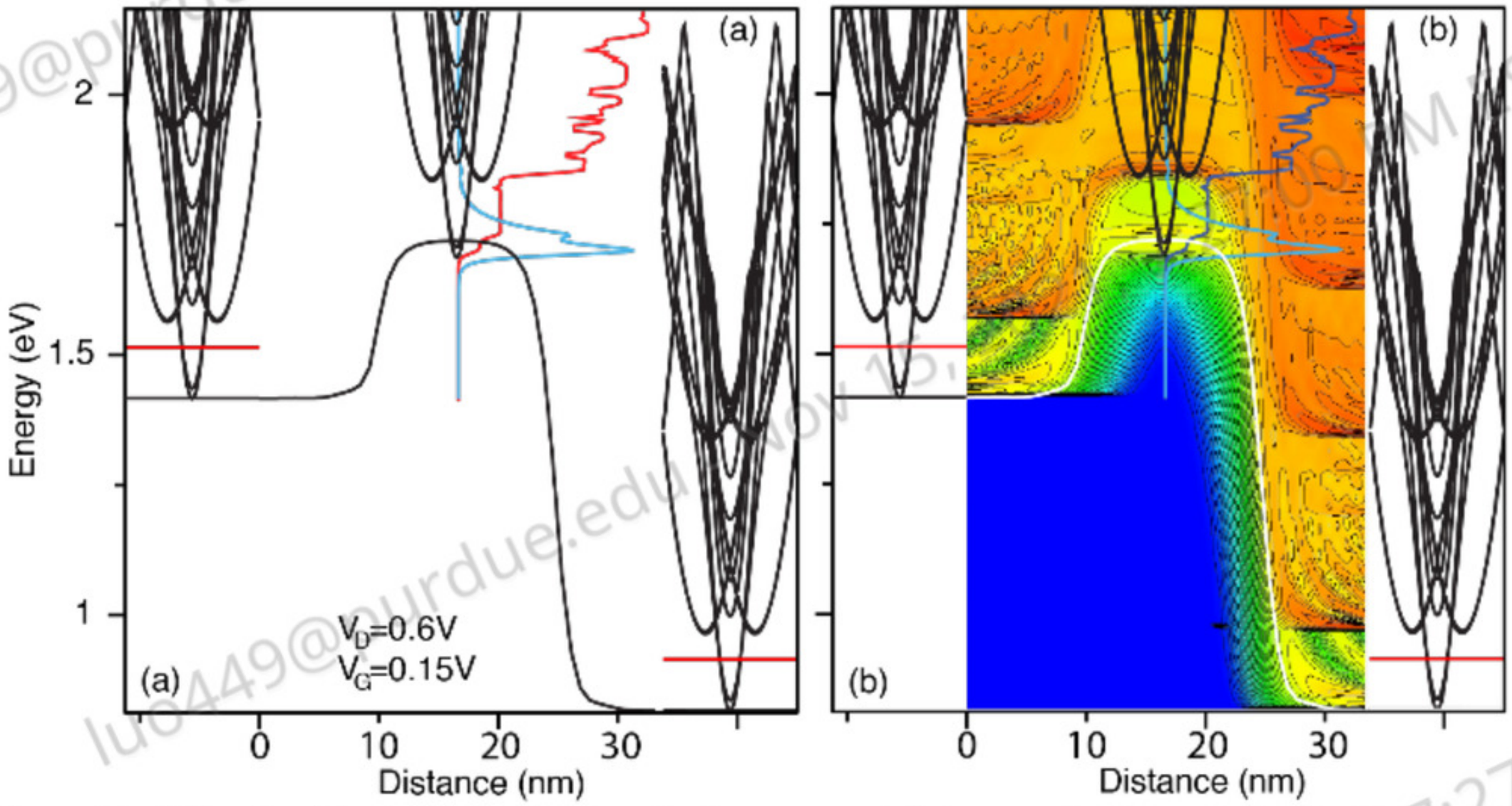

Figure 14: (a) Localized bandedge, bandstructure in the source and drain, normalized current density (blue line), and transmission coefficient (red line) for the [110] nanowire similar to Figure 11(b). The specific gate bias here is different – it is further in the OFF state at 0.15V. The source/drain bandstructure is also superposed into the middle of the channel and aliged with the second set of bands. (b) Same as (a) but the local density of states (DOS) is superposed in the quantum device simulation region.

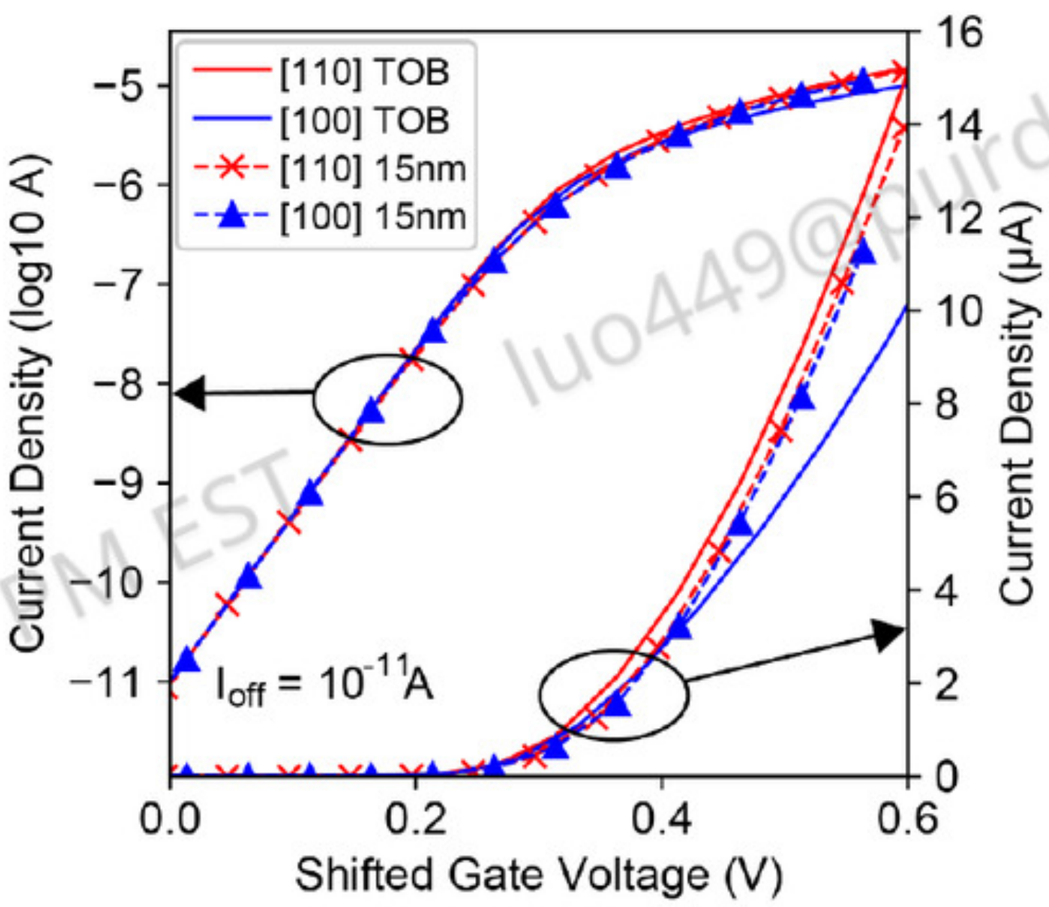

Figure 15: Comparison of the TOB and full 3D quantum transport calculations. Solid lines are the results with the analytical top-of-the-barrier (TOB) model shown in Figure 10 for [110] and [100] wires. The dashed lines with labels correspond to the full 3D OMEN quantum transport results shown in Figure 11(b) for a 15nm gate length wire. All curves are threshold voltage shifted to the same off current at 10^{-11}A.

1.7.6. Full Quantum Transport in Short 5nm Nanowires: 3D representation

The full quantum transport simulations presented above provide significant insights into the device physics of the 15nm gate length nanowires with 2.5 nm cross sections. In this section we consider a gate length of 5nm and keep all other device parameters such as doping source/drain extension, gate oxide etc. the same. Figure 16(a) depicts the drain current as a function of applied gate voltage for a source drain-voltage of 0.6V. Just as before for the 15nm gate length wire, here also for 5nm gate length the [100] wire delivers the highest current, followed by the [110] wire and the [111] has the lowest current for the same workfunction specifications. However, Figure 16(a) shows that the three different wire orientations show very different subthreshold swings (SS), far worse than the ideal 60mV/dec. The high-current [110] wire shows the worst SS at 119mV/dec. For zero gate voltage its OFF current is 2 orders of magnitude higher than the [111] OFF current. The ON-OFF swing for the high current [110] wire in the 0.6V gate sweep is less than 3 orders of magnitude. For typical circuits the typical requirements is a ON-OFF swing of 4 or 5 orders of magnitude. *While the [110] wire delivers the highest current, it fails to turn off effectively.*

Figure 16(b) depicts the same performance curves as (a) except thresholds shifted to an OFF current of 5×10^{-11}A. The [110] wire is shifted out of range as it does not deliver such low OFF current with the chosen workfunction in the simulation. Possibly some

workfunction configuration could turn off the [110] wire enough to reach 5×10^{-10}A. The [100] and [111] wires roughly perform the same in the threshold region at an SS of 81mV/dec and 79mV/dec, respectively. The [100] wire delivers about twice as much ON current compared to the [111] wire.

Figure 17 explains the origin of the very different OFF current behaviors of the three different wires. Figure 17 depicts the local conduction band edge and density of states for the three different nanowires at zero gate voltage and 0.6V source-drain voltage. In each of the three wire source regions the Fermi level is indicated with a horizontal red line and an energy window of $5k_BT$ above the Fermi level is indicated with a gray box. The [110] has the highest Fermi level and the [111] wire has the lowest Fermi level due to the very different effective masses, as discussed above. The normalized current density as a function of energy is indicated as a blue line and the red line indicates the transmission coefficient.

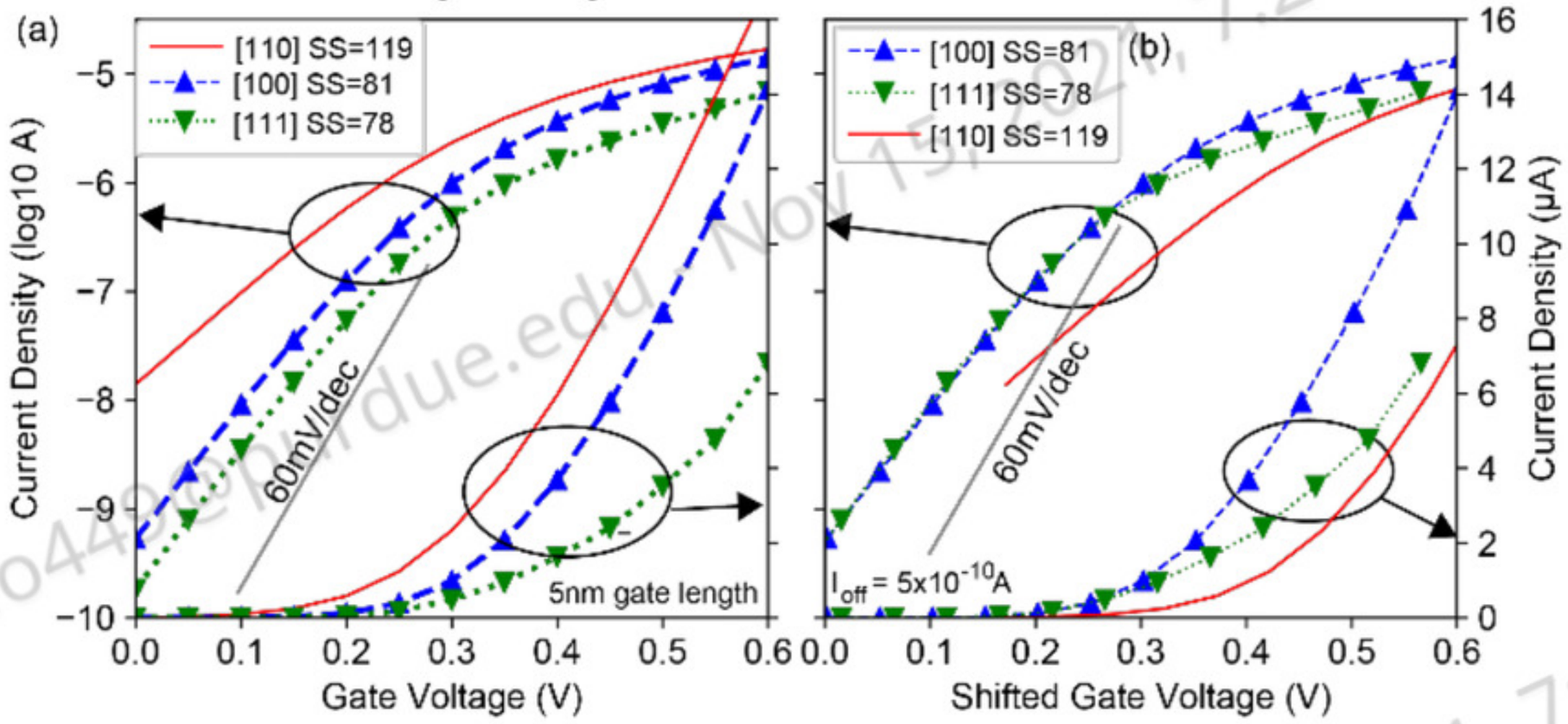

Figure 16: I_d -V_g performance curves of the 2.5nm diameter [100], [110], and [111] nanowire transistors with 5nm gate length. Source-Drain bias is 0.6V. (a) Simulation results with specific gate work function value of 4.05eV. The [110] wire delivers the highest current and the [111] wire delivers the smallest current. The [110] high current wire fails to turn off effectively and has a subthreshold swing of 119mV/dec. The ON/OFF current swing is less than 3 orders of magnitude. (b) Performance curves of (a) shifted to a common OFF current 5×10^{-10}A . The [100] wire delivers the best performance.

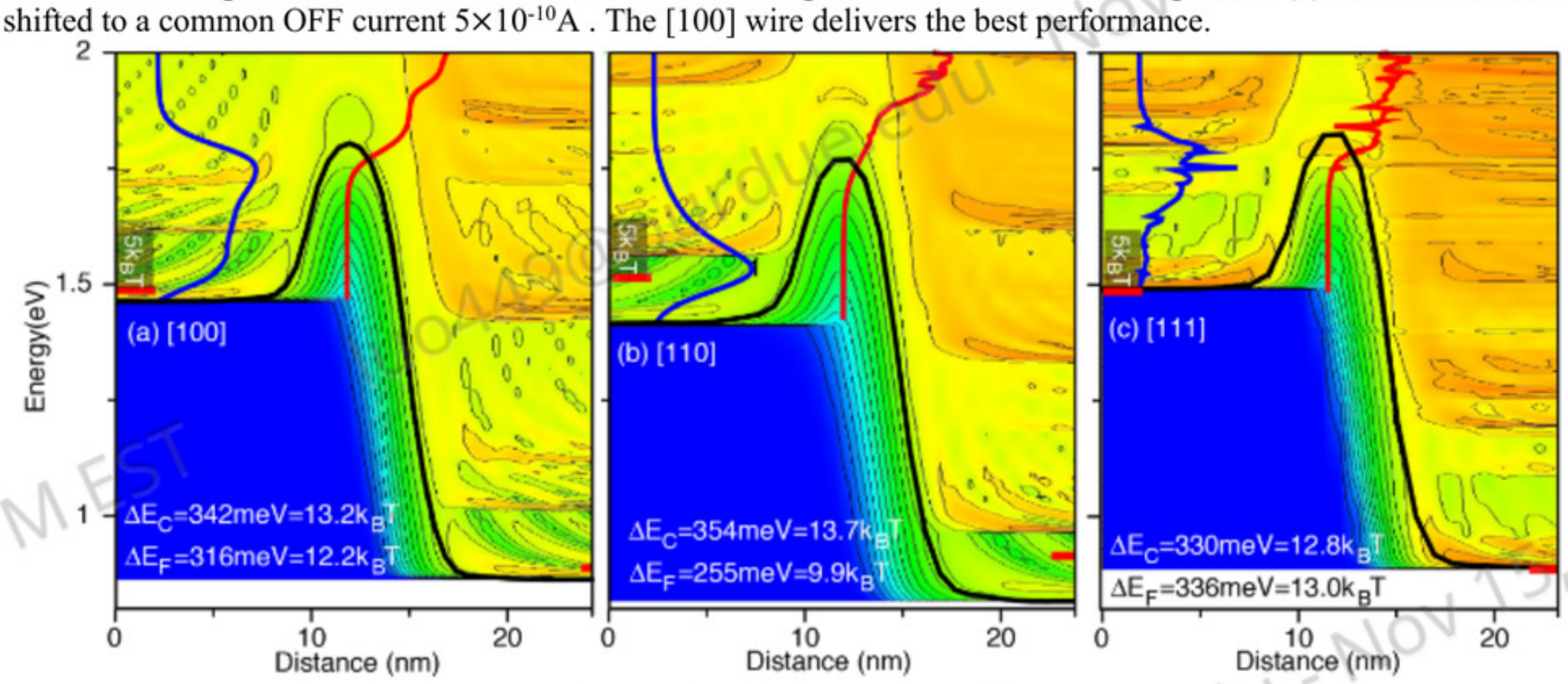

Figure 17: Local conduction band edge and density of states for the three different nanowires at zero gate voltage and 0.6V source-drain voltage. In the wire source regions the Fermi level is indicated with a horizontal red line and an energy window of $5k_BT$ above the Fermi level is indicated with a gray box. The blue line indicates the normalized current density as a function of energy and the red line indicates the transmission coefficient. ΔE_C and ΔE_F indicate in each figure the energy difference between the source conduction band and the Fermi level with respect to top of the barrier in the channel. There is a significant density of states available under the nominal conduction band edge resulting in significant tunneling. The [110] wire (middle) has virtually all of its normalized current flow within the $5k_BT$ energy range around the Fermi level. The [100] wire some current flow within that $5k_BT$ energy range, but most of the current flows above. The [111] wire has virtually all its current flow above the $5k_BT$ window.

There is a significant density of states available under the nominal conduction band edge resulting in significant tunneling. The [110] wire (middle) with the very high Fermi level has virtually all of its normalized current flow within the $5k_BT$ energy range around the Fermi level. The central barrier height measured against the Fermi level in the source

is $9.9k_BT$. The barrier is nominally high enough if this were a classical transport problem. Clearly the OFF current in the [110] wire with its light effective mass is completely carried by tunneling.

The [100] wire shows some normalized current flow within that $5k_BT$ energy range, but most of the current flows above. Since the availability of injected carriers above the $5k_BT$ window is exponentially lowered one can clearly see why the [100] wire OFF current in much lower that the [110] wire OFF current. The [111] wire has virtually all its current flow above the $5k_BT$ window. In both [100] and [111] wires the exponential carrier fall-off above the Fermi function is compensated by the exponential increase in the tunneling below the nominal barrier in the gate region which results in some tunneling current flow. In these two wires, [100] and [111] the OFF current is dominated by tunneling, but due to the heavier transport effective masses the tunneling currents are significantly smaller.

1.7.7. Convergence issues in High-Bias Coherent Transport Simulations

In figure 11 the dashed lines labeled "not conv" indicate that several high bias points are not converged for the [111] wire. Such high bias convergence issues are not unusual for coherent quantum transport simulations for such gate lengths and we will describe some of the fundamental issues and limits of the simulation approach. We focus on the drain current versus gate voltage characteristics of the [111] wire in a bias point range of 0.3V to 0.6V as depicted in the insert of Figure 18(a). The main component of Figure 18(a) depicts the electrostatic potential in the middle of the wire as referenced to the local bandstructure in the source similar to Figure 12(c) which depicted a single gate bias of 0.3V. Here we plot the band edge for gate biases ranging from 0.3V to 0.6V in steps of 0.05V. The lowering of the barrier as a function of gate voltage is well visualized. At a gate bias of 0.5V the local band edge develops an additional, step-like feature in the middle of the channel, which migrates closer to the source side for further increased gate voltages. The formation of the "unreasonable looking" potential band edge in the middle of the channel is a symptom related to the sharp increase in dI/dV in the inset current-voltage characteristic.

The top left of Figure 18 indicates the transmission coefficient (red) and current density injected from the source at a gate bias of 0.3V. Next to it a gray box indicates a carrier distribution in the source that reaches as high as $10k_BT$. For the gate voltage of 0.3V one can see that the barrier in the channel still limits the distribution of the current flow. That is, the current turns on about 0.1eV or $4k_BT$ above the Fermi level in the source. At 0.6V gate bias the barrier has virtually completely vanished and the current (as indicated in the top right of the figure) now flows above the source conduction band edge. As expected there is no current flow under the conduction band edge. The bottom right of Figure 18 indicates the supply function of carriers in the drain. Clearly no electrons can reach from the drain into the channel at the applied source drain voltage of 0.6V. All carriers in the channel must come from the source. The red vertical arrow between labeled "No Supply" between the source and drain supply functions indicates that there are no carriers that can be injected in that energy range from either source or drain coherently.

The transport effective masses of the [111] wire are much larger than the ones in the [100] and [110] wires and therefore the injection of carriers from the source into the channel is not as intense in the [111] wire. The [111] wire cannot supply enough charge into the channel to "oppose" the electrostatic pull down from the gate and the drain.

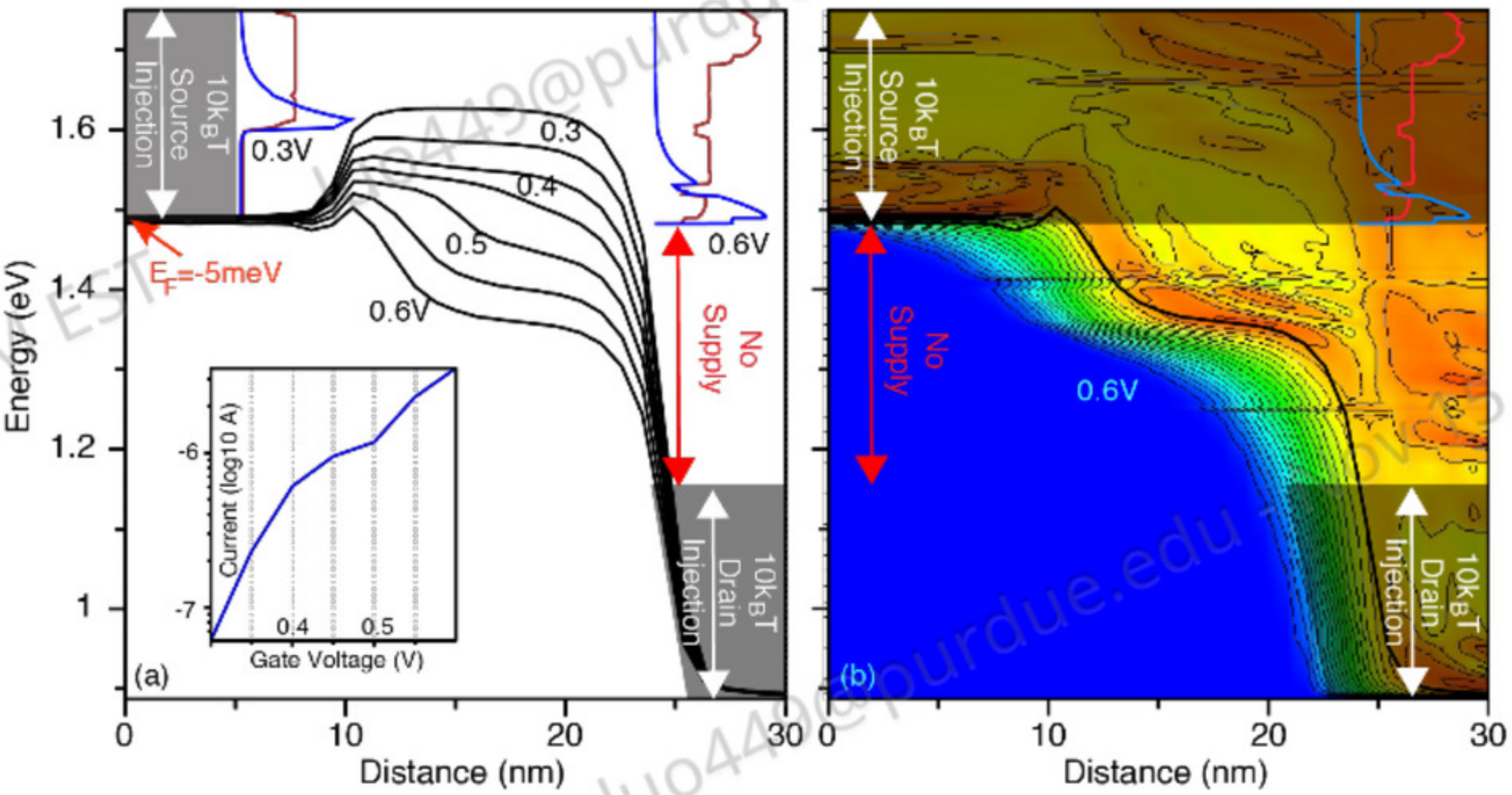

Figure 18: (a) insert: drain vs. gate current voltage characteristic of the [111] 15nm gate nanowire. A non-smooth curve is visible with a dip at 0.5V. The main part shows the conduction band edge in the [111] nanowire for different gate biases ranging from 0.3V to 0.6V. Top left and bottom right shaded regions indicate the available Fermi see of electrons to a height of $10k_BT$ above the Fermi level. Blue lines at the top right and left indicate the energy distribution of the current density for 0.3V (left) and 0.6V (right). (b) Similar to (b) for a gate voltage of 0.6V with the local density of states superposed.

Figure 18(b) considers the gate bias point 0.6V in more detail with a superposition of the local density of states under the local band edge diagram of (a). The source and drain injection energy ranges as well as the current density and transmission coefficient are included. The local density of states is numerically well-resolved, and the channel broadened quasi bound states that were also visible in Figure 14(b). Again, there is no indication that the quantum mechanical components of this simulation have any issues of energy or spatial resolution.

One can now argue that the Poisson solution is not really stable and causes the step-like feature in the channel potential. Numerical experiments on this very device show that neither tightening the Poisson convergence criteria nor increasing the momentum or energy resolution in the quantum transport simulation change the convergence of the potential or add more numerical detail. The Poisson solution "converges" to the same potential and overall current every time. Physical intuition tells us that the potential does not "look right". This electrostatic potential is not a result of bad numerical stability but due to incompleteness of the physical model. In a more realistic simulation would need to include energy loss mechanisms in the channel and a more detailed model of the source contact.

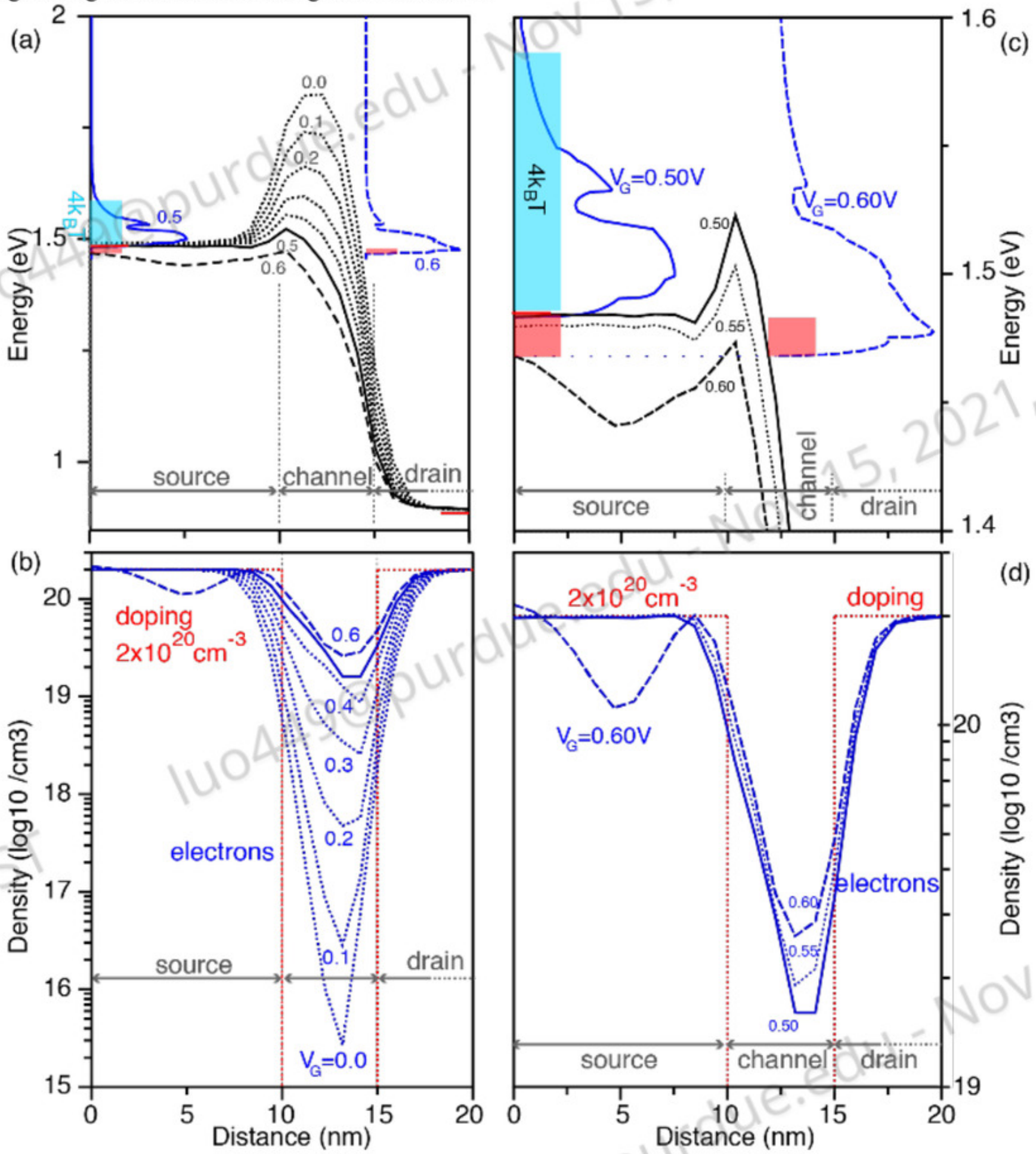

Figure 19: (a) Conduction band edge (black lines) in the [111] nanowire for gate biases ranging from 0V to 0.6V. Blue line indicates the energy distribution of the normalized current density (blue line). Fermi levels in source and drain are red lines. An energy range of $4k_BT$ above the Fermi level is indicated with a blue shaded region. Only a 5nm section of the 10nm drain region is shown. (b) Electron density (blue line) and doping density (red dashed line) in the wire for the same gate voltages as (a). (c) Zoomed in data of (a) for the gate bias $V_G = 0.5$V and $V_G = 0.6$V. A dotted line guiding the view of the potential evolution at $V_G = 0.55$V is added as a dotted line. The conduction band edge shows a potential dip in the source at $V_G = 0.6$V. There is a significantly increased energy range of injection from the source (red shaded area). (d) Zoomed in data of (b) for the gate bias $V_G = 0.5$V and $V_G = 0.6$V. A dotted line guiding the view of the potential evolution at $V_G = 0.55$V is added as a dotted line.

The intuitive view to "fix" the issue of Figure 18 is to include inelastic electron scattering mechanisms [152,153] such that the electrons can lose energy and populate the states in the channel, and then the potential will float up, restoring the expected smooth evolution of the electrostatic potential in the channel. In fact, OMEN has shown that scattering adds an additional potential drop into the channel and the source extension region. However, in light of the relatively short channel of 15nm and an expected high degree of ballistic/coherent it is the authors' opinion that a more sophisticated treatment of the channel will not provide the real physical solutions.

In our opinion, the real issue is the treatment of the source contact. The model assumption is that the Fermi level at the numerical start of the device is constant with regard to the equilibrium, zero bias condition. It is assumed that the contact can supply unlimited carriers, even at perfectly unit transmission in all available channels. A more realistic representation of that source injection would include the details of the electron reservoir and its ultimately connected metal contact. A comprehensive model would represent in more detail the ability of the "contact" to inject carriers into the device. A very nice depiction on how scattering self-consistently modifies the electrostatic potential in the gate region as well as the highly doped spacer has been given in [152].

The failure of the coherent quantum transport model was relatively obvious for the 15nm gate length [111] nanowire, as the current showed "unusual" behavior. The natural question now emerges by something similar was not observed in the 5nm gated [111] wire? The *OMENwire* simulator returns "converged" results and the current voltage characteristics do not reveal any trouble.

Intuitively one can argue that the 5nm gate region is already so short that the gate and drain voltage combined will pull very hard on the channel potential that the drop should almost be linear. That intuition is actually about right as depicted in Figure 19(a) where the expected electrostatic potential is shown throughout the nanowire for different gate potentials ranging from 0V to 0.6V. Nicely visible is the increased asymmetry of the potential in the gate region, where the peak of the potential shifts from the middle of the channel more and more to the source. This effect is called drain-induced-barrier-lowering (DIBL).

The low gate voltage data sets from $V_G = 0$V to $V_G = 0.4$V are drawn in dotted lines. The last "reasonably looking" gate voltage data sets at $V_G = 0.5$V are shown in a solid line and the data for $V_G = 0.6$V is denoted with long dashed lines to draw the viewer's attention. All the potential profiles "look" reasonable except for the last bias point of 0.6V which we will discuss below.

The charge filling of the channel as a function of gate voltage is depicted in Figure 19(b) for different gate voltages. At zero gate voltage the channel is virtually empty with a smooth exponential decay at the abrupt doping profile change. This corresponds of course well to the high gate potential barrier in Figure 19(a) for $V_G = 0$V. As the gate voltage is increased the charge in the channel increases linearly on a logarithmic scale which corresponds to the linear subthreshold current increase on a logarithmic scale (Figure 16). Everything looks reasonable about these charge profiles except the one for the gate voltage $V_G = 0.6$V.

The right column of Figure 19 shows the zoomed-in data sets of the left column for the gate voltages $V_G = 0.5$V (solid line) and $V_G = 0.6$V (long dashed line). A data set at $V_G = 0.55$V is added as a dotted line to guide the eye in the evolution as a function of gate voltage. The data at $V_G = 0.5$V all look very reasonable. At $V_G = 0.5$V the source potential is still flat and pegged to the Fermi level (Figure 19(c)). Current flow (blue line) sets in as expected above the conduction band edge basically tunneling through the remaining tiny barrier. The electron density (Figure 19(d)) for $V_G = 0.5$V is converged to the doping profile level in the source region.

As the gate potential is increased above $V_G = 0.5$V all the device internal quantities begin to show unusual behaviors. At $V_G = 0.55$V the electrostatic potential in the source region begins to drop slightly (on this scale) under the potential values that correspond to an equilibrium source. At $V_G = 0.6$V the electrostatic potential at the injection point drops precipitously and the overall source potential develops a physically very unreasonable dip. The current continues to increase since now there is current flow in a potential range that previously had no injection (see the red energy range). The charge density in Figure 19(d) shows a very unphysical dip in the source region. Clearly something is starting to go "wrong" in the simulation that should require some careful attention.

Figure 20 depicts (black line with circles) the energy difference between the Fermi level and the conduction band edge at the coherent injection point of the source (at the 0nm coordinate). As discussed above the Fermi level is about 5nm under the conduction band edge for $V_G = 0.0$V. As the gate voltage is increased the Fermi level remains roughly constant and then begins to increase. The Poisson solution in *OMENwire* requires more and more charge in the source to balance the doping level and the transport throughout the whole device. Let us look at a very trivial and incomplete model based on semiclassical charge where

$$N_{class} = N_c \exp\left[-\left(E_C - E_F\right)/k_B T\right] \qquad (32)$$

At E_C-E_F=-5.51meV for $V_G = 0.0$V and a balanced doping of $N_D = 2 \times 10^{20}\text{cm}^{-3}$ we can take $N_C = 2.48 \times 10^{20}\text{cm}^{-3}$. Now we define

$$\Delta n\left(V_G\right) = N_c \left\{\exp\left[-\left(E_C\left(V_G\right) - E_F\right)/k_B T\right] - 1\right\} \qquad (33)$$

And plot $\Delta n(V_G)$ in Figure 20 with a blue line. While there is some numerical noise for gate voltages below $V_G = 0.2$V we numerically observe the extra charge in the source injection point grows exponentially over 3 orders of magnitude. At a bias of $V_G = 0.55$V the additional charge approaches reaches 1/2 of the Fermi sphere in the source and at $V_G = 0.6$V the additional charge exceeds the doping in the source.

Figure 20 clearly demonstrates that the source injection is driven to a point where the assumption that it can provide "infinite" charge breaks down. As the needed charge to balance the Poisson equation approaches that of

the doping level, the physical model breaks down completely and results in erroneous charge and potential profiles.

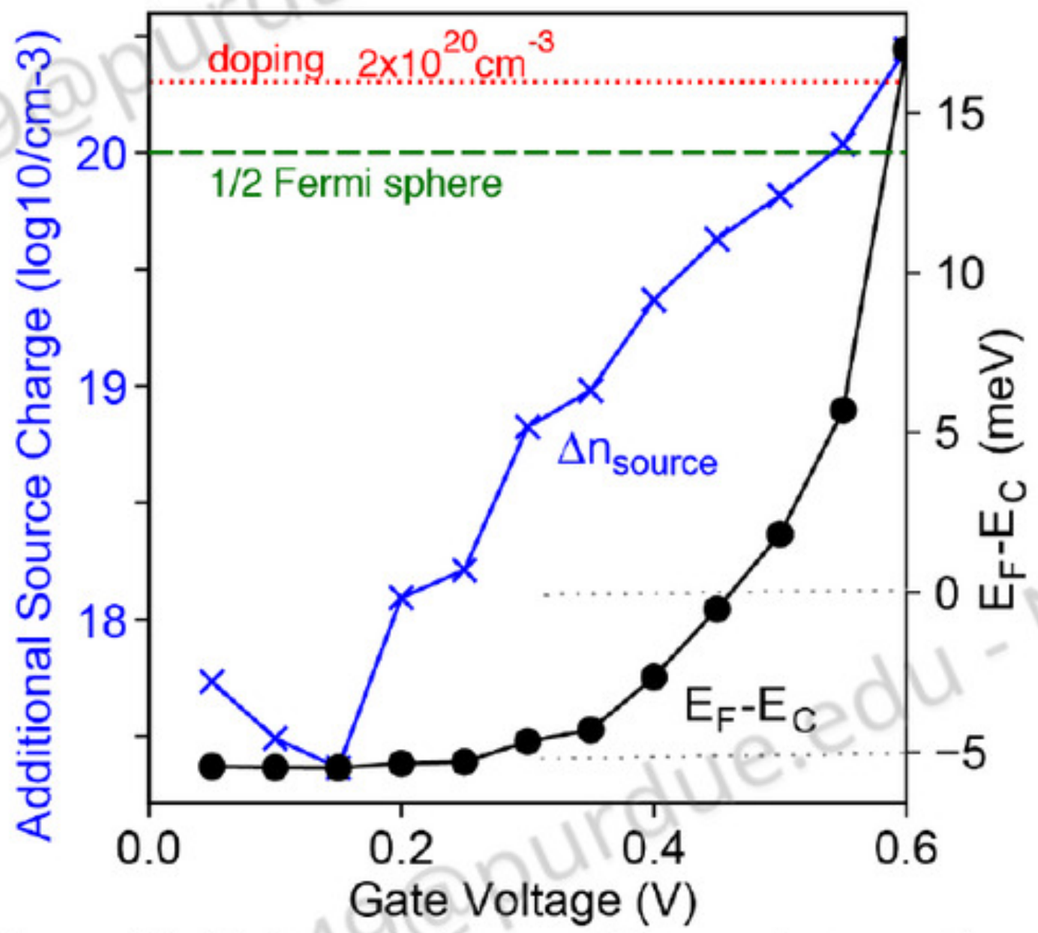

Figure 20: Right scale: Energy difference between the Fermi level and the conduction band edge at the coherent injection point at site 0nm of the [111] 5nm gate length nanowire. Left scale: additional charge needed at the coherent injection point to balance the charge throughout the device. At high gate biases that additional charge approaches and then exceeds that of half the forward moving Fermi sphere. The source cannot possibly provide enough charge moving forward through the device.

We argue the point here that the erroneous high gate bias results are not caused by a bad numerical implementation, but by the fundamental model assumption. We believe that a more comprehensive model of the contact regions of the highly doped semiconductor as well as the connected metal are needed in the limits of extreme bias operation which would deplete the numerical contact regions.

We conclude this section with an advice of caution – look at the numerical data carefully even as you use sophisticated tools!

1.7.8. Short Channel Devices: A New Design Paradigm with new Requirements

Effective mass and bandgaps are design parameters: The discussions above have shown that at the nanometer-scale effective masses and bandgaps *can be designed by geometry and crystal direction* [66,154]. Strain alters the atomic arrangements and can be used to tune the bandstructure further [66,154-157]. The self-consistently calculated charge distribution in nanowires shows that the bandstructure is affected by charge filling [154]. Silicon nanowires can be designed to have effective masses that are competitive and even lower than the traditional III-V materials. We note here that especially under scaling into the 5nm regime that III-V materials show significantly increased masses due to their non-parabolic bandstructure [19,152]. This typically reduces the perceived benefits of III-V materials even more [157-159].

III-V material versus Silicon materials: The discussions in the previous sections demonstrated that for a "long" 15nm channel a light effective mass material can in principle deliver the highest current as derived from the top-of-the-barrier model. This seems to confirm the long-standing expectation that lighter effective masses as found typically in III-V materials will outperform typical Silicon devices. However, with smart crystal orientation choice and strain engineering the effective masses in Silicon can be tuned as discussed above. The III-V material insertion into the Silicon process flow is extremely expensive and Silicon has been able to achieve competitive performance.

Fast materials with light effective masses are not desirable: For short channel devices such as the 5nm gate length nanowire discussed here we can clearly demonstrate that low effective mass materials are not desirable at all. At such short gate lengths, the potential barrier becomes very small, electrons principally tunnel through it and a heavy mass is desired to suppress that tunneling. [156-160]

Tight-binding captures critical device physics: Here we have focused on the more intuitive electron transport physics. We note here that a nanowire electron dispersion begins to look just as complicated as a hole dispersion. All the methodologies presented here can be explored for hole transport and have been implemented in OMEN and NEMO5. Necessary physical details such as spin-orbit coupling, strain dependence, geometry and crystal direction dependence can be captured efficiently with only a few model assumptions. Transport in hole nanowires and device design as a function of geometry and gating has been studied [155,161-163].

Tight-binding combined with coherent and incoherent transport models: The combinations of TB with QTBM and TB with NEGF have now been established as the state-of-the-art device modeling basis sets and approaches. Software packages based on this baseline are built by industrial and by academic research groups and commercialization efforts are underway to release these methods in standard device TCAD packages. The methodology is stable and reliable enough that they can be released to the scientific community. The *OMENwire* and *Band Structure Lab* codes used here are available on nanoHUB as discussed below.

Roadmap Modeling changed from Analytical Device models to Full 3D Quantum Transport Models: The 2013 IRTS roadmap [164] in its Process Integration, Device and Structure (PIDS) chapter, relies for its device model performance projection on a compact model using a computer aided design (CAD) tool called MASTAR [165] The subthreshold swing (SS) enters this model as a phenomenological parameter, rather than a characteristic predicted by a physics based model. Full 3D atomistic modeling quantum transport modeling showed that such a compact model can lead to erroneous predictions [166] and physics-based model results were adopted to calibrate predictions in the subsequent ITRS release.

Physics-based contact models require further development: In the sections above we have also demonstrated and explained in detail how the state-of-the-art transport models break down, when driven to extreme conditions. Under very high bias conditions the source may no longer supply enough charge to satisfy the channel

transport ability. More sophisticated models are needed to model the heavy doping semiconductor source/reservoir as well as the metal contact to that reservoir region. Several efforts have been reported in treating the occupation limitations of of the injecting source regions in 1D geometries with an effective mass model [167,168]. It is our opinion that the traditional NEGF perturbative scattering methods may provide some help when used in the semiconductor device region. However, we do also believe that the source depletion and treatment of the contacts is a first order effect that must be treated not as a perturbation but by a fundamental approach.

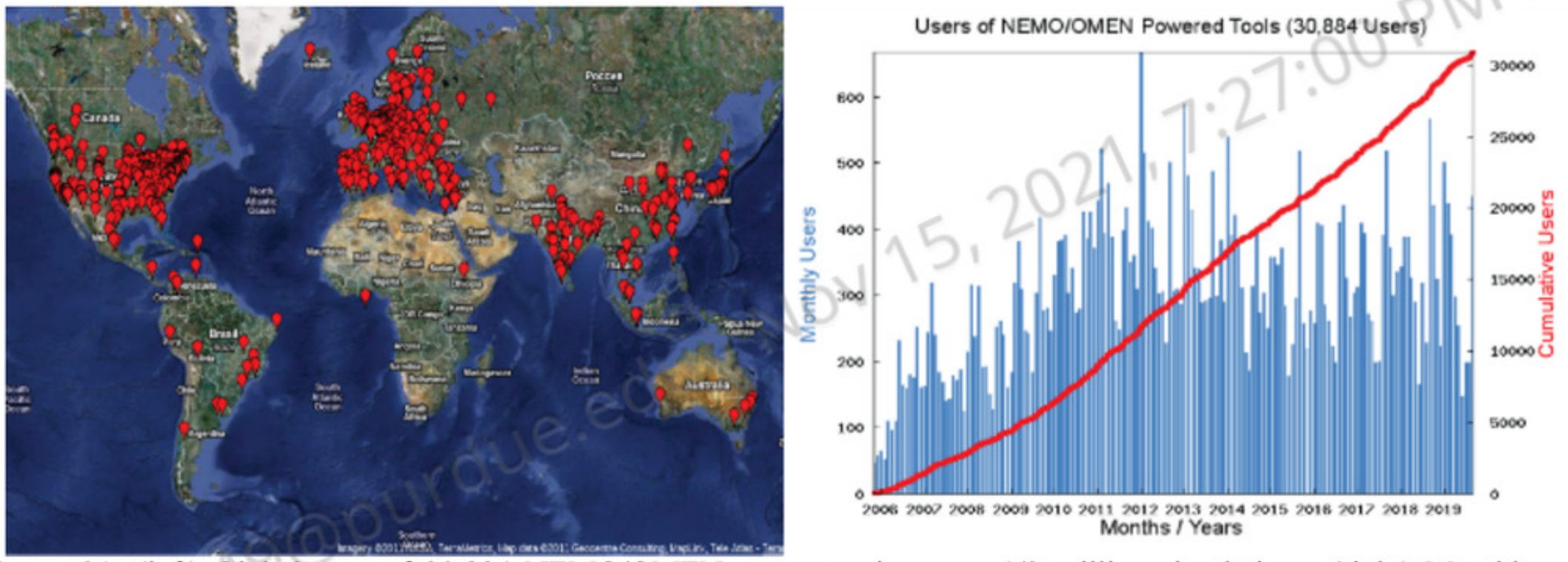

Figure 21: (left) Global map of 30,884 NEMO/OMEN users running over 1/2 million simulations. (right) Monthly and cumulative usage statistics.

1.8. Beyond Device Physics Advancements: Reaching the World

Nanoelectronic devices such as quantum dots and nanowires that have been subject to intense research in the past 25 years are finding widespread adoption in today's technologies. As these devices are inserted into real products their design and optimization now require accurate physics-based models that can explore the design space through simulation. The approaches are literally used today, in the year 2019, to advance, develop, and integrate the 7nm and 5nm transistors that will soon be deployed for billions of users around the world in commercial products. This view and understanding of modeling transport at the nanometer scale truly enable products that impact the globe. However, transferring such modeling and simulation into the education and training of engineers who will shape the future requires more: 1) building upon, rather than re-building every time on the results of prior researchers, 2) duplicating the results of prior researchers, and 3) pushing the recent research knowledge into the education and training of future engineers.

nanoHUB.org offers the opportunity to researchers to share their simulation tools and results to the broader community. Anyone can contribute their own simulation tools and results into nanoHUB to enable duplication of research results and to infuse these results into further use. We utilize this capability extensively in this work with nanoHUB tools *Bandstructure Lab* [108], *FETtoy* [150], and *OMENwire* [151] *where any reader can now duplicate the results shown here*.

nanoHUB analytics have shown that simulation tools will be adopted in a median time of 6 months into the first classroom by 6 months. This constitutes a rather rapid change in curricula compared to writing a new text book, which requires 4-6 years. We have identified over 35,000 students at over 1,800 classes at 180 institutions who have used nanoHUB for formal instruction in a classroom [169]. *nanoHUB can support education!*

As of September 2021 [170] 2,500+ citations of nanoHUB in the literature demonstrate that these tools are used in subsequent research. Over 30% of these papers include experimental data or an experimentalist as a co-author. nanoHUB does not just serve the computational folks, but also experimental work. The 2,500+ citations generated over 52,000 secondary citations resulting in an h-index of 103. *nanoHUB can support research!*

The NEMO/OMEN team has deployed 9 tools into nanoHUB over the years who have been used by 30,700 users as depicted in Figure 21. These users have run over 1/2 million simulations. The tools have been used in over 380 structured education settings or classrooms. These 9 tools have been cited in the literature 107 times. All the data and figures involving the nanowire simulations in this work are directly generated with nanoHUB tools such as *Bandstructure Lab* [108] and *OMENwire* [151].

1.9. Conclusion

We have seen that the empirical tight-binding method is well-suited to modern nanodevice modeling. It is physics-based, and naturally incorporates geometry at the atomic level (e.g., strain, disorder). It can successfully model both semiconductors and metals, and is efficient since it excludes core electrons; it handles varying geometries and materials easily. Its parameters represent interactions between orbitals centered on the same and neighboring atoms, and may be determined by either fitting to bulk bandstructures or by projection from *ab initio* Hamiltonians and wavefunctions. The method offers great flexibilty in terms of basis orbitals and the extent of their interactions.

This flexibility in basis is responsible for one of tight-binding's strengths: its numerical efficiency. Choosing a somewhat larger atomic basis usually allows for a reduced

range of interactions, often to first- or second-near neighbors. When the extent of interactions is limited in this way the Hamiltonian matrix becomes sparser, resulting in better time and memory efficiency. As a result, multimillion atom electronic structure calculations are possible on very small computer clusters, and realistic, extended nanowire and ultra-thin body transistor calculations are possible on moderately sized computer clusters.

Like all models, TB has challenges and limitations. Parameterization from experimental results or *ab initio* calculations can require significant effort. When projecting out of *ab initio* wavefunctions, the resulting tight-binding basis functions are not uniquely determined, so there remains some art in choosing their form. Likewise, for new materials (especially two-dimensional ones), determination of a basis which correctly reproduces the important bandstructure features is not easily automated; it still requires significant expertise. Finally, because the parameters depend on the inter-atomic potentials, parameters are generally not transferrable between dissimilar materials systems. While these issues are different from those faced by other full-band methods, they are no worse.

Today the method has been widely adopted in numerical simulation tools. QTBM and NEGF were established as state-of-the-art by 2002; somewhat later the underlying materials models transitioned to tight-binding bases. The TB approach coupled to QTBM or NEGF is now in wide use by academics, industrial researchers, and software vendors. Many such tools are fully accessible to researchers, educators, and students on nanoHUB. The computational intensity for realistic nanodevice models is sufficiently low that NEMO/OMEN tools can be used on nanoHUB by anyone in the world.

ACKNOWLEDGEMENTS

The tight-binding and NEMO developments described here have been developed by the two authors since about 1992 originally in graduate School (Stanford (TB) and Purdue (GK)) and then in various teams centered at Texas Instruments (1994-1998), NASA Jet Propulsion Laboratory (1998-2003), and Purdue (2004-present). The core developers and theory contributors were Roger Lake, R. Chris Bowen, Mathieu Luisier, Tillmann Kubis, and Michailo Povolotskyi. At Purdue dozens of students contributed to the final version of NEMO5. NEMO5 is now commercialized with SILVACO and used by Intel to explore and design nanotransistors. Funding for this work came from many agencies such as NSF, DARPA, NRO, ARO, SRC.

REFERENCES

[1] E. E. Mendez, F. A. Agullo-Rueda, J. M. Hong: Temperature dependence of the electronic coherence of GaAs-GaAlAs superlattices, Appl. Phys. Lett. **56**, 2545-2547 (1990).

[2] I. N. Stranski, L. Krastanow, Lubomir: Zur Theorie der orientierten Ausscheidung von Ionenkristallen aufeinander, Abhandlungen der Mathematisch-Naturwissenschaftlichen Klasse IIb. Akademie der Wissenschaften Wien **146**, 797–810 (1938).

[3] S. Ahmed, N. Kharche, R. Rahman, M. Usman, S. Lee, H. Ryu, H. Bae, S. Clark, B. Haley, M. Naumov, F. Saied, M. Korkusinski, R. Kennel, M. McLennan, T. B. Boykin, G. Klimeck: Multimillion Atom Simulations with NEMO 3-D: Encyclopedia of Complexity and System Science, vol. 6: R. A. Meyers (ed.), (Springer, New York, 2009) pp. 5745-5783.

[4] G. Lansbergen, R. Rahman, C.. Wellard, J. Caro, N. Collaert, S. Biesemans, J. Woodall, G. Klimeck, L. Hollenberg, S. Rogge: Gate induced quantum confinement transition of a single dopant atom in a Si FinFET, Nature Phys. **4**, 656-661 (2008).

[5] M. Usman, H. Ryu, J. Woodall, D. Ebert, G. Klimeck: Moving towards nano-TCAD through multi-million atom quantum dot simulations matching experimental data, IEEE Trans. Nanotech. **8**, 330 – 344 (2009).

[6] C. Auth, et. al.: A 22nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors, 2012 Symposium on VLSI Technology (VLIST) Tech. Digest, (IEEE,2012), pp. 131-132.

[7] C.-H. Jan, et. al.: A 22nm SoC platform technology featuring 3-D tri-gate and high-k/metal gate, optimized for ultra low power, high performance and high density SoC applications, 2012 International Electron Devices Meeting (IEDM) Tech. Digest, (IEEE, 2012), pp. 3.1.1-3.1.4.

[8] R. Xie, et. al.: A 7nm FinFET Technology Featuring EUV Patterning and Dual Strained High Mobility Channels, 2016 International Electron Devices Meeting (IEDM) Tech. Digest, (IEEE, 2016), pp. 2.7.1-2.7.4.

[9] R. C. Bowen, Y. Wang: Enhanced PMOS Via Transverse Stress, U. S. Patent No. 7,268,399 B2 (2007).

[10] S. Thompson, et al.: A 90 nm logic technology featuring 50nm strained silicon channel transistors, 7 layers of Cu interconnects, low k ILD, and 1 um2 SRAM Cell, 2002 International Electron Devices Meeting (IEDM) Tech. Digest, (IEEE, 2002), pp. 61-64.

[11] See for example, E. Artacho, T. Beck, E. Hernandez: Special Issue: Current trends in electronic structure: Real-space, embedding and linear scaling techniques, Phys. Stat. Solidi (b) **243**, 971-972 (2006) and other articles in this issue..

[12] D. Sanchez-Portal, P. Ordejon, and E. Canadell: Computing the properties of materials from first principles with SIESTA, In: *Principles and Applications of density functional method in inorganic chemistry II, Structure and Bonding*, **113**, 103-170 (Springer, Berlin, Heidelberg, 2004).

[13] C. K. Skylaris, P. D. Haynes, A. A. Mostofi, M. C. Payne: Using ONETEP for accurate and efficient *O(N)* density functional calculations, J. Phys. Condens. Matt. **17**, 5757-5770 (2005).

[14] D. J. Singh, L. Nordstrom: *Planewaves, Pseudopotentials, and the LAPW Method* (Springer, Berlin, Heidelberg 2006).

[15] L. W. Wang, J. N. Kim, J.N., A. Zunger: Electronic structures of [100]-faceted self-assembled pyramidal InAs/GaAs quantum dots, Phys. Rev. B **59**. 5678-5687 (1999).

[16] P. Pieecuch, K. Kowalski, I. Pimienta, M. J. McGuire: Recent advances in electronic structure theory: Method of moments of coupled-cluster equations and renormalized coupled-cluster approaches, Int. Rev. Phys. Chem. **21**. 527-655 (2002).

[17] A. J. Williamson, J. C. Grossman, R. Q. Hood, A. Puzder, G. Galli: Quantum Monte Carlo calculations of Nanostructure Optical Gaps: Application to Silicon Quantum Dots, Phys. Rev. Lett. **89**, 196803 (2002).

[18] F. Aryasetiawan, O. Gunnarsson: The GW method, Rep. Prog. Phys. **61**, 237-312 (1998).
[19] R. C. Bowen, G. Klimeck, R. K. Lake, W. R. Frensley, T. Moise: Quantitative simulation of a resonant tunneling diode, J. Appl. Phys. **81**, 3207-3213 (1997).
[20] T. B. Boykin, G. Klimeck, M. A. Eriksson, M. Friesen, S. N. Coppersmith, P. von Allmen, F. Oyafuso, S. Lee: Valley splitting in strained silicon quantum wells, Appl. Phys. Lett. **84**,115-117 (2004).
[21] T. B. Boykin, G. Klimeck, M. Friesen, S. N. Coppersmith, P. von Allmen, F. Oyafuso, S. Lee:Valley splitting in low-density quantum-confined heterostructures studied using tight-binding models, Phys. Rev. **B** 70, 165325 (2004).
[22] T. B. Boykin, M. Luisier, A. Schenk, N. Kharche, G. Klimeck: The electronic structure and transmission characteristics of disordered AlGaAs nanowires, IEEE Trans. Nanotechnol. **6**, 43-47 (2007).
[23] N. Kharche, M. Prada, T. B. Boykin, G. Klimek: Valley splitting in strained silicon quantum wells modeled with 2 degrees miscuts, step disorder, and alloy disorder, Appl. Phys. Lett. **90**, 092109 (2007).
[24] G. Klimeck, S. Ahmed, N. Kharche, M. Korkusinski, M. Usman, M. Prada, T. B. Boykin: Atomistic Simulation of Realistically Sized Nanodevices Using NEMO 3-D: Part II – Applications, IEEE Trans. Elect. Dev. **54**, 2090-2099 (2007).
[25] G. Klimeck, F. Oyafuso, T. B. Boykin, R. C. Bowen, P. von Allmen: Development of a nanoelectronic 3-D (NEMO 3-D) simulator for multimillion atom simulations and its application to alloyed quantum dots, J. Comp. Mod. Eng. Sci. (CMES) **3**, 601-642 (2002).
[26] M. Korkusinski, G. Klimeck: Atomistic simulations of long-range strain and spatial asymmetry molecular states of seven quantum dots, J. Phys. Conf. Ser. **38**, 75-78 (2006).
[27] S. Lee, O. L. Lazarenkova, P. von Allmen, F. Oyafuso, G. Klimeck: Effect of wetting layers on the strain and electronic structure of InAs self-assembled quantum dots, Phys. Rev. B **70**, 125307 (2004).
[28] S. W. Lee, P. von Allmen, F. Oyafuso, G. Klimeck, K. B. Whaley: Effect of electron-nuclear spin interactions for electron-spin qubits localized in InGaAs self-assembled quantum dots, J. Appl. Phys. **97**, 043706 (2005).
[29] G. C. Liang, J. Xiang, N. Kharche, G. Klimeck, C. M. Lieber, M. Lundstrom: Performance analysis of a Ge/Si core/shell nanowire field-effect transistor, Nano Lett. **7**, 642-646 (2007).
[30] F. Oyafuso, G. Klimeck, R. C. Bowen, T. B. Boykin: Atomistic Electronic Structure Calculations of Unstrained Alloyed Systems Consisting of a Million Atoms, J. Comp. Electr., **1** , 317-321 (2002).
[31] F. Oyafuso, G. Klimeck, R. C. Bowen, T. B. Boykin, P. von Allmen: Disorder Induced Broadening in Multimillion Atom Alloyed Quantum Dot Systems, Phys. Stat. Sol. (c), 0004. 1149-1152 (2003).
[32] R. Rahman, C. J. Wellard, F. R. Bradbury, M. Prada, J. H. Cole, G. Klimeck, L. C. L. Hollenberg: High precision quantum control of single donor spins in Si, Phys. Rev. Lett. **99**, 036403 (2007).
[33] S. Lee, J. Kim, L. Jonsson, J. W. Wilkins, G. W. Bryant, G. Klimeck: Many-body levels of optically excited and multiply charged InAs nanocrystals modeled by semiempirical tight-binding, Phys. Rev. **B 66**, 235307 (2002).
[34] G. Klimeck, S. Ahmed, H. Bae, N. Kharche, S. Clark, B. Haley, S. Lee, M. Naumov, H. Ryu, F. Saied, M. Prada, M. Korkusinski, T. B. Boykin: Atomistic Simulation of Realistically Sized Nanodevices Using NEMO 3-D: Part I - Models and Benchmarks, IEEE Trans. Elect. Dev. **54**, 2079-2089 (2007).
[35] J. C. Slater, G. F. Koster: Simplified LCAO Method for the Periodic Potential Problem, Phys. Rev. **94**, 1498-1524 (1954).
[36] J. Cerda, F. Soria: Accurate and transferable extended Hückel-type tight-binding parameters, Phys. Rev. B **61**, 7965-7971 (2000).
[37] P. O. Löwdin: On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals , J. Chem. Phys. **18**, 365-375 (1950).
[38] P. Vogl, H. P. Hjalmarson, J. D. Dow: A Semi-empirical tight-binding theory of the electronic structure of semiconductors, J. Phys. Chem. Solids **44**, 365-378 (1983).
[39] T. B. Boykin: Improved fits of effective masses at Γ in the spin-orbit, second-near-neighbor sp3s* model: results from analytic expressions, Phys. Rev. B **56**, 9613-9618 (1997).
[40] J.-M. Jancu, R. Scholz, F. Beltram, F. Bassani: Empirical spds* tight-binding calculation for cubic semiconductors: General method and material parameters, Phys. Rev. B **57**, 6493-6507 (1998).
[41] T. B. Boykin, G. Klimeck, F. Oyafuso: Valence band effective mass expressions in the sp3d5s* empirical tight-binding model applied to a Si and Ge parameterization, Phys. Rev. B **69**, 115201 (2004)
[42] W. A. Harrison: *Elementary Electronic Structure* (World Scientific, New Jersey, 1999).
[43] T. B. Boykin, G. Klimeck, R. C. Bowen, R. Lake: Effective mass reproducibility of the nearest-neighbor sp3s* models: analytic results, Phys. Rev. B **56**, 4102-4107 (1997).
[44] M. Graf, P. Vogl: Electromagnetic fields and dielectric response in empirical tight-binding theory, Phys. Rev. B **51**, 4940-4949 (1995).
[45] T. B. Boykin: Incorporation of incompleteness in the kp perturbation theory, Phys Rev. B **52**, 16317-16320 (1995).
[46] G. Klimeck, R. C. Bowen, T. B. Boykin, C. Salazar-Lazaro, T. Cwik, A. Stoica: Si tight-binding parameters from genetic algorithm fitting, Superlatt. Microstruct. **27**, 77-88 (2000).
[47] Y. Tan, M. Povolotskyi, T. Kubis, Y. He, Z. Jiang, G. Klimeck, T. B. Boykin: Empirical tight-binding parameters for GaAs and MgO with explicit basis through DFT mapping, J. Comp. Electr. **12**, 56-60 (2013).
[48] Y. Tan, M. Povolotskyi, T. Kubis, T. B. Boykin, G. Klimeck: Tight-binding analysis of Si and GaAs ultrathin bodies with subatomic wave-function resolution, Phys. Rev. B. **92**, 085301 (2015).
[49] Y. Tan, M. Povolotskyi, T. Kubis, T. B. Boykin, G. Klimeck: Transferable tight-binding model for strained group IV and III-V materials and heterostructures, Phys. Rev. B **94**, 045311 (2016).
[50] Y.-M. Niquet, D. Rideau, C. Tavernier, H. Jaouen, X. Blase: Onsite matrix elements of the tight-binding

Hamiltonian of a strained crystal: Application to silicon, germanium, and their alloys, Phys. Rev. B **79**, 245201 (2009).
[51] T. B. Boykin, G. Klimeck, R. C. Bowen, F. Oyafuso: Diagonal parameter shifts due to nearest-neighbor displacements in empirical tight-binding theory, Phys. Rev. B **66**, 125207 (2002).
[52] T. B. Boykin, M. Luisier, M. Salmani-Jelodar, G. Klimeck: Strain-induced, off-diagonal, same-atom parameters in empirical tight-binding theory suitable for [110] uniaxial strain applied to a silicon parameterization, Phys. Rev. B **81**, 125202 (2010).
[53] T. Shishidou, T. Oguchi: k·p formula for use with linearized augmented plane waves, Phys. Rev. B **78**, 245107 (2018).
[54] D. J. Chadi: Spin-orbit splitting in crystalline and compositionally disordered semiconductors, Phys. Rev. B **16**, 790-796 (1977).
[55] T. B. Boykin, P. Vogl: Dielectric response of molecules in empirical tight-binding theory, Phys. Rev. B **65**, 035202 (2001).
[56] T. B. Boykin, R. C. Bowen, G. Klimeck: Electromagnetic coupling and gauge invariance in the empirical tight-binding method, Phys. Rev. B **63**, 245314 (2001).
[57] B. A. Foreman: Consequences of local gauge symmetry in empirical tight-binding theory Phys. Rev. B **66**, 165212 (2002).
[58] R. Peierls: Zur Theorie des Diamagnetismus von Leitungselektronen, Z. Phys. **80**, 763-791 (1933).
[59] T. B. Boykin: Tight-binding-like expressions for the continuous-space electromagnetic coupling Hamiltonian Am. J. Phys. **69**, 793-798 (2001).
[60] Y.-C. Chang: Complex band structures of zinc-blende materials, Phys. Rev. B **25**, 605-619 (1982).
[61] Y.-C. Chang, J. N. Schulman: Complex band structures of crystalline solids: An eigenvalue method, Phys. Rev. B **25**, 3975-3986 (1982).
[62] J. N. Schulman, Y.-C. Chang: Band mixing in semiconductor superlattices, Phys. Rev. B **31**, 2056-2068 (1985).
[63] R. C. Bowen, W. R. Frensley, G. Klimeck, R. K. Lake: Transmission resonances and zeros in multiband models, Phys. Rev. B **52**, 2754-2765 (1995).
[64] T. B. Boykin: Generalized eigenproblem method for surface and interface states: the complex bands of GaAs and AlAs, Phys. Rev. B **54**, 8107-8115 (1996).
[65] T. B. Boykin: Tunneling calculations for systems with singular coupling matrices: results for a simple model, Phys. Rev. B **54**, 7670-7673 (1996).
[66] M. Luisier, A. Schenk, W. Fichtner, G. Klimeck: Atomistic simulation of nanowires in the sp3d5s* tight-binding formalism: From boundary conditions to strain calculations, Phys. Rev. B **74**, 205323 (2006).
[67] R. Tsu, L. Esaki: Tunneling in a finite superlattice , Appl. Phys. Lett. **22**, 562-564 (1973).
[68] T. B. Boykin, J. P. A. van der Wagt, J. S. Harris, Jr.: Tight-binding model for GaAs/AlAs resonant tunneling diodes, Phys. Rev. B **43**, 4777-4784 (1991).
[69] J. N. Schulman, Y.-C. Chang: Reduced Hamiltonian method for solving the tight-binding model of interfaces, Phys. Rev. B **27**, 2346-2354 (1983).
[70] D. Z. Y. Ting, E. T. Yu, T. C. McGill: Multiband treatment of quantum transport in interband tunnel devices, Phys. Rev. B **45**, 3583-3592 (1992).
[71] G. Grosso, S. Moroni, G. P. Parravicini: Electronic structure of the InAs-GaSb superlattice studied by the renormalization method, Phys. Rev. B **40**, 12328-12337 (1989).
[72] T. B. Boykin, J. S. Harris, Jr.: X-valley tunneling in single AlAs barriers ,J. Appl. Phys. **72**, 988-992 (1992).
[73] T. B. Boykin, M. Luisier, G. Klimeck: Multi-band transmission calculations for nanowires using an optimized renormalization method, Phys. Rev. B **77**, 165318 (2008).
[74] M. Luisier, G. Klimeck, A. Schenk, W. Fichtner, T. B. Boykin: A parallel sparse linear solver for nearest-neighbor tight-binding problems. In: E. Lunque, T. Maragalef, T. Benitez Europar 2008. *Lecture Notes in Computer Science* **5168**, pp. 790-800 (Springer, Berlin, Heidelberg, 2008).
[75] L. P. Kadanoff, G. Baym: *Quantum Statistical Mechanics,* Frontiers in Physics Lecture Note Series, (W.A. Benjamin, New York, 1962).
[76] L. V. Keldysh: Diagram technique for non-equilibrium processes, Sov. Phys. JETP **20**, 1018 (1965).
[77] R. Bertoncini, A. M. Kirman, D. K. Ferry: Airy-coordinate Green's-function technique for high-field transport in semiconductors, Phys. Rev. B **40**, 3371-3374 (1989); Airy-coordinate technique for nonequilibrium Green's-function approach to high-field quantum transport, Phys. Rev. B **41**, 1390-1400 (1990).
[78] S. Datta: A simple kinetic equation for steady-state quantum transport, J. Phys. Condens. Matt. **2**, 8023-8052 (1990).
[79] S. Datta: Nanoscale Device Simulation: The Green's Function Method, Superlatt. Microstruct. **28,** 253-278 (2000).
[80] S. Datta: Non-Equilibrium Green's Function (NEGF) Formalism: An elementary Introduction, 2002 International Electron Devices Meeting (IEDM) Tech. Digest, (IEEE, 2002), pp. 703-706.
[81] S. Datta: Electrical resistance: an atomic view, Nanotechnol. **15**, S433-S451 (2004).
[82] S. Datta, *Electronic Transport in Mesoscopic Systems*, (Cambridge UP, New York, 1997).
[83] S. Datta, *Quantum Transport: Atom to Transistor*, (Cambridge UP, New York, 2005).
[84] S. Datta: *A New Perspective on Transport*, (World Scientific, New Jersey, 2012).
[85] S. Datta: *Lessons from Nanoelectronics: A New Perspective on Transport - Part A: Basic Concepts*, (World Scientific, New Jersey, 2017).
[86] S. Datta, *Lessons From Nanoelectronics: A New Perspective On Transport - Part B: Quantum Transport*, (World Scientific, New Jersey, 2017).
[87] S. Datta: nanoHUB-U: Fundamentals of Nanoelectronics - Part A: Basic Concepts, 2nd Edition, https://nanohub.org/courses/FON1.
[88] S. Datta, nanoHUB-U: Fundamentals of Nanoelectronics - Part B: Quantum Transport, 2nd Edition, https://nanohub.org/courses/FON2.
[89] R. Lake, G. Klimeck, R. C. Bowen, D. Jovanovic: Single and multiband modeling of quantum electron transport through layered semiconductor devices, J. Appl. Phys. **81**, 7845-7869 (1997).

[90] C. S. Lent and D. J. Kirkner: The quantum transmitting boundary method, J. Appl. Phys. **67**, 6353-6359 (1990).
[91] R. C. Bowen: Full Bandstructure Modelingof Quantum Transport in Nano-Scaled Devices, (Ph.D. Thesis, University of Texas at Dallas, 1996).
[92] R. Haydock, V. Heine V, M. J. Kelly: Electronic structure based on the local atomic environment for tight-binding bands, J. Phys. C: Solid State Phys. **5** 2845-2858 (1972); Electronic structure based on the local atomic environment for tight-binding bands II, J. Phys. C: Solid State Phys. **8** 2591-2605 (1975),
[93] M. P. Lopez Sancho, J. M. Lopez Sancho and J. Rubio: Quick iterative scheme for the calculation of transfer matrices: application to MO(100), J. Phys. F: Met. Phys. **14**, 1205-1215 (1984).
[94] S. Park, H.-H. Park, M. Salmani-Jelodar, S. Steiger, M. Povolotskyi, T. Kubis, G. Klimeck: Contact Modeling and Analysis of InAs HEMT Transistors, Proc. IEEE Nanotechnology Materials and Devices Conference (IEEE NMDC 2011), (IEEE, Pistcataway, NJ, 2011), pp. 376-379.
[95] M. Luisier, T. B. Boykin, G. Klimeck, W. Fichtner: Atomistic nanoelectronic device simulations with sustained performances up to 1.44 PFlop/s, SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, Seattle, WA, Nov. 2011 (IEEE, Pistcataway, NJ 2011).
[96] G. Klimeck, R. Lake, R. C. Bowen, W. Frensley, T. Moise: Quantum Device Simulation with a Generalized Tunneling Formula, Appl. Phys. Lett., **67**, 2539-2541 (1995).
[97] P. Long, J. Huang, Z. Jiang, G. Klimeck, M. Rodwell, M. Povolotskyi: Performance degradation of superlattice MOSFETs due to scattering in the contacts, J. Appl. Phys. **120**, 224501 (2016).
[98] T. Kubis, Y. He, R. Andrawis, Gerhard Klimeck: General Retarded Contact Self-energies in and beyond the Non-equilibrium Green's Function Method, J. Physics: Conf. Ser. Volume 696, 012019 (2016).
[99] Y. He, Y. Wang, G. Klimeck, T. Kubis: Non-equilibrium Green's functions method: Non-trivial and disordered leads, Appl. Phys. Lett. **105**, 213502 (2014).
[100] T. Ameen, H. Ilatikhameneh, J. Huang, M. Povolotskyi, R. Rahman, G. Klimeck: Combination of Equilibrium and Nonequilibrium Carrier Statistics Into an Atomistic Quantum Transport Model for Tunneling Heterojunctions, IEEE Trans. Elect. Dev. **64**, 2512 - 2518, (2017).
[101] P. Long, J. Huang, M. Povolotskyi, P. Sarangapani, G. Valencia-Zapata, T. Kubis, M. Rodwell, G. Klimeck: Atomistic modeling trap-assisted tunneling in hole tunnel FETs, J. Appl. Phys. **123**, 174504 (2018).
[102] G. Klimeck: Quantum and semi-classical transport in RTDs in NEMO 1-D, J. Comp. Elect. **2**, 177-182 (2003).
[103] J. Huang, M. Povolotskyi, H. Ilatikhameneh, T. Ameen, R. Rahman, M. Rodwell, P. Long, G. Klimeck: A Multiscale Modeling of Triple-Heterojunction Tunneling FETs, IEEE Trans. Elect. Dev. **64**, 2728 – 2735 (2017).
[104] S. Steiger, M. Povolotskyi, H.-H. Park, T. Kubis, G. Klimeck: NEMO5: A Parallel Multiscale Nanoelectronics Modeling Tool, IEEE Trans. Nanotech. **10**, 1464 – 1474 (2011).
[105] J. Fonseca, T. Kubis, M. Povolotskyi, B. Novakovic, A. Ajoy, G. Hegde, H. Ilatikhameneh, Z. Jiang, P. Sengupta, Y. Tan, G. Klimeck: Efficient and realistic device modeling from atomic detail to the nanoscale**,** J. Comp. Electr. **12**, 592-600 (2013).
[106] M. Kuroda, Z. Jiang, M. Povolotskyi, G. Klimeck, D. Newns, G. Martyna: Anisotropic strain in SmSe and SmTe: Implications for electronic transport, Phys. Rev. B **90**, 245124 (2014).
[107] F. Oyafuso, G. Klimeck, P. von Allmen, T. B. Boykin, R. C. Bowen: Strain Effects in large-scale atomistic quantum dot simulations, Phys. Stat. Sol. (b) **239**, 71-79 (2003).
[108] S. Mukherjee, K. Miao, A. Paul, N. Neophytou, R. Kim, J. Geng, M. Povolotskyi, T. C. Kubis, A. Ajoy, B. Novakovic, J. Fonseca, H. Ilatikhameneh, S. Steiger, M. McLennan, M. Lundstrom, G. Klimeck: Band Structure Lab, https://nanohub.org/resources/bandstrlab, DOI: 10.4231/D3Z02Z95M (2015).
[109] S. Li, S. Ahmed, G. Klimeck, E. Darve: Computing entries of the inverse of a sparse matrix using the FIND algorithm, J. Comp. Phys., **227**, 9408-9427 (2008).
[110] S. Cauley, M. Luisier, V. Balakrishnan, G. Klimeck, C.-K. Koh: Distributed non-equilibrium Green's function algorithms for the simulation of nanoelectronic devices with scattering, J. Appl. Phys. **110** , 043713 (2011).
[111] S. Cauley, V. Balakrishnan, G. Klimeck, C.-K. Koh: A two-dimensional domain decomposition technique for the simulation of quantum-scale devices, J. Comp. Phys. **231**, 1293–1313 (2012).
[112] U. Hetmaniuk, Y. Zhao, M. P. Anantram: A nested dissection approach to modeling transport in nanodevices: Algorithms and applications, Int. J. Num. Meth Eng. **95**, 587-607 (2013).
[113] Y. Zhao, U. Hetmaniuk, S.R. Patil, J. Qi, M.P. Anantram: Nested dissection solver for transport in 3D nano-electronic devices, J. Comp. Electr. **15**, 708-720 (2016).
[114] Y. Ahn, M. Shin: Efficient Atomistic Simulation of Heterostructure Field-Effect Transistors, IEEE J. Electr. Dev. Soc. **7**, 668-676, (2019).
[115] E. Polizzi and N. B. Abdallah: Subband decomposition approach for the simulation of quantum electron transport in nanostructures, J. Comp. Phys. **202**, 150-180 (2005).
[116] J. Wang, E. Polizzi, and M. Lundstrom: A three-dimensional quantum simulation of silicon nanowire transistors with the effective-mass approximation, J. Appl. Phys. **96**, 2192-2203 (2004).
[117] S. Jin, Y. J. Park, and H. S. Min: A three-dimensional simulation of quantum transport in silicon nanowire transistor in the presence of electron-phonon interactions, J. Appl. Phys. **99**, 123719 (2006).
[118] H.-H. Park, L. Zeng, M. Buresh, S. Wang, G. Klimeck, S. R. Mehrotra, C. Heitzinger, B. P. Haley (2014), "Nanowire," https://nanohub.org/resources/nanowire.
[119] M. Shin: Full-quantum simulation of hole transport and band-to-band tunneling in nanowires using the k·p method, J. Appl. Phys. **106**, 054505 (2009).
[120] J. Z. Huang, W. C. Chew, J. Peng, C.-Y. Yam, L. J. Jiang, and G.-H. Chen: Model Order Reduction for Multiband Quantum Transport Simulations and its Application to p-Type Junctionless Transistors, IEEE Trans. Elec. Dev. **60**, 2111-2119 (2013).
[121] J. Z. Huang, L. Zhang, W. C. Chew, C.-Y. Yam, L. J. Jiang, G.-H. Chen, and M. Chan: Model Order Reduction for Quantum Transport Simulation of Band-To-Band

Tunneling Devices, IEEE Trans. Elec. Dev. **61**, 561-568 (2014).
[122] J. Huang, L. Zhang, P. Long, M. Povolotskyi, G. Klimeck: Quantum Transport Simulation of III-V TFETs with Reduced-Order k·p Method, Ch. 6; L. Zhang, M. Chan (Eds.): Tunneling Field Effect Transistor Technology:(Springer International, Switzerland, 2016), pp.151-180.
[123] J. Guo, S. Datta, M. Lundstrom, and M. Anantram: Toward Multiscale Modeling of Carbon Nanotube Transistors, Int. J. Multiscale Comp. Eng. **2**, 257-277 (2004).
[124] G. Fiori, G. Iannaccone, and G. Klimeck: Coupled Mode Space Approach for the Simulation of Realistic Carbon Nanotube Field-Effect Transistors, IEEE Trans. Nanotech. **6**, 475-480 (2007).
[125] R. Grassi, A. Gnudi, E. Gnani, S. Reggiani, and G. Baccarani: Mode Space Approach for Tight-binding Transport Simulation in Graphene Nanoribbon FETs, IEEE Trans. Nanotech. **10**, 371-378 (2011).
[126] M. Luisier: Quantum transport beyond the effective mass approximation (Doctoral Thesis ETH Zurich 2007).
[127] U. Hetmaniuk, D. Ji, Y. Zhao, M. P. Anantram: A Reduced-Order Method for Coherent Transport Using Green's Functions, IEEE Trans. Electr. Dev. **62**, 736-742 (2015).
[128] G. Milnikov, N. Mori, Y. Kamakura: Equivalent transport models in atomistic quantum wires, Phys. Rev. B **85**, 035317 (2012).
[129] A. Afzalian, J. Huang, H. Ilatikhameneh, J. Charles, D. Lemus, J. Bermeo, S. Rubiano, T. Kubis, M. Povolotskyi, G. Klimeck, M. Passlack, Y.-C. Yeo: Mode space tight-binding model for ultra-fast simulations of III-V nanowire MOSFETs and heterojunction TFETs, Proc. Int. Workshop on Computational Electronics (IWCE 2015) West Lafayette, Indiana USA, 2015(IEEE, Piscataway, NJ 2015) pp. 1-3
[130] M. Shin, W. J. Jeong, and J. Lee: Density functional theory based simulations of silicon nanowire field effect transistors, J. Appl.Phys. **119**, 154505 (2016).
[131] W. J. Jeong, J. Seo, and M. Shin, in Simulation of Semiconduc-tor Processes and Devices (SISPAD), 2016 International Conference on (IEEE, 2016) pp. 81
[132] J. Huang, H. Ilatikhameneh, M. Povolotskyi, G. Klimeck: Robust Mode Space Approach for Atomistic Modeling of Realistically Large Nanowire Transistors, J. Appl. Phys. **123**, 044303 (2018).
[133] S. Lee, F. Oyafuso, P. von Allmen, G. Klimeck: Boundary conditions for the electronic structure of finite-extent, embedded semiconductor nanostructures, Phys. Rev. B **69**, 045316 (2004).
[134] Y. He, Y. Tan, Z. Jiang, M.l Povolotskyi, G. Klimeck, T. Kubis: Surface Passivation in Empirical Tight-binding, IEEE Trans. Elec. Dev. **63**, 954-958 (2016).
[135] G. Klimeck, F. Oyafuso, R. C. Bowen, T. B. Boykin, T. Cwik, E. Huang, E. Vinyard: 3-D Atomistic Nanoelectronic Modeling on High Performance Clusters: Multimillion Atom Simulations, Superlattices and Microstructures **31**, 171-179 (2002).
[136] G. Klimeck, I. Woo, M. Usman, D. S. Ebert: "Self-Assembled Quantum Dot Wave Structure," (2011), https://nanohub.org/resources/10689.
[137] https://engineering.purdue.edu/gekcogrp/research-group/DanielMejia/
[138] T. B. Boykin, G. Klimeck: Practical Application of Zone-Folding Concepts in Tight-Binding calculations, Phys. Rev. B **71**, 115215 (2005).
[139] T. B. Boykin, N. Kharche, G. Klimeck, M. Korkusinski: Approximate bandstructures of semiconductor alloys from tight-binding supercell calculations, J. Phys.: Condens. Matter **19**, 036203 (2007) .
[140] T. B. Boykin, N. Kharche, G. Klimeck: Brillouin-zone unfolding of perfect supercells having nonequivalent primitive cells illustrated with a Si / Ge tight-binding parameterization, Phys. Rev. B **76**, 035310 (2007).
[141] N. Kharche, M. Luisier, T. B. Boykin, G. Klimeck: Electronic Structure and Transmission Characteristics of SiGe Nanowires, J. Comp. Elect. **7**, 350-354 (2008).
[142] A. Rahman, J. Guo, S. Datta, M. S. Lundstrom: Theory of Ballistic Nanotransistors, IEEE Trans. Elec. Dev. **50**, 1853-1864 (2003).
[143]N. Neophytou, A. Paul, M. Lundstrom, G. Klimeck: Simulation of nanowire transistors: Atomistic vs. Effective Mass Models, J. Comp. Electron. **7**, 363-366 (2008)
[144] Y. Liu, N. Neophytou, T. Low, G. Klimeck, M. Lundstrom: A Tight-binding Study of the Ballistic Injection Velocity for Ultrathin-body SOI MOSFETs, IEEE Trans. Elect. Dev. **55**, 866-871 (2008).
[145] Y. Liu, N. Neophytou, G. Klimeck, M. Lundstrom: Band-Structure Effects on the Performance of III-V Ultrathin-body SOI MOSFETs, IEEE Trans. Elect. Dev. **55**, 1116-1122 (2008).
[146] N. Neophytou, A. Paul, M. Lundstrom, G. Klimeck: Bandstructure Effects in Silicon Nanowire Electron Transport, IEEE Trans. Elect. Dev. **55**, 1286-1297 (2008)
[147] N. Neophytou, A. Paul, G. Klimeck: Bandstructure Effects in Silicon Nanowire Hole Transport, IEEE Trans. Nanotech.**7**, 710-719 (2008)
[148] G. Klimeck, N. Neophytou: Design Space for Low Sensitivity to Size Variations in [110] PMOS Nanowire Devices: The Implications of Anisotropy in the Quantization Mass, Nano Lett. **9**, 623–630 (2009)
[149] Á.Szabó , 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. Electr. Dev. 60, 2353-2360 (2013).
[150] Anisur Rahman, , Jing Guo, Md. Sayed Hasan, Yang Liu, Akira Matsudaira, Shaikh S. Ahmed, Supriyo Datta, Mark Lundstrom (2015), "FETToy," https://nanohub.org/resources/fettoy. (DOI: 10.4231/D38S4JQ3J).
[151] S.G. Kim, M. Luisier, B. P. Haley, A. Paul, S. R. Mehrotra, G. Klimeck, H. Ilatikhameneh: OMEN Nanowire" (2017), https://nanohub.org/resources/omenwire.
[152] M. Luisier, G. Klimeck: Atomistic Full-Band Simulations of Si Nanowire Transistors: Effects of Electron-Phonon Scattering, Phys. Rev. B**80**, 155430 (2009).
[153] J. Charles, P. Sarangapani, R. Golizadeh-Mojarad, R. Andrawis, D. Lemus, X. Guo, D. Mejia, J. Fonseca, M. Povolotskyi, T. Kubis, G. Klimeck: Incoherent transport in NEMO5: realistic and efficient scattering on phonons, J. Comp. Electr. **15**, 1123-1129 (2016).
[154] N. Neophytou, A. Paul, M. Lundstrom, G. Klimeck: Bandstructure Effects in Silicon Nanowire Electron Transport, IEEE Trans. Elect. Dev. **55**, 1286-1297 (2008).

[155] N. Neophytou, S.G. Kim, G. Klimeck, H. Kosina: On the bandstructure velocity and ballistic current of ultra-narrow silicon nanowire transistors as a function of cross section size, orientation, and bias, J. Appl. Phys., **107**, 113701 (2010).
[156] S. Mehrotra, S.G. Kim, T. Kubis, M. Povolotskyi, M. Lundstrom, G. Klimeck: Engineering Nanowire n-MOSFETs at Lg < 8nm, IEEE Trans. Elect. Dev. **60**, 2171-2177 (2013).
[157] M. Salmani-Jelodar, S. Mehrotra, H. Ilatikhameneh, G. Klimeck: Design Guidelines for Sub-12 nm Nanowire MOSFETs, IEEE Trans. Nanotech. **14**, 210-213 (2015).
[158] Y. Liu, N. Neophytou, G. Klimeck, M. Lundstrom: Band-Structure Effects on the Performance of III-V Ultrathin-body SOI MOSFETs, IEEE Trans. Elect. Dev. **55**, 1116-1122 (2008).
[159] S. Park, Y. Liu, N. Kharche, M. Salmani-Jelodar, G. Klimeck, M. Lundstrom, M. Luisier: Performance Comparisons of III-V and strained-Si in Planar FETs and Non-planar FinFETs at Ultra-short Gate Length (12nm), IEEE Trans. Elect. Dev. **59**, 2107-2114 (2012).
[160] S. Sylvia, H.-H. Park, M. Khayer, K. Alam, G. Klimeck, R. Lake: Material Selection for Minimizing Direct Tunneling in Nanowire Transistors, IEEE Trans. Elect. Dev. **59**, 2064-2069 (2012).
[161] N. Neophytou, A. Paul, G. Klimeck: Bandstructure Effects in Silicon Nanowire Hole Transport, IEEE Trans. Nanotech. **7**, 710-719 (2008).
[162] G. Klimeck, N. Neophytou: Design Space for Low Sensitivity to Size Variations in [110] PMOS Nanowire Devices: The Implications of Anisotropy in the Quantization Mass, Nano Lett. **9**, 623–630 (2009).
[163] A. Paul, S. Mehrotra, M. Luisier, G. Klimeck: Performance Prediction of Ultra-scaled SiGe/Si Core/Shell Electron and Hole Nanowire MOSFETs, IEEE Elect. Dev. Lett. **31**, 278-280 (2010).
[164] Publications of International Technology Roadmap for Semiconductors (itrs), 2013 ed. (http://www.itrs.net).
[165] T. Skotnicki et al., MASTAR 4.0 user manual (2011).
[166] M. Salmani-Jelodar, S. Kim, K. Ng, G. Klimeck: Transistor roadmap projection using predictive full-band atomistic modeling, Appl. Phys. Lett. **105**, 083508 (2014).
[167] W. Potz: Self-consistent model of transport in quantum well tunneling structures, J. of Applied Physics, **66,** 2458-2466 (1989)
[168] T. Kubis and P. Vogl: Assessment of approximations in nonequilibrium Green's function theory, Physical Review B, 83, 195304 (2011).
[169] K. Madhavan, M. Zentner, G. Klimeck: Learning and research in the cloud, Nature Nanotech. **8**, 786–789 (2013).
[170] https://nanohub.org/citations