Efficient Modeling of Electron Transport with Plane Waves

Maarten L. Van de Put, Akash A. Laturia, Massimo V. Fischetti, and William G. Vandenberghe
Department of Materials Science and Engineering
The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, Texas 75080, USA

***Abstract*—We present a method to simulate ballistic quantum transport in one-dimensional nanostructures, such as extremely scaled transistors, with a channel of nanowires or nanoribbons. In contrast to most popular approaches, we develop our method employing an accurate plane-wave basis at the atomic scale while retaining the numerical efficiency of a localized (tight-binding) basis at larger scales. At the core of our method is a finite-element expansion, where the finite element basis is enriched by a set of Bloch waves at high-symmetry points in the Brillouin zone of the crystal. We demonstrate the accuracy and efficiency of our method with the self-consistent simulation of ballistic transport in graphene nanoribbon FETs.**

I. Introduction

To assess the potential of highly scaled transistors, such as those presented in this work, predictive computational modeling is essential. In these devices, the short channel results in mostly ballistic conduction through the channel, while the overall dimensions necessitate the proper treatment of quantum effects such as tunneling and confinement. Therefore, a ballistic quantum transport model at the atomistic level is required. Historically, a wide variety of different techniques have been used to numerically implement such a transport model. Most current methods employ a variant of the tight-binding (TB) approximation since it offers excellent numerical performance and scalability [1]–[3]. On the other hand, for highly accurate *ab initio* calculations, an expansion on the plane-wave basis is often used [4], [5]. In contrast to the TB method, a plane-wave basis offers straightforward access to the real space information that is required to deal with position-dependent interactions as present in most electronic devices. However, transport calculations using plane waves in combination with the envelope expansion are known not to scale well to large structures [6].

To obtain the accuracy of plane waves with improved computational efficiency, we develop a new, physically grounded, numerical approach. In our approach, the computationally intensive solution of the crystal Hamiltonian in the plane wave basis is separated from the calculation of the envelope that determines the global transport properties of the device. In particular, we show that we can solve the crystal Hamiltonian with high accuracy using plane waves, leading to a fine spatial resolution, while retaining TB-like efficiency in our transport calculations.

Our paper is structured as follows. In Section II, we provide the theoretical details of our method. In Section III, the numerical implementation is discussed. In Section IV, we demonstrate our method by simulating a graphene nanoribbon transistor. Finally, we conclude in Section V.

II. Method

Here, we consider only homogeneous, one-dimensional atomic structures composed of a repeated supercell, as illustrated in Fig. 1 for a graphene nanoribbon (GNR).

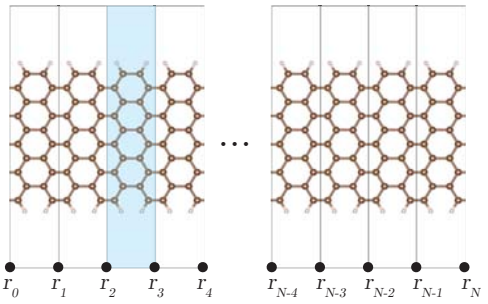

Figure 1. An illustration of the decomposition of an atomistic structure in supercells. The shaded supercell is repeated to make a structure. Locations for the nodes r_i are also indicated.

For such structures, we decompose the Hamiltonian as

$$\mathrm{H} = \mathrm{H}^{\mathrm{c}} + V^{\mathrm{e}}\,,$$

where H^{c} denotes the intrinsic crystal Hamiltonian, including, and $V^{\mathrm{e}}(\mathbf{r})$ is an extrinsic potential, *i.e.*, the additional self-consistent Hartree potential induced by the external bias. In the commonly adopted envelope function approximation, the extrinsic potential is assumed to be slowly varying and only interacts with the, coarsely discretized, envelope of the wavefunction. While we do not require the slowly-varying assumption, we will also use the decomposition of the Hamiltonian to discretize the wavefunctions and Hamiltonian.

As the first step in our method, we solve the $\mathbf{k}$-dependent crystal Hamiltonian of a single supercell ($\mathrm{H}^{\mathrm{c}}_{\mathbf{k}} = \mathrm{e}^{-\mathrm{i}\mathbf{k}\cdot\mathbf{r}}\mathrm{H}^{\mathrm{c}}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$) in a plane-wave basis, yielding the periodic part of the Bloch waves $u_{n\mathbf{k}}(\mathbf{r})$ and their eigenenergies $\epsilon_{n\mathbf{k}}$. In this work, we use the empirical pseudopotential method to obtain the Bloch waves [7], [8]. However, our methodology can straightforwardly be applied to solutions determined from first principles, *e.g.*, from density functional theory (DFT).

Next, we approximate the wavefunction in the whole device. We discretize the structure using a partition-of-unity [9], constructed using linear finite-element shape functions $f_i(\mathbf{r})$

978-1-5386-6790-3/18/$31.00 ©2018 IEEE

that extend over two supercells with nodes $\mathbf{r}_i$ centered between the supercells, as indicated in Fig. 1. To capture the atomic structure, the shape functions are enhanced with the Bloch waves $u_{n\mathbf{k}}(\mathbf{r})\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\mathbf{r}}$, positioned relative to $\mathbf{r}_i$. Formally, the wavefunctions are discretized as follows

$$\psi(\mathbf{r}) = \sum_{in\mathbf{k}} c_{in\mathbf{k}}\, f_i(\mathbf{r})\, u_{n\mathbf{k}}(\mathbf{r})\, \mathrm{e}^{\mathrm{i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}_i)}\,. \tag{1}$$

Specifically, the enhancement Bloch-waves are taken at the Γ-point and the Brillouin zone edge, while the band indices n are chosen to include all valence bands and a few (~ 10) conduction bands. Thanks to the compact support of the shape functions in $\mathbf{r}$, the expansion in (1) results in a nearest-neighbors coupling between adjacent nodes. In tight-binding terminology, the nodes correspond to atomic sites, while the Bloch waves (modulated by the shape functions) replace the localized orbitals.

Finally, applying the Galerkin method, the expansion (1) transforms the Schrödinger equation into a sparse generalized eigenvalue problem in the coefficient vector $\mathbf{c} = \{c_{in\mathbf{k}}\}$,

$$\left[\mathrm{T} + \mathrm{P} - \mathrm{Q} + \frac{\mathrm{M}\epsilon + \epsilon\mathrm{M}}{2} + \mathrm{V}^{\mathrm{e}}\right]\mathbf{c} = E\mathrm{M}\mathbf{c}\,, \tag{2}$$

where T is the kinetic energy matrix, P and Q are inter- and intra-node momentum coupling matrices, M is a mass matrix containing the basis function overlaps and ϵ is a diagonal matrix containing the eigenenergies associated to the Bloch waves. The exact form of each matrix element is provided in the Appendix. The matrices are of size $N \times N$ where $N = N_{\mathrm{nodes}}\, N_{\mathrm{Bloch}}$, with N_{nodes} the number of nodes (one more than supercells) and N_{Bloch} the number of Bloch waves taken in the expansion basis. Since only adjacent nodes interact, the matrices are very sparse with dense blocks of $N_{\mathrm{Bloch}} \times N_{\mathrm{Bloch}}$.

III. Implementation

To study the ballistic quantum transport in a system with open contacts, we use the quantum transmitting boundary method (QTBM) [10]. We calculate the contact self-energies $\Sigma_{\mathrm{L/R}}$ directly from the complex bandstructure of each contact and truncate Eq. (2) accordingly. Furthermore, since we are only interested in ballistic transport, we can gain some efficiency over the commonly used NEGF technique by directly injecting the $N_{\mathrm{L/R}}$ contact eigenmodes. For a given energy, the coefficients are then obtained by solving the system of equations

$$\left[\mathrm{T} + \mathrm{P} - \mathrm{Q} + \frac{\mathrm{M}\epsilon + \epsilon\mathrm{M}}{2} + \mathrm{V}^{\mathrm{e}} - E\mathrm{M} - \Sigma\right]\mathbf{c} = \mathrm{I}_{\mathrm{inject}}\,,$$

using LU factorization. Note that the injection matrix $\mathrm{I}_{\mathrm{inject}}$ is only of shape $N \times (N_{\mathrm{L}} + N_{\mathrm{R}})$.

In Fig. 2, an overview of the implemented solver is given. For the calculation of observables and more specifically the electron density, an integration over energies is required. To guarantee a threshold accuracy, we use a parallel implementation of the adaptive Simpson integration method to determine the energies at which (2) is solved. Using the definition of the expansion in Eq. (1), we convert the coefficients to a real-space wavefunction which yields real space observables with sub-atomic accuracy, as illustrated in Fig. 4. The free electron density, obtained in this way, is used in the self-consistent solution of the Poisson equation to determine $V_{\mathrm{e}}(\mathbf{r})$ under the application of bias and gating potentials. For efficiency, the Poisson equation is discretized using the finite-element method and solved using an iterative multi-grid method [11]. Self-consistency is obtained using the self-consistent Newton method, accelerated using the Direct Inversion of the Iterative Subspace (DIIS) method [12].

Finally, a separate integration is performed to extract the current through the device to arbitrary precision. By repeating this procedure for several bias points, we extract the transfer and output characteristics of the device.

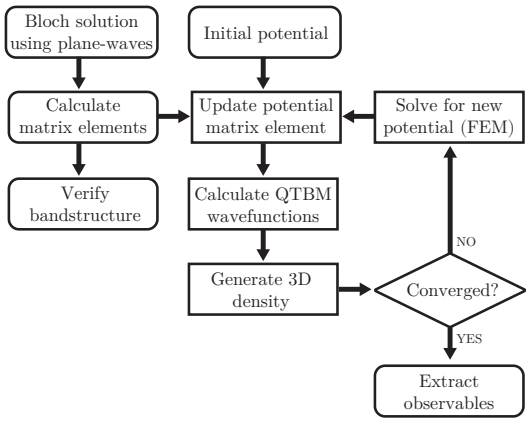

Figure 2. A high-level flow chart of the implemented solver. Elements with rounded corners indicate a single evaluation, while straight corners indicate updates within the self-consistent loop.

IV. Results

To demonstrate our method, we apply it to a highly scaled GNR field-effect transistor (FET) consisting of 80 supercells, containing a total of 2880 atoms, as depicted in Fig. 3. The device is 34 nm long with a channel width of approximately 2 nm. The gate is placed in a gate-all-around configuration with equivalent oxide thickness (EOT) of 1 nm and a gate length of 5 nm. For this device, we use a computational

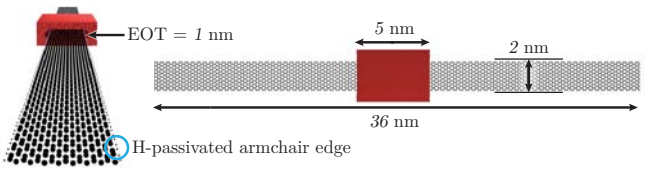

Figure 3. A depiction of the simulated GNR FET device, the nanoribbon is approximately 2 nm wide and 36 nm long for the transport calculations. The gate is 5 nm long indicated in red. The source and drain regions are uniformly doped in a 3.34 Å thick layer with 10^{20} cm^{-3} donors, while the channel under the gate is doped with with 10^{20} cm^{-3} acceptors. The simulations are done at room temperature (300 K)

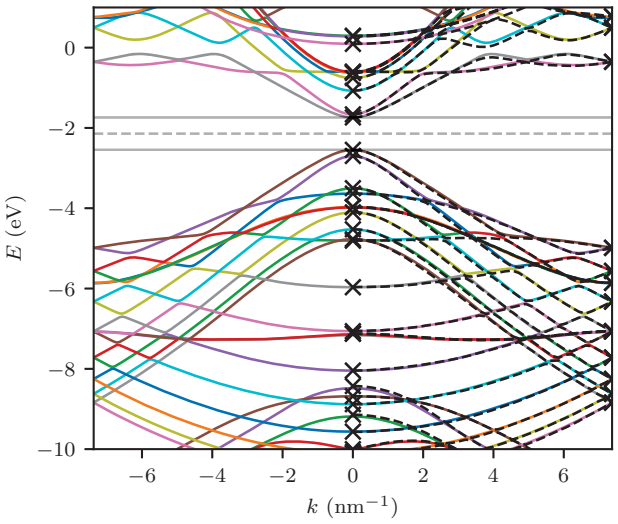

Figure 5. Reconstruction of the bandstructure of a GNR in the entire Brillouin zone from Bloch waves at the Γ-point in the center and Brillouin zone edge (indicated with ×). The band structure is shown in a region around the Fermi level (horizontal dashed line). The average error between the reconstruction (solid color lines) and the correct bandstructure (black dashed lines) in this range is around 25 meV.

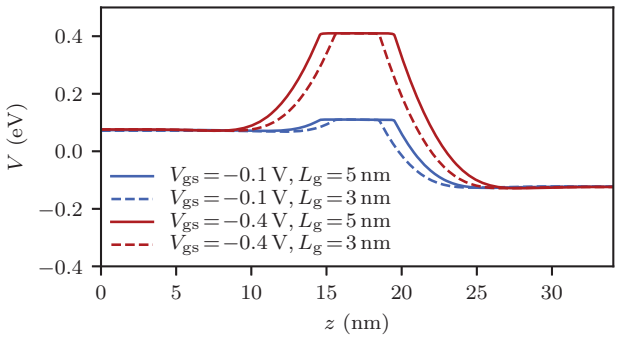

Figure 6. The potential energy, as seen by electrons, in the transport direction for different gate bias and gate length, averaged in the x and y directions. In this plot, the source-drain bias is fixed at 0.2 V, while the gate bias is set to −0.1 V (threshold) or −0.4 V (off-state).

basis consisting of the Bloch waves of all 66 valence bands and 10 conduction bands, evaluated at the center (Γ-point) and edge (Z-point) of the first Brillouin zone, for a total of 152 basis functions. These Bloch waves are calculated using the plane-wave empirical pseudopotential method using a fast solver demonstrated in our previous work [7]. To verify the accuracy of this basis expansion, we reconstruct the bandstructure from the Bloch-basis. As shown in Fig. 5, the GNR bandstructure obtained using these 152 basis functions is in good agreement with the reference bandstructure calculated between 13029 and 13238 plane waves, depending on the k-point. This is a difference of two orders of magnitude in required computational power.

Next, we find the self-consistent solution of the complete, open system shown in Fig. 3. In Fig 4, we show the final density, the potential energy and electric field in the off-state ($V_{gs} = -0.4\,V$ and $V_{ds} = 0.2\,V$). The x and y averaged potential is shown in Fig. 6 for reference. Note that these quantities are fully three dimensional and resolved at well below the atomic scale at minimal computational cost thanks to the expansion on the Bloch waves, which are computed on a plane-wave basis. The heavily reduced computational burden results in a self-consistent solution for the GNR FET at a particular bias from a flat potential in 60 minutes on a single CPU core, using less than 2 GB of memory. This highlights the potential for this method to scale to much larger structures.

Finally, in Fig 7, we show the device transfer characteristics $I_{ds}(V_{gs})$ for the GNR FET with gate lenghts of 3 nm and 5 nm. We sweep the gate voltage (V_{gs}) from −0.4 V to 0.3 V, at different bias points V_{ds}. As seen in Fig 7, the GNR FETs with short gates (3 nm) show poor sub-threshold performance, even with the excellent gate control offered by a gate-all-around configuration in combination with an atomically flat nanoribbon. This is a known issue for GNR FETs that is caused by high source-to-drain tunneling through the barrier of these extremely short gate devices [13]. The GNR with a longer gate, at 5 nm, has a much improved sub-

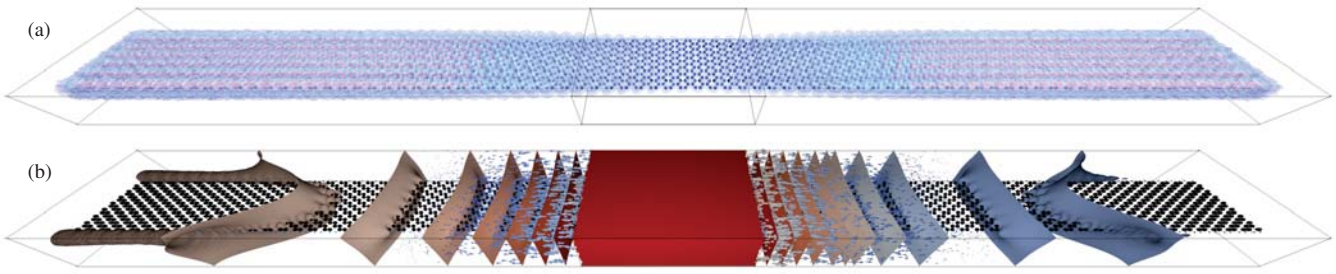

Figure 4. 3D renders of (a) the free electron density and (b) the potential energy iso-surfaces and electric field of the device shown in Fig 3. The device in the off-state ($V_{gs} = -0.4$ V and $V_{ds} = 0.2$ V). Red is used to indicate a higher value while blue indicates a lower value. The atomic positions are indicated as small balls and the electric field is illustrated with oriented and scaled arrows. The simulation domain and region of the gate are indicated with thin black lines.

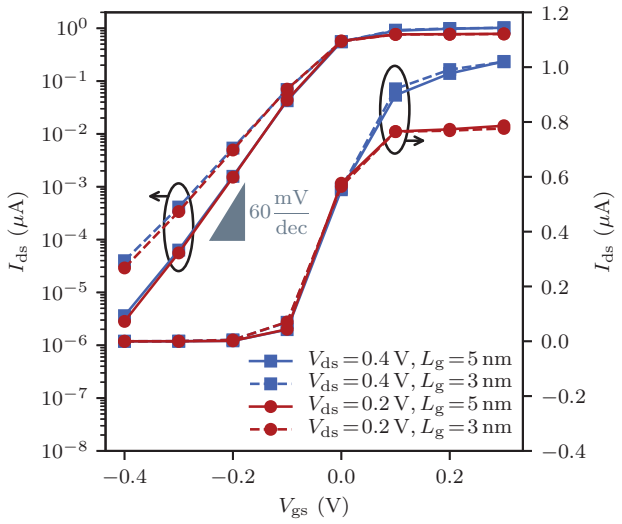

Figure 7. The device characteristics: source-drain current $I_{\rm ds}$ with respect to gate bias $V_{\rm gs}$ for different source-drain biases $V_{\rm ds}$ on a log scale (left axis) and a linear scale (right axis). For the logarithmic scale, the theoretical limit of $60\,{\rm mV/decade}$ sub-threshold slope is indicated for reference.

threshold performance with a minimum sub-threshold slope of $\sim 69\,{\rm mV/dec}$ at $V_{\rm ds} = 0.4\,{\rm V}$. The high source-to-drain leakage in the off-state in short channel GNR FETs prohibits their use in highly scaled, low power electronics, even in the limit of ideal ballistic transport. To correctly describe this deterioration of the sub-threshold behavior, a proper quantum mechanical treatment, as presented here, is warranted.

V. Conclusions

We presented a new technique to model full-band ballistic quantum transport using plane-waves by expanding on a Bloch wave basis at select high-symmetry points in the first Brillouin zone. This expansion results in a reduction of the computational cost, both in processing and memory, of two orders of magnitude, while retaining more than sufficient accuracy. We demonstrated our technique by self-consistently simulating a realistic GNR FET with over 2000 atoms. Furthermore, we have shown that our technique, being based on plane-wave calculations, provides direct access to sub-atomically resolved 3D densities.

Acknowledgements

This work is partially supported by the National Science Foundation, NSF grant 1710066.

Appendix

For completeness, we provide the matrix elements used in (2) (in atomic units for notational convenience):

$$
\begin{aligned}
{\rm M}_{in\mathbf{k};i'n'\mathbf{k}'} &= \int {\rm d}^3r\, f_i(\mathbf{r})\phi^*_{in\mathbf{k}}(\mathbf{r}) f_{i'}(\mathbf{r})\phi_{i'n'\mathbf{k}'}(\mathbf{r})\,,\\
{\rm T}_{in\mathbf{k};i'n'\mathbf{k}'} &= \frac{1}{2}\int {\rm d}^3r \big[\nabla f_i(\mathbf{r})\big]\phi^*_{in\mathbf{k}}(\mathbf{r})\cdot\big[\nabla f_{i'}(\mathbf{r})\big]\phi_{i'n'\mathbf{k}'}(\mathbf{r})\,,\\
{\rm P}_{in\mathbf{k};i'n'\mathbf{k}'} &= \frac{1}{4}\int {\rm d}^3r \big[\nabla f_i(\mathbf{r})\big]\phi^*_{in\mathbf{k}}(\mathbf{r})\cdot f_{i'}(\mathbf{r})\big[\nabla\phi_{i'n'\mathbf{k}'}(\mathbf{r})\big]\\
&\quad + \frac{1}{4}\int {\rm d}^3r\, f_i(\mathbf{r})\big[\nabla\phi^*_{in\mathbf{k}}(\mathbf{r})\big]\cdot\big[\nabla f_{i'}(\mathbf{r})\big]\phi_{i'n'\mathbf{k}'}(\mathbf{r})\,,\\
{\rm Q}_{in\mathbf{k};i'n'\mathbf{k}'} &= \frac{1}{4}\int {\rm d}^3r \big[\nabla f_i(\mathbf{r})\big]\cdot\big[\nabla\phi^*_{in\mathbf{k}}(\mathbf{r})\big] f_{i'}(\mathbf{r})\phi_{i'n'\mathbf{k}'}(\mathbf{r})\\
&\quad + \frac{1}{4}\int {\rm d}^3r f_i(\mathbf{r})\phi^*_{in\mathbf{k}}(\mathbf{r})\big[\nabla f_{i'}(\mathbf{r})\big]\cdot\big[\nabla\phi_{i'n'\mathbf{k}'}(\mathbf{r})\big]\,,\\
\epsilon_{in\mathbf{k};i'n'\mathbf{k}'} &= \delta_{ii'}\delta_{nn'}\delta_{\mathbf{k}\mathbf{k}'}E_{n\mathbf{k}}\,.
\end{aligned}
$$

where the shifted Bloch waves are $\phi_{in\mathbf{k}}(\mathbf{r}) = u_{n\mathbf{k}}(\mathbf{r}){\rm e}^{{\rm i}\mathbf{k}\cdot(\mathbf{r}-\mathbf{r}_i)}$.

References

[1] S. Steiger, M. Povolotskyi, H.-H. Park, T. Kubis, and G. Klimeck, "Nemo5: A parallel multiscale nanoelectronics modeling tool," *IEEE Transactions on Nanotechnology*, vol. 10, no. 6, pp. 1464–1474, 2011.

[2] T. B. Boykin, M. Luisier, and G. Klimeck, "Multiband transmission calculations for nanowires using an optimized renormalization method," *Physical Review B*, vol. 77, no. 16, p. 165318, 2008.

[3] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, "Density-functional method for nonequilibrium electron transport," *Physical Review B*, vol. 65, no. 16, p. 165401, 2002.

[4] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," *Physical review B*, vol. 54, no. 16, p. 11169, 1996.

[5] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo *et al.*, "Quantum espresso: a modular and open-source software project for quantum simulations of materials," *Journal of physics: Condensed matter*, vol. 21, no. 39, p. 395502, 2009.

[6] J. Fang, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, "Pseudopotential-based electron quantum transport: Theoretical formulation and application to nanometer-scale silicon nanowire transistors," *Journal of Applied Physics*, vol. 119, no. 3, p. 035701, 2016.

[7] M. L. Van de Put, W. G. Vandenberghe, B. Sorée, W. Magnus, and M. V. Fischetti, "Inter-ribbon tunneling in graphene: An atomistic bardeen approach," *Journal of Applied Physics*, vol. 119, no. 21, p. 214306, 2016.

[8] M. V. Fischetti, S. J. Aboud, Z.-Y. Ong, J. Kim, S. Narayanan, and C. E. Sachs, "Pseudopotential-based study of electron transport in low-dimensionality nanostructures," *ECS Transactions*, vol. 58, no. 7, pp. 229–234, 2013.

[9] I. Babuska and J. M. Melenk, "The partition of unity method," *International journal for numerical methods in engineering*, vol. 40, no. 4, pp. 727–758, 1997.

[10] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method," *Journal of Applied Physics*, vol. 67, no. 10, pp. 6353–6359, 1990.

[11] W. N. Bell, L. N. Olson, and J. B. Schroder, "PyAMG: Algebraic multigrid solvers in Python v3.0," 2015, release 3.2. [Online]. Available: https://github.com/pyamg/pyamg

[12] A. J. Garza and G. E. Scuseria, "Comparison of self-consistent field convergence acceleration techniques," *The Journal of chemical physics*, vol. 137, no. 5, p. 054110, 2012.

[13] J. Fang, S. Chen, W. G. Vandenberghe, and M. V. Fischetti, "Theoretical study of ballistic transport in silicon nanowire and graphene nanoribbon field-effect transistors using empirical pseudopotentials," *IEEE Transactions on Electron Devices*, vol. 64, no. 6, pp. 2758–2764, 2017.