ARTICLE IN PRESS

Computer Physics Communications xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computer Physics Communications

journal homepage: www.elsevier.com/locate/cpc

Scalable atomistic simulations of quantum electron transport using empirical pseudopotentials

Maarten L. Van de Put [*], Massimo V. Fischetti, William G. Vandenberghe

Department of Materials Science and Engineering, The University of Texas at Dallas, 800 W. Campbell Rd., Richardson, TX 75080, USA

ARTICLE INFO

Article history:
Received 4 March 2019
Received in revised form 7 June 2019
Accepted 9 June 2019
Available online xxxx

Keywords:
Quantum electron transport
Empirical pseudopotentials
Bloch waves
Finite elements

ABSTRACT

The simulation of charge transport in ultra-scaled electronic devices requires the knowledge of the atomic configuration and the associated potential. Such “atomistic” device simulation is most commonly handled using a tight-binding approach based on a basis-set of localized orbitals. Here, in contrast to this widely-used tight-binding approach, we formulate the problem using a highly accurate plane-wave representation of the atomic (pseudo)-potentials. We develop a new approach that separately deals with the intrinsic Hamiltonian, containing the potential due to the atomic configuration, and the extrinsic Hamiltonian, related to the external potential. We realize efficient performance by implementing a finite-element like partition-of-unity approach combining linear shape functions with Bloch-wave enhancement functions. We match the performance of previous tight-binding approaches, while retaining the benefits of a plane wave based model. We present the details of our model and its implementation in a full-fledged self-consistent ballistic quantum transport solver. We demonstrate our implementation by simulating the electronic transport and device characteristics of a graphene nanoribbon transistor containing more than 2000 atoms. We analyze the accuracy, numerical efficiency and scalability of our approach. We are able to speed up calculations by a factor of 100 compared to previous methods based on plane waves and envelope functions. Furthermore, our reduced basis-set results in a significant reduction of the required memory budget, which enables devices with thousands of atoms to be simulated on a personal computer.

1. Introduction

The numerical study of electron transport in solid-state transistors provides an important contribution to the improvement of future electronic devices. To keep ahead of technological progress, the methods used to predict electron transport behavior have shifted from simplified quasi-classical methods to advanced quantum mechanical descriptions. Historically, this evolution has been driven by the continual reduction of the length-scales to dimensions at which the classical limit is no longer appropriate. More recently, novel materials have been considered to improve the performance of electronic devices. For example, atomically thin monolayers, such as graphene, [1] phosphorene, [2] and transition-metal dichalcogenides, [3,4] and their ribbons, [5–8] are being actively investigated as possible replacements of silicon as the channel material in field-effect transistors. These materials have caused an additional shift from transport models based on bulk-material properties towards the comprehensive modeling of the atomic structure of the material. Whereas an atomistic description of quantum electron transport is widely applicable to different materials and device structures, atomistic resolution comes at a significant computational expense.

The atomistic calculation of the electronic structure starts by selecting an appropriate set of basis functions to discretize the problem. Two popular approaches, each at one end of the spectrum, are the Linear Combination of Atomic Orbitals (LCAO), which is closely related to the picture of chemical bonding, and plane-wave based methods which form a natural basis for the physics of periodic crystals. The most commonly used approximation of LCAO is the tight-binding (TB) approximation in which the interaction of the localized orbitals is short range, often only nearest neighbor (NN) orbitals being taken to overlap [9,10]. However, as remarked by Slater [11], in the interstitial region, away from the ionic cores, the wavefunction in a crystalline solid is plane-wave like. Due to the lack of non-bound states (*i.e.*, “scattering” or “traveling” wavefunctions) in the tight-binding basis, its accuracy is limited when describing higher energy valence and conduction states where electrons are located in the interstitial region. On the other hand, the plane-wave basis is a complete set whose accuracy can be carefully controlled by changing its truncation through a cutoff of the kinetic energy. However, to describe the core-states accurately, a high energy-cutoff is needed to obtain a sufficiently fine spatial resolution in the region close to the

[*] Corresponding author.
E-mail address: maarten.vandeput@utdallas.edu (M.L. Van de Put).

https://doi.org/10.1016/j.cpc.2019.06.009
0010-4655/© 2019 Elsevier B.V. All rights reserved.

ionic cores, a region that is more easily described by localized orbitals. For this reason, all-electron calculations often feature hybrid methods, using plane waves to describe the interstitial regions, augmented with a localized basis to capture the core states [12,13].

For the purposes of electron transport, we are interested in an accurate representation of the highest valence and lowest conduction states, which are, as discussed before, best captured by a plane-wave basis. However, plane waves are, by definition, not localized and interactions between all plane waves need to be considered; resulting in dense linear algebra formulations that have a high computational burden compared to the sparse linear algebra that results from tight-binding methods. For this reason the tight-binding approach is currently the most commonly used method to study quantum electron transport; using either a pre-defined set of orbitals with empirical parameters, *e.g.*, the well known $sp^3d^5s^*$ set, or using maximally localized Wannier functions to calculate the local orbitals from first-principles [9,14–16]. Commercial tight-binding transport simulators have already been developed to complement Technology Computer Aided Design (TCAD) in the semiconductor industry [17]. More limited investigations of plane-wave based transport have been undertaken academically, both based on *ab-initio* pseudopotentials [18,19] and empirical pseudopotentials [6,7]. In addition to the high accuracy of these plane-wave methods, they allow us to probe locally or disturb the interstitial region with impurities and local fields, for example. However, plane-wave methods have been applied only to relatively small atomic structures (up to thousands of atoms) due to their computational burden, and even for these small systems they require expensive high-performance computing infrastructure.

In this paper, we develop a method that combines the computational benefits of the tight-binding approach, while maintaining the versatility and accuracy of plane-wave methods to represent the real-space wavefunctions throughout the atomic structure. To achieve this goal, we turn to the Bloch waves of the crystal as an alternative basis to plane waves and tight-binding orbitals. Our approach generalizes mode-space [20,21] approaches that use the eigenmodes on cross-sections of the structure as a basis, instead of Bloch waves. In a similar spirit Bloch waves have been used in hybrid classical-quantum treatments of electronic transport in carbon nanotubes [22]. On the other hand, the benefits of using Bloch waves have been described for non-atomistic quantum models in the context of the linear combination of bulk bands (LCBB) method [23–26] and a recently developed empirical pseudopotential method for confined nanostructures [27]. In contrast to these methods, our method relies on an expansion on the Bloch-waves of the atomic structure. This enables the full quantum-mechanical treatment of atomistic nano-structures that do not have a bulk crystal counterpart or whose electronic structure is dissimilar to the bulk material, *e.g.*, carbon nanotubes, graphene nanoribbons, and extremely small silicon nanowires. In addition, the atomistic nature of our method provides access to the atomic positions which enables the study of lattice defects and impurities in a straightforward way.

We focus on transport through nanostructures featuring one-dimensional transport, *i.e.*, where the carriers are sufficiently confined such that they have only one degree of freedom. To describe the electronic structure of these nanostructures, we adopt the atomistic empirical pseudopotential approximation [6–8,28,29]. Note that we make the distinction between bulk and atomistic empirical pseudopotential methods. In the bulk empirical pseudopotential method, it is sufficient to know the values of the pseudopotential only at discrete reciprocal lattice vectors (form-factors). In our method, which we call the atomistic empirical pseudopotential method, the pseudopotential $V(\mathbf{q})$ is given as a function of a wave vector $\mathbf{q}$ in reciprocal space, yielding a more general method. Care must still be taken when transferring the pseudopotential from one system to another since one cannot expect, *a-priori*, that different atomic configurations can be described by a non self-consistent pseudopotential. However, there are known cases, such as the set of pseudopotentials for carbon nanostructures, introduced by Kurokawa [30], that show unexpected good performance for a wide range of atomic structures, including the graphene nanoribbons, we study as an example of a one-dimensional nanostructure in this work.

Our paper is structured as follows. In Section 2, we discuss the models for the atomic and electronic structure and develop the theory of our Bloch-wave basis. Section 3 details the calculation of the electronic properties in an open system with contacts. In Section 4, we explain the self-consistent procedure, coupling the electrostatics with the electron density in the system. Section 5 shows the application of our method to an armchair graphene-nanoribbon transistor, including verification of the accuracy and computational efficiency of our approach. Finally, we conclude in Section 6

2. Theoretical model

2.1. Model Hamiltonian

To model electron transport in nanoscaled devices, two length-scales should be considered: (1) The atomic (∼ Å) scale, which defines the electronic structure of the charge-carrying quasi-particles (electrons and holes), intrinsic to the material; (2) the device scale (∼ nm), determined by extrinsic factors such as applied fields, contacts and doping. For our purposes, we assume that the complex quasi-particle dynamics in a device is well-described by an effective single-particle Schrödinger equation of the form,

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + \left[V^{\mathrm{c}}(\mathbf{r}) + V^{\mathrm{e}}(\mathbf{r})\right]\psi(\mathbf{r}) = E\psi(\mathbf{r})\,, \tag{1}$$

where $V^{\mathrm{c}}(\mathbf{r})$ describes the intrinsic crystal potential, and the extrinsic potential $V^{\mathrm{e}}(\mathbf{r})$ captures the variations of the potential at the device length-scale. In our case, the crystal potential is given by local atomistic empirical pseudopotentials of each atom α,

$$V^{\mathrm{c}}(\mathbf{r}) = \sum_{\alpha} V^{\alpha}(|\mathbf{r} - \mathbf{R}_{\alpha}|)\,, \tag{2}$$

where $V^{\alpha}(r)$ represents the radial empirical pseudopotential of atom α, centered at location $\mathbf{R}_{\alpha}$. As will be highlighted later on, our method is not limited to this specific form of the crystal potential, and could be extended to non-local, and even *ab-initio* pseudopotentials. However, in this paper, we will limit our discussion to local empirical pseudopotentials of the form specified in Eq. (2).

Various existing computational models discretize Eq. (1) by introducing an appropriate basis-set to capture the smallest atomic scale. In tight-binding (TB) methods, a limited set of atomic orbitals is used to capture the atomic scale, while on-site potential variations are used to capture the extrinsic potential [9,31,32]. In plane-wave based pseudopotential methods, the envelope-function approach has been used to capture the extrinsic potential variations [7,33]. In both approaches, the total Hamiltonian in Eq. (1), including the extrinsic potential that varies only at the device scale, is solved on the basis set that is used to capture the small atomic scale (atomic orbitals or plane-waves). This is acceptable for the TB method that scales linearly and features a small basis set of N_{orbitals} and $\mathcal{O}(N_{\mathrm{bands}}N_{\mathrm{orbitals}})$ complexity, thanks to their nearest-neighbor interactions [9,32]. Plane wave methods, on the other hand, are severely restricted by their large number of plane waves ($N_{\mathbf{G}}$) that scales with the volume of the structure rather than the number of electron. Efficient plane-wave methods, using the Fast Fourier Transform (FFT) algorithm, reduce the complexity of plane-wave algorithms to $\mathcal{O}(N_{\text{bands}} N_{\mathbf{G}} \log N_{\mathbf{G}})$, albeit with a rather large pre-factor [28,34]. However, the lack of periodic boundary conditions in the transport direction (z), induced by the extrinsic potential, prohibits the use of the FFT algorithm in the transport direction, increasing the complexity to $\mathcal{O}(N_{\text{bands}}[N_{G_z}^2 + N_{\mathbf{G}} \log N_{\mathbf{G}_{xy}}])$. The large basis set, combined with sub-optimal scaling, necessitates a different approach for transport calculations that use plane-wave pseudopotentials.

Instead of treating the intrinsic crystal Hamiltonian and the extrinsic potential with a single method, we propose an alternative approach, where the atomic and device scales are decoupled. First, we determine the Bloch wave solutions of the intrinsic crystal Hamiltonian, and in a second step, we solve the Hamiltonian of the entire device. This approach allows us to simulate systems that are currently inaccessible using plane-wave based atomistic pseudopotentials [6,7].

Fig. 1 shows a typical target structure, featuring one-dimensional electron transport, which is assumed to be in the z-direction. The structure consists of a supercell that is periodically repeated N_{block} times in the transport direction. The periodic supercell completely captures the atomic configuration of the one-dimensional crystal. Extensions to inhomogeneous systems, where the supercell changes throughout the structure are possible, but left for future work.

2.2. Bloch-wave expansion

Our method is constructed around an expansion of the wavefunctions on a Bloch-wave basis. At a high level, our method proceeds as follows: we separate the device in its supercells, we calculate the Bloch waves in each supercell and "stitch" them together using finite-elements. Fig. 2 shows the different ingredients for the basis, which we detail in this section.

The first ingredient of our method is the Bloch-waves of the atomic structure, as illustrated in the first panel of Fig. 2. For a single repeated supercell, we compute the solution of the intrinsic crystal Hamiltonian with periodic boundaries,

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V^{\mathrm{c}}(\mathbf{r})\right]\left[u_{nk}(\mathbf{r})e^{\mathrm{i}kz}\right] = \epsilon_{nk}\, u_{nk}(\mathbf{r})e^{\mathrm{i}kz}.$$

The solutions are the Bloch functions $u_{nk}(\mathbf{r})e^{\mathrm{i}kz}$, with band index n and wave vector k in the direction of transport. The Bloch-wave solutions are obtained to high precision using the appropriate plane-wave basis, where computational efficiency is realized using FFTs [28].

The second ingredient is a one-dimensional finite element (FE) discretization in the transport direction which will "stitch" together the supercells and allow for the capture of any extrinsic fields. The finite element discretization uses the supercells as elements, with nodes z_i located on the interface between the supercells along the transport direction, as shown in Fig. 1. The FE shape functions $f_i(\mathbf{r})$, as shown in the second panel of Fig. 2, are the standard linear FE 'hat' shape functions which obey $f_i(\mathbf{r}_j) = \delta_{ij}$.

The last panel of Fig. 2 shows the product of the FE shape functions $f_i(\mathbf{r})$ and the node-centered Bloch-waves, defined as:

$$\phi_{ink}(\mathbf{r}) = u_{nk}(\mathbf{r})e^{\mathrm{i}k(z-z_i)}\,. \tag{3}$$

The products $f_i(\mathbf{r})\phi_{ink}(\mathbf{r})$ form the Bloch-wave basis-functions on which the wavefunction is expanded,

$$\psi(\mathbf{r}) = \sum_{ink} c_{ink}\, f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,. \tag{4}$$

The shape functions $f_i(\mathbf{r})$ capture the overall, global variation of the wavefunction, much like the slowly varying envelope functions commonly used. Note that the shape functions $f_i(\mathbf{r})$ also serve to localize the basis functions within the two elements around the node. The explicit inclusion of the wave vector k in the node-centered Bloch-waves allows for the expansion on more than one (high-symmetry) point of the reciprocal lattice. Particularly, in Section 5.1, we demonstrate that a basis built using Bloch waves at the zone-center (Γ) and zone-edge (X) yields an accurate description throughout the entire Brillouin zone.

The expansion presented in Eq. (4) is a specific application of the Partition-of-Unity Method (PUM) [35–37]. In the PUM, a set of overlapping patches $\{\Omega_i\}$ is defined which form an open cover of the complete coordinate space Ω, covering the device. In our case, a patch Ω_i is defined as the union of the two supercells touching the node z_i. Adopting the PUM terminology, a shape function $f_i(\mathbf{r})$ takes on the role of a patch function that is only supported on the patch Ω_i. The set of patch (shape) functions $\{f_i(\mathbf{r})\}$ satisfies $\forall \mathbf{r} \in \Omega : \sum_i f_i(\mathbf{r}) = 1$ and is therefore called a partition-of-unity on the full domain Ω. The PUM allows for the further enhancement of each patch with a set of functions $\{\phi_{ink}(\mathbf{r})\}$ that span an appropriate subspace of the solution space on the patch $\{\phi_{ink}(\mathbf{r})|\phi_{ink}(\mathbf{r}) \subset \mathrm{H}^1(\Omega_i)\}$. In other words, the linear combination of $\phi_{ink}(\mathbf{r})$ should be a good approximation of the solution on the patch. In our case, the node-centered Bloch-waves $\phi_{ink}(\mathbf{r})$ take on the role of enhancement functions, capturing the solution on the atomic scale within the supercell. The wavefunction function is then well approximated in the solution space of the full domain $\{\psi(\mathbf{r})|\psi(\mathbf{r}) \subset \mathrm{H}^1(\Omega)\}$ by an expansion on the patches, as defined in Eq. (4), where the expansion coefficients c_{ink} are to be determined numerically. Note that the partition of unity formed by $f_i(\mathbf{r})$ enforces continuity of the solution independent of the enhancement functions (node-centered Bloch-waves) $\phi_{ink}(\mathbf{r})$.

In general, the enhancement functions in the PUM expansion can be patch-dependent and only have to capture the local variations of the solution space. In this paper, we consider homogeneous structures (as shown in Fig. 1) with a single repeated supercell and reuse a single set of Bloch waves for every patch. However, in general, our method can be extended to use patch-dependent Bloch waves that are able to capture variations in the atomic structure.

2.3. Matrix equations

Inserting the expression for the wavefunction in Eq. (4), into the Schödinger equation (1), we determine a linear system of equations for the expansion coefficients c_{ink}. Following the Galerkin method, we convert the Schödinger equation into a weak form, multiplying it by a test function $\bar{\psi}(\mathbf{r})$ and integrating it over the full domain Ω,

$$-\frac{\hbar^2}{2m}\int_{\Omega} \mathrm{d}^3r\, \bar{\psi}(\mathbf{r})\mathrm{H}^{(\mathrm{c})}\psi(\mathbf{r}) + \int \mathrm{d}^3r\, \bar{\psi}(\mathbf{r})[V^{\mathrm{e}}(\mathbf{r}) - E]\psi(\mathbf{r}) = 0\,. \tag{5}$$

This weak form is equivalent to the Schrödinger equation when the test functions $\bar{\psi}(\mathbf{r})$ span the full solution space. The complex conjugate of the wavefunctions forms a natural choice for the test functions in Eq. (5). After expansion, Eq. (5) becomes

$$\sum_{\substack{i\,k\,n\\ i'k'n'}} \bar{c}_{i'n'k'}\left[\mathrm{H}^{\mathrm{c}}_{i'n'k',ink} + \mathrm{V}^{\mathrm{e}}_{i'n'k',ink} - E\,\mathrm{M}_{i'n'k',ink}\right] c_{ink} = 0\,, \tag{6}$$

where we have introduced the matrix elements

$$\mathrm{M}_{i'n'k',ink} = \int_{\Omega} \mathrm{d}^3r\, f_{i'}^{*}(\mathbf{r})\phi_{i'n'k'}^{*}(\mathbf{r})f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,, \qquad \text{(overlap / "mass")} \tag{7}$$

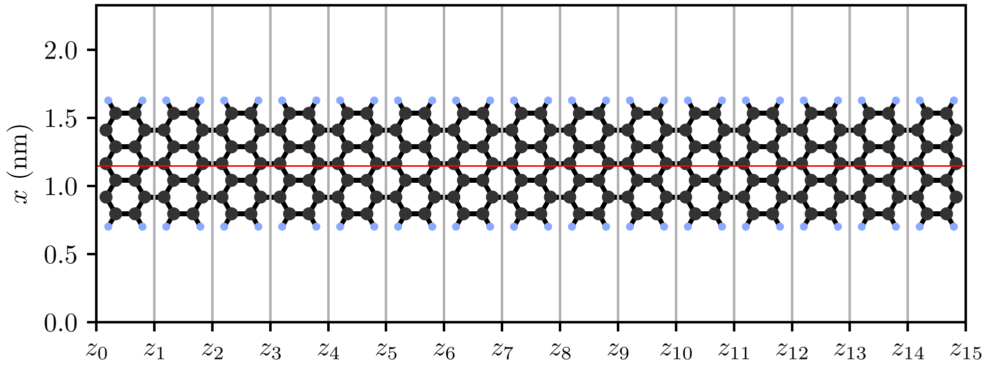
x (nm)
2.0
1.5
1.0
0.5
0.0
z_0 z_1 z_2 z_3 z_4 z_5 z_6 z_7 z_8 z_9 z_{10} z_{11} z_{12} z_{13} z_{14} z_{15}

Fig. 1. A top-view of an armchair graphene nanoribbon, where carbon (black) and hydrogen (blue) atom positions are indicated with spheres and where black lines represent chemical bonds. Electron transport proceeds in the z-direction, where node positions z_i indicate the boundaries between repeated supercells. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

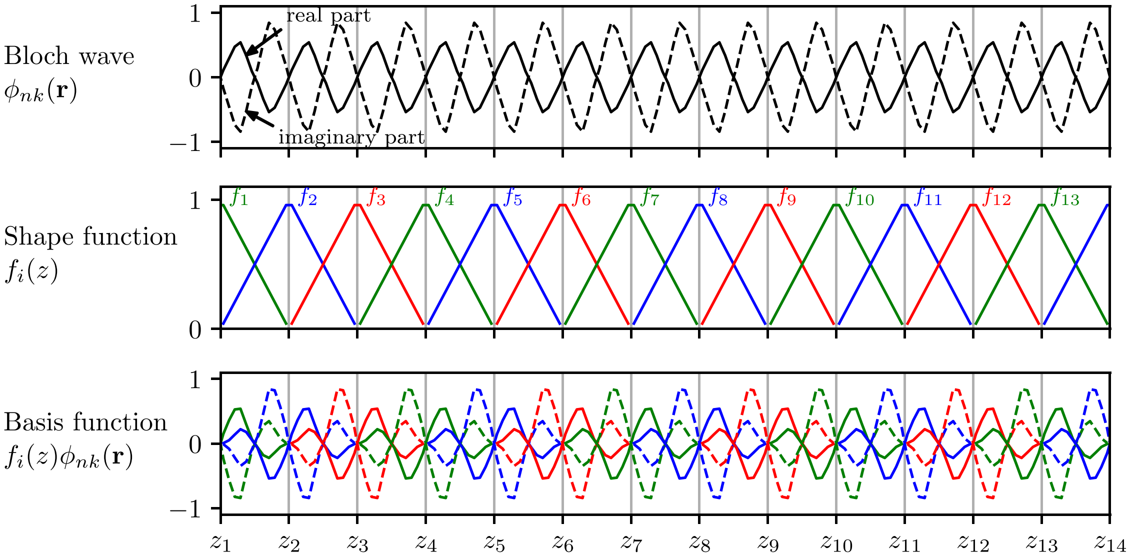

Fig. 2. An illustration of the components of the basis set used to expand the wavefunction for the armchair graphene nanoribbon shown in Fig. 1. The Bloch wave of the 32nd band ($n = 31$) at the Γ-point ($k = 0$) is shown along a cut-line through the middle of the ribbon. The triangular shape functions, forming a partition of unity, are shown for all nodes. The local basis functions that are shown correspond to the Bloch wave in the first panel, *i.e.*, $n = 31$ and $k = 0$, and are plotted along the same cut-line. The Bloch-wave and basis-function units are arbitrary.

$$\mathrm{H}^{\mathrm{c}}_{i'n'k',ink} = \int_{\Omega} \mathrm{d}^3 r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\mathrm{H}_{\mathrm{c}}(\mathbf{r})f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,, \qquad \text{(crystal Hamiltonian)} \tag{8}$$

$$\mathrm{V}^{\mathrm{e}}_{i'n'k',ink} = \int_{\Omega} \mathrm{d}^3 r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})V_{\mathrm{e}}(\mathbf{r})f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,. \qquad \text{(extrinsic potential)} \tag{9}$$

Note that using the complex conjugates of the Bloch basis as the test functions preserves the Hermiticity of the discretized Hamiltonian and overlap matrices.

The direct evaluation of the crystal Hamiltonian matrix elements in Eq. (8) requires the use of the crystal potential. While this is fairly easy for the case of the local empirical pseudopotential approximation, the evaluation of the crystal Hamiltonian in, *e.g.*, *ab-initio* methods, can be more cumbersome or computationally expensive. To make our model independent of the intricacies to evaluate the crystal Hamiltonian, we avoid the direct use of the crystal potential itself by substituting the eigenvalues of the crystal Hamiltonian in Eq. (8) (the full details are given in Appendix),

$$\mathrm{H}^{\mathrm{c}}_{i'n'k',ink} = \frac{\epsilon_{ink} + \epsilon_{i'n'k'}}{2}\,\mathrm{M}_{i'n'k',ink} + \mathrm{T}_{i'n'k',ink} + \mathrm{P}_{i'n'k',ink}\,, \tag{10}$$

where ϵ_{ink} is the eigenvalue of the corresponding Bloch wave $\phi_{ink}(\mathbf{r})$ and two new matrix elements have been defined as:

$$\mathrm{T}_{i'n'k',ink} = \frac{\hbar^2}{4m}\int_{\Omega} \mathrm{d}^3 r\, \nabla\big[f^*_{i'}(\mathbf{r})f_i(\mathbf{r})\big]\cdot\nabla\big[\phi^*_{i'n'k'}(\mathbf{r})\phi_{ink}(\mathbf{r})\big] \tag{11}$$

$$+\frac{\hbar^2}{2m}\int_{\Omega} \mathrm{d}^3 r\, \big[\nabla f^*_{i'}(\mathbf{r})\big]\phi^*_{i'n'k'}(\mathbf{r})\cdot\big[\nabla f_i(\mathbf{r})\big]\phi_{ink}(\mathbf{r})\,, \qquad \text{(kinetic energy)} \tag{12}$$

$$\mathrm{P}_{i'n'k',ink} = -\frac{\hbar^2}{m}\int_{\Omega} \mathrm{d}^3 r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\big[\nabla f_i(\mathbf{r})\big]\cdot\big[\nabla\phi_{ink}(\mathbf{r})\big] + \text{h.c.}\,, \qquad \text{(momentum coupling)} \tag{13}$$

where h.c. has been used to indicate the Hermitian conjugate of the previous term, swapping indices *ink* and $i'n'k'$.

Since Eq. (6) has to hold for all test functions, *i.e.*, all coefficients c^*_{ink}, we write,

$$\sum_{ink}\left[T_{i'n'k',ink} + V_{i'n'k',ink} + \frac{(\epsilon_{ink} + \epsilon_{i'n'k'})}{2}M_{i'n'k',ink} + P_{i'n'k',ink}\right]c_{ink} = E\sum_{ink} M_{i'n'k',ink}\,c_{ink}\,. \tag{14}$$

This generalized eigenvalue problem can be written in matrix form as $\mathrm{H}\mathbf{c} = E\mathrm{M}\mathbf{c}$. Note that, apart from the extrinsic potential, all the matrix elements depend only on the properties of the material, not on those of the device, and are independent of changes of the extrinsic potential. Thanks to the shape functions, only elements for which i and i' are equal or refer to nearest-neighbor nodes are non-zero. The matrices H and M have a block tridiagonal form. For example, the Hamiltonian matrix is written as:

$$\mathrm{H} = \begin{bmatrix} \ddots & & & & & & \iddots \\ & \mathrm{H}_{i-1,i-2} & \mathrm{H}_{i-1,i-1} & \mathrm{H}_{i-1,i} & 0 & 0 & \\ & 0 & \mathrm{H}_{i,i-1} & \mathrm{H}_{i,i} & \mathrm{H}_{i,i+1} & 0 & \\ & 0 & 0 & \mathrm{H}_{i+1,i} & \mathrm{H}_{i+1,i+1} & \mathrm{H}_{i+1,i+2} & \\ \iddots & & & & & & \ddots \end{bmatrix} \tag{15}$$

where each block $\mathrm{H}_{ii'}$ (and $\mathrm{M}_{ii'}$ for the overlap matrix) is a square matrix with size equal to the number of basis functions used in a supercell N_{basis}. Correspondingly, the solution vector $\mathbf{c}$ combines the column vectors $\mathbf{c}_i$ that contain the expansion coefficients for slice i.

3. Open system

Having obtained a suitable discretization of the atomic structure, we now turn to the calculation of the electronic transport properties in devices. We consider an open system with injecting and absorbing contacts on either side of the device, here referred to as source (s) and drain (d). Both contacts are considered infinite reservoirs which inject electrons in thermodynamic equilibrium and absorb all incident waves. We employ the quantum transmitting boundary condition method (QTBM) [38] to model the contacts and calculate the extended states that are injected from each contact.

3.1. Contact self-energies

The calculation of contact self-energies using iterative and direct approaches (as used here) is already well established in literature [32,39–41]. Nonetheless, we will detail the procedure here. Our reasons for this are twofold; (1) our basis, being non-orthogonal, introduces additional complexity that, to our knowledge, has not been previously described for the direct approach, and (2) our numerical approach avoids some numerical errors in calculating the self-energies directly. We note that this procedure can be applied to calculate the self-energies for other non-orthogonal bases, for example in non-orthogonal Gaussian-type tight-binding [31] and projector-augmented wave methods [13].

We calculate the self-energies $\Sigma_{\mathrm{s/d}}$, associated with the truncation of the block matrices in Eq. (15) at the open contacts using a direct, non-iterative, method. For this purpose, we calculate the so-called complex band structure at the source or drain node $i \in \{\mathrm{s}, \mathrm{d}\}$, for a given energy E, as the solution of the non-linear eigenvalue problem

$$\left[\mathrm{H}_i(\lambda) - E\mathrm{M}_i(\lambda)\right]\mathbf{c}_i = 0\,, \tag{16}$$

where the eigenvalues $\lambda = \mathrm{e}^{ik\,\Delta z}$ are the phase difference between the edge node i and its nearest neighbor inside the contact $i+1$ for the drain ($i-1$ for the source), with $\Delta z = z_i - z_{i+1}$. The polynomial matrices are given by

$$\begin{aligned} &\mathrm{H}_i(\lambda) = \lambda^{-1}\mathrm{H}_{i,i-1} + \mathrm{H}_{i,i} + \lambda\mathrm{H}_{i,i+1} \quad \text{and} \\ &\quad \mathrm{M}_i(\lambda) = \lambda^{-1}\mathrm{M}_{i,i-1} + \mathrm{M}_{i,i} + \lambda\mathrm{M}_{i,i+1}\,. \end{aligned} \tag{17}$$

Eq. (16) represents a second-order, generalized eigenvalue equation. This can be solved readily by linearizing the second-order eigenvalue problem to a first order problem of double the rank. To avoid excessive numerical round-off errors in the calculation of the eigenvalues λ, we linearize Eq. (16) using the symmetric scheme from Ref. [42],

$$\begin{aligned} &\begin{bmatrix} \mathrm{H}_{i,i} - E\mathrm{M}_{i,i} & \mathrm{H}_{i,i-1} - E\mathrm{M}_{i,i-1} \\ \mathrm{H}_{i,i+1} - E\mathrm{M}_{i,i+1} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix} \\ &\quad = \lambda \begin{bmatrix} -\left(\mathrm{H}_{i,i+1} - E\mathrm{M}_{i,i+1}\right) & 0 \\ 0 & \mathrm{H}_{i,i+1} - E\mathrm{M}_{i,i+1} \end{bmatrix} \begin{bmatrix} \mathbf{d}_i \\ \mathbf{c}_i \end{bmatrix}, \end{aligned} \tag{18}$$

where $\mathbf{d}_i = \lambda\mathbf{c}_i$, and the left-hand-side is a Hermitian matrix, since $\mathrm{H}_{i,i+1} = \mathrm{H}^{\dagger}_{i,i-1}$ and $\mathrm{M}_{i,i+1} = \mathrm{M}^{\dagger}_{i,i-1}$.

Eq. (18) is solved to machine precision using a direct linear eigenvalue solver and admits $2N_{\mathrm{basis}}$ solution pairs $(\lambda_\nu, \mathbf{c}_\nu)$. Based on the phase factors λ_ν, we sort them into two sets of size N_{basis} each, the in-flowing and out-flowing solutions. To determine flow-direction, we calculate the group velocity v_ν of each eigenvector $\mathbf{c}_\nu$ using a generalization of the Hellmann–Feynman theorem [43],

$$v_\nu = \frac{1}{\hbar}\frac{\partial E_\nu}{\partial k} = \frac{1}{\hbar}\frac{\left\langle \mathbf{c}_\nu \right| \frac{\partial}{\partial k}\mathrm{H}_i(\lambda_\nu) \left| \mathbf{c}_\nu \right\rangle - E_\nu \left\langle \mathbf{c}_\nu \right| \frac{\partial}{\partial k}\mathrm{M}_i(\lambda_\nu) \left| \mathbf{c}_\nu \right\rangle}{\left\langle \mathbf{c}_\nu \right| M_i(\lambda_\nu) \left| \mathbf{c}_\nu \right\rangle}\,. \tag{19}$$

The set of solutions with an out-flow (in-flow) condition is split into purely traveling waves with $|\lambda_\nu| = 1$ and $v_\nu > 0$ ($v_\nu < 0$), and evanescent modes where $|\lambda_\nu| < 1$ ($|\lambda_\nu| > 1$). In practical implementations, a tolerance should be used to determine the traveling waves, *i.e.*, $|\lambda_\nu| = 1 \pm \varepsilon$. Thanks to the increased accuracy of the symmetric linearization of Eq. (16), we obtained a drastic improvement in the accuracy of $|\lambda_\nu|$ and all traveling modes satisfy $|\lambda_\nu| = 1$ to machine precision in all our tests. In the envelope-function approximation, a necessary additional step is the removal of spurious solutions [7]. However, our method does not admit spurious traveling solutions within (or below) the energy range spanned by the Bloch waves in the basis set, negating the need for additional filtering.

Before proceeding, care must be taken to correctly normalize the traveling wavefunctions in each contact. In the infinitely long contacts, the wavefunctions for different values of the crystal momentum k_z are orthonormal, $\int_{\Omega_{\mathrm{s/d}}} \mathrm{d}^3r\, \psi^*_{k_z}(\mathbf{r})\psi_{k'_z}(\mathbf{r}) = \delta(k_z - k'_z)$, where the domain $\Omega_{\mathrm{s/d}}$ spans the entire infinite contact. When the integration domain is reduced to a single supercell Ω_{sc}, the normalization condition for the wavefunction becomes

$$\int_{\Omega_{\mathrm{sc}}} \mathrm{d}^3r\, \psi^*_{k_z}(\mathbf{r})\psi_{k_z}(\mathbf{r}) = \frac{L_z}{2\pi}\,, \tag{20}$$

where L_z is the length of the supercell along the transport direction. In terms of our wavefunction expansion, the condition is straightforward:

$$\left\langle \mathbf{c}_\nu \right| \mathrm{M}_i(\lambda_\nu) \left| \mathbf{c}_\nu \right\rangle = \frac{L_z}{2\pi}\,. \tag{21}$$

This normalization condition is applied immediately upon identification of the running modes we obtain after solving the complex band structure in Eq. (16).

For each contact node $i \in \{\mathrm{s}, \mathrm{d}\}$, we define a Bloch matrix, $\mathrm{B}_i = [\mathbf{c}_1, \ldots, \mathbf{c}_\nu, \ldots, \mathbf{c}_N]$, whose columns are the out-flow eigenvectors $\mathbf{c}_\nu$ of the respective contact. The contact self-energy of the contact-node $i \in \{\mathrm{s}, \mathrm{d}\}$ is built by projecting the wavefunction in the device on the out-flowing waves,

$$\Sigma_i = \left[\mathrm{H}'_i - E\mathrm{M}'_i\right]\mathrm{B}_i \Lambda \mathrm{B}_i^{-1}\,, \tag{22}$$

where $\Lambda_{i,\mathrm{out}}$ is a diagonal matrix with elements given by the out-flow λ_i, while the $N_{\mathrm{basis}} \times N_{\mathrm{basis}}$ matrices H'_i and M'_i correspond to the truncated matrices just outside the simulation domain, *e.g.*,

$\mathrm{H}_i' = \mathrm{H}_{i,i-1}$ for the source contact. The effect of the projection can be understood as follows: B^{-1} converts the wavefunction into the coefficients of each mode, Λ propagates the coefficients to the next node by multiplying each mode with $e^{\mathrm{i}k_z\Delta z}$ and B converts the coefficients back into its wavefunction form. Finally, we define Σ, a matrix of the size of the system ($N_{\mathrm{basis}}N_{\mathrm{block}} \times N_{\mathrm{basis}}N_{\mathrm{block}}$) that contains the two contact self-energy matrices Σ_{s} and Σ_{d} at their respective positions on the diagonal, and is zero otherwise.

3.2. Extended states

Using the contact self-energies, we calculate the extended states of the open system by solving directly for the coefficients c_{ink} of the wavefunction,

$$\big[E\mathrm{M} - \mathrm{H} - \Sigma\big]\mathbf{c} = \mathrm{B}\,, \tag{23}$$

where the right-hand-side matrix B has N_{mode} columns that each represent the injection of a single eigenmode from one of the contacts. For each in-flowing mode γ in each contact node $i \in \{\mathrm{s}, \mathrm{d}\}$, with coefficients $\mathbf{c}_{i,\gamma}$ and phase $\lambda_{i,\gamma}$, we obtain

$$\mathrm{B}_{i,\gamma} = \big[(\mathrm{H}_i' - E\mathrm{M}_i')\lambda_{i,\gamma} - \Sigma_i\big]\mathbf{c}_{i,\gamma}\,, \tag{24}$$

with B zero everywhere else. Having calculated the coefficients for all injected modes γ from all contacts by solving Eq. (23) at a certain energy, the expansion in Eq. (4) is used to express the wavefunctions in the real-space basis:

$$\psi_\gamma(\mathbf{r}) = \sum_{ink} c_{\gamma,ink} f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,. \tag{25}$$

The label γ is used to identify both the originating contact (s/d) and individual injected mode index.

Rather than following the procedure described above, we could also use the popular nonequilibrium Green's function (NEGF) approach and solve for the Green's function in our Bloch wave basis $\mathrm{G} = [E\mathrm{M} - \mathrm{H} - \Sigma]^{-1}$. NEGF can be implemented efficiently by using an appropriate recursive technique, calculating only the diagonals and off-diagonals of the Green's function [44–46]. Such a recursive Green's function approach would, in our case, reduce the computational complexity from the inversion of the entire Hamiltonian, $\mathcal{O}(N_{\mathrm{blocks}}^2 \times N_{\mathrm{basis}}^2)$, to the inversion of the individual blocks of size N_{basis}, *i.e.*, $\mathcal{O}(N_{\mathrm{blocks}} \times N_{\mathrm{basis}}^2)$. However, in general, the number of traveling modes N_{mode} using wavefunctions is much smaller than the number of basis vectors N_{basis} at a single node. Therefore the QTBM based on wave functions, with a complexity of $\mathcal{O}(N_{\mathrm{blocks}} \times N_{\mathrm{basis}} \times N_{\mathrm{modes}})$, is more efficient than solving for the Green's function, as already noted by Bruck et al. [14]. Both approaches are identical when considering ballistic transport [47].

3.3. Density

The full electron density of the open system is formally given by

$$n(\mathbf{r}) = \int \mathrm{d}E \sum_\nu g_\nu(E)|\psi_{E\nu}(\mathbf{r})|^2 f_{\mathrm{FD}}(E - \mu_\nu)\,, \tag{26}$$

where $g_\nu(E)$ represents the density of states (including spin degeneracy) of the injecting contact of mode ν, calculated from the velocity determined from the generalized Hellmann–Feynman theorem (Eq. (19)), and $f_{\mathrm{FD}}(E-\mu_\nu)$ is the Fermi–Dirac distribution, where μ_ν is the electrochemical potential in the contact of mode ν. In the evaluation of the integral over energy E, singularities of the type $1/\sqrt{E - E_{\mathrm{singularity}}}$ are encountered in the density of states at local band-extrema, *i.e.*, where $\mathrm{d}E/\mathrm{d}k = 0$. Since the location of these singularities is *a-priori* unknown and the evaluation of the wavefunctions $\psi_{E\nu}(\mathbf{r})$ is computationally expensive, we have adopted an adaptive Simpson technique for the numerical evaluation of the integral to a specified numerical tolerance. In our tests, the Simpson method provides an accurate error estimate, which gives a reliable accuracy for our results.

In a naive implementation of Eq. (26), the wavefunctions $\psi_{E\nu}(\mathbf{r})$ are evaluated directly using the expansion defined in Eq. (4). This step is computationally expensive, as the Bloch-wave grid, with $N_{\mathbf{r}}$ points, is generally very fine. However, during the adaptive Simpson integration, we can compute an estimated average density on the N_{nodes} nodes, instead of on all $N_{\mathrm{nodes}} \times N_{\mathbf{r}}$ points in space. We call this the node density,

$$\langle n\rangle^{\mathrm{node}}[z_i] = \int \mathrm{d}E \sum_\nu \langle n\rangle_{E,\nu}^{\mathrm{node}}[z_i] f_{\mathrm{FD}}(E - \mu_\nu)\,, \tag{27}$$

where the local density of states of the nodes is simply given by

$$\langle n\rangle_{E,\nu}^{\mathrm{node}}[z_i] = g_\nu(E)\sum_{nk}|c_{ink}|^2\,. \tag{28}$$

This evaluation of the node density comes at virtually no cost. Note that to interpret $\langle n\rangle_{\mathrm{node}}[z_i]$ as an estimate of average of the real density, the Bloch-waves need to be normalized in a specific way,

$$\int_{\Omega_{\mathrm{sc}}} \mathrm{d}^3r\,|u_{nk}|^2 = V_{\mathrm{sc}}\,, \tag{29}$$

where Ω_{sc} covers the supercell and V_{sc} is its volume. With this normalization, the coefficients c_{ink} have units $[\sqrt{m^{-2}}]$ and the normalized Bloch-waves are dimensionless weights that average to unity in a single cell. Since all coefficients c_{ink} are normalized with respect to the mass matrix M upon injection (see Eq. (21)), the wavefunctions remain properly normalized.

By storing all the integration energies E, weights $w_{E\nu}$ and coefficients $c_{E\nu,ink}$ when computing the node density, we can efficiently reconstruct the complete real-space density in one step, avoiding the costly naive evaluation of Eq. (26). To achieve this, we compute the matrix elements of the density matrix, expressed in the Bloch-basis:

$$n_{i'n'k',ink} = \sum_{E,\nu} w_{E\nu} g_{E\nu} c_{E\nu,i'n'k'}^* c_{E\nu,ink} f_{\mathrm{FD}}(E - \mu_\nu)\,. \tag{30}$$

Thanks to the locality of the shape functions in the Bloch-basis, only the matrix-elements for $i = i'$ and nearest neighbors i, i' need to be computed for the evaluation of the density

$$n(\mathbf{r}) = \sum_{i'n'k',ink} n_{i'n'k',ink}(\mathbf{r})\psi_{i'n'k'}^*(\mathbf{r})\psi_{ink}(\mathbf{r})\,. \tag{31}$$

Careful analysis of the operations involved in each procedure shows that the matrix-based approach is faster if the number of Bloch-waves used in the basis is lower than or equal to the number of injected states, *i.e.*, $N_{\mathrm{basis}} < N_{\mathrm{waves}}$.

Fig. 3 shows an example of the Simpson integration with a fixed potential as shown in Fig. 3(a). The selected integration energies are indicated in Fig. 3(a), showing adaptive refinements near the band extrema of the contacts, as expected. The converged y-averaged density in Fig. 3(b), clearly shows the subatomic resolution of the reconstructed density. We also show that the node density in Fig. 3(c) matches very well the xy-averaged density inside the ribbon.

Since we use the node density in the Simpson integration, the error estimate that is used for the refinement is based on the node density. In theory, we cannot guarantee that a specified tolerance for the node density, using the error estimate on the node density, is an exact measure of the error of the complete

Please cite this article as: M.L. Van de Put, M.V. Fischetti and W.G. Vandenberghe, Scalable atomistic simulations of quantum electron transport using empirical pseudopotentials, Computer Physics Communications (2019), https://doi.org/10.1016/j.cpc.2019.06.009.

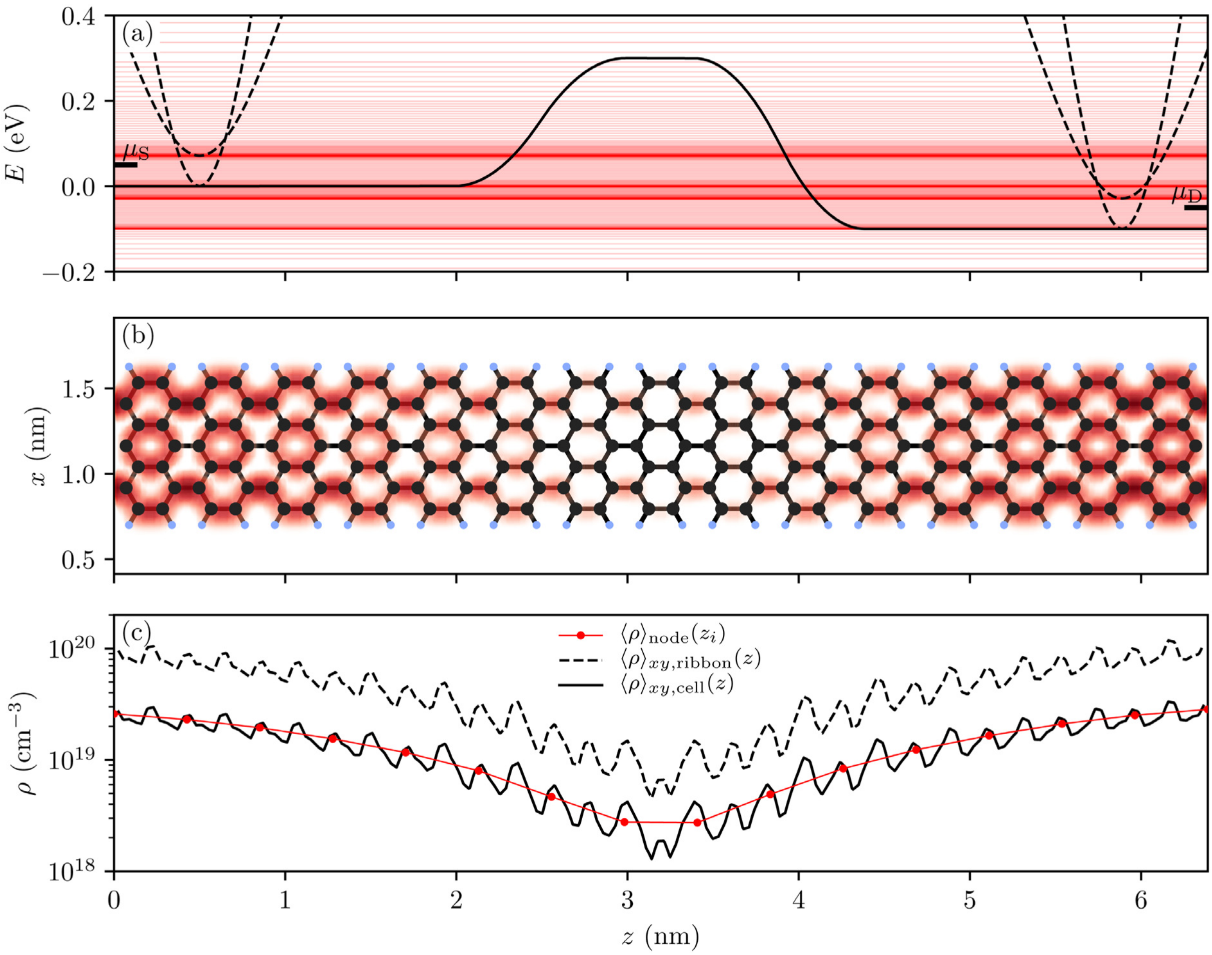

Fig. 3. An illustration of the calculation of the density. (a) Band-diagram showing the variation of the bottom of the conduction band w.r.t. the z coordinate and the chemical potentials of the source and drain contacts $\mu_{s/d}$ (50 meV above their respective conduction band). Each horizontal (red) line is an energy of injection (289 total), determined by the adaptive Simpson integrator for a tolerance of 10^{14} cm^{-3}. For reference, the band structure in the source and drain regions is shown as dashed lines. (b) The resulting free electron density in the ribbon, averaged over the y-direction, shown as pseudo-color. (c) The same electron density, averaged over the x–y plane of the full cell ($\langle\rho\rangle_{xy,\mathrm{cell}}$), averaged over an x–y plane inside the ribbon only ($\langle\rho\rangle_{xy,\mathrm{ribbon}}$), and the 'node averaged' density ($\langle\rho\rangle_{\mathrm{node}}$) as explained in the text. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

density at every point. We verify that this is not an issue in practice by determining the accuracy of the calculation of the density in Fig. 3. We first request an absolute tolerance of the integrated density of 10^{14} cm^{-3} and then perform a more accurate calculation with a tolerance of 10^{11} cm^{-3} (three orders of magnitude smaller). A comparison of the two results shows a root-mean-squared error of 1.4×10^{13} cm^{-3} for the node density, close to the requested value, and a difference of 33×10^{13} cm^{-3} for the complete real-space density. As expected, the error on the density is underestimated by the error on the node density due to the atomic variations. In practice, we account for this discrepancy by selecting a node-density tolerance at least two orders of magnitude smaller than the required charge density.

3.4. Transmission and ballistic current

The transmission probability of each state is calculated by taking the ratio of the injected and transmitted current density. For a mode γ injected from the source contact (s), the transmission probability to the drain contact (d) is calculated from the Bloch matrices and group velocity as

$$T_{\mathrm{sd}} = \frac{J_{\mathrm{d,out}}}{J_{\mathrm{s,inj}}} = \frac{\sum_\mu \left| \left[\mathrm{B}_{\mathrm{d}}^{-1}\right]_{\mu\gamma} \mathbf{c}_{\mathrm{d},\gamma} \right|^2 v_{\mathrm{d},\mu}}{v_{\mathrm{s},\gamma}}, \tag{32}$$

assuming the injected coefficients are properly normalized, *i.e.*, $|\mathbf{c}^{\mathrm{inj}}_{\mathrm{s},\gamma}| = 1$. As a sanity check, we explicitly calculate the reflection coefficient T_{ss} with Eq. (32) after first removing the injected part $\mathbf{c}^{\mathrm{inj}}_{\mathrm{s},\gamma}$ from the coefficients $\mathbf{c}_{\mathrm{s},\gamma}$, $T_{\mathrm{sd}} + T\mathrm{sd} = 1$.

The ballistic current from source (s) to drain (d) is calculated from the transmission coefficients as

$$I_{\mathrm{sd}} = \int \mathrm{d}E \sum_\nu g_\nu(E)\, T_{\mathrm{sd},E\nu}\ \mathrm{sgn}(v_{E\nu}) f_{\mathrm{FD}}(E - \mu_\nu)\,, \tag{33}$$

where $\mathrm{sgn}(v_{E\nu})$ gives the sign of the velocity of the injected state, *i.e.*, $+1$ (-1) for states ν originating from the source (drain). The integral in Eq. (33) is evaluated using the same adaptive Simpson method used to calculate the density. A separate integration of the current, rather than using the states obtained in the density simulation, is advised since the energies that are associated with a high current do not necessarily align with those that are responsible for the density. Furthermore, since the current integration is free of singularities, the integration converges quickly, with fewer evaluations than required in the density simulation for the same relative accuracy.

4. Self-consistency

In realistic electronic devices, external potentials are applied by gates and fixed charges are associated with ionized doping. To account for all these effects, as well as the mean-field interaction of the electron charge, we adopt the Hartree approximation. The extrinsic potential is found self-consistently with the electron density by solving the non-linear Poisson equation,

$$\nabla\cdot[\epsilon(\mathbf{r})\nabla V(\mathbf{r})] = \rho[\mathbf{r};V] + \rho_{\mathrm{doping}}(\mathbf{r})\,, = \rho_{\mathrm{net}}[\mathbf{r};V]\,, \tag{34}$$

where $\rho[\mathbf{r};V] = -en[\mathbf{r};V]$ is the free-electron charge for a given potential $V(\mathbf{r})$, $\rho_{\mathrm{doping}}(\mathbf{r})$ represents the fixed charge density originating from the ionized dopants, and $\rho_{\mathrm{net}}[\mathbf{r};V]$ is the net charge density.

To allow for the general application of boundary conditions and shapes for the gates and doping profiles the density and potential are discretized on a linear tetrahedral finite-elements mesh, forming the Poisson domain $\Omega_{\mathrm{Poisson}}$. At the edges of the Poisson domain, *i.e.*, for $\mathbf{r} \in \partial\Omega_{\mathrm{Poisson}}$, we impose Neumann boundary conditions $\nabla V(\mathbf{r})\cdot\hat{\mathbf{n}}(\mathbf{r}) = 0$, where $\hat{\mathbf{n}}$ is the normal to the edge. The electrostatic control of the device by gates is included by applying Dirichlet boundary conditions to their domains. For a single gate at a fixed potential V_{g} with domain Ω_{g} the Dirichlet condition is $V(\mathbf{r}\in\Omega_{\mathrm{g}}) = V_{\mathrm{g}}$. A high quality tetrahedral mesh, covering the Poisson domain, and conforming to the gates is automatically generated.

However, since the density, constructed using Eq. (26), is given on a uniform grid corresponding to the Fourier transform of the plane-wave components of the Bloch waves, the density needs to be interpolated to the tetrahedral finite-elements mesh. To avoid unnecessary approximation and the introduction of errors, the finite-element mesh explicitly includes all points of the uniform Bloch-wave mesh where the Bloch waves that comprise the basis set have non-negligible values. Specifically, a point $\mathbf{r}_l$ from the uniform grid is included in the tetrahedral mesh if

$$|u_{nk}(\mathbf{r}_l)|^2 > 10^{-3} \times \max_{\{n,k,\mathbf{r}\}} |u_{nk}(\mathbf{r})|^2\,, \tag{35}$$

for any band n and wave-vector k in the basis-set. To cover the entire Poisson domain, additional mesh points are generated automatically using the TetGen library [48]. This procedure yields a coarser global mesh with a gradual transition to the fine mesh points determined by the Bloch waves. This way of constructing the mesh is equivalent to an adaptively refined mesh in regions of high (expected) density. Since the potential is calculated on the tetrahedral mesh, the calculation of the matrix elements of the extrinsic potential $\mathrm{V}^{\mathrm{e}}_{i'n'k',ink}$ requires an interpolation of the tetrahedral mesh to the uniform Bloch-wave mesh. By sharing points between the tetrahedral mesh and the uniform mesh, we introduce a significant amount of additional bookkeeping. However, doing so we combine the sub-atomic resolution of Bloch-waves with the ability of the mesh to comply to general boundary conditions.

In addition to the flexibility in applying boundary conditions, a tetrahedral mesh provides a significant decrease in computational burden compared to the uniform Bloch-wave mesh that could otherwise be used, since the number of points in the tetrahedral mesh is significantly lower than those of the Bloch waves, $N_{\mathrm{tetra}} \ll N_{\mathbf{r}} \times N_{\mathrm{blocks}}$. The penalty we incur consists in the burden of interpolation whenever we transition between the meshes. For this purpose, we use a linear interpolation that matches the linear shape functions used in the finite-element representation of the linear Poisson equation. However, the interpolation burden is limited because the values on the points that are shared between the two meshes, which account for the majority of points in all our test structures, do not require interpolation.

Using linear shape functions on the tetrahedra, we arrive at the FE representation of the non-linear Poisson equation

$$\mathrm{D}\mathbf{V} = \mathrm{M}\left\{\boldsymbol{\rho}[\mathbf{V}] + \boldsymbol{\rho}_{\mathrm{doping}}\right\} = \mathrm{M}\boldsymbol{\rho}_{\mathrm{net}}[\mathbf{V}]\,, \tag{36}$$

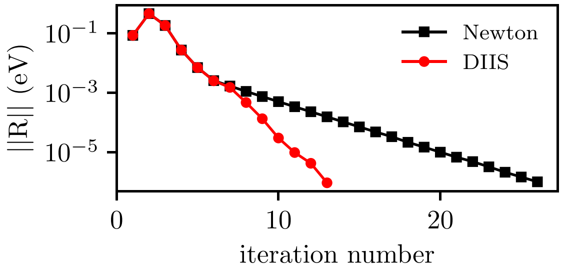

Fig. 4. Convergence rate of self-consistent procedure for the device shown in Fig. 6 in the off-state ($V_{\mathrm{g}} = -0.2$ V and $V_{\mathrm{ds}} = 0.2$ V) from a uniform starting potential. The l^2-norm of the residual is shown for the semi-classical Newton iteration, as well as the accelerated DIIS method. The convergence criterion is set to 10^{-6} eV.

where M is the mass-matrix and D represents the $\nabla\cdot[\epsilon(\mathbf{r})\nabla]$ operator. The non-linear Poisson equation is solved using a Newton–Raphson method, which, for iteration $p+1$ reads:

$$\mathbf{V}^{p+1} = \mathbf{V}^{p} - \left[\mathrm{J}^{p}\right]^{-1}\left[\mathrm{D}\mathbf{V}^{p} - \mathrm{M}\boldsymbol{\rho}_{\mathrm{net}}\left[\mathbf{V}^{p}\right]\right], \tag{37}$$

where the Jacobian is given by $\mathrm{J}^{p} = \mathrm{D} - \mathrm{M}\mathrm{J}^{p}_{\rho}$. We approximate the Jacobian for the free charge density J^{p}_{ρ} with a semi-classical diagonal matrix, calculated by varying the local chemical potential. In practice, we calculate it by evaluating the free-density in Eq. (26) with the derivative of the Fermi–Dirac distribution instead of the Fermi–Dirac distribution itself, *i.e.*,

$$\begin{aligned}\tilde{J}^{p}_{\rho}(\mathbf{r},\mathbf{r}') &= \delta(\mathbf{r},\mathbf{r}')\frac{\delta\rho[\mathbf{r},V^{p}(\mathbf{r})]}{\delta\mu(\mathbf{r})}\\ &= \delta(\mathbf{r},\mathbf{r}')\int \mathrm{d}E\sum_{\nu} g_{\nu}(E)\left|\psi^{p}_{E\nu}(\mathbf{r})\right|^{2}\frac{\partial f_{\mathrm{FD}}}{\partial E}(E-\mu_{\nu})\,. \end{aligned}\tag{38}$$

Inserting the semi-classical approximation of the Jacobian in Eq. (37) and rearranging yields a linear Poisson equation for each iteration $p+1$ of the Newton–Raphson procedure:

$$\left(\mathrm{D} - \mathrm{M}\mathrm{J}^{p}_{\rho}\right)\mathbf{V}^{p+1} = \mathrm{M}\left(\boldsymbol{\rho}_{\mathrm{net}}\left[\mathbf{V}^{p}\right] - \mathrm{J}^{p}_{\rho}\mathbf{V}^{p}\right). \tag{39}$$

The linear Poisson equation is a simple elliptic partial differential equation which is efficiently and accurately solved using the algebraic multi-grid (AMG) method [49]. Fig. 4 shows the convergence behavior of a self-consistent calculation starting from a flat potential. After an initial period, our Newton method, with a semi-classical approximation of the Jacobian, converges linearly.

To further accelerate the convergence of the self-consistent procedure, we use the Direct Inversion of the Iterative Subspace (DIIS) technique, commonly known as Pulay mixing in computational chemistry [50]. In the DIIS technique, the residual $\mathrm{R}^{p+1} = \mathbf{V}^{p+1} - \mathbf{V}^{p}$ and the new solution $\mathbf{V}^{p+1}$ are added to the previous solutions and form the iterative subspace, which is used to predict a new vector $\tilde{\mathbf{V}}^{p+1}$. Following the analysis in Ref [51], we have implemented the DIIS technique using a least-squares approach based on the Singular-Value Decomposition (SVD) that improves the resolution of components of the iterative subspace when the residuals are almost linearly dependent. This linear dependence occurs naturally when self-consistency is almost reached, and the tolerance for convergence is set very low. However, to avoid a spurious linear-dependence that could degrade the convergence behavior, we limit the range of the iterative subspace. Typically, only the last 5 iterations are kept. In practice, both the bare Newton and the accelerated DIIS method exhibit a linear convergence when solving the non-linear Poisson equation. However, as demonstrated in Fig. 4, the DIIS method accelerates convergence by a factor of two, which is well worth the additional complexity of implementation.

Please cite this article as: M.L. Van de Put, M.V. Fischetti and W.G. Vandenberghe, Scalable atomistic simulations of quantum electron transport using empirical pseudopotentials, Computer Physics Communications (2019), https://doi.org/10.1016/j.cpc.2019.06.009.

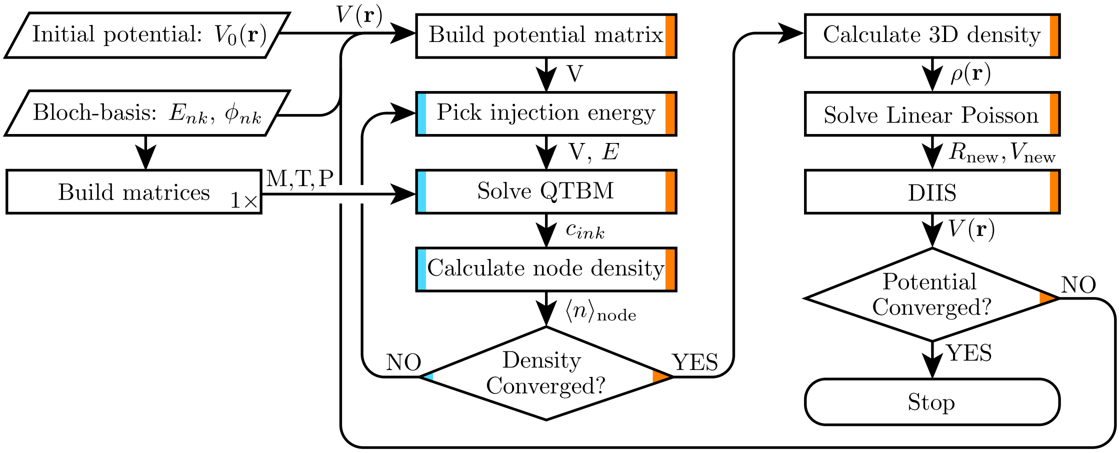

Fig. 5. Flowchart of the self-consistent procedure explained in the text. The inputs are an estimated initial potential and the Bloch-basis calculated using empirical pseudopotentials. The loop responsible for the adaptive Simpson-integration is indicated with blue shading on the left side, while the self-consistent loop is indicated with orange shading on the right hand side of the box. Note that the M, T, and P matrices are built a single time (1×). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 5 gives an overview of the entire self-consistent procedure for a typical simulation. Upon convergence, other quantities, such as the electronic current, can be calculated using the converged potential.

5. Results: Graphene nanoribbon

We demonstrate the Bloch wave method presented in Section 2 using an armchair graphene nanoribbon (aGNR) field-effect transistor (FET), as shown in Fig. 6. The aGNR is 25 carbon atoms wide (3.8 nm) and is terminated by hydrogen at the armchair edge. The complete device is 17 nm long, with a channel length of 5 nm, and contains a total of 2160 atoms (2000 carbon and 160 hydrogen).

5.1. Electronic structure

We calculate the band structure of the 3.8 nm wide armchair GNR shown in Fig. 6 and show the results in Fig. 7. We first calculate the electronic structure using the plane-wave empirical pseudopotential method. We use the local pseudopotentials from Ref. [30] for both the carbon and hydrogen ions. Local pseudopotentials have been used extensively for carbon compounds and have been shown to accurately reproduce the band structure of graphene as well as its nanoribbons [6,8,28,30,52,53]. The resulting Schrödinger equation is solved in a plane-wave basis, using the fast Fourier transform for efficient evaluation, as described in Ref. [28]. In particular, we calculate both the eigenenergies, ϵ_{nk}, and wavefunctions, *i.e.*, the Bloch waves $\phi_{nk}(\mathbf{r})$, for the lowest $N = 112$ bands (102 valence bands and 10 conduction bands) at 20 wave vectors, equally spaced from the first Brillouin zone (BZ) center ($\varGamma$-point) to its edge (Z-point). The resulting band structure is shown in Fig. 7(a) as dotted lines.

To verify if this basis is capable of describing the electronic structure of the ribbon, we use the Bloch-wave expansion to reconstruct the band structure throughout the entire first BZ. The Bloch waves, $\phi_{nk}(\mathbf{r})$, at the $\varGamma$-point and Z-point (112 bands for each point) are used as a basis in our Bloch wave expansion of the wavefunction in Eq. (4). The procedure is a straightforward application of Bloch's theorem: For a given wave vector k', we calculate the expansion coefficients c_{ink} at a single node i by enforcing periodicity to the neighboring nodes with a phase-difference given by the wave vector k',

$$c_{ink} = c_{jnk} \mathrm{e}^{\mathrm{i}k'(z_i - z_j)}.$$

Exploiting this periodicity reduces the matrix equation, given in Eq. (14), to a generalized eigenvalue problem of size ($N_{\text{Bloch}} \times N_{\text{Bloch}}$),

$$\mathrm{H}_i(k')\mathbf{c}_i = E_{k'}\mathrm{M}_i(k')\mathbf{c}\,, \tag{40}$$

with:

$$\mathrm{H}_i(k') = H_{i,i} + \sum_{\langle i,j\rangle} H_{i,j}\, \mathrm{e}^{\mathrm{i}k'(z_i - z_j)} \quad \text{and}$$

$$\mathrm{M}_i(k') = M_{i,i} + \sum_{\langle i,j\rangle} M_{i,j}\, \mathrm{e}^{\mathrm{i}k'(z_i - z_j)},$$

where the matrix $\mathrm{H}_{i,j}$ ($M_{i,j}$) corresponds to a block of the matrix in Eq. (14), containing the elements coupling node i to node j. $\langle i, j\rangle$ is a reminder that the coupling between nodes is nearest-neighbors only. Note that for the band structure, no external potential is applied and all external potential matrix elements vanish, *i.e.*, $V_{ink} = 0$, and thus all nodes i are equivalent.

Fig. 7(a) shows the Bloch waves selected as the basis-set, and the band structure reconstructed using the Bloch wave method as solid lines. In effect, we interpolate the band structure from the $\varGamma$-point and Z-point to the full first BZ. Comparing the reconstructed band structure to the plane-wave calculation shows a good match. Fig. 7(b) quantifies the error in the reconstruction, showing the absolute energy difference, $\varDelta E$, between the Bloch-wave reconstructed band structure and the plane-wave values at the wave vectors of the plane-wave calculation. As expected, the error increases in the upper conduction bands, where it is more likely that our selected Bloch-wave basis-functions do not span the solution space for every wave vector. By selecting two k-points, the eigenvalue problem in Eq. (40) admits $N_{\text{Bloch}} = 2 \times N$ solutions, where N is the number of bands included for each k-point. Since the basis has been constructed using all of the lowest energy Bloch-waves, the additional N spurious solutions at each k-point have associated eigenergies outside of our desired spectrum. In particular, up to 1 eV above the Fermi level the Bloch-wave reconstruction matches the plane-wave results very well, showing a root mean squared error of 24 meV. This energy range is more than adequate for transport purposes. Moreover, if more accuracy is needed at a higher energy, one can readily increase the basis to cover those higher energies, albeit at an increased computational cost.

Having verified the Bloch-wave method's accuracy, we verify our earlier computational claims. Fig. 8 shows the time required to calculate the eigenvalues for a single wave vector using (a)

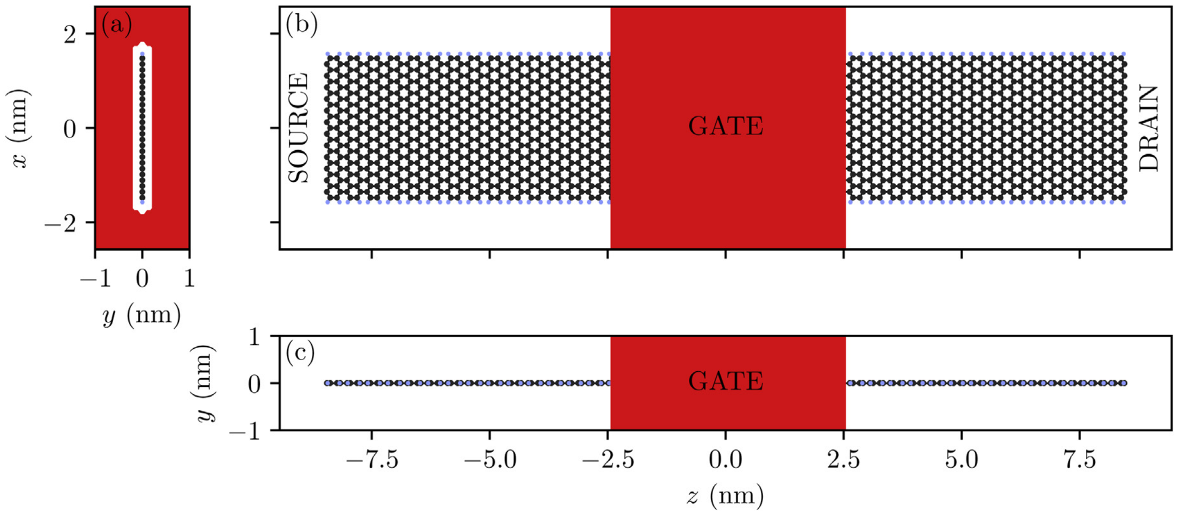

Fig. 6. (a) Front, (b) top, and (c) side view of a depiction of the aGNR FET under study. The armchair graphene-nanoribbon (aGNR) is 3.8 nm wide (25 carbon atoms). The simulated device is built from 40 repetitions of a single supercell, totaling 2160 atoms and a device length of approximately 17 nm. The gate (shaded region) is centrally located in an all-around configuration, is 5 nm long, and has an oxide thickness equivalent to 1 nm of SiO$_2$. The source and drain terminals are assumed to be infinite extensions of the aGNR. The ribbon is uniformly n-type doped, except for the channel under the gate which is assumed to be p-type. Carbon (dark gray) and hydrogen (light blue) atom locations are indicated with spheres. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

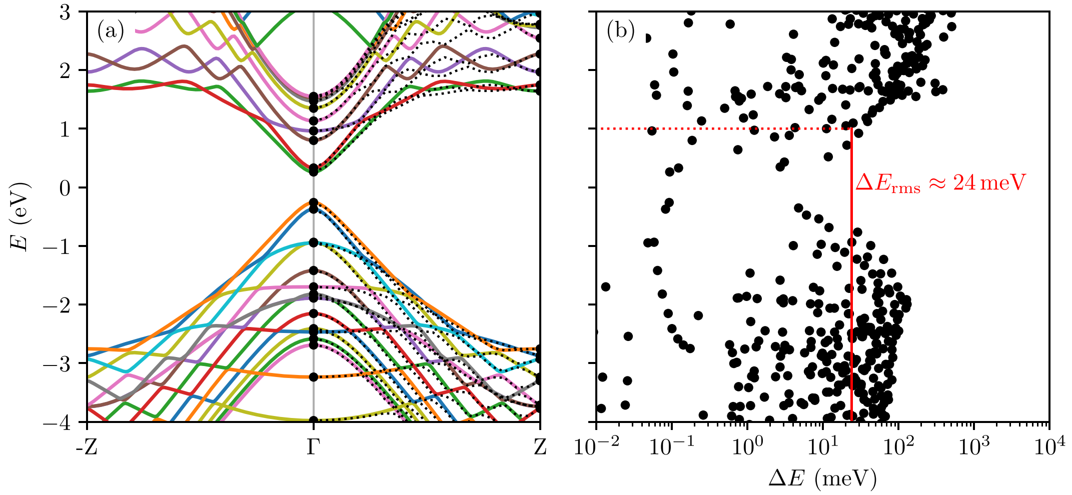

Fig. 7. Reconstruction of the band structure using the Bloch-wave basis. (a) Band structure of the graphene nanoribbon shown in Fig. 6 near the Fermi level (0 eV), calculated using the plane-wave empirical pseudopotential method (dashed) and reconstructed from the Bloch waves at the Γ-point and Z-point (dots) (solid lines). (b) The absolute energy difference, ΔE, between the full plane-wave method and the Bloch wave reconstruction. The root mean squared value of the energy difference, ΔE_{rms}, up to 1 eV is indicated on the graph.

the plane-wave method and (b) the Bloch-wave reconstruction for different ribbon widths, as indicated by the number of carbon atoms from one edge to the other. Fig. 8 also shows the size of the basis set used for both methods. The basis size corresponds to the number of plane-waves, N_G, for the plane-wave empirical pseudopotential method (a), and the number of Bloch waves N_{Bloch} for the Bloch-wave method (b). In both methods, the basis set increases linearly with the ribbon size, scaling with the supercell length in (a) and scaling with the number of atoms (valence electrons) in (b). However, for the range of GNR-widths shown here, the Bloch-wave basis-set is 100 times smaller than the plane-wave basis. The most immediate effect of this reduction of basis-set size is a hundred-fold reduction in the required memory to store the coefficients c_{ink} instead of all the plane-wave components of the wavefunction. Therefore, we are able to avoid the single most limiting factor for the scaling of the envelope function approach to plane-wave based electron transport calculations [6,7]. Note that, using the expansion in Eq. (4), we can obtain the real-space representation of the calculated coefficients c_{ink} when needed, as we demonstrate in Section 5.2.

The reduction of the size of the basis set also translates directly to a decrease of computing time. For example, the band-structure calculations for the 2 nm-wide 25-aGNR band-structure, shown in Fig. 7, take 25 seconds using the plane-wave method and only 5 ms using the Bloch wave method. While this speedup of a factor of 5000 does not include the construction of the various overlap matrices required by the Bloch-method, these are pre-calculated only once. Therefore, we expect equivalent performance gains for transport simulations. Finally, comparing the computation time for different widths in Fig. 8, both methods show the scaling behavior expected from their computational complexity. The plane-wave method is bounded by the $\mathcal{O}(N_G \log N_G)$ complexity of the FFT algorithm, while the Bloch-wave method behaves in line with the $\mathcal{O}(N^{\approx 2.376})$ complexity of the matrix products, as implemented in the optimized Basic Linear Algebra Subprograms (BLAS) [54].

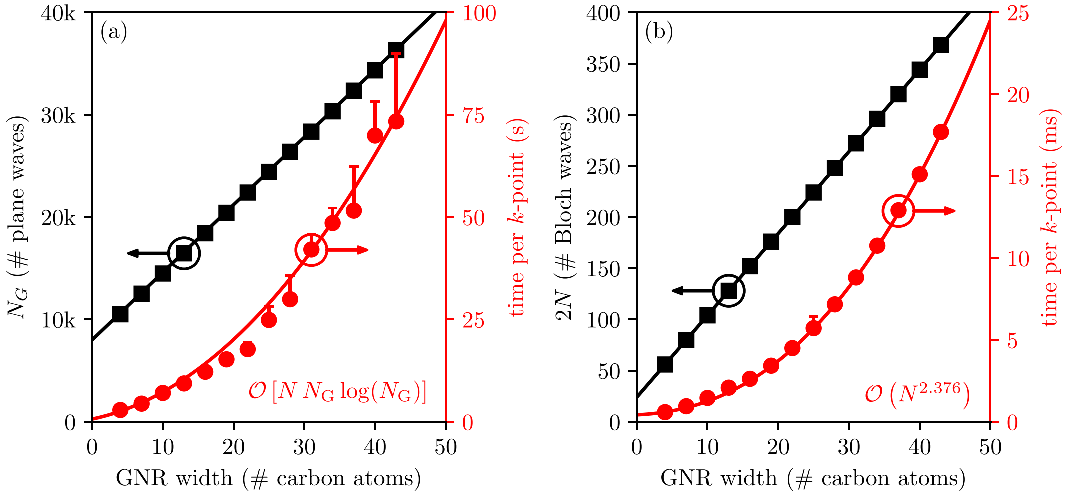

Fig. 8. Computational time for the calculation of the band structure for various ribbon widths. (a) Using the plane-wave empirical pseudopotential method, with a basis of N_G plane-waves. (b) Using our Bloch wave method, with N Bloch waves taken at the first Brillouin-zone center and its edge. Note the different scales used in (a) and (b). The best computational time, out of seven runs, is indicated with a dot, while the range of the timing is shown with a bar. The ideal scaling behavior for each case is indicated and a fit is shown as a continuous curve. The basis set for the plane-wave method in (a) is 100 times larger than the Bloch wave basis used in (b). The timing shows an even greater speed-up than expected from the basis-set size alone.

5.2. Transport: aGNR FET

The electron transport through the aGNR FET, shown in Fig. 6, is calculated using the self-consistent procedure described in Section 2. For our purposes, we apply a 0.2 V bias between source and drain, $V_{\rm ds}$. We then vary the gate potential, $V_{\rm g}$, from −0.7 V to 0.3 V, calculate the potential self-consistently, and obtain the current through the device. The work-function of the gate is set to the electron-affinity of the aGNR.

Fig. 9(a) shows the obtained transfer-characteristics of the device. Fig. 9(b) shows the corresponding band-profiles for different gate biasses, obtained self-consistently. Under forward bias ($V_{\rm g} > 0$), the device operates as a conventional FET. The sub-threshold and linear regimes are clearly visible in the figure. The sub-threshold slope is about 160 mV/dec. As already described for smaller ribbons [6], this poor slope is caused by source-to-drain tunneling through the barrier in the channel, induced by the gate. These tunneling rates grow as the bandgap becomes smaller as the width of the ribbon increases.

For the simulated ribbon, the bandgap is only 0.52 eV. This small bandgap leads to interesting ambipolar behavior: the current increases if the gate is operated in reverse bias ($V_{\rm g} < 0$) due to band-to-band tunneling. Looking at the band-alignment for, *e.g.*, $V_{\rm g} = -0.6$ V, in Fig. 9(b), it is clear that band-to-band tunneling is possible from the source to the channel region, and once more towards the drain. Thanks to the blocking of carriers from the high energy tail of the injected Fermi–Dirac distribution, the tunneling current increases at a steeper slope than in the forward regime. This operating principle leading to the steep slope is the same as that of a Tunnel FET (TFET) [55,56].

Note that, in this device, the behavior of the normal mode of operation as well as the reverse biased gate operation is based on quantum mechanics. Our proposed method, offering an efficient full-band quantum mechanical transport solver for general atomistic structures, is naturally capable of dealing with these effects and provides an invaluable tool in the study of exotic materials and devices.

6. Conclusions

We have presented a numerical method for the atomistic calculation of quantum electron transport in nanoscaled structures using empirical pseudopotentials. Our method is highly efficient; we have shown a reduction of the size of the required computational basis by two orders of magnitude compared to the conventional plane-wave methods. This efficiency is achieved by treating differently the two length-scales in the system. A partition-of-unity captures the large-scale behavior of the system and admits only nearest-neighbor coupling, resulting in excellent scalability. The atomic scale, meanwhile, is captured by an expansion based on Bloch waves of the atomic structure at high symmetry points. The Bloch-waves are computed to high accuracy using a Fourier-based plane-wave approach before starting the transport calculations. Our method approximates the computational efficiency of tight-binding and mode-space approaches while retaining the advantages of the plane-wave method, which features a natural real space representation with sub-atomic resolution.

We solve the electronic states in our open system using the well known quantum-transmitting boundary method (QTBM) and update self-consistently the Hartree potential from the three-dimensional density. The density is obtained by adaptively integrating the individual wave-functions' densities. Notably, we systematically control and estimate the numerical error at each stage in our method by using iterative solvers and adaptive integration methods. We are thus assured that the accuracy of our results is limited by the physical approximations made, and not by the numerical errors.

We have demonstrated the accuracy and efficiency of our method by calculating the ballistic transport properties of a graphene nanoribbon transistor. In this test case, the reconstruction of the band structure from our Bloch-wave basis is accurate to 24 meV when compared to the full-plane wave calculation, while being three orders of magnitude faster more efficient. Comparing different widths of nanoribbon shows that our method scales as expected, with a significantly improved performance compared to previous plane-wave approaches. A hundred-fold reduction in the size of the basis set results in a similar reduction in the memory requirements, which severely limit the device-size that can be handled by previous plane-wave envelope-function approaches. As a demonstration, we have simulated transport in a 3 nm wide nanoribbon transistor. We have observed a significant deterioration of the sub-threshold behavior due to source-to-drain tunneling through the potential barrier induced by the gate bias. In reverse bias, we observe significant ambipolar current due to band-to-band tunneling through the small bandgap.

Please cite this article as: M.L. Van de Put, M.V. Fischetti and W.G. Vandenberghe, Scalable atomistic simulations of quantum electron transport using empirical pseudopotentials, Computer Physics Communications (2019), https://doi.org/10.1016/j.cpc.2019.06.009.

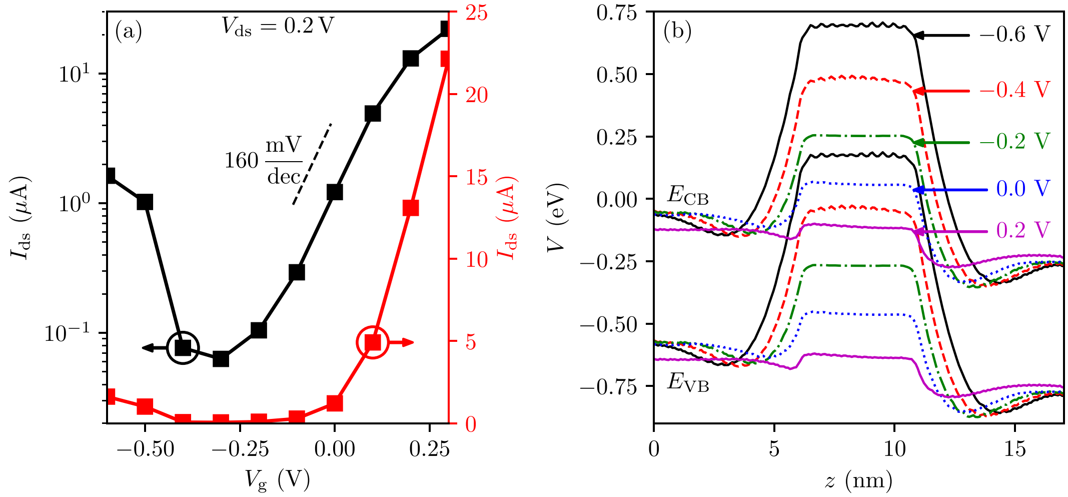

Fig. 9. Simulation results for device in Fig. 6. (a) Transfer-characteristics, showing source–drain current $I_{\rm ds}$ on a logarithmic (left) and linear (right) scale for different gate potentials $V_{\rm g}$. (b) Band-edge profile along z, through the middle of the ribbon, showing the approximate position of the conduction band minimum $E_{\rm CB}$ and valence band maximum $E_{\rm VB}$, for different gate potentials, as indicated.

This reaffirms the need for a quantum mechanical treatment of the transport in nanostructured devices in the 'intermediate' nanoscale, between bulk crystalline behavior and few-atom devices. Our presented method provides an efficient and flexible basis for such studies.

Finally, while we have only illustrated our method using empirical pseudopotentials, our approach is generally applicable to any formulation that can provide the Bloch waves in a real-space basis. Of particular interest might be the various *ab-initio* methods based on plane-waves, for which electron transport calculations are prohibitively expensive. Additional avenues of research include: extending our approach to heterogeneous structures with variations of the atomic structure, optimizing the selection of the included Bloch waves, and higher order finite-element shape-functions.

Acknowledgment

This material is based in part on work supported by the National Science Foundation under Award Number 1710066. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Appendix. Matrix element of the crystal Hamiltonian

We derive an expression for the matrix elements for the crystal Hamiltonian that avoids explicit knowledge of the crystal potential,

$$\mathrm{H}^{\rm c}_{i'n'k',ink} = \int_\Omega \mathrm{d}^3r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\left[-\frac{\hbar^2}{2m}\nabla^2 + V^{\rm c}_i(\mathbf{r})\right]\left[f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\right]. \tag{A.1}$$

To remove the crystal potential, we start from the known Schrödinger equation for the Bloch waves in a supercell

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V^{\rm c}_i(\mathbf{r})\right]\phi_{ink}(\mathbf{r}) = \epsilon_{ink}\,\phi_{ink}(\mathbf{r}). \tag{A.2}$$

We left-multiply by $f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})f_i(\mathbf{r})$ and integrate over all of space, yielding

$$\int_\Omega \mathrm{d}^3r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})f_i(\mathbf{r})\left[-\frac{\hbar^2}{2m}\nabla^2 + V^{\rm c}_i(\mathbf{r})\right]\phi_{ink}(\mathbf{r})$$

$$= \mathrm{M}_{i'n'k',ink}\,\epsilon_{ink}\,, \tag{A.3}$$

where we have defined the overlap matrix element as

$$\mathrm{M}_{i'n'k',ink} = \int_\Omega \mathrm{d}^3r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,. \tag{A.4}$$

Comparing Eq. (A.3) to the matrix element of the crystal Hamiltonian in Eq. (A.1) we obtain

$$\mathrm{H}^{\rm c}_{i'n'k',ink} = \mathrm{M}_{i'n'k',ink}\,\epsilon_{ink} + \mathrm{T}^{(\rm r)}_{i'n'k',ink} + \mathrm{P}^{(\rm r)}_{i'n'k',ink}\,, \tag{A.5}$$

where we have defined additional matrix elements representing kinetic energy and momentum-coupling

$$\mathrm{T}^{(\rm r)}_{i'n'k',ink} = -\frac{\hbar^2}{2m}\int_\Omega \mathrm{d}^3r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\left[\nabla^2 f_i(\mathbf{r})\right]\phi_{ink}(\mathbf{r})\,, \tag{A.6}$$

$$\mathrm{P}^{(\rm r)}_{i'n'k',ink} = -\frac{\hbar^2}{m}\int_\Omega \mathrm{d}^3r\, f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\left[\nabla f_i(\mathbf{r})\right]\cdot\left[\nabla\phi_{ink}(\mathbf{r})\right], \tag{A.7}$$

where the subscript (r) is a reminder that the matrix elements are non-Hermitian and the operators they contain act only to the right. Similarly, we can define "left" matrix elements

$$\mathrm{T}^{(\rm l)}_{i'n'k',ink} = \left(\mathrm{T}^{(\rm r)}_{ink,i'n'k'}\right)^*$$

$$= -\frac{\hbar^2}{2m}\int_\Omega \mathrm{d}^3r\, \left[\nabla^2 f^*_{i'}(\mathbf{r})\right]\phi^*_{i'n'k'}(\mathbf{r})f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,, \tag{A.8}$$

$$\mathrm{P}^{(\rm l)}_{i'n'k',ink} = \left(\mathrm{P}^{(\rm r)}_{ink,i'n'k'}\right)^*$$

$$= -\frac{\hbar^2}{m}\int_\Omega \mathrm{d}^3r\, \left[\nabla f^*_{i'}(\mathbf{r})\right]\cdot\left[\nabla\phi^*_{i'n'k'}(\mathbf{r})\right]f_i(\mathbf{r})\phi_{ink}(\mathbf{r})\,, \tag{A.9}$$

that satisfy

$$\mathrm{H}^{\rm c}_{i'n'k',ink} = \epsilon_{i'n'k'}\,\mathrm{M}_{i'n'k',ink} + \mathrm{T}^{(\rm l)}_{i'n'k',ink} + \mathrm{P}^{(\rm l)}_{i'n'k',ink}\,, \tag{A.10}$$

Combining the expressions of Eqs. (A.5) and (A.10) yields

$$\mathrm{H}^{\rm c}_{i'n'k',ink} = \frac{\epsilon_{ink} + \epsilon_{i'n'k'}}{2}\,\mathrm{M}_{i'n'k',ink} + \mathrm{T}_{i'n'k',ink} + \mathrm{P}_{i'n'k',ink}\,, \tag{A.11}$$

in which $\mathrm{T}_{i'n'k',ink} = \left(\mathrm{T}^{(\rm r)}_{i'n'k',ink} + \mathrm{T}^{(\rm l)}_{i'n'k',ink}\right)/2$ and $\mathrm{P}_{i'n'k',ink} = \left(\mathrm{P}^{(\rm r)}_{i'n'k',ink} + \mathrm{P}^{(\rm l)}_{i'n'k',ink}\right)/2$ are both Hermitian matrix elements.

For the correct preservation of probability current across nodes, and in particular if we intend to use linear shape functions for $f_i(\mathbf{r})$, we should further use integration by parts in the derivation of the kinetic matrix elements,

$$\begin{aligned}\mathrm{T}^{(\mathrm{r})}_{i'n'k',ink} = \frac{\hbar^2}{2m}\int_{\Omega} \mathrm{d}^3 r \, \Big\{ & \big[\nabla f^*_{i'}(\mathbf{r})\big]\phi^*_{i'n'k'}(\mathbf{r}) \cdot \big[\nabla f_i(\mathbf{r})\big]\phi_{ink}(\mathbf{r}) \\ & + f^*_{i'}(\mathbf{r})\big[\nabla \phi^*_{i'n'k'}(\mathbf{r})\big] \cdot \big[\nabla f_i(\mathbf{r})\big]\phi_{ink}(\mathbf{r}) \\ & + f^*_{i'}(\mathbf{r})\phi^*_{i'n'k'}(\mathbf{r})\big[\nabla f_i(\mathbf{r})\big] \cdot \big[\nabla \phi_{ink}(\mathbf{r})\big]\Big\} , \end{aligned} \tag{A.12}$$

where we have omitted the vanishing boundary term. Combining this result with its Hermitian conjugate, and grouping appropriate terms, yields a compact form for the Hermitian kinetic energy matrix elements,

$$\begin{aligned}\mathrm{T}_{i'n'k',ink} = \frac{\hbar^2}{2m}\int_{\Omega} \mathrm{d}^3 r \, \Big\{ & \frac{1}{2}\nabla\big[f^*_{i'}(\mathbf{r})f_i(\mathbf{r})\big] \cdot \nabla\big[\phi^*_{i'n'k'}(\mathbf{r})\phi_{ink}(\mathbf{r})\big] \\ & + \big[\nabla f^*_{i'}(\mathbf{r})\big]\phi^*_{i'n'k'}(\mathbf{r}) \cdot \big[\nabla f_i(\mathbf{r})\big]\phi_{ink}(\mathbf{r}) \Big\} . \end{aligned} \tag{A.13}$$

References

[1] F. Schwierz, Nature Nanotechnol. 5 (7) (2010) 487–496, http://dx.doi.org/10.1038/nnano.2010.89, URL http://dx.doi.org/10.1038/nnano.2010.89http://www.nature.com/articles/nnano.2010.89.

[2] G. Gaddemane, W.G. Vandenberghe, M.L. Van de Put, S. Chen, S. Tiwari, E. Chen, M.V. Fischetti, Phys. Rev. B 98 (2018) 115416, http://dx.doi.org/10.1103/PhysRevB.98.115416.

[3] V. Giacometti, B. Radisavljevic, A. Radenovic, J. Brivio, A. Kis, Nature Nanotechnol. 6 (3) (2011) 147–150, http://dx.doi.org/10.1038/nnano.2010.279.

[4] A. Laturia, M.L. Van de Put, W.G. Vandenberghe, npj 2D Mater. Appl. 2 (1) (2018) 6, http://dx.doi.org/10.1038/s41699-018-0050-x, URL http://www.nature.com/articles/s41699-018-0050-x.

[5] J.P. Llinas, A. Fairbrother, G. Borin Barin, W. Shi, K. Lee, S. Wu, B. Yong Choi, R. Braganza, J. Lear, N. Kau, W. Choi, C. Chen, Z. Pedramrazi, T. Dumslaff, A. Narita, X. Feng, K. Müllen, F. Fischer, A. Zettl, P. Ruffieux, E. Yablonovitch, M. Crommie, R. Fasel, J. Bokor, Nature Commun. 8 (1) (2017) 8–13, http://dx.doi.org/10.1038/s41467-017-00734-x.

[6] J. Fang, S. Chen, W.G. Vandenberghe, M.V. Fischetti, IEEE Trans. Electron Devices 64 (6) (2017) 2758–2764, http://dx.doi.org/10.1109/TED.2017.2695960.

[7] J. Fang, W.G. Vandenberghe, B. Fu, M.V. Fischetti, J. Appl. Phys. 119 (3) (2016) http://dx.doi.org/10.1063/1.4939963.

[8] M.V. Fischetti, S. Narayanan, J. Appl. Phys. 110 (8) (2011) http://dx.doi.org/10.1063/1.3650249.

[9] J.E. Fonseca, T. Kubis, M. Povolotskyi, B. Novakovic, A. Ajoy, G. Hegde, H. Ilatikhameneh, Z. Jiang, P. Sengupta, Y. Tan, G. Klimeck, J. Comput. Electron. 12 (4) (2013) 592–600, http://dx.doi.org/10.1007/s10825-013-0509-0.

[10] A. García, J.D. Gale, J.M. Soler, D. Sánchez-Portal, E. Artacho, P. Ordejón, J. Junquera, J. Phys.: Condens. Matter 14 (11) (2002) 2745–2779, http://dx.doi.org/10.1088/0953-8984/14/11/302.

[11] J.C. Slater, Phys. Rev. 51 (1937) 846–851, http://dx.doi.org/10.1103/PhysRev.51.846.

[12] M. Weinert, G. Schneider, R. Podloucky, J. Redinger, J. Phys. Condens. Matter 21 (8) (2009) http://dx.doi.org/10.1088/0953-8984/21/8/084201.

[13] P.E. Blöchl, Phys. Rev. B 50 (24) (1994) 17953–17979, http://dx.doi.org/10.1103/PhysRevB.50.17953, https://link.aps.org/doi/10.1103/PhysRevB.50.17953.

[14] S. Brück, M. Calderara, M.H. Bani-Hashemian, J. VandeVondele, M. Luisier, J. Chem. Phys. 147 (7) (2017) http://dx.doi.org/10.1063/1.4998421.

[15] M. Luisier, Chem. Soc. Rev. 43 (13) (2014) 4357–4367, http://dx.doi.org/10.1039/c4cs00084f.

[16] J. Maassen, M. Harb, V. Michaud-Rioux, Y. Zhu, H. Guo, Proc. IEEE 101 (2) (2013) 518–530, http://dx.doi.org/10.1109/JPROC.2012.2197810, URL http://ieeexplore.ieee.org/document/6287539/.

[17] K. Stokbro, D.E. Petersen, S. Smidstrup, A. Blom, M. Ipsen, K. Kaasbjerg, Phys. Rev. B 82 (2010) 075420, http://dx.doi.org/10.1103/PhysRevB.82.075420.

[18] A. Garcia-Lekue, M.G. Vergniory, X.W. Jiang, L.W. Wang, Prog. Surf. Sci. 90 (3) (2015) 292–318, http://dx.doi.org/10.1016/j.progsurf.2015.05.002.

[19] H. Joon Choi, J. Ihm, Phys. Rev. B 59 (3) (1999) 2267–2275, http://dx.doi.org/10.1103/PhysRevB.59.2267, URL https://link.aps.org/doi/10.1103/PhysRevB.59.2267.

[20] E. Polizzi, N.B. Abdallah, J. Comput. Phys. 202 (1) (2005) 150–180, http://dx.doi.org/10.1016/j.jcp.2004.07.003, URL http://linkinghub.elsevier.com/retrieve/pii/S0021999104002712.

[21] R. Venugopal, Z. Ren, S. Datta, M.S. Lundstrom, D. Jovanovic, J. Appl. Phys. 92 (7) (2002) 3730–3739, http://dx.doi.org/10.1063/1.1503165.

[22] C. Jourdana, P. Pietra, SIAM J. Sci. Comput. 36 (3) (2014) B486–B507, http://dx.doi.org/10.1137/130926353.

[23] L.W. Wang, A. Franceschetti, A. Zunger, Phys. Rev. Lett. 78 (14) (1997) 2819–2822, http://dx.doi.org/10.1103/PhysRevLett.78.2819, URL http://link.aps.org/doi/10.1103/PhysRevLett.78.2819.

[24] X.W. Jiang, S.S. Li, L.W. Wang, Solid-State Electron. 68 (2012) 56–62, http://dx.doi.org/10.1016/j.sse.2011.09.015.

[25] X.W. Jiang, S.S. Li, J.B. Xia, L.W. Wang, J. Appl. Phys. 109 (5) (2011) http://dx.doi.org/10.1063/1.3556430.

[26] D. Esseni, P. Palestri, Phys. Rev. B 72 (16) (2005) 1–14, http://dx.doi.org/10.1103/PhysRevB.72.165342.

[27] M.G. Pala, D. Esseni, Phys. Rev. B 97 (12) (2018) 1–14, http://dx.doi.org/10.1103/PhysRevB.97.125310.

[28] M.L. Van de Put, W.G. Vandenberghe, B. Sorée, W. Magnus, M.V. Fischetti, J. Appl. Phys. 119 (21) (2016) 214306, http://dx.doi.org/10.1063/1.4953148, URL http://aip.scitation.org/doi/10.1063/1.4953148.

[29] J. Kim, M.V. Fischetti, J. Appl. Phys. 110 (2011) (2011) 033716, http://dx.doi.org/10.1063/1.3615942, URL http://link.aip.org/link/JAPIAU/v110/i3/p033716/s1{&}Agg=doi.

[30] Y. Kurokawa, S. Nomura, T. Takemori, Y. Aoyagi, Phys. Rev. B 61 (19) (2000) 12616–12619, http://dx.doi.org/10.1103/PhysRevB.61.12616, URL https://link.aps.org/doi/10.1103/PhysRevB.61.12616.

[31] M.J. Mehl, D.A. Papaconstantopoulos, Phys. Rev. B 54 (7) (1996) 4519–4530, http://dx.doi.org/10.1103/PhysRevB.54.4519.

[32] J.Z. Huang, W. Cho Chew, Y. Wu, L. Jun Jiang, J. Appl. Phys. 112 (1) (2012) http://dx.doi.org/10.1063/1.4732089.

[33] M. Fischetti, B. Fu, S. Narayanan, J. Kim, Semiclassical and Quantum Electronic Transport in Nanometer-Scale Structures: Empirical Pseudopotential Band Structure, Monte Carlo Simulations and Pauli Master Equation, Springer, 2011, pp. 183–247, http://dx.doi.org/10.1007/978-1-4419-8840-9, Ch. 3.

[34] G. Kresse, J. Furthmüller, Phys. Rev. B 54 (16) (1996) 11169–11186, http://dx.doi.org/10.1103/PhysRevB.54.11169, URL https://link.aps.org/doi/10.1103/PhysRevB.54.11169.

[35] J. Melenk, I. Babuška, Comput. Methods Appl. Mech. Engrg. 139 (1996) 289–314, http://dx.doi.org/10.1016/S0045-7825(96)01087-0.

[36] T. Strouboulis, I. Babuška, K. Copps, Comput. Methods Appl. Mech. Engrg. 181 (1–3) (2000) 43–69, http://dx.doi.org/10.1016/S0045-7825(99)00072-9, URL http://linkinghub.elsevier.com/retrieve/pii/S0045782599000729.

[37] I. Babuška, U. Banerjee, J.E. Osborn, Int. J. Comput. Methods 01 (01) (2004) 67–103, http://dx.doi.org/10.1142/S0219876204000083, URL http://www.worldscientific.com/doi/abs/10.1142/S0219876204000083.

[38] C.S. Lent, D.J. Kirkner, J. Appl. Phys. 67 (10) (1990) 6353–6359, http://dx.doi.org/10.1063/1.345156.

[39] H.H.B. Sørensen, P.C. Hansen, D.E. Petersen, S. Skelboe, K. Stokbro, Phys. Rev. B 79 (20) (2009) 1–10, http://dx.doi.org/10.1103/PhysRevB.79.205322.

[40] S. Tsukamoto, T. Ono, K. Hirose, S. Blügel, Phys. Rev. E 95 (3) (2017) 1–12, http://dx.doi.org/10.1103/PhysRevE.95.033309.

[41] H.H.B. Sørensen, P.C. Hansen, D.E. Petersen, S. Skelboe, K. Stokbro, Phys. Rev. B 77 (15) (2008) 1–12, http://dx.doi.org/10.1103/PhysRevB.77.155301.

[42] N.J. Higham, D.S. Mackey, N. Mackey, F. Tisseur, SIAM J. Matrix Anal. Appl. 29 (1) (2005) 143–159, http://dx.doi.org/10.1137/050646202, URL http://eprints.ma.man.ac.uk/74/.

[43] R.P. Feynman, Phys. Rev. 56 (4) (1939) 340–343, http://dx.doi.org/10.1103/PhysRev.56.340.

[44] S. Li, E. Darve, J. Comput. Phys. 231 (4) (2012) 1121–1139, http://dx.doi.org/10.1016/j.jcp.2011.05.027, URL https://arxiv.org/pdf/1104.0623.pdf.

[45] A. Kuzmin, M. Luisier, O. Schenk, Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 8097 LNCS, 2013, pp. 533–544, http://dx.doi.org/10.1007/978-3-642-40047-6{_}54, URL http://www.ics.inf.usi.chhttp://www.iis.ee.ethz.ch.

[46] K. Kazymyrenko, X. Waintal, Phys. Rev. B 77 (11) (2008) 115119, http://dx.doi.org/10.1103/PhysRevB.77.115119, URL https://link.aps.org/doi/10.1103/PhysRevB.77.115119.

[47] G. Mil'nikov, N. Mori, Y. Kamakura, Phys. Rev. B 85 (3) (2012) 1–11, http://dx.doi.org/10.1103/PhysRevB.85.035317.

[48] H. Si, AMC Trans. Math. Softw. 41 (2) (2015) 11, http://dx.doi.org/10.1007/3-540-29090-7_9.

[49] W.N. Bell, L.N. Olson, J.B. Schroder, PyAMG: Algebraic Multigrid Solvers in Python v3.0, release 3.2 URL https://github.com/pyamg/pyamg, 2015.

[50] P. Pulay, J. Comput. Chem. 3 (4) (1982) 556–560, http://dx.doi.org/10.1002/jcc.540030413.

[51] R. Shepard, M. Minkoff, Mol. Phys. 105 (19–22) (2007) 2839–2848, http://dx.doi.org/10.1080/00268970701691611.

[52] M.V. Fischetti, W.G. Vandenberghe, Advanced Physics of Electron Transport in Semiconductors and Nanostructures, Graduate Texts in Physics, Springer International Publishing, Cham, 2016, http://dx.doi.org/10.1007/978-3-319-01101-1, URL http://link.springer.com/10.1007/978-3-319-01101-1.

[53] J. Fang, W.G. Vandenberghe, M.V. Fischetti, Phys. Rev. B 94 (4) (2016) http://dx.doi.org/10.1103/PhysRevB.94.045318.

[54] K. Goto, R.A. van de Geijn, ACM Trans. Math. Software 34 (3) (2008) 1–25, http://dx.doi.org/10.1145/1356052.1356053, URL http://portal.acm.org/citation.cfm?doid=1356052.1356053.

[55] D. Verreck, A.S. Verhulst, M. Van de Put, B. Sorée, W. Magnus, A. Mocuta, N. Collaert, A. Thean, G. Groeseneken, J. Appl. Phys. 118 (13) (2015) 134502, http://dx.doi.org/10.1063/1.4931890, URL http://aip.scitation.org/doi/10.1063/1.4931890.

[56] Y. Balaji, Q. Smets, C.J. Lockhart De La Rosa, A.K.A. Lu, D. Chiappe, T. Agarwal, D.H. Lin, C. Huyghebaert, I. Radu, D. Mocuta, G. Groeseneken, IEEE J. Electron Devices Soc. 6 (January) (2018) 1018–1055, http://dx.doi.org/10.1109/JEDS.2018.2815781.

Please cite this article as: M.L. Van de Put, M.V. Fischetti and W.G. Vandenberghe, Scalable atomistic simulations of quantum electron transport using empirical pseudopotentials, Computer Physics Communications (2019), https://doi.org/10.1016/j.cpc.2019.06.009.