Electron density and transport in top-gated graphene nanoribbon devices: First-principles Green function algorithms for systems containing a large number of atoms

Denis A. Areshkin and Branislav K. Nikolić
Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716-2570, USA

(Received 7 October 2009; published 26 April 2010)

The recent fabrication of graphene nanoribbon (GNR) field-effect transistors poses a challenge for first-principles modeling of carbon nanoelectronics due to many thousand atoms present in the device. The state of the art quantum transport algorithms, based on the nonequilibrium Green function formalism combined with the density-functional theory (NEGF-DFT), were originally developed to calculate self-consistent electron density in equilibrium and at finite bias voltage (as a prerequisite to obtain conductance or current-voltage characteristics, respectively) for small molecules attached to metallic electrodes where only a few hundred atoms are typically simulated. Here we introduce combination of two numerically efficient algorithms which make it possible to extend the NEGF-DFT framework to device simulations involving large number of atoms. Our first algorithm offers an alternative to the usual evaluation of the equilibrium part of electron density via numerical contour integration of the retarded Green function in the upper complex half-plane. It is based on the replacement of the Fermi function $f(E)$ with an analytic function $\tilde{f}(E)$ coinciding with $f(E)$ inside the integration range along the real axis, but decaying exponentially in the upper complex half-plane. Although $\tilde{f}(E)$ has infinite number of poles, whose positions and residues are determined analytically, only a finite number of those poles have non-negligible residues. We also discuss how this algorithm can be extended to compute the nonequilibrium contribution to electron density, thereby evading cumbersome real-axis integration (within the bias voltage window) of NEGFs which is very difficult to converge for systems with large number of atoms while maintaining current conservation. Our second algorithm combines the recursive formulas with the geometrical partitioning of an arbitrary multiterminal device into nonuniform segments in order to reduce the computational complexity of the retarded Green function evaluation by extracting only its submatrices required for electron density and transmission function. We illustrate fusion of these two algorithms into the NEGF-DFT-type code by computing charge transfer, charge redistribution and conductance in zigzag-GNR|variable-width-armchair-GNR|zigzag-GNR two-terminal device covered with a gate electrode made of graphene layer as well. The total number of carbon and edge-passivating hydrogen atoms within the simulated central region of this device is $\simeq 7000$. Our self-consistent modeling of the gate voltage effect suggests that rather large gate voltage $\simeq 3$ eV might be required to shift the band gap of the proposed AGNR interconnect and switch the transport from insulating into the regime of a single open conducting channel.

DOI: 10.1103/PhysRevB.81.155450 PACS number(s): 73.63.−b, 71.15.−m, 72.80.Vp, 85.35.−p

I. INTRODUCTION

The recent discovery of graphene[1,2]—a single layer of graphite representing first truly two-dimensional crystal[3]—has opened new avenues for carbon nanoelectronics.[4,5] The limits on continued scaling of present silicon-based electronics are set by the fundamental physical effects (such as quantum tunneling of carriers through the gate insulator and through the body-to-drain junction; dependence of the subthreshold behavior on temperature; and discrete doping effects) where the most detrimental one is power dissipated in various leakage mechanisms.[6] This is especially dangerous for minimal field-effect transistor (FET) dimensions and oxide thicknesses. Following the discovery of carbon nanotubes (CNTs), which are rolled up sheets of graphene, the exploration of carbon nanoelectronics over the past decade as a strong contender to aging silicon technology has been centered around semiconducting CNTs as the new type of channel for FET that also makes possible unconventional transistor designs.[4]

Single-wall CNTs bring their unique features into nanoelectronics arena, such as ballistic transport or diffusion with very long mean free paths, high mobility at room temperature due to suppressed electron-acoustic-phonon scattering, current carrying capacities of the order of 10^9 A/cm^2, and one of the largest known specific stiffness.[4] However, full integration of CNTs into complex high-performance nanoelectronic devices has been thwarted by several unresolved issues, such as: (i) electronic inhomogeneity where random mixture of semiconducting and metallic CNT (due to uncontrolled distribution of diameters and chirality in current synthesis methods) degrade device performance; (ii) difficulty in aligning and patterning through standard lithography methods suitable for high-volume production because of CNTs not being flat; and (iii) extreme sensitivity to minute changes in their local chemical environment.[7]

Graphene shares many of the features of CNT, offering large critical current densities[8] and intrinsic mobility limit $\simeq 2\times 10^5$ cm^2/Vs at room temperature being higher than any of the known inorganic semiconductors.[9] Such high mobility promises near-ballistic transport and ultrafast switching. Thus, from its inception,[8] application of graphene in FET devices has been a major experimental endeavor.[10,11]

However, all graphene-FETs fabricated with wide sheets[10,11] have poor ratio of on-state current I_{on} to off-state current I_{off} due to the bulk graphene samples behaving as a zero-gap semiconductor. Nevertheless, recent breakthrough fabrication (via chemical derivation,[12] STM tip drawing[13] or CNT unrolling[14,15]) of sub-10-nm-wide graphene nanoribbons (GNRs), *all of which are semiconducting*, has led to the development of GNRFETs (Ref. 16) with $I_{\rm on}/I_{\rm off}$ ratio up to $\simeq 10^6$ which is suitable for logic devices.

Moreover, unusual band structure of graphene has generated a plethora of proposals to create devices that have no analog in silicon-based electronics. The new functionality brought by the GNR electronic structure,[17] such as "valley valves[18]" or difference in transmission properties of reflectionless 120° and highly reflective 60° turns made of GNRs with zigzag edges,[19] can only be captured by quantum transport analysis. At the same time, equilibrium interatomic charge transfer and chemical doping by different atoms[20–22] or atomic groups[23] that passivate GNR edges require to model explicitly atomistic structure and corresponding charge density within the device. These tasks are beyond the scope of popular tight-binding models[18,24,25] (projected onto the basis of single p_z orbital per carbon atom), or even simpler continuous Weyl Hamiltonian describing massless Dirac fermions as low-energy quasiparticles close to the charge neutrality point.[3] Furthermore, in the nonequilibrium state driven by the finite bias voltage one has to compute self-consistently charge redistribution and the corresponding electric potential in order to keep the gauge invariance[26] of the *I-V* characteristics[27] intact.

Finally, virtually every experiment on graphene employs gate electrodes to move the Fermi level away from the charge neutrality point or shift conduction from electron to hole carriers, so that self-consistent computation of the *inhomogeneous* charge distribution[28–30] induced by the gate voltage and its highly nontrivial effects on the band structure of GNRs (Refs. 28, 30, and 31) is necessary to understand device performance (rather than using unrealistic constant shift of the on-site potential to simulate the presence of the gate electrode in the simple tight-binding models[18]).

Thus, the prime candidate capable of handling all of these issues within a unified quantum transport framework[32,33] is the nonequilibrium Green function (NEGF) formalism[34] combined with the density functional theory (DFT) in standard approximation schemes[35] (such as LDA, GGA, or B3LYP) for its exchange-correlation potential. The sophisticated algorithms[36–45] developed to implement the NEGF-DFT framework over the past decade can be encapsulated by the iterative self-consistent loop,[34]

$$n^{\rm in}(\mathbf{r}) \Rightarrow {\rm DFT} \rightarrow \mathbf{H}_{\rm KS}[n(\mathbf{r})] \Rightarrow {\rm NEGF} \rightarrow n^{\rm out}(\mathbf{r}). \qquad (1)$$

The loop starts from the initial input electron density $n^{\rm in}(\mathbf{r}) \Rightarrow$ employs some standard DFT code[35] (typically in the basis set of finite-range orbitals for the valence electrons which allows for faster numerics and unambiguous partitioning of the system into "central region" and the semi-infinite ideal leads) to get the single particle Kohn-Sham Hamiltonian $\mathbf{H}_{\rm KS}[n(\mathbf{r})] = -\hbar^2\nabla^2/2m + V^{\rm eff}(\mathbf{r})$ $[V^{\rm eff}(\mathbf{r}) = V_H(\mathbf{r}) + V_{\rm xc}(\mathbf{r}) + V_{\rm ext}(\mathbf{r})$ is the DFT mean-field potential due to other electrons where $V_H(\mathbf{r})$ is the Hartree, $V_{\rm xc}(\mathbf{r})$ is the exchange-correlation, and $V_{\rm ext}(\mathbf{r})$ is the external potential contribution$]$ $\Rightarrow$ inversion of $\mathbf{H}_{\rm KS}[n(\mathbf{r})]$ yields the retarded Green function $\mathbf{G}^r(E)$ whose integration over energy determines the density matrix via NEGF-based formula,

$$\boldsymbol{\rho} = -\frac{1}{\pi}\int_{-\infty}^{+\infty} dE\, {\rm Im}[\mathbf{G}^r(E)] f(E-\mu_R) - \frac{1}{\pi}\int_{-\infty}^{+\infty} dE \mathbf{G}^r(E)\cdot {\rm Im}[\boldsymbol{\Sigma}_L(E)]\cdot \mathbf{G}^a(E)[f(E-\mu_L) - f(E-\mu_R)] = \boldsymbol{\rho}_{\rm eq} + \boldsymbol{\rho}_{\rm neq}. \qquad (2)$$

The matrix elements $n^{\rm out}(\mathbf{r}) = \langle \mathbf{r}|\boldsymbol{\rho}|\mathbf{r}\rangle$ are the new electron density as the starting point of the next iteration. This procedure is repeated until the convergence criterion $\|\boldsymbol{\rho}^{\rm out} - \boldsymbol{\rho}^{\rm in}\| < \delta$ is reached, where $\delta \ll 1$ is a tolerance parameter.

The representation of the retarded Green function in the local orbital basis requires to compute the inverse matrix

$$\mathbf{G}^r(E) = [E - \mathbf{H}_{\rm KS}[n(\mathbf{r})] - \boldsymbol{\Sigma}(E)]^{-1}. \qquad (3)$$

The advanced Green function matrix is defined as $\mathbf{G}^a(E) = [\mathbf{G}^r(E)]^\dagger$. The non-Hermitian matrix $\boldsymbol{\Sigma}(E) = \boldsymbol{\Sigma}_L(E) + \boldsymbol{\Sigma}_R(E)$ is the sum of the retarded self-energy matrices introduced by the "interaction" with the left $[\boldsymbol{\Sigma}_L(E)]$ and the right $[\boldsymbol{\Sigma}_R(E)]$ leads. These self-energies determine escape rates of electrons from the central region into the semi-infinite ideal leads, so that an open quantum system can be viewed as being described by the (non-Hermitian) Hamiltonian $\mathbf{H}_{\rm open} = \mathbf{H}_{\rm KS}[n(\mathbf{r})] + \boldsymbol{\Sigma}(E)$.

The NEGF postprocessing of the converged result of DFT calculations makes it possible to obtain the current through a two-terminal device in terms of the Landauer-type formula[34]

$$I(V_{ds}) = \frac{2e}{h}\int_{-\infty}^{+\infty} dE T(E,V_{ds})[f(E-\mu_L) - f(E-\mu_R)]. \qquad (4)$$

This integrates the self-consistent transmission function

$$T(E,V_{ds}) = {\rm Tr}\{\boldsymbol{\Gamma}_R(E,V_{ds})\mathbf{G}^r_{S,1}\boldsymbol{\Gamma}_L(E,V_{ds})\mathbf{G}^a_{1,S}\}, \qquad (5)$$

for electrons injected at energy E to propagate from the left to the right electrode under the source-drain applied bias voltage $\mu_L - \mu_R = eV_{ds}$. Here $\mathbf{G}^r_{S,1}$ is the submatrix of $\mathbf{G}^r$ whose elements $\langle S|\hat{G}^r|1\rangle$ connect orbitals in the first lead supercell (layer denoted as 1) of the extended central region "sample+portion of the electrodes" to the last lead supercell (layer denoted as S) of the simulated region.

The matrices $\boldsymbol{\Gamma}_{L,R}(E) = i[\boldsymbol{\Sigma}_{L,R}(E) - \boldsymbol{\Sigma}^\dagger_{L,R}(E)] = -2\,{\rm Im}\,\boldsymbol{\Sigma}_{L,R}(E)$ account for the level broadening due to the coupling to the leads.[34] A usual assumption about the leads is that the effect of the bias voltage can be taken into account by a rigid shift of their electronic structure, so that $\boldsymbol{\Sigma}_{L,R}(E,V_{ds}) = \boldsymbol{\Sigma}_{L,R}(E \mp eV_{ds}/2,0)$ and $\boldsymbol{\Gamma}_{L,R}(E,V_{ds}) = \boldsymbol{\Gamma}_{L,R}(E \mp eV_{ds}/2,0)$ are computed in equilibrium and then the shift $\pm eV_{ds}/2$ is applied to their electronic structure to mimic the applied bias. The energy window for the integral in Eq. (4) is defined by the difference of Fermi functions $f(E-\mu_L) - f(E-\mu_R)$ of macroscopic reservoirs into which semi-infinite ideal leads terminate. The formula (4) is valid only for coherent transport, i.e., assuming absence of dephasing[47] due electron-phonon or electron-electron interactions (beyond those captured by the mean-field treatment[45,46]).

Thus, the most demanding computational task of the NEGF-DFT framework is the self-consistent evaluation of the density matrix $\boldsymbol{\rho}$ whose different algorithmic steps have the following[32] *computational complexity*[48] in terms of the number of atoms N (Ref. 49): (i) the computation $n^{\text{in}}(\mathbf{r}) \rightarrow V^{\text{eff}}(\mathbf{r})$ of the effective potential for $\mathbf{H}_{\text{KS}}[n(\mathbf{r})]$ has complexity $O(N \log N)$; (ii) the second step, $V^{\text{eff}}(\mathbf{r}) \rightarrow \mathbf{H}_{\text{KS}}[n(\mathbf{r})]$, has complexity $O(N)$; (iii) computation of all elements of the retarded Green function, $\mathbf{H}_{\text{KS}}[n(\mathbf{r})] \rightarrow \mathbf{G}^r$, requires $O(N^3)$ operations; (iv) $\mathbf{G}^r \rightarrow \boldsymbol{\rho}$ scales as $O(N)$; and (v) the final step $\boldsymbol{\rho} \rightarrow n^{\text{out}}(\mathbf{r})$ also has complexity $O(N)$. Obviously, the bottleneck is set by the retarded Green function computation. Since NEGF-DFT computational codes[36–42,44] are developed and tested for small molecules attached to metallic electrodes (where they are successful when coupling between the molecule and the electrodes is strong enough to diminish Coulomb blockade effects[33]), they typically evaluate all elements of $\mathbf{G}^r$ by inverting through Eq. (3) the Hamiltonian of the extended molecule region. Because this has to be done repeatedly through self-consistent loop [Eq. (1)], the number of atoms in the extended central region "molecule+portion of the electrodes" that can be simulated is limited to few hundreds. This bottleneck also prevents realistic modeling of single or multiple[50] gate electrodes—instead of an additional layer of atoms covering portion of the central region, one typically employs a uniform electric field in the direction perpendicular to the transport.[51,52]

A more subtle reason for the failure of conventionally implemented NEGF-DFT codes when applied to systems containing large number of atoms is the integration in the second term $\boldsymbol{\rho}_{\text{neq}}$ in Eq. (2) which must be performed along the real axis since the integrand is not analytic anywhere in the complex plan. Although this integration is restricted by the Fermi functions to a segment of the order of the applied bias voltage, a very fine integration grid must be used to capture locations of subband edges (introduced by semi-infinite leads) and broadened molecular orbitals where sharp peaks in the integrand occur. This problem is exacerbated in devices containing large number of atoms where the increasing number of such sharp peaks—due to van Hove singularities in the density of states of the leads or quasibound states present when different contacts throughout the device are not perfectly transparent—can make it virtually impossible to converge $\boldsymbol{\rho}_{\text{neq}}$.

The present approach in NEGF-DFT algorithms to deal with this issue is to move the line of integration slightly into the complex plane. However, this effectively adds small imaginary part $i\eta$ to the Hamiltonian $\mathbf{H}_{\text{open}}$ which, therefore, does not conserve current. For example, direct application of this procedure to experimental graphene devices, such as 100 nm long GNRFET of Ref. 16, would lead to substantial difference between the total current in the left and the right leads. This issue is rarely discussed in the usual NEGF-DFT treatment of transport through relatively short molecules where such violation of current conservation is small.

Some recent attempts to solve it, such as locating the peaks due to quasibound states and patching the nonequilibrium density matrix integral,[53,54] cannot be applied to large systems with many such peaks. The peaks can be broadened by physical dephasing mechanisms due to electron-electron[45,46] or electron-phonon interactions,[47] but this drastically changes the NEGF-DFT approach by requiring additional and computationally very expensive self-consistent loops to calculate extra self-energy functionals[34,45,46] due to interactions within the device for which the sparsity of the Hamiltonian matrix $\mathbf{H}_{\text{open}}$ becomes irrelevant.

Recent efforts[50,53–58] to replace some of the algorithms within the NEGF part of the NEGF-DFT scheme, such as unfavorable computational complexity of the brute force matrix inversion[55–57] or the real-axis integration[53,54] in $\boldsymbol{\rho}_{\text{neq}}$, have still not led to self-consistent electron density and transport calculations for systems composed of more than about a thousand of atoms.[50] Here we introduce modified NEGF-DFT scheme which is based on our novel algorithm for the integrations in Eq. (2) combined with the partitioning the nanostructure of arbitrary shape into slices containing much smaller number of atoms. The Green function matrices of these slices, needed to obtain the electron density within the slice, are computed recursively with much more favorable computational complexity than $O(N^3)$. The number of iteration steps within the self-consistent loop is further reduced, in the case of nanodevices in equilibrium or in quasiequilibrium situations (e.g., due to by nonzero gate voltage and zero or linear response bias voltage), via modified Broyden mixing scheme for input and output charge density. We demonstrate the capability of our computational code, termed CANNES (carbon nanoelectronics simulator), to treat multiterminal structures containing large number of atoms by computing the self-consistent electron density and conductance in the presence of the gate voltage in a graphene nanodevice whose extended central region is composed of $\simeq$7000 carbon and hydrogen atoms.

The paper is organized as follows. Sec. II elaborates on the "pole summation" algorithm for computing integrals in $\boldsymbol{\rho}$. In Sec. III we demonstrate efficiency of our approach by setting up a *three-terminal* FET-type device whose source and drain electrodes are made of zigzag graphene nanoribbon (ZGNR) while its channel is an armchair GNR (AGNR) of variable width and with sizable energy gap. The third electrode is gate modeled as a rectangularly shaped layer of carbon atoms covering the FET channel. The dangling bonds of all-graphene layers are terminated by hydrogen atoms. The DFT part of the calculation is carried out using the self-consistent environment-dependent tight-binding model (SC-EDTB) with four orbitals per carbon atom and one orbital per hydrogen atom, which is specifically tailored to simulate eigenvalue spectra, electron densities and Coulomb potential distributions for carbon-hydrogen nanostructures.[59,60] The combination of "pole summation" algorithm with the recursive Green function formulas allows us to compute in Sec. III intricate electric potential distribution in the space around ZGNR-AGNR-ZGNR FET device, as well as to demonstrate how much voltage has to be applied on the gate electrode to push the device from the off-state due to the gap of AGNR into an on-state enabled by a single transport channel crossing the Fermi level. The computed source-drain conductance as a function of the gate voltage also demonstrates that even at zero gate voltage there is a difference between the non-self-consistent and self-consistent conductance, where the latter takes into account charge transfer between different atomic species or different segments of the device. We conclude in Sec. IV.

II. SELF-CONSISTENT ALGORITHMS FOR ELECTRON DENSITY

We start by rewriting the equilibrium contribution to the density matrix (2),

$$\boldsymbol{\rho}_{\text{eq}}(\mu,T) = -\frac{1}{\pi}\int_{E_{\min}}^{+\infty} dE\ \text{Im}[\mathbf{G}^r(E)]f(\mu,T,E), \tag{6}$$

in the form which emphasizes its dependence on the chemical potential μ and temperature T, as well as that the lower limit of integration is the lowest energy at which $\text{Im}[\mathbf{G}^r(E_{\min})] \neq 0$. As long as the end-point $E_{\min}$ is selected[37,44] below the bottom of the valence band edge, there is no further contribution to the integrand, and thus the expression is exact. Although this looks obvious, it is important to point out that if the value $|E_{\min}|$ is too small, and there are states left outside of the contour, the corresponding states will not be included in the integration. This causes charge to erroneously disappear from the system, which typically initiates an avalanche effect, pushing the energy levels even further out, and even more charge is lost, until the system is totally void of electrons. When this occurs, the calculation will actually converge trivially, but to a physically incorrect solution.

Since diagonal matrix elements of $\mathbf{G}^r(E)$ are a rapidly varying function of energy, a direct integration along the real axis would be rather ineffective since its numerical accuracy is not sufficient to achieve convergence of the self-consistent electron density. Instead, present NEGF-DFT computational codes[36,37,44] deform the integration contour into the upper complex half-plane $\text{Im}[E]>0$, where the retarded Green function is much smoother. This is allowed since $\mathbf{G}^r(E)$ is analytic in the upper complex half-plane (all of its poles are slightly displaced below the real axis).

The thick white line in Fig. 1 designates typically chosen[36,37,40,44] integration contour. It consists of a semicircular part SC and a horizontal line L parallel to the real axis on the right which is positioned to enclose specific number N_{poles} of the Fermi function poles $z^{(n)}$ while ensuring that SC and L are sufficiently far away from the real axis so that the Green function is smooth over both of these two segments [the main variation of the integrand on L comes from the Fermi function $f(E)$ which, therefore, can be used as a weight function in the quadrature[37,44]]. The final expression for $\boldsymbol{\rho}_{\text{eq}}$ obtained in this procedure (using the Cauchy residue theorem for the closed contour $SC+L+$vertical segment from L to the real axis+portion of the real axis) is

$$\boldsymbol{\rho}_{\text{eq}} = -\frac{1}{\pi}\text{Im}\left[\int_{SC+L} dz\mathbf{G}^r(z)f(\mu,T,z) - 2\pi i k_B T\sum_{n}^{N_{\text{poles}}}\mathbf{G}^r(z^{(n)})\right], \tag{7}$$

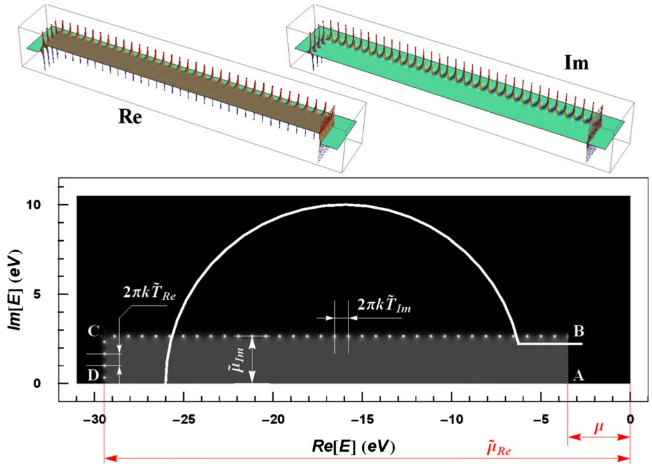

FIG. 1. (Color online) The density plot of the absolute value of $\tilde{f}(E)$ in the upper complex half-plane. Lighter color denotes greater value of $|\tilde{f}|$. Solid black corresponds to zero, while gray color inside the dotted rectangle represents unity. White dots denote the poles with their size being roughly proportional to the absolute value of the residue. Poles running along AB, BC, and CD edges of the rectangle correspond to $z^{(n)}$, $\tilde{z}_{\text{Im}}^{(n)}$, and $\tilde{z}_{\text{Re}}^{(n)}$, respectively. Thick white curve denotes the integration contour traditionally used in NEGF-DFT computational codes (Refs. 37 and 44). Top insets are three-dimensional (3D) plots of $\text{Re}[\tilde{f}]$ and $\text{Im}[\tilde{f}]$ in the upper complex half-plane.

where the smoothness of $\mathbf{G}^r(E)$ on $SC+L$ contour is exploited to perform the approximate integration in the first term by using a quadrature with a small number of points.[37,44]

Obviously, it would be highly advantageous to be able to compute integral in Eq. (6) precisely and without worrying about proper selection of parameters for positioning SC and L, via a simple summation over a finite set of complex energies akin to the second term of Eq. (7). Here we introduce such an algorithm which makes possible virtually exact evaluation of $\boldsymbol{\rho}_{\text{eq}}$ by "pole summation." This algorithm is discussed separately for high temperatures (and/or valence electrons) in Sec. II A and for low temperatures (and/or core electrons) in Sec. II B.

A. High temperature and/or valence electrons

The algorithm for equilibrium density matrix computation discussed in this Section can be used when the inequality

$$(\mu - E_{\min})/k_B T \lesssim 10^3, \tag{8}$$

is satisfied. If Eq. (8) is not satisfied, a slightly more elaborate algorithm described in the next Sec. II B is needed. Let us define the desired precision through the non-negative number p, such that the magnitude of the relative error is $\delta \leq e^{-p}$. In most cases the machine precision roughly corresponds to $p=30$, while the practical range of p is usually between 21 and 27.

We start by introducing a function $\tilde{f}$

$$\tilde{f}(\mu,\tilde{\mu}_{\mathrm{Re}},\tilde{\mu}_{\mathrm{Im}},T,\tilde{T}_{\mathrm{Re}},\tilde{T}_{\mathrm{Im}},E) = f(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},E) \times [f(\mu,T,E) - f(\tilde{\mu}_{\mathrm{Re}},\tilde{T}_{\mathrm{Re}},E)], \tag{9}$$

where all its arguments except E are limited to real domain and satisfy the following inequalities (k_B is the Boltzmann constant and $i^2=-1$):

$$\tilde{T}_{\mathrm{Re}} > 0, \quad \tilde{T}_{\mathrm{Im}} > 0, \tag{10a}$$

$$\tilde{\mu}_{\mathrm{Re}} \le E_{\mathrm{min}} - pk_B\tilde{T}_{\mathrm{Re}}, \tag{10b}$$

$$\tilde{\mu}_{\mathrm{Im}} \ge pk_B\tilde{T}_{\mathrm{Im}}. \tag{10c}$$

The choice of parameters given by Eq. (10) guarantees that for real $E \ge E_{\mathrm{min}}$ the function $\tilde{f}$ deviates from f by no more than δ. Therefore the replacement of f with $\tilde{f}$ in the integrand of Eq. (6) will result in the relative error less than δ. In the following we assume that $p \ge 21$ so that $\delta \le 10^{-9}$.

Thus, for all practical purposes we can state that (all arguments except E are omitted for brevity)

$$\boldsymbol{\rho}_{\mathrm{eq}} = -\frac{1}{\pi}\mathrm{Im}\left[\int_{-\infty}^{+\infty} dE\mathbf{G}^r(E)\tilde{f}(E)\right]. \tag{11}$$

The poles and residues of the first term in the product on the right-hand side of Eq. (9) are given by

$$\tilde{z}^{(n)}_{\mathrm{Im}} = i\tilde{\mu}_{\mathrm{Im}} + \pi k_B\tilde{T}_{\mathrm{Im}}(2n+1), \tag{12a}$$

$$\mathrm{Res}[f(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},z)]_{z=\tilde{z}^{(n)}_{\mathrm{Im}}} = -ik_B\tilde{T}_{\mathrm{Im}}. \tag{12b}$$

where n is an integer. Similarly, the poles and residues of $f(\mu,T,E)$ in the second term are

$$z^{(n)} = \mu + \pi ik_BT(2n+1), \tag{13a}$$

$$\mathrm{Res}[f(\mu,T,z)]_{z=z^{(n)}} = -k_BT, \tag{13b}$$

and for $f(\tilde{\mu}_{\mathrm{Re}},\tilde{T}_{\mathrm{Re}},E)$ they are

$$\tilde{z}^{(n)}_{\mathrm{Re}} = \tilde{\mu}_{\mathrm{Re}} + \pi ik_B\tilde{T}_{\mathrm{Re}}(2n+1), \tag{14a}$$

$$\mathrm{Res}[f(\tilde{\mu}_{\mathrm{Re}},\tilde{T}_{\mathrm{Re}},z)]_{z=\tilde{z}^{(n)}_{\mathrm{Re}}} = -k_B\tilde{T}_{\mathrm{Re}}. \tag{14b}$$

Inequalities (10) provide sufficient freedom to prevent the coincidence of the poles $z^{(j)}$, $\tilde{z}^{(m)}_{\mathrm{Im}}$, and $\tilde{z}^{(n)}_{\mathrm{Re}}$ ($\forall$ j, m, and n). Thus, $\tilde{f}$ only has first-order poles with residues given by

$$\mathrm{Res}[\tilde{f}(z)]_{z=\tilde{z}^{(n)}_{\mathrm{Im}}} = -ik\tilde{T}_{\mathrm{Im}} \times [f(\mu,T,\tilde{z}^{(n)}_{\mathrm{Im}}) - f(\tilde{\mu}_{\mathrm{Re}},\tilde{T}_{\mathrm{Re}},\tilde{z}^{(n)}_{\mathrm{Im}})], \tag{15a}$$

$$\mathrm{Res}[\tilde{f}(z)]_{z=z^{(n)}} = -k_BTf(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},z^{(n)}), \tag{15b}$$

$$\mathrm{Res}[\tilde{f}(z)]_{z=\tilde{z}^{(n)}_{\mathrm{Re}}} = k_BTf(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},\tilde{z}^{(n)}_{\mathrm{Re}}). \tag{15c}$$

In the upper complex half-plane the residues [Eq. (15a)] decay exponentially if $\mathrm{Re}(\tilde{z}^{(n)}_{\mathrm{Im}})$ lies outside the interval $[\tilde{\mu}_{\mathrm{Re}},\mu]$, and the residues [Eqs. (15b) and (15c)] decay exponentially if the imaginary component of the poles $z^{(n)}$ or $\tilde{z}^{(n)}_{\mathrm{Re}}$ exceeds $\tilde{\mu}_{\mathrm{Im}}$. Thus, for any given p only the limited number of poles $\{Z_j\}$, $j \in \{1,N_{\mathrm{pole}}\}$ have non-negligible residues.

If one replaces the real-axis integration in Eq. (11) by the integration along the semicircular contour of the sufficiently large radius in the upper complex half-plane, the contour contribution to the integral is zero, and the contribution from the poles is solely from $\{Z_j\}$. The integral (11) is computed as the sum over all nonzero residues,

$$\boldsymbol{\rho}_{\mathrm{eq}} = -\frac{1}{\pi}\mathrm{Im}\left[\sum_{j=1}^{N_{\mathrm{pole}}} 2\pi i\ \mathrm{Res}[\tilde{f}(z)]_{z=Z_j}\mathbf{G}^r(Z_j)\right], \tag{16}$$

where the set $\{Z_j\}$ is comprised of only those $\{\tilde{z}^{(n)}_{\mathrm{Im}}\}$, $\{z^{(n)}\}$, and $\{\tilde{z}^{(n)}_{\mathrm{Re}}\}$ poles which satisfy

$$|f(\mu,T,\tilde{z}^{(n)}_{\mathrm{Im}}) - f(\tilde{\mu}_{\mathrm{Re}},-\tilde{T}_{\mathrm{Re}},\tilde{z}^{(n)}_{\mathrm{Im}})| \ge e^{-p}, \tag{17a}$$

$$|f(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},z^{(n)})| \ge e^{-p}, \tag{17b}$$

$$|f(i\tilde{\mu}_{\mathrm{Im}},i\tilde{T}_{\mathrm{Im}},\tilde{z}^{(n)}_{\mathrm{Re}})| \ge e^{-p}, \tag{17c}$$

respectively, in order to keep the relative error below e^{-p}.

For values of E_{min} and T obeying the inequality (8) and $21 \le p \le 30$ the number of relevant poles N_{pole} is moderate. For example, it is safe to chose $E_{\mathrm{min}}=-27$ eV for valence electrons in a hydrocarbon system (note that this value for E_{min} is measured from the vacuum level). Then, at room temperature the ratio [Eq. (8)] is around 700, and for $p=21$ the minimal number of required poles for parameters satisfying Eq. (10) equals 76. Decreasing p down to machine precision raises the minimal number of poles to 96.

Figure 1 shows the density plot of $\tilde{f}$ corresponding to $p=21$ and $E_{\mathrm{min}}=-27$ eV used to compute self-consistent electron within the graphene nanodevice example of Sec. III. The minimal number of poles N_{pole} is obtained as follows. We consider $\tilde{T}_{\mathrm{Im}}$ and $\tilde{T}_{\mathrm{Re}}$ as free parameters, and the minimum allowed $\tilde{\mu}_{\mathrm{Re}}$ and $\tilde{\mu}_{\mathrm{Im}}$ are obtained from equalities in constraints imposed by Eq. (10). Then, the number of poles $z^{(n)}$ is approximately twice the value of $\tilde{\mu}_{\mathrm{Im}}$ divided by the interpole distance

$$N_{AB} = \frac{2\tilde{\mu}_{\mathrm{Im}}}{2\pi k_BT}, \tag{18}$$

and the approximate numbers of poles along the lines CB and DC in Fig. 1 are

$$N_{CB} = \frac{\mu - \tilde{\mu}_{\mathrm{Re}} + pk_B\tilde{T}_{\mathrm{Re}} + pk_BT}{2\pi k_B\tilde{T}_{\mathrm{Im}}}, \tag{19a}$$

$$N_{DC} = \frac{2\tilde{\mu}_{\mathrm{Im}}}{2\pi k_B\tilde{T}_{\mathrm{Re}}}, \tag{19b}$$

respectively. The optimal values of $\tilde{T}_{\mathrm{Im}}$ and $\tilde{T}_{\mathrm{Re}}$ are obtained by minimizing $N_{\mathrm{pole}}=N_{AB}+N_{CD}+N_{DC}$ in the space of these two parameters. A small $\tilde{T}_{\mathrm{Re}}$ and $\tilde{\mu}_{\mathrm{Im}}$ adjustment, subject to

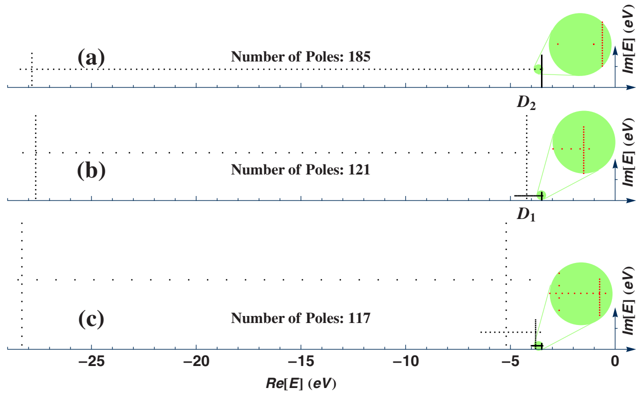

FIG. 2. (Color online) The poles with nonzero residues for the same system shown in Fig. 1 but at $k_BT=0.003$ eV: (a) poles of $\widetilde{f}$; (b) poles of $\widetilde{F}^{(2)}$; and (c) poles of $\widetilde{F}^{(3)}$. Circular zoomed-out regions depict dense pole arrangement at energies close to the chemical potential μ.

constraints [Eq. (10)], is made afterward to place the line CD right in between the two poles on lines AB and DC (cf. Figs. 1 and 2). This is done to ensure that the poles are not too close to each other, otherwise a large numerical errors may occur.

B. Low temperature and/or full core simulations

The minimum number of poles N_{pole} is scaled by the temperature and the energy interval $\mu-\widetilde{\mu}_{\text{Re}}$. In order to reduce N_{pole}, it is desirable to have as large spacing between the poles $\widetilde{z}_{\text{Im}}^{(n)}$ as possible. According to Eq. (10c), increasing $\widetilde{T}_{\text{Im}}$ for the given p means the increase in $\widetilde{\mu}_{\text{Im}}$. The increase in $\widetilde{\mu}_{\text{Im}}$ in turn increases the length of the segment AB, and hence the number of poles $z^{(n)}$ to be summed. On the other hand, reducing the number of $z^{(n)}$ (i.e., decreasing $|AB|=\widetilde{\mu}_{\text{Im}}$), will bring the line BC closer to the real axis, so to prevent deviation of $\widetilde{f}$ from f on the real axis requires to decrease $\widetilde{T}_{\text{Im}}$. The latter increases the number of poles $\widetilde{z}_{\text{Im}}^{(n)}$ along the line BC.

The simple solution to this problem is to break the interval between $\widetilde{\mu}_{\text{Re}}$ and μ into several subintervals, and apply the scheme presented in Sec. II A to each subinterval. For example, if the original interval is split into two subintervals, one needs to replace $\widetilde{f}$ with $\widetilde{F}^{(2)}$, which is the sum of two functions $\widetilde{f}$

$$\begin{aligned}&\widetilde{F}^{(2)}(\mu,\widetilde{\mu}_{\text{Re}_{1,2}},\widetilde{\mu}_{\text{Im}_{1,2}},T,\widetilde{T}_{\text{Re}_{1,2}},\widetilde{T}_{\text{Im}_{1,2}},E)\\&=\widetilde{f}(\mu,\widetilde{\mu}_{\text{Re}_1},\widetilde{\mu}_{\text{Im}_1},T,\widetilde{T}_{\text{Re}_1},\widetilde{T}_{\text{Im}_1},E)\\&\quad+\widetilde{f}(\widetilde{\mu}_{\text{Re}_1},\widetilde{\mu}_{\text{Re}_2},\widetilde{\mu}_{\text{Im}_2},\widetilde{T}_{\text{Re}_1},\widetilde{T}_{\text{Re}_2},\widetilde{T}_{\text{Im}_2},E),\end{aligned}\tag{20}$$

where $T<\widetilde{T}_{\text{Re}_1}<\widetilde{T}_{\text{Re}_2}$; $\widetilde{\mu}_{\text{Re}_2}<\widetilde{\mu}_{\text{Re}_1}<\mu$; and $\widetilde{\mu}_{\text{Im}_1}<\widetilde{\mu}_{\text{Im}_2}$. The parameters $\widetilde{\mu}_{\text{Re}_{1,2}}$, $\widetilde{\mu}_{\text{Im}_{1,2}}$, $\widetilde{T}_{\text{Re}_{1,2}}$, and $\widetilde{T}_{\text{Im}_{1,2}}$ ensure the required precision by satisfying the constraints similar to Eq. (10),

$$\widetilde{\mu}_{\text{Re}_2}\leq E_{\text{min}}-pk_B\widetilde{T}_{\text{Re}_2},\tag{21a}$$

$$\widetilde{\mu}_{\text{Im}_1}\geq pk_B\widetilde{T}_{\text{Im}_1},\quad \widetilde{\mu}_{\text{Im}_2}\geq pk_B\widetilde{T}_{\text{Im}_2}.\tag{21b}$$

Figure 2(b) illustrates these concepts. Poles forming the left (smaller) and the right (bigger) rectangles are associated respectively with the first and the second term in Eq. (20). The poles running along the line D_1D_2 are the same for the first and second term in Eq. (20).

The minimization of the total number of poles N_{pole} is performed analogously to Eqs. (18) and (19). For $\widetilde{F}^{(2)}$ the optimization parameters are $\widetilde{T}_{\text{Re}_1}$, $\widetilde{T}_{\text{Re}_2}$, $\widetilde{T}_{\text{Im}_1}$, and $\widetilde{T}_{\text{Im}_1}$. The starting point for the conjugate gradient minimization is $\widetilde{T}_{\text{Re}_1}=10\times T$ and $\widetilde{\mu}_{\text{Im}_2}=10\times\widetilde{\mu}_{\text{Im}_1}$, so that the optimized parameters fit this order of magnitude relationship. Indeed, the size of the integration intervals in Figs. 2(b) and 2(c) increases by an order of magnitude from right to left. For this reason N_{pole} grows logarithmically with increasing ratio $(\mu-E_{\text{min}})/k_BT$. That is, depending on p, approximately 30 to 40 extra poles are required for each decade of this ratio increase (i.e., per order of magnitude in temperature reduction).

C. Approximate real-axis integration of nonanalytic functions

The concepts presented in Sec. II A allow for efficient and exact evaluation of the $\mathbf{G}^r(E)$ moments in the interval bounded by two Fermi functions. This property can be used for systematic approximation of $\mathbf{G}^a(E)$ with the function $\widetilde{\mathbf{G}}^a(E)$ such that $\widetilde{\mathbf{G}}^a(E)\approx\mathbf{G}^a(E)$ on the real axis, and which is analytic in the upper complex half-plane. This approximation can be used to transform the nonanalytic integrands to analytic functions.

Obvious applications of this idea to NEGF-DFT framework would be the computation of nonequilibrium contribution $\boldsymbol{\rho}_{\text{neq}}$ to the density matrix in Eq. (2). Because the functions $\mathbf{G}^r(E)$ and $\mathbf{G}^a(E)$ in the integrand of $\boldsymbol{\rho}_{\text{neq}}$ are nonanalytic below and above the real axis, respectively, the integrand is nonanalytic function in the entire complex energy plane. Thus, no integration contour deformation akin to Fig. 1 can be exploited to avoid direct integration along the real axis to obtain $\boldsymbol{\rho}_{\text{neq}}$. On the other hand, such direct integration along the real axis is computationally expensive due to the need for very fine integration grids.[53,54] As discussed in Sec. I, integration may not even converge when the integrand becomes too spiky with numerous closely spaced sharp peaks for devices containing large number of atoms.

Let us divide the interval $[\mu_R,\mu_L]$ into M subintervals of equal size $\Delta\mu$

$$\mu_0=\mu_R,\quad \mu_M=\mu_L,\quad \mu_m=\mu_R+m\Delta\mu,\tag{22}$$

where we assume for simplicity that $\Delta\mu=2k_BT$. Then $\boldsymbol{\rho}_{\text{neq}}$ in Eq. (2) can be rewritten as

$$\begin{aligned}\boldsymbol{\rho}_{\text{neq}}=\sum_{m=1}^{M}\int_{-\infty}^{+\infty}dE\,\mathbf{G}^r(E)\cdot\text{Im}[\boldsymbol{\Sigma}(E)]\cdot\mathbf{G}^a(E)\\\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)].\end{aligned}\tag{23}$$

For each interval $[\mu_{m-1},\mu_m]$ in the sum (23) we approximate $\mathbf{G}^r(E)$ by the power expansion with respect to the deviation from the center of the interval $\xi_m=(\mu_{m-1}+\mu_m)/2$

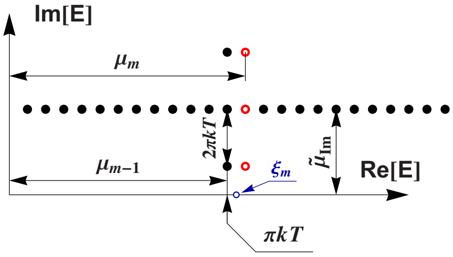

FIG. 3. (Color online) Poles of the function $\tilde{f}(\mu_m,\mu_{m-1},\tilde{\mu}_{\rm Im},T,T,\tilde{T}_{\rm Im},z)$ used to evaluate the integral in Eq. (26). For a chosen precision set by $p=23$, the contribution from 28 poles has to be summed. Three poles of $f(\mu_m,T,z)$ are marked with the red empty circles. The values of the retarded Green function at most of the poles shown are reused to compute matrices $\mathbf{M}_n^{(d)}$ in Eq. (25) for $n\neq m$, so that the average number of Green functions to be computed per interval equals 3.

$$\mathbf{G}^r_m(E)\approx\tilde{\mathbf{G}}^r_m(E)=\sum_{\kappa=0}^{K}\mathbf{g}_m^{(\kappa)}\times(E-\xi_m)^{\kappa}, \qquad (24)$$

where $\mathbf{g}_m^{(\kappa)}$ are constant matrices. We require that the moments $\mathbf{M}_m^{(d)}$ up to order K for $\tilde{\mathbf{G}}^r_m$ and $\mathbf{G}^r_m$ coincide

$$\begin{aligned}\mathbf{M}_m^{(d)}&\equiv\int_{-\infty}^{+\infty}dE\mathbf{G}^r(E)(E-\xi_m)^d\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)]\\&=\int_{-\infty}^{+\infty}dE\sum_{\kappa=0}^{K}\mathbf{g}_m^{(\kappa)}\times(E-\xi_m)^{\kappa+d}\\&\quad\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)], \qquad (25)\end{aligned}$$

where $d\subset[0,K]$.

The first integral in Eq. (25) can be computed accurately as

$$\begin{aligned}&\int_{-\infty}^{+\infty}dE\mathbf{G}^r(E)(E-\xi_m)^d\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)]\\&\quad=\int_{-\infty}^{+\infty}dE\mathbf{G}^r(E)(E-\xi_m)^d\times\tilde{f}(\mu_m,\mu_{m-1},\tilde{\mu}_{\rm Im},T,T,\tilde{T}_{\rm Im},E).\end{aligned} \qquad (26)$$

Figure 3 shows the poles of $\tilde{f}$ from Eq. (26) for the case $p=23$, $\tilde{\mu}_{\rm Im}=3\pi k_BT$, and $\tilde{T}_{\rm Im}=T/\pi$. Even though the number of poles to be summed per every moment equals 28, the number of points per integration interval $\Delta\mu$ at which $\mathbf{G}^r(E)$ needs to be calculated is 3 because the values of $\mathbf{G}^r(E)$ at different poles are reused in computation of the moments at different intervals. Thus, $\mathbf{M}_m^{(d)}$ is computed similarly to Eq. (16), with the only difference being that $\mathbf{G}^r(Z_j)$ is now replaced by $\mathbf{G}^r(Z_j)(Z_j-\xi_m)^d$.

Because matrices $\mathbf{g}_m^{(\kappa)}$ do not depend on energy, the integrals in the second term of Eq. (25)

$$\Upsilon_\kappa\equiv\int_{-\infty}^{+\infty}dE(E-\xi_m)^{\kappa}\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)], \qquad (27)$$

can be computed analytically. Here we provide example solution of this problem for $K=2$ (the solutions for $K>2$ are similar to this). The integrals Υ_κ are nonzero when integer κ is even. For example, assuming $\Delta\mu=2k_BT$ they are

$$\Upsilon_0=2k_BT,\quad \Upsilon_2=\frac{2}{3}(k_BT)^3(1+\pi^2),$$

$$\Upsilon_4=\frac{2}{15}(k_BT)^5(3+10\pi^2+7\pi^4). \qquad (28)$$

Then, to satisfy Eq. (25) for $d=0,1,2$, matrices $\mathbf{g}_m^{(\kappa)}$ should be chosen as

$$\mathbf{g}_m^{(0)}=\frac{M_m^{(2)}\Upsilon_2-M_m^{(0)}\Upsilon_4}{\Upsilon_2^2-\Upsilon_0\Upsilon_4}, \qquad (29a)$$

$$\mathbf{g}_m^{(1)}=\frac{M_m^{(1)}}{\Upsilon_2}, \qquad (29b)$$

$$\mathbf{g}_m^{(2)}=\frac{M_m^{(2)}\Upsilon_0-M_m^{(0)}\Upsilon_2}{-\Upsilon_2^2+\Upsilon_0\Upsilon_4}. \qquad (29c)$$

The analytic continuation of $\tilde{\mathbf{G}}^a(E)$ into the upper complex half-plane is simply

$$\tilde{\mathbf{G}}_m^a(z)=\sum_{\kappa=0}^{2}[\mathbf{g}_m^{(\kappa)}]^{\dagger}\times(z-\xi_m)^{\kappa}. \qquad (30)$$

Then Eq. (23) becomes

$$\boldsymbol{\rho}_{\rm neq}=\frac{1}{2i}\sum_{m=1}^{M}(\boldsymbol{\Omega}_m-\boldsymbol{\Omega}_m^{\dagger}), \qquad (31)$$

where

$$\begin{aligned}\boldsymbol{\Omega}_m=&\int_{-\infty}^{+\infty}dE\mathbf{G}^r(E)\cdot\boldsymbol{\Sigma}(E)\cdot\tilde{\mathbf{G}}^a(E)\\&\times[f(\mu_m,T,E)-f(\mu_{m-1},T,E)]. \qquad (32)\end{aligned}$$

The integrand in Eq. (32) is now analytic in the upper-half-plane and can be evaluated through our "pole summation" algorithm discussed in Secs. II A and II B. The integration precision is controlled by varying $\Delta\mu$, although the condition $\Delta\mu>k_BT$ should be satisfied. Otherwise the interval size becomes smaller than the "overlap" between adjacent intervals due to the Fermi smear, and further reduction of $\Delta\mu$ does not lead to the precision improvement.

The algorithm presented in this section is actually more computationally expensive than the usually implemented[37,40,44] real-axis integration to get $\boldsymbol{\rho}_{\rm neq}$ since for every interval one needs to compute the retarded Green function at three different points instead of one, as shown in Fig. 3. Nonetheless, the benefit of this approach is in systematic approximation by exact match of the Green function moments which can evade insufficiently fine integration grid or, most importantly, uncontrolled usage[37,43,44] of the real-axis infinitesimal $\mathbf{H}_{\text{open}}+i\eta$ that leads to serious current nonconservation in long devices beyond molecular electronics scale. For example, a very large system poorly coupled to its contacts may have several sharp peaks within 10 meV interval. None of the adaptive real-axis integration methods[53,54] can properly account for these peaks if the integration step equals 10 meV, while the moments-matching algorithm has capability to capture the contribution from these peaks to the integral.

The current implementation of moments matching technique is not perfect, though it is reasonably fast and precise. Even though our method allowed 1.5 times larger step for the same integration precision, the test runs on large systems have shown that it worked twice as slower than the traditional real-axis integration offset by a small imaginary constant $i\eta$. The slowdown was due to the large number of matrices to be summed and extra operations required on these matrices.

Another systematic problem of the current implementation is in the following: when one matches the moments on just one interval (limited by two sets of vertical points in Fig. 3) and assumes power expansion of the Green function in the vicinity of the interval center, there is a good chance that outside the interval (but in the range where limiting Fermi functions are not small enough) the Green function severely deviates from its true value (e.g., the imaginary part of the diagonal elements becomes positive so that the DOS becomes negative). For that reason one should be cautious about expanding $\mathbf{G}^a(E)$ in power series beyond the first order. Nevertheless, these technical issues do not undermine the basic idea of local expansion of $\mathbf{G}^a(E)$ with analytic functions through moments matching by pole summation. We believe that it is possible to substantially enhance this method by using analytic functions other than x^n and by extending the base for moments matching, e.g., to simultaneously match moments on one, two and three adjacent intervals. This approach would require more "basis functions" and will lead to a set of coupled linear equations, which must be solved for the entire real-axis integration interval simultaneously to obtain coefficients for each local expansion of $\tilde{\mathbf{G}}_m^a(z)$.

III. EXAMPLE: FIRST-PRINCIPLES MODELING OF TOP-GATED GNR-BASED NANOELECTRONIC DEVICES

From the very outset, the discovery of graphene has been intimately connected to attempts to fabricate carbon-based planar FETs.[8] Since FETs produced using micron-size graphene sheets as channels have poor $I_{\text{on}}/I_{\text{off}} \lesssim 10$ ratio, the pursuit of FETs suitable for digital electronics applications has shifted toward fabrication of GNRs with large band gaps[12] $\simeq 0.4$ eV. Their band gap can be engineered by transverse quantum confinement effects in the case of AGNR (where the gap is additionally affected by the increased hopping integral between the p_z orbitals on carbon atoms around the armchair edge caused by slight changes in atomic bonding length in the presence of edge passivating hydrogen[61]) or by staggered sublattice potential arising due to nonzero spin polarization around zigzag edges of ZGNR.[27,61–64]

The very recent experiments[12–16] have demonstrated that all sub-10-nm-wide GNRs are semiconducting. Since band gaps due to edge magnetic ordering in ZGNR are easily destroyed at room temperature,[63] by finite current under nonequilibrium bias voltage conditions,[27] or by impurities and vacancies along the edge,[64] we assume that AGNRs are essential ingredient to introduce sizable band gap in graphene nanodevices operating at room temperature, as confirmed also by recent tunneling spectroscopy.[65]

The fabricated GNRFETs thus far have utilized metallic source and drain electrodes where Schottky barrier (SB) is introduced at the contact between metallic electrode (typically Pd with high work function) and GNR, so that the current is modulated by carrier tunneling probability through SB at contacts. On the other hand, planar structure of graphene is envisaged to make possible all-graphene electronic circuits patterned from either a single graphene plane or multiple planes separated by layers of insulating material.[19]

Any *all-graphene* circuit concept will require both active FETs and passive elements for wiring individual circuit elements. Although ZGNR can be expected to be metallic at room temperature, the wiring based on them is nontrivial issue because only few specific ZGNR patterns have close to ideal conductance and can transmit electron flux without losses.[19] Furthermore, at finite bias voltage ZGNRs can open a band gap if they are mirror symmetric with respect to the midplane between the two zigzag edges.[20]

A. Three-terminal device setup

Our FET-type device setup, based on the combination of ZGNR source and drain metallic electrodes and semiconducting AGNR channel in between them, is shown in Fig. 4 The source and drain have different widths and are modeled as semi-infinite ideal ZGNRs leads. The size of the AGNR band gap is an oscillating function of the ribbon width. The width variation causing AGNR to switch between small and large gaps equals to just a single C-C bond length, which was found to greatly affect the transfer characteristics (i.e., current I vs gate voltage V_{gs} at fixed source-drain bias V_{ds}) in the recent study[25] of several FET concepts with AGNR channels. Because cutting graphene with atomic precision in order to obtain uniform device performance is currently not an option, the variable-width AGNR seems to be the simplest realistic path toward making a short semiconducting fragment. Above the semiconducting "active region" we place a graphene rectangle, which is assumed to have no electrical contact with the ZGNR-AGNR-ZGNR structure below it. This may be achieved by placing boron-nitride insulating layer in between. In fact, recent experiments have fabricated graphene devices with a top gate separated from the graphene layer by an air gap design which does not decrease the mobility of charge carriers under the gate.[67]

We note that the recent analysis[25] (using NEGF for simple p_z-orbital tight-binding model that is self-consistently

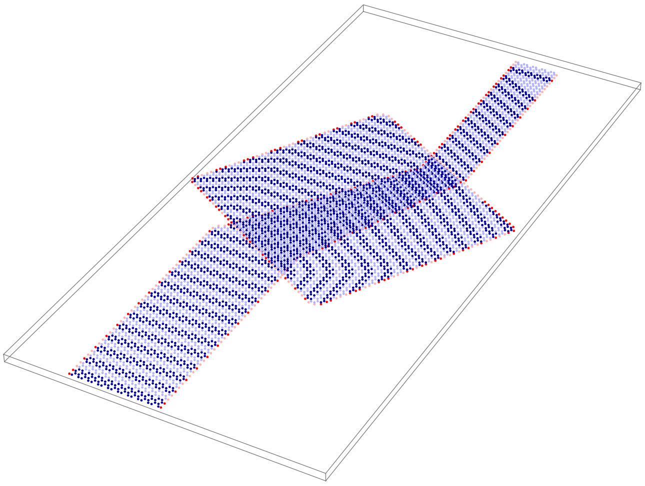

FIG. 4. (Color online) Graphical depiction of the atomic structure of simulated nanodevice composed of two narrow graphene layers. The lower graphene layer contains two unidirectional ZGNRs of different width, which act as the source and drain metallic electrodes, sandwiching semiconducting AGNR of variable width as the FET channel. The top graphene layer plays the role of a gate electrode, covering all of the AGNR channel region, and has the shape of a rectangle that is sufficiently large to have negligible band gap. The interlayer distance is 3.35 Å, which corresponds to the interlayer spacing in graphite (Ref. 66). The hydrogen atoms (red dots) passivate edges of both layers, whose internal carbon atoms (blue) form defect free finite-size honeycomb lattice. Dark and light colored transverse segments, which have variable shape as one moves from the source to the drain electrode, are used to mark odd and even slices of the partitioned system. Each slice $i=1,\ldots,S$ is described by the Hamiltonian matrix $\mathbf{H}_{i,i}$, all of which are stored in computer memory together with matrices $\mathbf{H}_{i,i+1}$ describing the coupling between adjacent slices i and $i+1$.

coupled to a three-dimensional Poisson solver for treating the electrostatics) of dual-gate Schottky barrier GNRFETs, with uniform width AGNR channels and several different types of graphene- or nongraphene-based source and drain electrodes, has singled out ZGNR-AGNR-ZGNR device concept as an optimal one with high enough $I_{\rm on}/I_{\rm off}$ ratio and advantageous features of ZGNR metallic contacts.

The usage of wide graphene sheets as the channel of FET is conceptually difficult because depending on the position of the Fermi level graphene possesses either electron or hole conductivity making it impossible to produce regions depleted of mobile charge carriers. At the same time, the concept of GNR devices allows to build both normally-off and normally-on transistors based solely on the device geometries.[18,19] One of the main benefits of graphene in nanoelectronics is its one-atom-thickness which leads to very low parasitic capacitance, and therefore allows terahertz cutoff frequencies for all-graphene devices and circuits.[2] So far both the experiments[16] and quantum transport simulations[24,25] have been focused on GNRFETs whose channel is long and narrow semiconducting GNR attached to metallic source and drain (such as Pd) contacts while being controlled by metallic top-gate shifting the band gap. Although such transistors play an important role in studying GNR properties, they compromise the main purpose of nanoelectronic devices—the speed. The parasitic gate-substrate or gate-source (drain) capacitances[28,30,31] for such hybrid metal-graphene structure are orders of magnitude higher than capacitance of the channel, and thus substantially decrease the transistor speed. Exploring all-graphene nanoelectronic devices to reach the optimal speed limit is one of the primary motivations for the design concept shown in Fig. 4.

B. System partitioning and the recursive Green function algorithm

The retarded Green function matrix $\mathbf{G}^r(E)$, as the central NEGF quantity in phase-coherent transport regime which yields electron density through Eq. (2) and current via Eq. (4), can be computed by direct matrix inversion in Eq. (3). However, the computational complexity $O(N^3)$ of this operation makes it virtually impossible for present NEGF-DFT codes (which typically perform this brute force operation) to be applied to systems containing large number of atoms.[32] Thus, first-principles simulation of transport in large systems can be accomplished only if relevant elements of $\mathbf{G}^r(E)$ can be obtained via algorithms that scale linearly with increasing length of assumed quasi-one-dimensional (Q1D) device geometry.[32]

In fact, since only a much smaller submatrix of $\mathbf{G}^r(E)$ determines transport properties given by Eq. (4), the recursive Green function algorithms[68] (in serial or parallel implementation[69]) have commonly been used to compute the submatrix $\mathbf{G}^r_{S,1}$ and obtain the transmission properties of mesoscopic devices.[68] They are based on using the Dyson equation, $\mathbf{G}^r_C=\mathbf{G}^r_0+\mathbf{G}^r_0\mathbf{V}\mathbf{G}^r_C$, to build the Green function slice by slice, so that the dimensions of the matrices that have to be inverted are strongly reduced ($\mathbf{G}^r_0$ is the Green function of some region of the device with one of the leads attached, $\mathbf{V}$ is the hopping matrix between that region and adjacent slice, and $\mathbf{G}^r_C$ is the Green function of the coupled system lead +region+slice).

This type of algorithms have also been extended[55,57,70–72] to obtain other submatrices of $\mathbf{G}^r$ needed to compute local quantities within the simulated region, such as $\mathbf{G}^r_{i,i}$ or $\mathbf{G}^r_{i,i+1}$ which define the electron density within slice i or spatial profile of local currents between slices i and $i+1$, respectively. Although it is often considered[56] that standard or extended recursive Green function algorithms can be applied only to Q1D two-terminal devices, some alternative approaches which invert smaller matrices than the full device Hamiltonian $\mathbf{H}_{\rm open}$ to build the Green function of multiterminal nanostructures of arbitrary geometrical shape have also been introduced recently.[56,57]

The key issue for a successful inclusion of the recursive Green function formulas into NEGF-DFT codes is not the specific set of equations, which is very similar in different approaches, but the ability to make a consistent partition of a system of arbitrary shape and with many attached electrodes into slices described by much smaller matrices $\mathbf{H}_{i,i}$. The full Hamiltonian matrix can then be written as

$$\mathbf{H}_{\mathrm{KS}} = \begin{pmatrix} \mathbf{H}_{1,1} & \mathbf{H}_{1,2} & 0 & 0 & \cdots & 0 \\ \mathbf{H}_{1,2}^{\dagger} & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \mathbf{H}_{i-1,i-1} & \mathbf{H}_{i-1,i} & 0 & \cdots & \vdots \\ \vdots & \mathbf{H}_{i-1,i}^{\dagger} & \mathbf{H}_{i,i} & \mathbf{H}_{i,i+1} & \cdots & \vdots \\ \vdots & 0 & \mathbf{H}_{i,i+1}^{\dagger} & \mathbf{H}_{i+1,i+1} & \cdots & \vdots \\ 0 & \cdots & \cdots & \cdots & \ddots & \mathbf{H}_{S-1,S} \\ 0 & 0 & 0 & \cdots & \mathbf{H}_{S-1,S}^{\dagger} & \mathbf{H}_{S,S} \end{pmatrix}. \tag{33}$$

since due to the finite range of basis functions in the transport direction the size of the slices can always be chosen so large that only neighboring ones are coupled through each other via the hopping matrices $\mathbf{H}_{i,i+1}$.

An example of the solution to this primarily *geometrical problem* is illustrated using the device setup in Fig. 4. Our algorithm here starts from the bitmap drawing of the device→converts the image into a finite-size honeycomb lattice→then attempts to partition the device within a loop until consistent set of slices is achieved across the whole device. The final result—a set of slices of nonuniform shape (in contrast to typical columns of sites orthogonal to the axis of the device when recursive algorithm is applied to two-terminal Q1D devices of simple shape[71,72])—is shown in Fig. 4 as dark and light colored segments of the honeycomb lattice. Each slice is described by a matrix $\mathbf{H}_{i,i}$ containing the interactions between atoms within the layer i $(i=1,\ldots,S)$. The size of the matrix $\mathbf{H}_{i,i}$ is $N_i \times N_i$, where N_i is the total number of atomic orbitals for all atoms in the slice i. These matrices are much smaller than $\mathbf{H}$, and are stored in memory at the beginning of the calculation together with matrices $\mathbf{H}_{i,i+1}$.

Starting from the set of matrices $\mathbf{H}_{i,i}$ and $\mathbf{H}_{i,i+1}$, we implement the simplest recursive Green function algorithm aimed at getting $\mathbf{G}_{i,i}^{r}$ from which we can compute the density matrix $\boldsymbol{\rho}_i$ of slice i by replacing $\mathbf{G}^{r}$ in Eq. (2) with $\mathbf{G}_{i,i}^{r}$. The retarded Green function $\mathbf{G}_{i,i}^{r}$ of each slices is given by

$$\mathbf{G}_{i,i}^{r}(E) = [E\mathbf{I}_{i,i} - \mathbf{H}_{i,i} - \boldsymbol{\Sigma}_{L}^{i,i}(E) - \boldsymbol{\Sigma}_{R}^{i,i}(E)]^{-1}. \tag{34}$$

where $\boldsymbol{\Sigma}_{L}^{i,i}(E)$ and $\boldsymbol{\Sigma}_{R}^{i,i}(E)$ are the self-energies due to the rest of the device on the left and on the right, respectively, attached to slice i ($\mathbf{I}_{i,i}$ is the unit matrix of the same size as $\mathbf{H}_{i,i}$).

The self-energies $\boldsymbol{\Sigma}_{L}^{i,i}(E)$ generated by the left side of the device attached to slide i are computed through the recursive formula which starts from the self-energy of the left semi-infinite ideal electrode

$$\boldsymbol{\Sigma}_{L}(E - eU_L) = \mathbf{H}_{0,1}^{\dagger} \cdot \mathbf{g}_{L}^{r}(E - eU_L) \cdot \mathbf{H}_{0,1}, \tag{35}$$

and proceeds through

$$\boldsymbol{\Sigma}_{L}^{1,1}(E) = \mathbf{H}_{1,2}^{\dagger} \cdot [E\mathbf{I}_{1,1} - \mathbf{H}_{1,1} - \boldsymbol{\Sigma}_{L}(E - eU_L)]^{-1} \cdot \mathbf{H}_{1,2}, \tag{36a}$$

$$\boldsymbol{\Sigma}_{L}^{2,2}(E) = \mathbf{H}_{2,3}^{\dagger} \cdot [E\mathbf{I}_{2,2} - \mathbf{H}_{2,2} - \boldsymbol{\Sigma}_{L}^{1,1}(E)]^{-1} \cdot \mathbf{H}_{2,3}, \tag{36b}$$

$$\vdots = \vdots$$

$$\boldsymbol{\Sigma}_{L}^{S-1,S-1}(E) = \mathbf{H}_{S-1,S}^{\dagger} \cdot [E\mathbf{I}_{S-1,S-1} - \mathbf{H}_{S-1,S-1} - \boldsymbol{\Sigma}_{L}^{S-2,S-2}(E)]^{-1} \cdot \mathbf{H}_{S-1,S}. \tag{36c}$$

Here $\mathbf{g}_{L}^{r}(E)$ is portion of the retarded Green function of the isolated lead connecting atoms in the edge principal layer that is coupled to the extended central region via $\mathbf{H}_{0,1}$. We note here that the usual simplification in NEGF-DFT codes is to treat the extended central region out of equilibrium while electronic structure of the ideal semi-infinite leads is computed in equilibrium, thereby ignoring the self-consistent response of the leads to the current. Although it has been pointed out[73] that this approximation can be incompatible with asymptotic charge neutrality, this is rarely taken into account. Instead of assuming that the equilibrium band structure of the leads is rigidly shifted by the bias voltage $\mp eV_{ds}/2$ applied between the macroscopic reservoirs to which they are attached, we employ $\mp eU_{L,R}$ satisfying $eV_{ds}/2 \geq eU_L > eU_R \geq -eV_{ds}/2$ as the shifts of the lead on-site energies, $\boldsymbol{\Sigma}_{L,R}(E,V_{ds}) = \boldsymbol{\Sigma}_{L,R}(E \mp eU_{L,R},0)$. In general nonequilibrium calculations at finite bias voltage, the potential $eU_{L,R}$ is adjusted after each iteration within the self-consistency loop if the total charge on slices 1 and S (obtained from Tr $\boldsymbol{\rho}_1$ and Tr $\boldsymbol{\rho}_S$ respectively) is found to deviate from the neutral state charge.

The same recursion starts from the right semi-infinite ideal electrode to generate the self-energies $\boldsymbol{\Sigma}_{R}^{i,i}(E)$, where the self-energy of the right semi-infinite ideal electrode,

$$\boldsymbol{\Sigma}_{R}(E - eU_R) = \mathbf{H}_{S,S+1} \cdot \mathbf{g}_{R}^{r}(E - eU_R) \cdot \mathbf{H}_{S,S+1}^{\dagger}, \tag{37}$$

and the Hamiltonian $\mathbf{H}_{S,S}$ of the first slice S on the right side of the extended central region are used to construct the starting equation of the recursion analogous to Eq. (36a).

After the self-consistency is reached, the transmission $T(E,V_{ds})$ in Eq. (4) is computed from the submatrix $\mathbf{G}_{S,1}^{r}$ obtained recursively via the Dyson equation by starting from the known retarded Green function $\mathbf{G}_{11}^{r}$ [Eq. (34)] of the first slice on the left,

$$\mathbf{G}_{i,1}^{r} = [E\mathbf{I}_{i,i} - \mathbf{H}_{i,i} - \boldsymbol{\Sigma}_{R}^{i,i}(E)]^{-1} \cdot \mathbf{H}_{i-1,i}^{\dagger} \cdot \mathbf{G}_{i-1,1}^{r}. \tag{38}$$

Thus, the computational complexity of the retarded Green function evaluation is reduced from $O(N^3)$ for the full matrix inversion to $3\bar{N}_i^3(S-1) + \bar{N}_i^3 S$ operations, where $\bar{N}_i$ is the average number of atoms within the slice i. This means that the time required to obtain all relevant submatrices $\mathbf{G}_{i,i}^{r}$ and $\mathbf{G}_{S,1}^{r}$ for the NEGF-DFT algorithm scales linearly $O(S)$ with increasing the length of the device (i.e., the number of slices S).

The recursive Green function algorithm helps to resolve only one of the two key problems in the application of NEGF-DFT to large devices. The other one discussed in Sec. I—numerous sharp peaks in the integrand of $\boldsymbol{\rho}_{\mathrm{neq}}$ that render real-axis integration nonconvergent—can be solved in principle by including the interactions[45,46] within the simulated region capable of washing out the quantum interference effects (that are, anyhow, seldom observed in devices at room temperature). For example, the inclusion of electron-electron correlation effects within the GW approximation was demonstrated[46] to broaden or remove sharp features in the NEGFs for test systems (such as a chain of gold atoms).

In the presence of such dephasing processes, one has to resort to the full NEGF formalism[34] whose core quantities are the retarded $\mathbf{G}^r$ and the lesser $\mathbf{G}^<$ Green function describing the density of available quantum-mechanical states and how electrons occupy those quantum states, respectively. Both Green functions can be obtained from the contour-ordered Green function defined for any two time values that lie along the Kadanoff-Baym-Keldysh time contour.[34] In addition to the retarded $\mathbf{\Sigma}_{\text{leads}}$ and the lesser $\mathbf{\Sigma}^<_{\text{leads}}$ self-energy due to attached electrodes, the full formalism requires to compute self-energy functionals due to many-body interactions within the sample, $\mathbf{\Sigma}_{\text{int}}$ and $\mathbf{\Sigma}^<_{\text{int}}$, while using conserving approximation[45] for their expression in terms of $\mathbf{G}^r$ and $\mathbf{G}^<$.

In the phase-coherent transport regime, $\mathbf{\Sigma}_{\text{int}}=0$ and $\mathbf{\Sigma}^<_{\text{int}}=0$, so that the lesser self-energy of noninteracting (i.e., mean-field or Kohn-Sham) quasiparticles can be expressed solely in terms of the retarded self-energies of the leads

$$\mathbf{\Sigma}^<_{\text{leads}}(E) = if(E-\mu_L)\mathbf{\Gamma}_L(E) + if(E-\mu_R)\mathbf{\Gamma}_R(E). \quad (39)$$

Then the Keldysh equation

$$\mathbf{G}^<(E) = \mathbf{G}^r(E)\cdot[\mathbf{\Sigma}^<_{\text{leads}}(E) + \mathbf{\Sigma}^<_{\text{int}}(E)]\cdot\mathbf{G}^a(E), \quad (40)$$

allows to eliminate $\mathbf{G}^<$ as independent NEGF and express the corresponding density matrix

$$\boldsymbol{\rho} = \frac{1}{2\pi i}\int dE\mathbf{G}^<(E), \quad (41)$$

using only $\mathbf{G}^r(E)$ and $\mathbf{\Sigma}_{\text{leads}}(E)$, as shown explicitly by Eq. (2).

On the other hand, even the simplest phenomenological NEGF models of dephasing, such as "momentum-conserving" choice $\mathbf{\Sigma}_{\text{int}}(E)=d\mathbf{G}^r(E)$ and $\mathbf{\Sigma}^<_{\text{int}}(E)=d\mathbf{G}^<(E)$ (d measures the "dephasing strength") proposed in Ref. 47, require to solve Eqs. (3) and (40) as a system of coupled matrix equations involving full size matrices in the Hilbert space of the simulated device region. For example, in the case of the dephasing model of Ref. 47, this means iterative solving of Eq. (3), with $\mathbf{G}^r_0(E)=[E-\mathbf{H}-\mathbf{\Sigma}^r_{\text{leads}}(E)]^{-1}$ as the initial guess, and then using converged $\mathbf{G}^r(E)$ to solve Eq. (40) as the Sylvester equation of matrix algebra. Obviously, in this case the sparse nature of $\mathbf{H}$ matrix in Eq. (33) and the corresponding recursive Green function formulas become irrelevant for reducing the time it takes to obtain all necessary NEGFs in a single step of the self-consistent loop [Eq. (1)].

More realistic description of interactions with the extended central region is far more computationally demanding.[45,46] Thus, the only route toward first-principles modeling of transport through large devices is to remain within the phase-coherent transport regime and develop algorithms that can resolve problems in the convergence of integration in $\boldsymbol{\rho}_{\text{neq}}$ along the real axis, as discussed in Sec. II C or by Refs. 53 and 54.

C. Quasinonequilibrium model

The DFT part of our simulation, which constructs the Hamiltonian of the central region as an input for NEGF post-processing to obtain the device transport properties, is performed by using the SC-EDTB model.[59,60] This model accounts for atomic polarization and interatomic charge transfer in a standard DFT-like fashion while making it possible to use a *minimal basis set* of four Gaussian orbitals per carbon and one orbital per hydrogen atom. The usage of such minimal basis set allows us to reduce the size of matrices $\mathbf{H}_{i,i}$ and $\mathbf{H}_{i,i+1}$ discussed in Sec. III B without loosing any of the important aspects of *ab initio* input about carbon-hydrogen systems. This makes SC-EDTB highly advantageous when treating systems with large number of atoms.

Conceptually, SC-EDTB can be viewed as the pseudopotential DFT scheme with each atom having its own atomic orbital basis set adjustable to the local atomic environment around this atom. It is a hybrid of the non-self-consistent environment-dependent tight-binding model[74] and a Gaussian-based DFT scheme. Such adaptive behavior adequately compensates for the low precision of the minimal orthogonal basis set. In practice, SC-EDTB implements the environment dependence as the parametrization of Hamiltonian matrix elements with respect to the atomic environment, rather than the parametrization of the atomic basis set. For example, the parametrized part of Hamiltonian matrix elements for the atom near the edge of the nanoribbon will be different from the respective matrix elements in the middle of the strip. Similarly, the in-plane Hamiltonian matrix elements for a single graphene layer will be different from the respective matrix elements in a graphene bilayer.

The SC-EDTB Hamiltonian matrix elements are the sums of parametrized adaptive "TB-like" and nonadaptive "true DFT" contributions. The former mainly accounts for the covalent bonding, while the latter describes interatomic charge transfer, atomic dipole polarization, and on-site variation of exchange potential. The extensive comparison of SC-EDTB with large basis set DFT calculations indicates that SC-EDTB produces more precise and transferable results than minimal basis set pseudopotential DFT schemes. At the same time, SC-EDTB is faster than minimal basis set pseudopotential DFT due to: (i) faster computation of matrix elements; (ii) unit overlap matrix (i.e., orthogonal basis set); and (iii) smaller number of components used for the description of electron density (SC-EDTB uses ten independent components, s^2, sp_x, sp_y, sp_z, p_x^2, p_y^2, p_z^2, p_xp_y, p_xp_z, p_yp_z, to describe the electron density at a given carbon atom). This allows us not only to capture the interatomic charge transfer, but also to account for the dipole polarization.

The compact description of electron density makes possible efficient combination of SC-EDTB with convergence acceleration schemes for both equilibrium and nonequilibrium cases, as discussed in the Appendix. The more detailed specification of electron density provided by standard DFT codes in local density (or some other) approximation[35] will decrease the computation efficiency, but will not affect the simulation of graphene devices whose operation is based on charge transfer at the scale larger than carbon-carbon bond length. To accommodate systems composed of tens of thousands atoms, the SC-EDTB part of our NEGF-DFT computational code also includes the possibility of multipole expansion of Coulomb potential and parallelization on distributed/shared memory systems.

The simulations of the gate voltage effect for the device in Fig. 4, presented in Sec. III D, were performed in the *linear response* regime where the bias voltage is vanishingly small, $\mu_L - \mu_R \to 0$ (entailing $eU_L - eU_R \to 0$). Despite 5 Å cutoff radius for the orbitals used in SC-EDTB, the coupling Hamiltonian matrix elements between the top and the bottom graphene layers of the system depicted in Fig. 4 have to be masked with zeros to mimic the presence of a real insulating layer in between. This causes the nonequilibrium density matrix (2) in the presence of the nonzero gate voltage V_{gs} to evolve into two *equilibrium* integrals [Eq. (6)],

$$\boldsymbol{\rho}^{\text{quasi}}_{\text{neq}}(\mu, V_{gs}, T) = -\frac{1}{\pi}\int_{-\infty}^{+\infty} dE\, \text{Im}[\mathbf{G}^r(E)] f(\mu, T, E) - \frac{1}{\pi}\int_{-\infty}^{+\infty} dE\, \text{Im}[\mathbf{G}^r_{\text{gate}}(E)]\{f(\mu + eV_{gs}, T, E) - f(\mu, T, E)\}. \quad (42)$$

Each of these two integrals is evaluated through our "pole summation" algorithm encoded by the formula (16). Here $\mathbf{G}^r_{\text{gate}}$ refers to the Green function matrix Eq. (3) computed for the whole device, but whose all elements associated with atoms in the lower source-channel-drain layer are masked with zeros. That is, only those matrix elements which correspond to the gate layer are allowed to be nonzero.

We assume that the self-consistency of the recursive Green function algorithm+Broyden mixing scheme (see Appendix) is reached when $\|\mathbf{n}^{\text{out}} - \mathbf{n}^{\text{in}}\| < 10^{-5}$, where the elements of the electron density vector $\mathbf{n}$ are extracted from the diagonal blocks of the corresponding $\overset{\text{out}}{\boldsymbol{\rho}}{}^{\text{quasi}}_{\text{neq}}$ and $\overset{\text{in}}{\boldsymbol{\rho}}{}^{\text{quasi}}_{\text{neq}}$ matrices [as discussed in Sec. III B, only their diagonal blocks are computed from recursively generated submatrices $\mathbf{G}^r_{i,i}(E)$ of the retarded Green function].

D. Results and discussion

We first assume zero gate voltage and plot in Fig. 5 the self-consistent Hartree potential[28] computed via the Poisson equation with net charge density due to charging of carbon atoms as the source term. The potential profiles are evaluated within the planes that are parallel to two graphene layers in Fig. 4 and positioned in the region between them. The inhomogeneous profiles are caused by charge transfer between hydrogen and carbon atoms. Furthermore, it is important to emphasize that there is approximately 100 meV difference between the Fermi levels of the wide μ_{wide} and narrow μ_{narrow} source and drain ZGNR electrodes, respectively, in the bottom graphene layer of the device in Fig. 4. This is caused by different ratios of carbon atoms to hydrogen atoms passivating the zigzag edges in GNRs of different widths. That is, the edge hydrogen atoms effectively dope the nanoribbon[20–22] where the level of doping depends on its size and geometry. To account for this, the equilibrium Fermi level of the whole setup $\mu = (\mu_{\text{wide}} + \mu_{\text{narrow}})/2$ used in Eq. (42) is assumed to be the average of μ_{wide} and narrow μ_{narrow}. Such compensation of the difference in the Fermi levels requires a small built-in electric field in our model. Room-temperature (T=300 K) operation is assumed in all figures in this section.

Then we apply voltage eV_{gs}=1 eV to the gate electrode in Fig. 6 and plot the full three-dimensional spatial profile of the electric potential. Further increase in the gate voltage to eV_{gs}=3 eV leads to potential (within a geometrical plane in between two graphene layers) shown in Fig. 7. The self-consistent atomistic level simulation captures the potential variation in the transverse direction of the GNRs, as well as possible modifications of the band structure of GNRs with increasing gate voltage.[28,30,31]

In both figures, we find that the chosen portion of metallic ZGNR electrodes attached to the AGNR channel to form the "extended central region,[37,40,44]" encompassing ≃7000 carbon and hydrogen atoms for self-consistent electron density and potential calculations, is actually not large enough (despite many ZGNR supercells included into the extended central region) to completely screen the effect of the applied electric field via the top gate electrode. This is signified by the color of the Coulomb potential at the boundaries (marked by horizontal white lines in Fig. 7) of the "extended central region" not being identical to the color of the uniform potential along the semi-infinite leads. The total uncompensated charge at the boundary is approximately 0.03 e for eV_{gs} =1 eV and 0.07 e for eV_{gs}=3 eV.

Another feature conspicuous in Fig. 7 is that the on-site potential shift experienced by carbon atoms in the lower layer is much smaller than expected from the applied bias voltage. This unusual screening capability of the insulating AGNR channel can be attributed to the presence of short segments of metallic AGNR due to either particular width of such segments (we do not relax the coordinates and edge bonds that would ensure that all three classes of AGNRs, defined by their width, are insulating[61]) or doping by evanescent modes[75] that decay from ZGNR electrodes into AGNR channel thereby generating metal induced gap states[76] (localized at the ZGNR|AGNR interface).[25] This is also reflected in the conductance of our device—to shift the band gap of variable-width AGNR by 0.5 eV and bring it into single channel conducting regime demands a rater large gate voltage $eV_{gs} \simeq 3$ eV (when compared to $eV_{gs} \simeq$ half-the-band-gap required to turn uniform semiconducting AGNR into a single channel conductor[28]), as shown by the source-drain conductance computed as the function of V_{gs} in Figs. 8(b)–8(d).

The metallic behavior of ZGNR electrodes is characterized by the nonzero density of states and finite (zero temperature) conductance at the Fermi level E_F. We note that in simple nearest-neighbor tight-binding models[18] the conductance of infinite ZGNR around the charge neutral (Dirac) point E_F=0 is quantized $G = G_Q$ ($G_Q = 2e^2/h$ is the conductance quantum for spin-degenerate transport) due to a single open conducting channel (i.e., transverse propagating mode) defined by the overlap of edge-localized wave functions.[3,17] On the other hand, in DFT description (that can be mimicked by single p_z-orbital tight-binding models which include third nearest-neighbor hopping[17]) more complicated subband

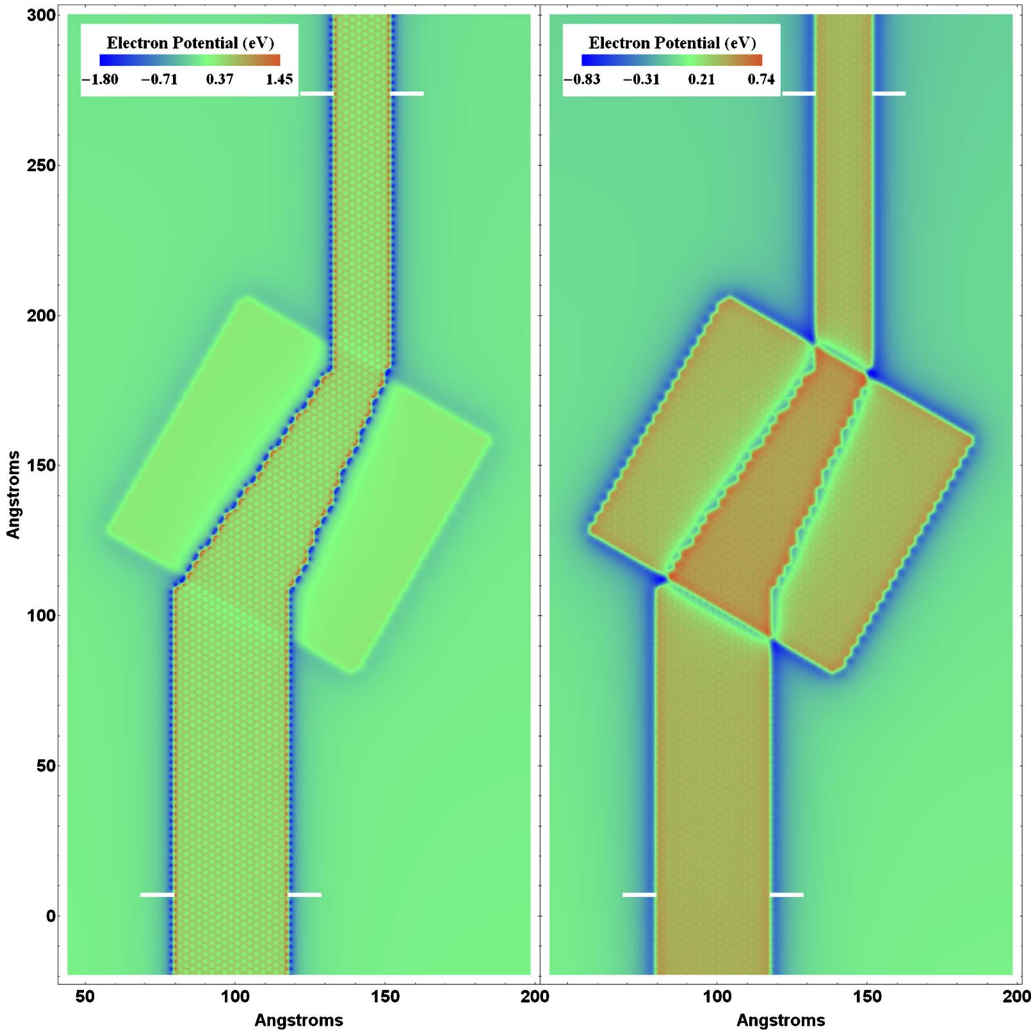

FIG. 5. (Color online) Contour plot of the Hartree potential for zero applied gate voltage ($V_{gs}=0$ V) in the planes which are 0.7 Å (left panel) and 0.5×3.35 Å (right panel) above the lower graphene layer of the system depicted in Fig. 4. White horizontal lines in the ZGNR electrode regions mark the boundaries of the extended central region "AGNR channel+portion of ZGNR electrodes" composed of $\simeq7000$ atoms (for which the retarded Green function is evaluated to obtain electron density and electric potential through the self-consistent loop).

structure of ZGNR leads to three open conducting channels[17] around $E_F=0$ and $G=3G_Q$ quantized conductance for semi-infinite source and drain ZGNR electrodes. This is confirmed in the context of our NEGF-DFT approach by Fig. 8(a).

Comparing Fig. 8(a) with Fig. 8(b), which are both obtained at $V_{gs}=0$ V, highlights the importance of self-consistent electron density computation, even in the absence of gate voltage effects. We find a marked difference in two panels between the position of the gap region [over which the transmission function $T(E,0)$ in Eq. (5) is zero] and conductance oscillations outside of it. The conductance in Fig. 8(a) was obtained without computing charge transfer effects, and it could be reproduced by popular non-self-consistent tight-binding models[17,18] without resorting to full NEGF-DFT formalism. The local charge transfer is due to the polarization of C-H bonds and slight system-wide charge redistribution is due to the different carbon to hydrogen ratios in different portions of the system. Both effects induce the change in position of the Fermi level with respect to the band gap and cannot be neglected when computing the transport properties of realistic nanodevices.

IV. CONCLUDING REMARKS

The modeling of realistic multiterminal graphene nanoelectronic devices requires quantum transport methods that can capture effects of its highly unusual electronic properties[3,17] and their dependence on detailed device geometry,[18,19] as well as charge transfer (in equilibrium) and charge redistribution (out of equilibrium) effects on atomistic scale. While quantum transport approaches based on simple predefined Hamiltonians[18] cannot handle all of these issues, the NEGF-DFT framework, which generates the self-consistent Hamiltonian of the device prior to the calculation of conductance or I-V characteristics, offers a proper methodology for first-principles modeling of electron transport involving accurate quantum-chemical description of atomic scale geometry.

However, NEGF-DFT simulations thus far have been limited[32] to rather small systems, such as short molecules connected to metallic electrodes. Here we address several obvious[32] and more subtle (Sec. I) impediments that have to be resolved to make possible the application of NEGF-DFT

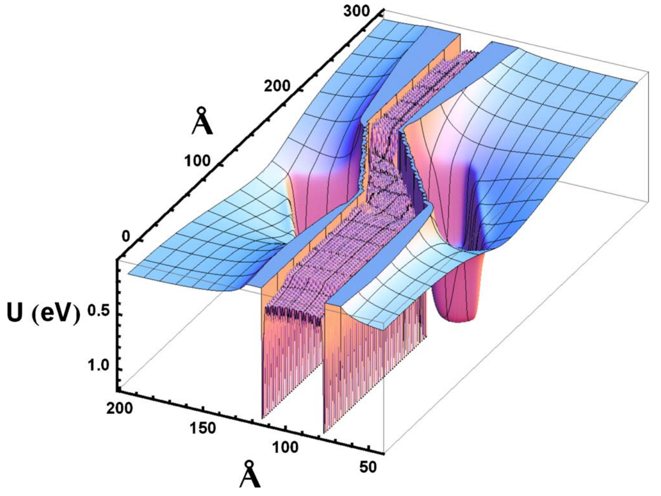

FIG. 6. (Color online) Contour plot of the Hartree potential in the plane 0.2 Å above the lower graphene layer when the applied gate voltage is $eV_{gs}=1$ eV. The semiconducting region is shifted by approximately 0.35 eV. The potential spikes pointing downward correspond to the hydrogen atoms. Positive potential spikes associated with carbon atoms in the C-H dipole pairs are truncated to make a clear view of the potential inside the conducting channel. Note that the potential axis points downward.

codes to devices containing many thousand atoms: (i) computational complexity of the retarded Green function calculation, as the main time limiting part of the simulation when full Hamiltonian matrix is inverted, should scale linearly with the system size; (ii) integration of NEGFs to get the equilibrium and nonequilibrium part of the density matrix has to be performed in a way (especially in the case of nonequilibrium contribution) which ensures convergence despite sharp peaks (due to assumed phase-coherent transport of noninteracting quasiparticles) along the real axis whose number increases substantially in large systems; and (iii) the convergence of the self-consistent loop, which repeatedly evaluates (i) and (ii), should be accelerated with proper mixing scheme of previous iterative steps that is compatible with solution of problems in (i) and (ii).

The algorithms presented here extend the NEGF-DFT methodology to systems containing large number of atoms through a combination of

(1) the "pole summation" algorithm for the exact integration of the retarded Green function in the expression for the equilibrium part of the density matrix offers an alternative to standard numerical contour integration by replacing the Fermi function $f(E)$ with the analytic function $\widetilde{f}(E)$, which coincides with $f(E)$ inside the integration range along the real axis but decays exponentially in the upper complex half-plane. Only a finite number N_{pole} of its poles, which can be found analytically, has non-negligible residues, so that $\boldsymbol{\rho}_{\mathrm{eq}}=\mathrm{Im}\Sigma_{j=1}^{N_{\mathrm{pole}}}\alpha_j\mathbf{G}^r(Z_j)$ where α_j are scalars given by simple analytical expressions in Eq. (16). The typical value of N_{pole} for valence electrons at room temperature is 80, and it increases with the temperature decrease with an approximate rate of 40 extra poles per order of magnitude in temperature reduction.

(2) Possible application of the "pole summation" algorithm to tackle the problem of difficult-to-converge integration of NEGFs along the real-axis (due to numerous sharp peaks in the integrand which would be impossible to locate and handle individually[53,54] for devices contains large number of atoms) to obtain $\boldsymbol{\rho}_{\mathrm{neq}}$ after its nonanalytic integrand in the entire complex plane is approximated with an analytic function in the upper complex plane, so that the same type of summation can be performed as in the case of $\boldsymbol{\rho}_{\mathrm{eq}}$ integral.

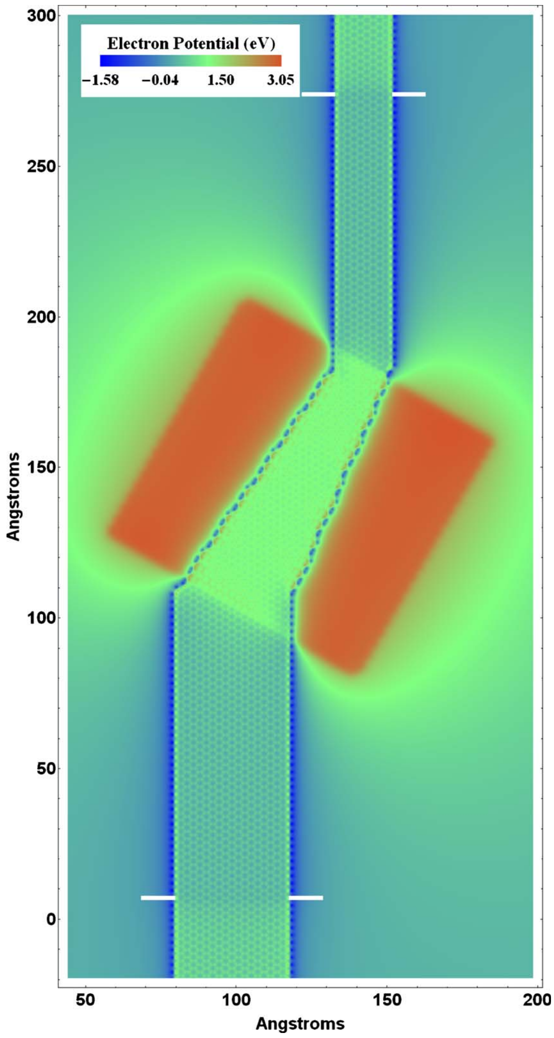

FIG. 7. (Color online) Contour plot of the Hartree potential for the applied gate voltage $eV_{gs}=3$ eV in the plane 0.7 Å above the lower graphene layer. White horizontal lines around the ZGNR electrodes mark the boundaries of the extended central region "AGNR channel+portion of ZGNR electrodes" composed of ≃7000 atoms.

(3) The recursive Green function formulas which, assuming proper geometrical decomposition of the lattice of the device into slices of irregular shape for arbitrary nanostructure geometry, makes it possible to reduce scaling of the required computing time from $O(N^3)$ for the full Hamiltonian matrix inversion in the single iteration of the self-consistent loop to linear scaling $O(S)$ [S is the number of slices in the transport direction] of the computation of only the diagonal blocks of the retarded Green function that yield the electron density within the slice.

In the case of equilibrium or quasiequilibrium (such as generated by nonzero gate voltage and zero or linear re-

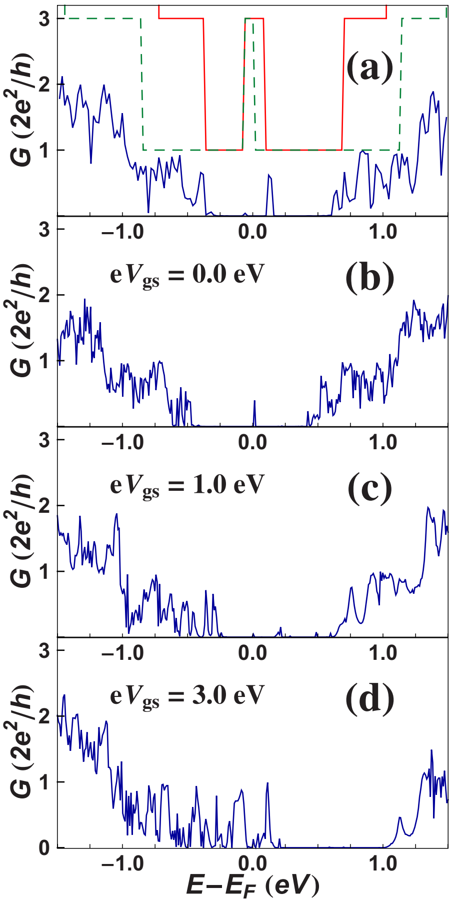

FIG. 8. (Color online) The non-self-consistent (a) and self-consistent (b)–(d) source-drain conductance (at linear response bias voltage V_{ds}) of the nanodevice depicted in Fig. 4 as a function of energy. The conductances are obtained in the absence (a), (b) or presence (c), (d) of the gate voltage V_{gs}, where charge redistribution is computed self-consistently in all three cases (b)–(d) [unlike in (a)]. The solid and dashed rectangular lines in panel (a) show the conductance quantization of the infinite source (wide nanoribbon, solid line) and drain (narrow nanoribbon, dashed line) ZGNR electrodes, respectively. The equilibrium Fermi level in the case of unbiased gate corresponds to $E-E_F=0$. The Fermi level of the source and drain electrodes in panels (a)–(d) corresponds to $E-E_F=0$.

sponse bias voltage) situations, we additionally accelerate convergence of the self-consistent loop for the density matrix by using the modified Broyden scheme discussed in Appendix, which is compatible with the recursive Green function algorithm and mixes input and output electron density from all previous iterations to generate input density for the next iteration step.

We illustrate the numerical efficiency of the combination of these algorithms for NEGF part of the calculation by integrating it with the DFT code (based on the minimal basis set—four localized orbitals per carbon atom and one per hydrogen—tailored for carbon-hydrogen systems) to simulate gate voltage effects in all-graphene FET-type device. Our simulated ZGNR|variable-width-AGNR|ZGNR device is composed of ≃7000 atoms and employs AGNR of variable width (kept below 10 nm) as a realistic semiconductor channel accessible to present nanofabrication technology.[12–15] The device does not require atomic precision in controlling the width and the corresponding band gap when uniform sub-10-nm wide AGNR are used, while exploiting advantageous[25] ZGNR source and drain electrodes. We also use square-shaped gate electrode covering the channel which is made of graphene as well. The self-consistent evaluation of the electron density and Coulomb potential is required to capture inhomogeneous charge distribution and modification of the GNR band structure with increasing gate voltage.[28,30,31] This reveals that rather large gate voltage is required to shift the band gap of variable-width AGNR channel and bring this type of top-gated GNRFET into a window of single open transverse propagating mode with low scattering and heat dissipation.

The computation of self-consistent electron density and electrostatic potential, as the crucial aspect of NEGF-DFT approach to quantum transport modeling, is indispensable to properly take into account gate voltage effects or to ensure the gauge invariance[26] of the *I*-*V* characteristics in far from equilibrium transport.[27] In addition, we also demonstrate notable difference between the zero-bias transmission (i.e., linear response conductance) of non-self-consistent and self-consistent modeling. This can be attributed to charge transfer effects between edge passivating hydrogen atoms and carbon atoms, where such edge doping also affects the position of the Fermi level of isolated GNRs of different size and geometry.

ACKNOWLEDGMENTS

Financial support from NSF under Grant No. ECCS 0725566 is gratefully acknowledged.

APPENDIX: BROYDEN MIXING SCHEME FOR CONVERGENCE ACCELERATION OF THE SELF-CONSISTENT LOOP

The recursive Green function algorithm discussed in Sec. III B drastically reduces the computational complexity of a single iteration step within the self-consistent loop [Eq. (1)]. Another important ingredient of algorithms that can handle systems with large number of atoms is to combine the recursive techniques with the convergence acceleration scheme based on proper mixing of quantities found in previous steps to produce the input for the next step.

The simplest mixing scheme takes certain fraction ε of the output electron density $\mathbf{n}_m^{\text{out}}$ from the previous step m and the remaining fraction $(1-\varepsilon)$ from the corresponding input $\mathbf{n}_m^{\text{in}}$ to produce input for the next step, $\mathbf{n}_{m+1}^{\text{in}}=(1-\varepsilon)\mathbf{n}_m^{\text{in}}+\varepsilon\mathbf{n}_m^{\text{out}}$. Finding the optimal value for the mixing parameter, typically $\varepsilon \sim 0.1-0.01$, depends on the nature of the system (such as, insulating vs metallic or isolated vs attached to semi-infinite leads). This can require few thousand iteration steps to satisfy the convergence criterion $\|\mathbf{n}_m^{\text{out}}-\mathbf{n}_m^{\text{in}}\|<10^{-5}$ we employ in our simulation.

The more sophisticated mixing schemes employ Pulay[45] or Broyden[77–79] algorithms to mix several previous steps, where the quantities mixed can be the density matrix or Hamiltonian and Green functions[45] (which can be more efficient for open multiterminal systems where the central region does not have a fixed number of electrons). For a small bias voltage, the self-consistency can be achieved by applying the Broyden convergence acceleration method which has two major advantages. First, the modified second Broyden method[78,79] is compatible with the recursive Green function method discussed in Sec. III B. Second, the Broyden method adds $O(N)$ extra operations, so that the single iteration is not slowed down. However, the reduction of the number of iterations achieved by the Broyden method is appreciable.

The Broyden method works well when the correlation between the electron density and the potential is local, i.e., when the local potential distortion results in a local self-consistent density change. On the other hand, in the case of nonlocal correlations the Broyden method performance rapidly deteriorates. The nonequilibrium electron density in the coherent ballistic approximation constitutes the perfect example when the Broyden method fails. The reason for this is that electron-potential correlations becomes completely nonlocal—the change of the potential at one contact can shut off the electron flux through the entire system and cause the system-wide electron density redistribution. Thus, in far-from-equilibrium cases other mixing schemes have to be used.[27,37]

In particular, the modified second Broyden method[78,79] is compatible with the recursive Green function method discussed in Sec. III B, and makes it possible to reduce the number of iteration steps to the order of ∼10. In this scheme, an input electron density for iteration $m+1$ is constructed from the set of input and output densities generated in *all* previous iterations.

$$\mathbf{n}_{m+1}^{\text{in}} = \mathbf{n}_m^{\text{in}} - \varepsilon \mathbf{F}_m - \sum_{j=2}^{m} \mathbf{W}_j \cdot [\mathbf{\Phi}_j]^T \cdot \mathbf{F}_m, \tag{A1a}$$

$$\mathbf{F}_m = \mathbf{n}_m^{\text{out}} - \mathbf{n}_m^{\text{in}}, \tag{A1b}$$

$$\mathbf{W}_i = -\varepsilon(\mathbf{F}_i - \mathbf{F}_{i-1}) + \mathbf{n}_i^{\text{in}} - \mathbf{n}_{i-1}^{\text{in}} - \sum_{j=2}^{i-1} \mathbf{W}_j \cdot [\mathbf{\Phi}_j]^T \cdot (\mathbf{F}_i - \mathbf{F}_{i-1}), \tag{A1c}$$

$$[\mathbf{\Phi}_i]^T = \frac{(\mathbf{F}_i - \mathbf{F}_{i-1})^T}{(\mathbf{F}_i - \mathbf{F}_{i-1})^T \cdot (\mathbf{F}_i - \mathbf{F}_{i-1})}. \tag{A1d}$$

Here $\mathbf{n}_m^{\text{in}}$, $\mathbf{n}_m^{\text{out}}$, $\mathbf{F}_m$, $\mathbf{W}_j$, and $\mathbf{\Phi}_j$ comprise a relatively small set of vectors to be stored in computer memory. The compatibility of this modified Broyden scheme with the recursive Green function algorithm of Sec. III B stems from the fact that only diagonal blocks of $\mathbf{G}^r$, required to construct vectors in Eq. (A1), are computed recursively without knowing the full Green function needed in some other mixing schemes.[27,45]

[1] A. K. Geim and K. S. Novoselov, Nature Mater. **6**, 183 (2007).

[2] A. K. Geim, Science **324**, 1530 (2009).

[3] A. H. C. Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, Rev. Mod. Phys. **81**, 109 (2009).

[4] P. Avouris, Phys. Today **62**(1), 34 (2009).

[5] M. Burghard, H. Klauk, and K. Kern, Adv. Mater. **21**, 1 (2009).

[6] R. W. Keyes, Rep. Prog. Phys. **68**, 2701 (2005).

[7] P. Avouris, Z. Chen, and V. Perebeinos, Nat. Nanotechnol. **2**, 605 (2007).

[8] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science **306**, 666 (2004).

[9] J.-H. Chen, C. Jang, S. Xiao, M. Ishigami, and M. S. Fuhrer, Nat. Nanotechnol. **3**, 206 (2008).

[10] I. Meric, M. Y. Han, A. F. Young, B. Ozyilmaz, P. Kim, and K. L. Shepard, Nat. Nanotechnol. **3**, 654 (2008).

[11] Y.-M. Lin, K. A. Jenkins, A. Valdes-Garcia, J. P. Small, D. B. Farmer, and P. Avouris, Nano Lett. **9**, 422 (2009).

[12] X. Li, X. Wang, L. Zhang, S. Lee, and H. Dai, Science **319**, 1229 (2008).

[13] L. Tapasztó, G. Dobrik, P. Lambin, and L. P. Biró, Nat. Nanotechnol. **3**, 397 (2008).

[14] L. Jiao, L. Zhang, X. Wang, G. Diankov, and H. Dai, Nature (London) **458**, 877 (2009).

[15] D. V. Kosynkin, A. L. Higginbotham, A. Sinitskii, J. R. Lomeda, A. Dimiev, B. K. Price, and J. M. Tour, Nature (London) **458**, 872 (2009).

[16] X. R. Wang, Y. J. Ouyang, X. L. Li, H. L. Wang, J. Guo, and H. J. Dai, Phys. Rev. Lett. **100**, 206803 (2008).

[17] A. Cresti, N. Nemec, B. Biel, G. Niebler, F. Triozon, G. Cuniberti, and S. Roche, Nano Res. **1**, 361 (2008).

[18] A. Rycerz, J. Tworzydło, and C. W. J. Beenakker, Nat. Phys. **3**, 172 (2007).

[19] D. Areshkin and C. White, Nano Lett. **7**, 3253 (2007).

[20] Z. Li, H. Qian, J. Wu, B.-L. Gu, and W. Duan, Phys. Rev. Lett. **100**, 206802 (2008).

[21] S. Dutta and S. K. Pati, J. Phys. Chem. B **112**, 1333 (2008).

[22] B. Biel, F. Triozon, X. Blase, and S. Roche, Nano Lett. **9**, 2725 (2009).

[23] G. Lee and K. Cho, Phys. Rev. B **79**, 165440 (2009).

[24] Y. Ouyang, Y. Yoon, and J. Guo, IEEE Trans. Electron Devices **54**, 2223 (2007).

[25] G. Liang, N. Neophytou, M. S. Lundstrom, and D. E. Nikonov, Nano Lett. **8**, 1819 (2008).

[26] T. Christen and M. Büttiker, Europhys. Lett. **35**, 523 (1996).

[27] D. A. Areshkin and B. K. Nikolić, Phys. Rev. B **79**, 205430 (2009).

[28] J. Fernández-Rossier, J. J. Palacios, and L. Brey, Phys. Rev. B **75**, 205441 (2007).

[29] P. G. Silvestrov and K. B. Efetov, Phys. Rev. B **77**, 155436 (2008).

[30] A. Shylau, J. Kłos, and I. Zozoulenko, Phys. Rev. B **80**, 205402 (2009).

[31] J. Guo, Y. Yoon, and Y. Ouyang, Nano Lett. **7**, 1935 (2007).

[32] K. Stokbro, J. Phys.: Condens. Matter **20**, 064216 (2008).

[33] M. Koentopp, C. Chang, K. Burke, and R. Car, J. Phys.: Con-

dens. Matter **20**, 083203 (2008).

[34] H. Haug and A.-P. Jauho, *Quantum Kinetics in Transport and Optics of Semiconductors*, 2nd ed. (Springer, Berlin, 2007).

[35] *A Primer in Density Functional Theory*, edited by C. Fiolhais, F. Nogueira, and M. Marques (Springer, Berlin, 2003).

[36] J. Taylor, H. Guo, and J. Wang, Phys. Rev. B **63**, 245407 (2001).

[37] M. Brandbyge, J.-L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, Phys. Rev. B **65**, 165401 (2002).

[38] Y. Xue, S. Datta, and M. A. Ratner, Chem. Phys. **281**, 151 (2002).

[39] J. J. Palacios, A. J. Pérez-Jiménez, E. Louis, E. San Fabián, and J. A. Vergés, Phys. Rev. B **66**, 035322 (2002).

[40] S.-H. Ke, H. U. Baranger, and W. Yang, Phys. Rev. B **70**, 085410 (2004).

[41] A. Pecchia and A. Di Carlo, Rep. Prog. Phys. **67**, 1497 (2004).

[42] F. Evers, F. Weigend, and M. Koentopp, Phys. Rev. B **69**, 235411 (2004).

[43] S. V. Faleev, F. Léonard, D. A. Stewart, and M. van Schilfgaarde, Phys. Rev. B **71**, 195422 (2005).

[44] A. R. Rocha, V. M. García-Suárez, S. Bailey, C. Lambert, J. Ferrer, and S. Sanvito, Phys. Rev. B **73**, 085414 (2006).

[45] K. S. Thygesen and A. Rubio, Phys. Rev. B **77**, 115333 (2008).

[46] P. Darancet, A. Ferretti, D. Mayou, and V. Olevano, Phys. Rev. B **75**, 075102 (2007).

[47] R. Golizadeh-Mojarad and S. Datta, Phys. Rev. B **75**, 081301(R) (2007).

[48] S. Mertens, Comput. Sci. Eng. **4**, 31 (2002).

[49] The size of relevant matrices $\mathbf{H}_{\mathrm{KS}}[n(\mathbf{r})]$ and $\mathbf{G}^r$ is actually $\Sigma_{s=1}^{N_{\mathrm{species}}} N_s \times N_{\mathrm{orbitals}}$ where N_{orbitals} is the number of localized valence electron orbitals per atom type s.

[50] Hans Henrik B. Sørensen, P. C. Hansen, D. E. Petersen, S. Skelboe, and K. Stokbro, Phys. Rev. B **79**, 205322 (2009).

[51] S.-H. Ke, H. U. Baranger, and W. Yang, Phys. Rev. B **71**, 113401 (2005).

[52] H. He, R. Pandey, and S. P. Karna, Nanotechnology **19**, 505203 (2008).

[53] R. Li, J. Zhang, S. Hou, Z. Qian, Z. Shen, X. Zhao, and Z. Xue, Chem. Phys. **336**, 127 (2007).

[54] H. J. Choi, M. L. Cohen, and S. G. Louie, Phys. Rev. B **76**, 155420 (2007).

[55] A. Pecchia, G. Penazzi, L. Salvucci, and A. Di Carlo, New J. Phys. **10**, 065022 (2008).

[56] K. Kazymyrenko and X. Waintal, Phys. Rev. B **77**, 115119 (2008).

[57] M. Wimmer and K. Richter, J. Comput. Phys. **228**, 8548 (2009).

[58] E. Polizzi, Phys. Rev. B **79**, 115112 (2009).

[59] D. A. Areshkin, O. A. Shenderova, J. D. Schall, S. P. Adiga, and D. W. Brenner, J. Phys.: Condens. Matter **16**, 6851 (2004).

[60] D. A. Areshkin, O. A. Shenderova, J. D. Schall, and D. W. Brenner, Mol. Simul. **31**, 585 (2005).

[61] Y.-W. Son, M. L. Cohen, and S. G. Louie, Phys. Rev. Lett. **97**, 216803 (2006).

[62] L. Pisani, J. A. Chan, B. Montanari, and N. M. Harrison, Phys. Rev. B **75**, 064418 (2007).

[63] O. V. Yazyev and M. I. Katsnelson, Phys. Rev. Lett. **100**, 047209 (2008).

[64] B. Huang, F. Liu, J. Wu, B.-L. Gu, and W. Duan, Phys. Rev. B **77**, 153411 (2008).

[65] K. A. Ritter and J. Lyding, Nature Mater. **8**, 235 (2009).

[66] L. A. Girifalco and R. A. Lad, J. Chem. Phys. **25**, 693 (1956).

[67] R. V. Gorbachev, A. S. Mayorov, A. K. Savchenko, D. W. Horsell, and F. Guinea, Nano Lett. **8**, 1995 (2008).

[68] D. K. Ferry, S. M. Goodnick, and J. Bird, *Transport in Nanostructures*, 2nd ed. (Cambridge University Press, Cambridge, England, 2009).

[69] P. Drouvelis, P. Schmelcher, and P. Bastian, J. Comput. Phys. **215**, 741 (2006).

[70] A. Cresti, R. Farchioni, G. Grosso, and G. P. Parravicini, Phys. Rev. B **68**, 075306 (2003).

[71] G. Metalidis and P. Bruno, Phys. Rev. B **72**, 235304 (2005).

[72] A. Lassl, P. Schlagheck, and K. Richter, Phys. Rev. B **75**, 045346 (2007).

[73] H. Mera, P. Bokes, and R. W. Godby, Phys. Rev. B **72**, 085311 (2005).

[74] M. S. Tang, C. Z. Wang, C. T. Chan, and K. M. Ho, Phys. Rev. B **53**, 979 (1996).

[75] P. Pomorski, C. Roland, and H. Guo, Phys. Rev. B **70**, 115408 (2004).

[76] Although metal induced gap states do not affect the transmission for long channel devices at linear response bias voltage, they are expected to contribute to tunneling currents, particularly in short channel devices. In fact, they can also affect longer channel devices by enhancing scattering processes under high source-drain bias voltage.

[77] K. Ohno, K. Esfarjani, and Y. Kawazoe, *Computational Materials Science: From Ab Initio to Monte Carlo Methods* (Springer, Berlin, 2000).

[78] D. Singh, H. Krakauer, and C. S. Wang, Phys. Rev. B **34**, 8391 (1986).

[79] S. Ihnatsenka, I. V. Zozoulenko, and M. Willander, Phys. Rev. B **75**, 235307 (2007).