Journal of Computational Electronics
https://doi.org/10.1007/s10825-023-02055-3

Full charge incorporation in ab initio simulations of two-dimensional semiconductor-based devices

Rutger Duflou[1,2] · **Michel Houssa**[1,2] · **Aryan Afzalian**[1]

Received: 3 March 2023 / Accepted: 11 May 2023

Abstract
Quantum transport simulations based on the non-equilibrium Green's function formalism require accurate integration of the charges in the system. We demonstrate our implementation of a full charge integration scheme, which automatically incorporates electronic screening effects and is predicted to incorporate interface charges more correctly than the simpler excess charge approach. We first show that under certain conditions the two approaches are equivalent, e.g., for single doping type purely semiconducting devices. We then demonstrate that for devices containing metals, the two approaches may sometimes demonstrate significantly different behavior.

Keywords Semiconductor device modelling · NEGF · 2D materials · Contact modelling

1 Introduction

Two-dimensional (2D) materials are interesting candidates for next-generation ultra-scaled devices. They are predicted to provide excellent electrostatic control, which reduces short channel effects [1, 2], and to be more resilient against device variation due to surface defects and roughness [3, 4]. ATOMOS [5], our quantum transport solver combining density functional theory (DFT) and the non-equilibrium Green's function formalism (NEGF), has been shown to be effective at screening material choices and device designs for devices based on 2D materials.

One of the major limitations of 2D materials is the difficulty in achieving high performing metallic contacts [1]. Top-contacted metals often result in high contact resistances due the emergence of a van der Waals gap related to no out-of-plane dangling bonds in the 2D material, while side-contacted metals suffer from limited contact area and difficulties in device fabrication [6]. To not only screen the intrinsic properties of 2D materials but also their interactions with metals, accurate simulations of metal-2D material interfaces are highly desirable.

In previous work, ATOMOS relied on an excess charge approach (ECA), which only assigns charge to electrons (holes) in the conduction (valence) band. It is presently unclear whether this approach can accurately simulate the behavior at metallic interfaces, where there is no clear conduction band or valence band. Previous work [7] has shown that a full charge approach (FCA), where charge is assigned to all electronic states, can be important for the correct simulation of nanodevices based on conventional 3D materials, especially at certain interfaces.

In Sect. 2, we recapitulate the NEGF formalism and discuss the two models, i.e., the ECA and FCA. In Sect. 3, we demonstrate some of the intricacies in employing the FCA. This is done using a WS_2 metal-oxide-semiconductor field-effect transistor (MOSFET) as a test case, as the two approaches are expected to be largely equivalent for problems without interfaces. In Sect. 4, we employ a 2D material $HfTe_2$ contacted HfS_2 transistor to evaluate the pros and cons of both models for the accurate simulation of semiconducting metallic interfaces.

Funded by the FWO as part of the PhD fellowship 1100321N.

✉ Rutger Duflou
rutger.duflou@imec.be

Michel Houssa
michel.houssa@kuleuven.be

Aryan Afzalian
aryan.afzalian@imec.be

[1] IMEC, Kapeldreef 75, 3001 Leuven, Belgium

[2] Semiconductor Physics Laboratory, KU Leuven, Celestijnenlaan 200 D, 3001 Leuven, Belgium

Published online: 08 June 2023

2 Excess charge versus full charge approach

2.1 The non-equilibrium Green's function formalism

The NEGF formalism is a quantum transport formalism which can naturally incorporate phenomena such as contacts at different potentials and scattering effects. Using a non-orthogonal basis set, the transport characteristics can be calculated from an effective single particle device Hamiltonian matrix H and overlap matrix S, usually obtained from DFT,

$$G = ((E + i\eta)S - H - \Sigma)^{-1} \tag{1}$$

$$G^{\dagger} = ((E - i\eta)S - H - \Sigma^{\dagger})^{-1} \tag{2}$$

$$G^{\lessgtr} = G\Sigma^{\lessgtr}G^{\dagger}. \tag{3}$$

Here, G, $G^{\dagger}$, $G^{<}$ and $G^{>}$ are the retarded, advanced, lesser and greater Green's function, respectively [8]. The diagonal elements of the latter two are related to the existence of filled and empty states at energy E.

Four types of self-energy Σ^{*} are used to incorporate the interaction with contacts and phonon scattering. Consider a slabbed description of the device, i.e., the device is divided into slabs which only interact with their directly neighbouring slabs. The Hamiltonian and overlap matrix can then be divided into blocks with indices corresponding to the slabs, $H_{k,l}$ and $S_{k,l}$. Hence, $H_{k,l} = S_{k,l} = 0$ if $|k - l| > 1$. The slab indices $k = 1, ..., N$ correspond to the actual device and indices $k = 0, -1, ...$ $(k = N + 1, N + 2, ...)$ correspond to a left (right) contact in equilibrium, described by the Fermi-Dirac statistics function, $f_1 = \left(1 + \exp(E_{f_1}/k_BT)\right)^{-1}$ $\left(f_2 = \left(1 + \exp(E_{f_2}/k_BT)\right)^{-1}\right)$. The retarded/advanced self-energy, $\Sigma^{(\dagger)}$, and lesser and greater self-energy, $\Sigma^{\lessgtr}$, are then only nonzero at the leftmost and rightmost slab. The expressions for the leftmost slab are given by [8],

$$\Sigma^{(\dagger)}_{1,1} = \tau_{1,0} g^{(\dagger)}_{0,0} \tau_{0,1} \tag{4}$$

$$\begin{aligned} \Sigma^{<}_{1,1} &= if_1\Gamma_{1,1} \\ &= -f_1(\Sigma_{1,1} - \Sigma^{\dagger}_{1,1}) \end{aligned} \tag{5}$$

$$\begin{aligned} \Sigma^{>}_{1,1} &= -i(1 - f_1)\Gamma_{1,1} \\ &= (1 - f_1)(\Sigma_{1,1} - \Sigma^{\dagger}_{1,1}), \end{aligned} \tag{6}$$

with $\tau_{i,j} = H_{i,j} - (E + i\eta)S_{i,j}$ and $g^{(\dagger)}$ equal to the retarded/advanced Green's function of the contact in equilibrium, before it is contacted to the device. The calculation of this quantity is easily achieved using the Sancho-Rubio algorithm [9].

In addition to the contacts, there is also a contribution to Σ due to phonon scattering. The orthogonal scattering self-energy is given by [10],

$$\Sigma^{<}_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^{<}(E \pm \hbar\omega_q)S \tag{7}$$

$$\Sigma^{>}_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^{>}(E \mp \hbar\omega_q)S \tag{8}$$

where we only consider local scattering and q and ω_q are the phonon wave vector and corresponding angular frequency, $N_q = \left(\exp(\hbar\omega_q/k_BT) - 1\right)^{-1}$ is the phonon occupation number, M_q are the electron–phonon matrix elements and V is the volume of the sample. The non-orthogonal scattering self-energies are then given by [11],

$$\Sigma^{\lessgtr}_{\text{scat}} = \frac{1}{2}\left(S\Sigma^{\lessgtr}_{\text{scat},\perp} + \Sigma^{\lessgtr}_{\text{scat},\perp}S\right) \tag{9}$$

$$\Sigma_{\text{scat}} = \frac{1}{2}\left(\Sigma^{>}_{\text{scat}} - \Sigma^{<}_{\text{scat}}\right). \tag{10}$$

The latter formula is an approximation which only considers the imaginary part of the scattering self-energy. The real part of the scattering self-energy typically results in a small shift of the energy levels and can safely be neglected [10].

From the Green's functions, it is possible to calculate macroscopic properties, such as electron and hole concentrations, n_k and p_k, and the currents between slabs, $I_{k,l}$ [8],

$$n_k = -\frac{in_s}{2\pi}\int_{-\infty}^{+\infty}(G^{<}S)_{k,k}\mathrm{d}E \tag{11}$$

$$p_k = \frac{in_s}{2\pi}\int_{-\infty}^{+\infty}(G^{>}S)_{k,k}\mathrm{d}E \tag{12}$$

$$I_{k,l} = \frac{qn_s}{2\pi\hbar}\int_{-\infty}^{+\infty}(\tau_{k,l}G^{<}_{l,k} - \tau_{l,k}G^{<}_{k,l})\mathrm{d}E. \tag{13}$$

where n_s is due to spin degeneracy.

The carrier concentrations give rise to a charge density ρ, which affects the effective one-particle Hamiltonian H through a change in potential. The NEGF formalism thus usually results in an iterative procedure where the charge density has to be calculated multiple times. The calculation of the charge density is, however, prohibitively expensive due to the large energy windows in (11) and (12), and thus requires an approximate approach.

2.2 The excess charge approach

In the ECA model, as implemented in ATOMOS, the charge density is calculated as the sum of the charge density due to electrons (and holes), calculated as filled (empty) states above (below) a certain charge neutrality level (CNL).

$$\rho = q(p - n) \tag{14}$$

with

$$n = -\frac{in_s}{2\pi}\int_{E_{CNL}}^{+\infty} G^{<} S \mathrm{d}E \tag{15}$$

$$p = \frac{in_s}{2\pi}\int_{-\infty}^{E_{CNL}} G^{>} S \mathrm{d}E \tag{16}$$

where the indices were dropped for the sake of clarity. The CNL is set to the Fermi level for metals and set to the middle of the band gap for semiconductors. The major benefit of the ECA is that one can significantly reduce the computational cost of the calculation by introducing a cutoff for the integral over energy. Due to the exponential decay of the Fermi-Dirac statistics function, the number of filled (empty) states at high (low) energy will be small and this part of the integral can safely be neglected, significantly reducing the number of energy points for which the Green's function must be calculated.

The disadvantage of this approach is twofold. First, the CNLs are usually taken from bulk simulations, but it is presently unclear whether these CNLs are also accurate for heterojunctions, such as semiconductor–metal interfaces at contacts [12]. Second, the ECA does not account for charge transfer due to the deformation of states. Filled states in the valence band that deform to screen electric fields do not contribute to the charge density, and hence, the screening is not included. One can account for this screening by introducing an experimental or DFT-based permittivity value, extracted from bulk material, in the Poisson solver [5].

2.3 The full charge approach

In the FCA, charge is attributed to all filled states, which allows for screening by electronic states and does not rely on CNLs extracted from bulk simulations. Total neutrality is imposed by also taking the charge on the atom cores in equilibrium into account.

$$\rho = q(n_{\text{cores}} - n_{\text{electrons}}) \tag{17}$$

with

$$n_{\text{electrons}} = -\frac{in_s}{2\pi}\int_{-\infty}^{+\infty} G^{<} S \mathrm{d}E \tag{18}$$

The disadvantage of this approach lies in the fact that it is impossible to reduce the energy window over which the Green's function must be integrated. A solution can be found by splitting the Green's function into an equilibrium part, $G^{<}_{\text{eq}}$, and a non-equilibrium part, $G^{<}_{\text{neq}}$ [7, 13]. Considering again two contacts, with broadening matrices Γ_1 and Γ_2 [14], at which carriers are injected according to the Fermi-Dirac statistics f_1 and f_2, respectively,

$$\begin{aligned} G^{<} &= G\Sigma^{<}G^{\dagger} = G\left(i\Gamma_1 f_1 + i\Gamma_2 f_2 + \Sigma^{<}_{\text{scat}}\right)G^{\dagger} \\ &= G\left(i\Gamma_1(f_1 - f_r) + i\Gamma_2(f_2 - f_r) + \Sigma^{<}_{\text{scat}}\right)G^{\dagger} \\ &\quad + iG(\Gamma_1 + \Gamma_2)G^{\dagger}f_r \\ &= G^{<}_{\text{neq}} + G^{<}_{\text{eq}}. \end{aligned} \tag{19}$$

Here, f_r is a reference Fermi-Dirac statistic function with corresponding Fermi level E_{f_r}, usually taken as the Fermi level at one of the contacts, E_{f_1} or E_{f_2}. The equilibrium part corresponds to the device in equilibrium with this reference level, i.e., states in the device are filled up according to f_r. The non-equilibrium part corresponds to the difference compared to this reference equilibrium. Neglecting the scattering part for now, $G^{<}_{\text{neq}}$ only depends on the differences of two Fermi-Dirac statistics functions $f_{1/2} - f_r$, which falls off rapidly at high and low energies. This allows for a cutoff of the energy integral similar to the ECA. The energy window of $G^{<}_{\text{eq}}$ cannot be reduced, but it can be shown that [13],

$$G^{<}_{\text{eq}} = \left(G^{\dagger} - G\right)f_r. \tag{20}$$

G ($G^{\dagger}$) is analytic in the upper (lower) complex plane which enables the use of the residue theorem, as demonstrated in Fig. 1.

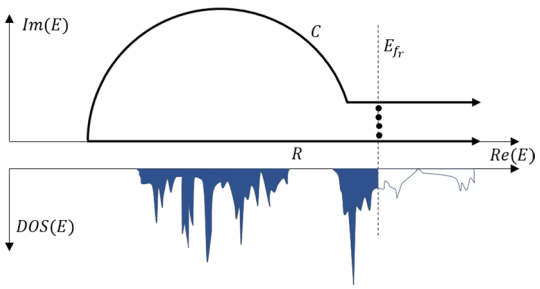

Fig. 1 The top part shows two contours in the complex plane, R and C, over which one can integrate the Green's function. Integration along R corresponds to real axis integration. Due to the analyticity of the Green's function, the two integrals are equivalent except for a summation over the poles of the Fermi-Dirac statistic, denoted by dots. The bottom part shows the DOS as a function of energy on the real axis. The Van Hove singularities require a fine grid for numerical integration. The fact that the total contour R+C is not closed is allowed due to the Fermi-Dirac function decaying exponentially at high energies (Color figure online)

$$\int_R GSf_\mathrm{r}\mathrm{d}E = \int_C GSf_\mathrm{r}\mathrm{d}E - 2\pi i k_B T \sum_{\text{poles}} G(E_p)S \tag{21}$$

The poles are the poles of the Fermi-Dirac statistic function f_r, with $E_p = E_{f_\mathrm{r}} + ik_bT\pi(2n+1)$ and $Res(f_\mathrm{r}, E_p) = -k_BT$ for $n \in \mathbb{N}$ [7]. The prohibitively high computational cost of integrating a Green's function over a large energy window is attributed to the presence of many Van Hove singularities, which require many grid points to be integrated accurately. On the complex contour, these Van Hove singularities disappear, resulting in a smooth function requiring a significantly lower number of grid points [7].

An additional benefit of (20) is that occupation of states is not governed by injection of carriers at the contacts. Hence, (20) also allows for occupation of bound states. These are states that are not connected to the contacts, and which would thus be empty in the ECA, irrespective of their energy [7]. The occupation of these bound states in the FCA depends on the reference level E_{f_r}. The issue of having the charge density depend on the choice of E_{f_r} can be resolved by making multiple choices of E_{f_r}, e.g., setting $E_{f_\mathrm{r}} = E_{f_c}$ for each contact c, and calculating the charge density for each of these choices, ρ_c. The total charge density is then obtained as the weighted average of all choices,

$$\rho = \sum_c w_c \rho_c. \tag{22}$$

Additionally, it is predicted that, for certain choices of the weights w_c, one can minimize the effect of the numerical integration error [15].

3 Challenges in employing the FCA

3.1 Implementation of the complex contour integration

The implementation of the FCA in ATOMOS uses a composite Gauss-Legendre quadrature with 5 grid points on each subinterval for the integral on the complex contour. It was found that this allows for an accurate integration for 20–40 subintervals or a total of 100–200 grid points on the complex contour. Additionally, (20) only requires G and $G^\dagger$, implying that we can forgo the expensive matrix multiplication in (3). This results in the calculation of the equilibrium charge density, ρ_eq, being computationally much less expensive than the calculation of its non-equilibrium counterpart, ρ_neq.

The computational cost can be further reduced by noting that the advanced Green's function is the Hermitian conjugate of the retarded Green's function at the complex conjugate of the energy, $G^\dagger(E) = G^H(E^*)$. The residue theorem requires a contour in the upper (lower) complex plane for the retarded (advanced) Green's function, which implies that one can use the Hermitian conjugate of the retarded Green's function calculated in the upper complex plane, for the contour integral of the advanced Green's function. In certain cases, the expression can be simplified even further. For real energies, G and $G^\dagger$ are each other's Hermitian conjugate. Consider now the case when the basis set is orthogonal and hence $S = I$,

$$\begin{aligned} n_{k,eq} &= -\frac{in_\mathrm{s}}{2\pi}\int_R \left(G^\dagger - G\right)_{k,k} f_\mathrm{r}\,\mathrm{d}E \\ &= -\frac{n_\mathrm{s}}{\pi}\mathrm{Im}\left(\int_R G_{k,k} f_\mathrm{r}\,\mathrm{d}E\right). \end{aligned} \tag{23}$$

Note that we take the imaginary part of the integral instead of the integral of the imaginary part. This is because $\mathrm{Im}\left(G_{k,k}\right)$ is not analytic in the upper complex plane, which would prevent the use of the residue theorem. It should also be noted that (23) only holds if $S = I$, as $G^\dagger S$ and GS are not Hermitian conjugates. However, a similar reasoning can be made when H and S are real, as is the case for matrices provided by certain localized orbital-based DFT packages, such as OpenMX [16] and CP2K [17]. If H and S are real and, hence, symmetric, then G and $G^\dagger$ are also symmetric as can be seen from (1), (2) and (4). Therefore, for real energies, $G_{k,l} = G_{l,k} = \left(G^\dagger_{k,l}\right)^*$ and,

$$\begin{aligned} (GS)_{k,k} &= \sum_l G_{k,l}S_{l,k} = \left(\sum_l G^\dagger_{k,l}S_{l,k}\right)^* \\ &= \left(G^\dagger S\right)^*_{k,k}. \end{aligned} \tag{24}$$

Therefore, for both the case where the overlap matrix and Hamiltonian are real or when the basis set is orthogonal, one has,

$$n_{k,eq} = -\frac{n_\mathrm{s}}{\pi}\mathrm{Im}\left(\int_R (GS)_{k,k} f_\mathrm{r}\,\mathrm{d}E\right). \tag{25}$$

3.2 Rescaling at the contacts

As (11) and (25) indicate, the calculation of the carrier concentration requires a matrix multiplication of the lesser Green's function and retarded Green's function with the overlap matrix. However, at the interfaces with the contacts this multiplication requires the respective Green's functions coupling the device and the contact. At the left contact,

$$(G^*S)_{1,1} = G^*_{1,0}S_{0,1} + G^*_{1,1}S_{1,1} + G^*_{1,2}S_{2,1} \tag{26}$$

where we use G^* as a generalized Green's function which can be either G or $G^<$. Note that the indices denote slab indices and the entities in (26) are therefore matrices. We cannot use

(26) as is, because we only calculate G^* within the device, and hence, $G^*_{1,0}$ is unknown. The same is true at the right contact for $G^*_{N,N+1}$. Indications on how to resolve this issue are provided in [8]. For the retarded Green's function, one can obtain an expression for $G_{1,0}$ using matrix algebra. More specifically, one considers the matrix G as the inverse of the matrix $((E - i\eta)S_{\text{tot}} - H_{\text{tot}})$ where S_{tot} and H_{tot} contain both the device and its contacts. $G_{1,0}$, the part of the inverse coupling the two, is then obtained by inversion by partitioning [18],

$$G_{1,0} = G_{1,1}\tau_{1,0}g_{0,0}. \tag{27}$$

For the lesser Green's function, an algebraic approach is not possible, and one has to start from the Dyson equation of the contour ordered Green's function[19] and employ the Langreth theorem to obtain [8],

$$G^<_{1,0} = G_{1,1}\tau_{1,0}g^<_{0,0} + G^<_{1,1}\tau_{1,0}g^\dagger_{0,0}. \tag{28}$$

Equation (28) requires the unperturbed Green's functions within the contacts, which are known. For real energies, $g^\dagger_{0,0}$ is just the Hermitian conjugate of $g_{0,0}$. $g^<_{0,0}$ is linked to the occupation of the edge states of the contact. Before contacting the contact and the device, the contact is in equilibrium according to the statistics function f_1. Using (20), we thus obtain,

$$g^<_{0,0} = \left(g^\dagger_{0,0} - g_{0,0}\right)f_1. \tag{29}$$

A similar derivation is possible for the right contact. It should be noted that these considerations are not limited to the FCA but are relevant for all NEGF simulations using non-orthogonal basis sets. However, the relative error introduced by neglecting this rescaling of the contact carrier concentration is small. This is shown for a homogeneous sheet of WS_2 in equilibrium, i.e., $f_1 = f_2$. The Hamiltonian and overlap matrix are obtained with OpenMX. The relative error on the carrier concentration at the contact interfaces when the contact rescaling is neglected, is about 10%, as denoted in Fig. 2, which does not influence the device behavior significantly in the ECA. In the FCA, the carrier concentration is much larger, but it is counteracted by the subtraction of n_{cores}, which is equally large and nearly identical to $n_{\text{electrons}}$. The small relative error on $n_{\text{electrons}}$ is therefore greatly amplified into a large relative error on the charge density in the FCA.

3.3 Explicit screening

A major benefit of the FCA is that screening of charges is automatically included, compared to the ECA, where a bulk permittivity value has to be introduced in the Poisson solver. To compare both models, we performed a self-consistent

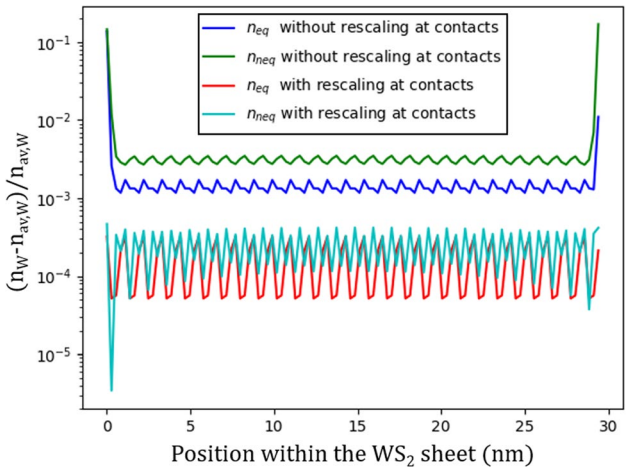

Fig. 2 The relative error on the carrier concentration on W atoms in a WS_2 sheet in equilibrium. The relative error is calculated as the relative difference of the carrier concentration with the average carrier concentration of all W atoms in the device. Both the equilibrium and the non-equilibrium contributions are listed. When Green's function rescaling with S at the contacts is taken into account correctly, the relative error is uniform in the device and never exceeds 10^{-3}. When the term $G^*_{1,0}S_{0,1}$ in (26) is neglected, peaks in the relative error as high as 10^{-1} appear at the contacts. Additionally, the relative error in the rest of the devices also increases because the average carrier concentrations is affected by this increase of the error at the contacts

simulation of an n-type MOS (nMOS) WS_2 transistor in OFF-state using both the ECA, with in-plane and out-of-plane relative permittivity values $(\varepsilon_{//}, \varepsilon_\perp) = (13.7, 6.3)$, extracted from DFT [20], and the FCA, with vacuum permittivity $(\varepsilon_{//}, \varepsilon_\perp) = (1.0, 1.0)$. The Hamiltonian matrix was obtained using Quantum Espresso [21] and Wannier90 [22] as detailed in [5]. The basis set obtained through Wannierization is orthogonal. For comparison, a similar device simulation was performed with a Hamiltonian matrix obtained using OpenMX.

The resulting carrier concentrations and bands are shown in Fig. 3. Figure 3a demonstrates how the carrier concentration drops in the channel region for both the ECA and FCA. For the FCA, an additional dipole formation can be distinguished, which screens the electric fields in the depletion regions. From the bands in Fig. 3b, it is, however, apparent that the FCA does not attain the same electrostatic behavior as the ECA with adjusted permittivity in the Poisson solver. Additionally, we find that the FCA screening depends on the DFT package used for the calculation of the device Hamiltonian.

The reason for this DFT package dependency is unclear at this point. One explanation could be that different DFT packages give rise to different band structures. The bands close to the Fermi level or band gap are equivalent, which is

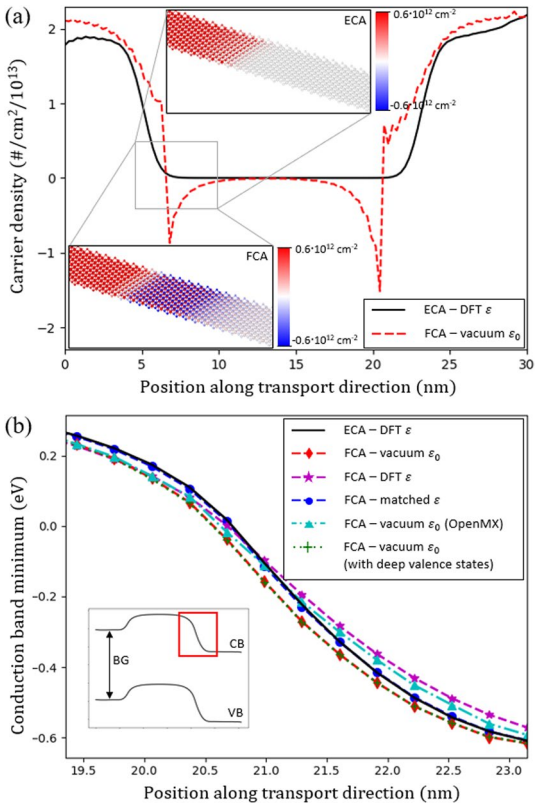

Fig. 3 Results on a WS_2 transistor in OFF-state. **a** shows the carrier concentrations as a function of the position. The insets show an atomic depiction of the depletion regions for both the ECA and the FCA, where the atom colors denote the carrier concentration. **b** shows part of the conduction band as a function of the position. The inset shows the complete conduction (CB) and valence band (VB) in the whole device. Different trials with the FCA are compared to an ECA reference using a macroscopic permittivity value in the Poisson solver. The FCA does give rise to screening but does not attain the same electrostatic control as the ECA. For accurate band matching, the FCA should be combined with an intermediate permittivity value. Additionally, a simulation using a Hamiltonian extracted with OpenMX, instead of Quantum Espresso with Wannier90, shows that the screening depends on the DFT-package used. The OpenMX simulations included more deep valence bands. To verify whether this influences results, more deep valence states were also included in the Wannierization process (Color figure online)

important for transport. For the Hamiltonian obtained with Quantum Espresso, these are the only bands included. The ultra-soft pseudo potentials used in Quantum Espresso gives rise to 13 valence bands for WS_2. During Wannierization only the 7 highest energy valence bands are included. Similarly, only the 4 lowest energy conduction bands are included during Wannierization for a total of 11 Wannier functions for each primitive cell. The OpenMX Hamiltonian typically includes more low-energy valence bands and high-energy conduction bands as no additional Wannierization process is needed. For WS_2, the OpenMX Hamiltonian corresponds to 12 valence bands and 19 conduction bands. The extra bands compared to Quantum Espresso are too far from the injection energies to influence transport, but could impact the electronic screening in the FCA. However, Fig. 3b shows that including more deep valence bands during Wannierization, for a total of 9 valence bands, does not affect the screening. Other parameters that could still explain the discrepancy in electronic screening are the different number of conduction band states, the difference in pseudopotentials, exchange-correlation functional or type of basis set that were used. A full investigation of the influence of the DFT package on electronic screening is, however, beyond the scope of this paper.

It should be noted that the FCA does not include ionic displacements and thus, generally, underestimates the screening, although this source of inaccuracy is expected to be small for WS_2 [20]. Due to these shortcomings of the FCA, neglect of ionic displacements and DFT package dependency, we propose the use of an intermediate permittivity value in the FCA that results in band matching with the ECA with the full permittivity value from DFT. The relative permittivity values that corresponded to optimal band matching were iteratively found to be $(\varepsilon_{//}, \varepsilon_{\perp}) = (1.5, 2.8)$ for the Hamiltonian obtained from Quantum Espresso, indicated by the blue line in Fig. 3b. For band matching with the Hamiltonian obtained from OpenMX, relative permittivity values smaller than 1.0 would be required which is unphysical. Another approach would therefore be to use a small or vacuum permittivity for the FCA and to search the matching permittivity value for the ECA. This is the approach we utilized in Sect. 4.

3.4 Scattering in the FCA

One of the major benefits of the NEGF formalism is that one can easily incorporate electron–phonon interactions. Scattering of charge carriers by interaction with phonons can heavily influence transport properties in devices. To our knowledge, a study investigating the influence of the FCA on scattering mechanisms has not been done. Indeed, it is not clear whether splitting the lesser Green's function into an equilibrium and non-equilibrium part is compatible with the calculation and incorporation of a scattering self-energy:

- The scattering self-energy in (8) depends on the lesser Green's function at energies shifted with the phonon energy. However, the lesser Green's function is unknown, as we only compute $G^{<}_{\text{neq}}$. The same is true for $G^{>}$ which is required for $\Sigma^{>}_{\text{scat}}$ to calculate Σ_{scat} in (10). However, provided we have a way to calculate $G^{>}$, we could use the identity $G - G^{\dagger} = G^{>} - G^{<}$ [8] to extract $G^{>}$ as well.

- We could, in principle, calculate $G^<$ for real energies in the same energy window as $G^<_{\text{neq}}$. However, this would amount to a doubling of the computational cost of the procedure as many energy points are required on the real energy axis.
- It is unclear on how to extend scattering to the equilibrium part. The calculation of G requires Σ_{scat} at complex energies. However, Σ_{scat} depends on both $\Sigma^<_{\text{scat}}$ and $\Sigma^>_{\text{scat}}$, and hence, depends on both $G^<$ and $G^>$ which are not analytic in the complex plane. Indeed, even the very derivation of (20) has only been done in the ballistic case [13].

Even though these concerns appear to prohibit the combination of the FCA with electron–phonon scattering, there is a rather simple solution for the simulation of semiconductor devices, where there is a band gap. Consider the case where the reference level, E_{f_r}, is in the band gap. The equilibrium part then corresponds to a completely filled valence band. The influence of the scattering self-energy is a shift in the energy of states due to the real part of the scattering self-energy [10] and a broadening of the states due to the imaginary part of the scattering self-energy, which can be interpreted as a finite lifetime of the states [23]. Both conserve the total mass of the states when integrated over energy. Especially considering the fact that only local scattering is taken into account here, we claim that scattering does not influence the charge density of a completely filled valence band. Additionally, the equilibrium part does not result in transport, and hence, the charge density is the only thing that is of interest for the equilibrium part. This claim therefore allows us to safely leave out the scattering self-energy in the equilibrium part of the Green's function. Indeed, this is exactly what was done in (19).

For the non-equilibrium part, we discern between n-type and p-type devices. Consider the case of an n-type device with E_{f_r} in the middle of the band gap. In this case, $f_r \approx 0$ in the conduction band and $G^<_{\text{neq}}$ is just equal to $G^<$. Practically, E_{f_r} does not have to be in the middle of the band gap as a few k_BT below the bottom of the conduction band is sufficient. We can then define,

$$G^>_{\text{neq}} = G - G^\dagger + G^<_{\text{neq}} \approx G - G^\dagger + G^< = G^>. \tag{30}$$

Equations (1)–(3) and (8)–(10) can then be used unaltered for the non-equilibrium parts.

It is tempting to make the same assumption for p-type devices. However, in the valence band, $f_r \approx 1$ when E_{f_r} is a few k_BT above the top of the valence band. Hence, in the ballistic case,

$$\begin{aligned} G^<_{\text{neq}} &= G\big(i\Gamma_1(f_1 - f_r) + i\Gamma_2(f_2 - f_r)\big)G^\dagger \\ &\approx -G\big(i\Gamma_1(1 - f_1) + i\Gamma_2(1 - f_2)\big)G^\dagger = G^>. \end{aligned} \tag{31}$$

$G^<_{\text{neq}}$ thus plays the role of $G^>$ in p-type devices when E_{f_r} is in the band gap. This is not surprising as $G^<_{\text{neq}}$ is the deviation from an equilibrium where the valence band is completely filled. $G^<_{\text{neq}}$ is thus related to the concentration of holes, just like $G^>$. It could therefore be argued that one should change the sign convention in (8) to the one in (9) for the calculation of $\Sigma^>_{\text{scat}}$ in p-type devices with the FCA. To then keep Σ_{scat} consistent with the ECA, we should also change (9), since,

$$\begin{aligned} G^>_{\text{neq}} &= G - G^\dagger + G^<_{\text{neq}} \\ &\approx G - G^\dagger + G^> = 2G^> - G^<. \end{aligned} \tag{32}$$

We can extend this reasoning to more general choices of E_{f_r} than E_{f_r} being in the middle of the band gap. We propose the following altered scattering equations in the FCA.

If $E_{f_r} < \min(E_{f_1}, E_{f_2})$

$$\Sigma^<_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^<_{\text{neq}}(E \pm \hbar\omega_q)S \tag{33}$$

$$\Sigma^>_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^>_{\text{neq}}(E \mp \hbar\omega_q)S \tag{34}$$

If $E_{f_r} > \max(E_{f_1}, E_{f_2})$

$$\Sigma^<_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^<_{\text{neq}}(E \mp \hbar\omega_q)S \tag{35}$$

$$\Sigma^>_{\text{scat},\perp}(E) = \sum_q \frac{|M_q|^2}{V}\left(N_q + \frac{1}{2} \pm \frac{1}{2}\right)G^>_{\text{neq}}(E \pm \hbar\omega_q)S + \Sigma^>_{\text{cor}} \tag{36}$$

with

$$\Sigma^>_{\text{cor}} = \sum_q \frac{|M_q|^2}{V}\big(2G^<_{\text{neq}}(E - \hbar\omega_q) - 2G^<_{\text{neq}}(E + \hbar\omega_q)\big)S. \tag{37}$$

It can easily be verified that (33)–(36) keep (10) consistent with the ECA. The importance of these corrections are demonstrated in Fig. 4, where both approaches are used to simulate a WS_2 MOSFET. Both approaches used the same Hamiltonian obtained using Quantum Espresso and Wannier90, and the intermediate permittivity value obtained in Sect. 3.3 was used for the FCA.

3.5 Choosing the reference level E_{f_r}

Section 3.4 showed that choosing the right reference energy level allows us to obtain strong arguments as to why and how electron–phonon scattering can be correctly combined with the FCA. These arguments relied heavily on the fact that the device was either n-type or p-type, such that the reference level could be put in the band gap throughout the

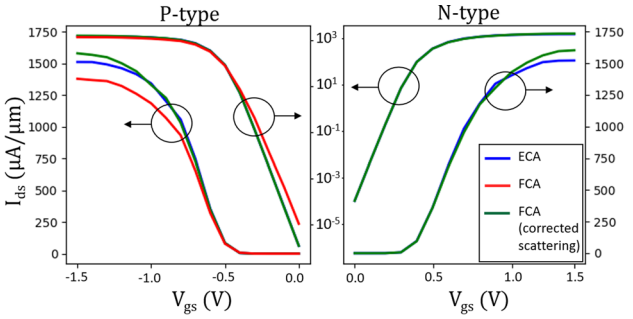

Fig. 4 IV curves for, respectively, a 14-nm channel WS_2 pMOS and nMOS transistor simulated using ATOMOS with both the FCA and ECA. For n-type devices, there is negligible difference in the current of the ECA and FCA. However, for p-type devices, this is only the case when the corrected scattering self-energies are used. When the naive approach of substituting in $G^{>}_{\mathrm{neq}}$ and $G^{<}_{\mathrm{neq}}$ in (8) and (9) is applied, the ON-current and the subthreshold slope are degraded compared to the ECA

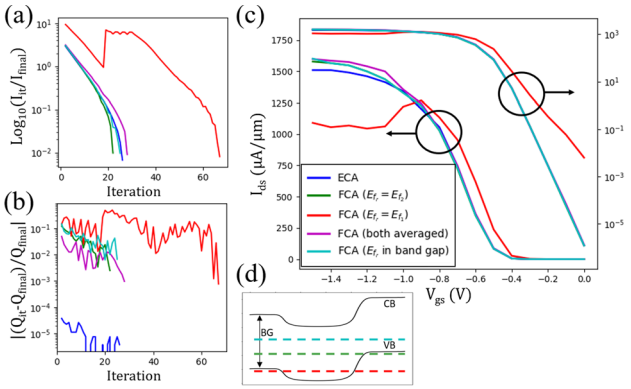

Fig. 5 Comparison of ATOMOS simulations of a 14 nm channel WS_2 pMOS transistor for different choices of E_{f_r}. **a** and **b** show the error in the current and charge, respectively, as a function of the iteration number for a single bias point. Due to the exponential behavior of the current, the error on the current is calculated as $\log_{10}(I_{\mathrm{it}}/I_{\mathrm{final}})$. **c** shows the IV curves on both a linear and log scale. **d** shows a schematic representation of the bands of the transistor as a function of the position along the transport direction. It also denotes the different choices of E_{f_r} with dashed lines, following the color codes in (**a**), (**b**) and (**c**) (Color figure online)

whole device. However, the reference level is conventionally not kept fixed at one level, but set equal to each contact Fermi level E_{f_i} once, after which the results are combined in a weighted average [13, 15]. The averaging resolves the arbitrariness of having to choose a E_{f_r} which might influence the charging of bound states, and, more importantly, is predicted to reduce numerical errors in the integration. Indeed, it can be shown that an appropriate choice of the weights can minimize the error of integration if one assumes that the error of integration follows a certain stochastic behavior and is dominated by the real axis integration [15]. The latter assumption is a reasonable assumption since the contour integral can be made extremely accurate by increasing the number of energy points without significantly increasing the computational cost.

To probe the influence of E_{f_r}, we performed ATOMOS simulations of the same p-type WS_2 transistor as in Sect. 3.4 for different choices of E_{f_r}. The results are shown in Fig. 5. It is clear that putting the reference level equal to the right contact Fermi level, $E_{f_r} = E_{f_2}$, putting it in the band gap or using the averaging procedure with E_{f_r} both equal to the left and right contact Fermi level, results in nearly identical IV curves and convergence behavior. Therefore, even though we only have strong arguments that the ECA and the FCA with scattering are equivalent when E_{f_r} is put in the band gap throughout the device, it appears that in some cases, it is also correct to use the averaging approach or to set E_{f_r} equal to the contact Fermi level closest to this band gap reference level.

This equivalence can be argued for by noting that the effect of scattering on transport in devices is most pronounced by its effect on states at the injection level of the source, i.e., the left contact. If this energy is more than a few k_BT away from the reference level, as is the case when $E_{f_r} = E_{f_2}$ and the source-drain bias is more than a few k_BT, then $G^{<}_{\mathrm{neq}} \approx G^{\lessgtr}$ for the states that contribute most to scattering, and the same reasoning as in the previous section applies.

When the reference level is put equal to the left contact Fermi level, $E_{f_r} = E_{f_1}$, significantly more iterations are required to achieve convergence, and the value that is obtained after convergence differs significantly from the result for the other choices of E_{f_r} and the ECA. The reason for this discrepancy is twofold. First, when the reference level is set to the injection level at the source, $G^{<}_{\mathrm{neq}}$ differs strongly from $G^{\lessgtr}$ and (33)–(36) in the previous section do not hold. Second, when E_{f_r} is put to E_{f_1}, there are filled states between E_{f_1} and E_{f_2} at the drain side, i.e., the right side of the device. These states are integrated with a real axis integration, which suffers from the presence of Van Hove singularities and which can hinder convergence.

On the other hand, when E_{f_r} is put to E_{f_2} the energy window between E_{f_1} and E_{f_2} contains very few states at the source, because most of this energy window is located in the band gap. This also explains why the averaging approach gives the same result as putting E_{f_r} equal to E_{f_2}. The weights in the procedure are chosen such that a larger weight is used for the charge density which has a smaller contribution of the real axis integral of ρ_{neq}, since this part is expected to bear a larger error [15]. Since the choice $E_{f_r} = E_{f_2}$ has fewer states integrated with the real axis integral, this contribution

will dominate the averaging approach, which explains the similarities in Fig. 5.

It should be noted that the averaging approach can be beneficial for devices which contains both n-type and p-type doped regions, to reduce the integration error in an automated way. However, even though the averaging approach takes about the same number of iterations, each iteration is twice as expensive since both ρ_{eq} and ρ_{neq} have to be calculated twice. In the case of purely p-type (n-type) devices, we can thus forgo this additional cost by setting E_{f_r} to $\max(E_{f_1}, E_{f_2})$ $(\min(E_{f_1}, E_{f_2}))$.

4 The importance of the FCA at metal interfaces

In the previous section, we used the ECA to test and verify the FCA using single doping type transistors without metallic contacts, since the two models are expected to be largely equivalent for such devices. For devices including strongly doped pn junctions or metallic interfaces, this is no longer the case and the FCA, being closer to reality, is expected to give more accurate results. In this section, we apply both the ECA and FCA for the simulation of a HfS_2 nMOS transistor with $HfTe_2$ metal contacts at the source and drain. We consider both a top contact and a lateral contact configuration.

4.1 Top contacts

The Hamiltonian elements of the top contact were extracted by simulating a bulk $HfTe_2$-HfS_2 bilayer with Quantum Espresso. Both layers were first relaxed individually with a variable cell relaxation. Both materials were then strained for 3.9% to have the lattices match and joined in a bilayer after which the atomic positions were relaxed again with fixed cell dimensions. The OptB86 functional was used in combination with ultrasoft pseudopotentials and the Grimme DFT-D3 van der Waals correction with a 30 Å vaccuum between cells. A 10×10×1 Monkhorst-Pack grid was used and the energy cut off for the plane wave basis was set to 70 Ry. The self-consistency convergence criterion and the atomic force convergence criterion were set to, respectively 10^{-6}, eV and 10^{-3} eV/Å. The Wannier90 package was then used to convert the plane wave basis to a localized basis, starting from 3 p orbitals on S and Te and 3 d orbitals on Hf.

The Hamiltonian elements provided by Wannier90 were then used to build a device Hamiltonian for the whole device, as demonstrated in Fig. 6a. The device dimensions depend on the crystal orientation. Two crystal orientations are considered, namely transport along the ΓK and ΓM direction. The corresponding device dimensions are provided in Table 1. The relative permittivity in the FCA was set to $(\varepsilon_{//}, \varepsilon_{\perp}) = (2.0, 2.0)$ and the matching relative permittivity

Table 1 Dimensions of the device with metal contacts

	L_{chan} (nm)	L_{metal} (nm)	L_{ext} (nm)	L_{int} (nm)
Top ΓK	14	6.8	9.4	4.5
Top ΓM	14	6.5	8	3.9
Lateral ΓM	14	11	10.6	/

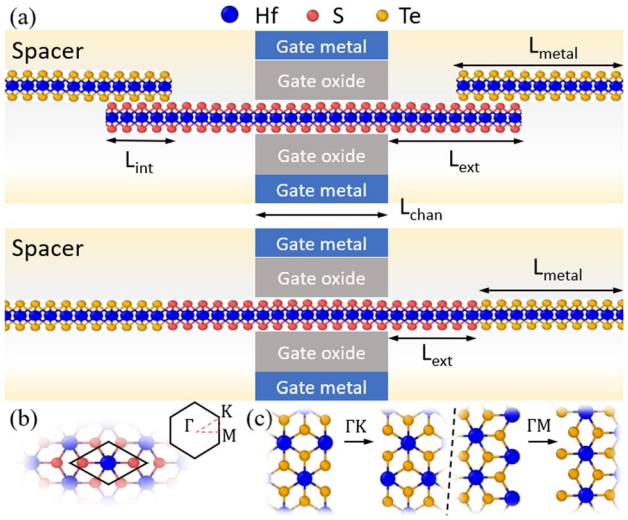

Fig. 6 **a** The device configuration of the nMOS HfS_2 transistor with $HfTe_2$ top and lateral contacts. The dimensions in the schematic are not to scale. The actual dimensions are listed in Table 1. **b** Crystal structure of HfS_2 with the corresponding high symmetry points in reciprocal space. **c** The edge of the metallic top layers at the source and drain side for both the device with transport along ΓK and ΓM

in the ECA was found to be $(\varepsilon_{//}, \varepsilon_{\perp}) = (9.0, 3.4)$. The channel oxide relative permittivy was set to 15.6, resulting in an effective oxide thickness of 0.5 nm. Finally, 10 k-points were used to capture half the Brillouin zone of the periodic direction in ATOMOS. The other half was obtained through symmetry.

The resulting IV curves are shown in Fig. 7. For comparison, also pure HfS_2 references are shown, where the explicit metallic contacts were removed, the semiconducting source-drain extensions were enlarged to have the same total device dimensions and perfect ohmic contacts were used just like in the sections above. The incorporation of metal top contacts into the device decreases the current by ~50% compared to the pure HfS_2 references. This decrease is attributed to the introduction of a van der Waals gap, which is consistent with our previous work in [24]. The results for the pure semiconductor references in the ECA and FCA are nearly identical, which is consistent with our earlier assertion that for single doping type pure semiconducting devices, the ECA and the FCA are equivalent. The same appears true for the transistor with top metal contacts with transport along the ΓK direction. However, when the transport direction is along ΓM,

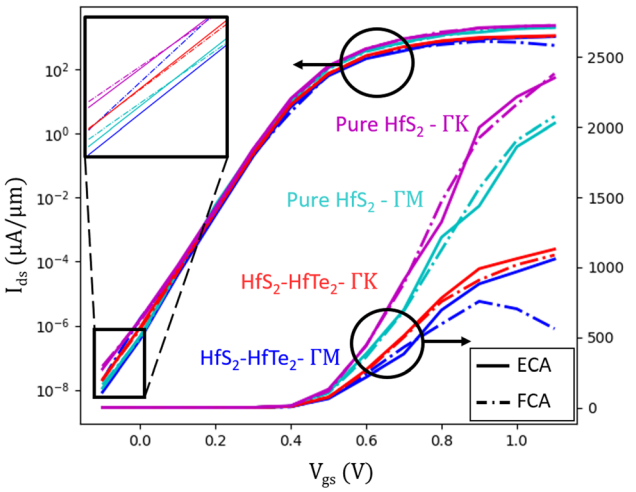

Fig. 7 IV curves of the nMOS HfS_2 transistor with $HfTe_2$ top contacts and pure HfS_2 references, simulated with both the ECA and FCA and both transport along the ΓK and ΓM direction. The IV curves are plotted both on a linear scale (right) and a logarithmic scale (left). The deep subthreshold regime on the log-scale is repeated in the inset for clarity

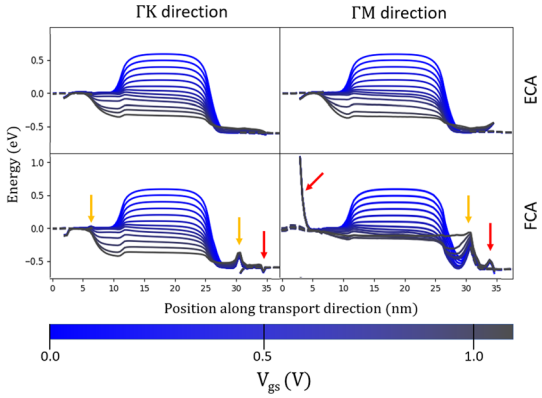

Fig. 8 Bottom of the conduction band of the nMOS HfS_2 transistor with $HfTe_2$ top contacts as a function of the position along the transport direction for different gate biases. The conduction bands are shown for both the ECA and FCA and for both transport along the ΓK and ΓM direction. The conduction bands are based on the middle of the HfS_2 parts, i.e., at the height of the Hf atoms, and are hence only shown for the regions with semiconducting material. The neutralitiy levels within the metal are also shown and denoted with a dashed line. Similarly, these neutrality levels are based on the potential in the middle of the $HfTe_2$ parts, i.e., at the height of the Hf atoms. The yellow (red) arrows denote potential peaks due to unpassivated $HfTe_2$ (HfS_2) edges (Color figure online)

there is a larger difference in both the ON-current and the subthreshold regime for the FCA and the ECA.

The origin for this discrepancy can be found in the band profile, shown in Fig. 8. In the ECA, there are no significant interface charges at the metal–semiconductor interfaces and no significant Schottky barrier is found. In the FCA, however, a strong peak in the potential can be seen, indicating the presence of significant interface charges. It should be noted that the ends of the metallic and semiconducting layers in Fig. 6a are not passivated, which implies that there are dangling bonds at the interface, and hence, there are large interface charges in the FCA. Indeed, this also explains the discrepancy between the source and drain side and the influence of the crystal orientation, as demonstrated in Fig. 6c. When the transport direction is along ΓK, the final Hf and Te atoms on both sides have a coordination of 4 and 2, respectively. When the transport direction is along ΓM, there are dangling Hf atoms with coordination 3 on the left side, and dangling Te atoms with a coordination of only 1 on the right. There are thus more dangling bonds when the transport direction is along ΓM, which complies with the interface charge and potential peaks in Fig. 8. Note that the FCA in combination with a transport direction along ΓM gives such large charges that the potential in the metal is non-flat, as indicated by the small peaks in the neutrality levels. This is attributed to the limited density of states at the Fermi level due to having a 2D metal and quantum effects limiting the screening capabilities of a single quantum state [25, 26].

It should be noted that the lack of passivation is not necessarily realistic. In a real device, the electronic states at the edge of a material will not correspond well with electronic bulk states which have been abruptly cut. The electronic states will rearrange themselves to a lower energy termination. Additionally, the edge atoms could be partially passivated by interaction with neighbouring oxide atoms. Therefore, correct capturing of these interface states by the FCA is not necessarily beneficial. In that regard, the ECA appears more robust, as these interface states appear to be screened away automatically.

4.2 Lateral contacts

The Hamiltonian and overlap matrix elements of the lateral contact were extracted by simulating a lateral heterojunction with 6.8 nm of $HfTe_2$ and 6.4 nm of HfS_2 with OpenMX. It was verified that the on-site energies in, respectively, the middle of the HfS_2 and $HfTe_2$ parts were constant, leading us to believe that these parts of the DFT cell can be used to extract bulklike properties. The same lattice constants were taken as in the top contact configuration for pure HfS_2 and $HfTe_2$ to limit the influence of the DFT package and the exchange-correlation functional. Both materials were again strained for 3.9% to have the lattices matched and joined in a heterojunction. The atomic positions were then relaxed again with fixed cell dimensions in the orthogonal direction. The dimension in the transport direction was allowed to vary freely. The GGA-PBE functional was used in combination

with the corresponding provided optimized basis sets and pseudopotentials [27, 28] and a 2×8×1 Monkhorst-Pack grid was used. The self-consistency convergence criterion and the atomic force convergence criterion were set to $2.7 \cdot 10^{-5}$ eV and $5 \cdot 10^{-3}$ eV/Å, respectively. The rest of the procedure was similar to the top contact case.

The relative permittivity value in the FCA was set to $(\varepsilon_{//}, \varepsilon_{\perp}) = (2.0, 2.0)$ and the matching relative permittivity in the ECA was found to be $(\varepsilon_{//}, \varepsilon_{\perp}) = (37.0, 5.5)$. Note that this is larger than the matching relative permittivity for HfS_2 extracted with Quantum Espresso and Wannier90, confirming our earlier observation that the explicit screening is DFT package dependent. Due to the increased computational cost of the larger basis set provided by OpenMX than by Wannier90, only the ΓM direction is considered here. One should be careful with choosing the core charges n_{cores}. A naive approach would be to set the n_{cores} of all atoms of the same type to the same value, namely the number of electrons on the atom in bulk in equilibrium. However, it should be noted that in our NEGF approach, we are attributing charge to a surplus number of electrons or holes compared to the DFT level. For the lateral junction, atoms at the interface can differ significantly from bulklike atoms and can thus carry a different equilibrium charge. Therefore, we set the value of n_{cores} for atoms at the interface to the Mulliken charge of the corresponding atom in the DFT simulation. Both approaches are equivalent for the case of a bilayer separated by a van der Waals gap, where both layers are considered to behave bulklike.

The IV curves of the HfS_2 nMOS transistor with lateral contacts are shown in Fig. 9 for both the ECA and FCA and for both Mulliken-charge-based n_{cores} and bulklike n_{cores} at the interface. Also pure HfS_2 nMOS transistors with ohmic contacts are shown for reference. The pure HfS_2 transistors show little difference between the ECA and FCA in ON-state, as is consistent with our discussion above. For the transistors with lateral contacts, the different models give different results. Both models achieve ON-currents which are slightly reduced compared to the pure HfS_2 reference due to the presence of a Schottky barrier. However, the reduction is more pronounced for the ECA, ~45% lower current in ON-state than the pure HfS_2 reference, than for the FCA, ~30% lower current. The relative difference in ON-current between the models is hence ~20%. The choice between Mulliken-charge-based n_{cores} and bulklike n_{cores} for the simulations with explicit metallic contacts, seems to be of minor importance, with an ON-current difference less than 5%. The decrease in ON-current is attributed to the different interface charge and interface screening and hence a difference in the interface potential profile. The interface potential profile influences the effective Schottky barrier height, which becomes an important factor in ON-state, when the channel barrier collapses.

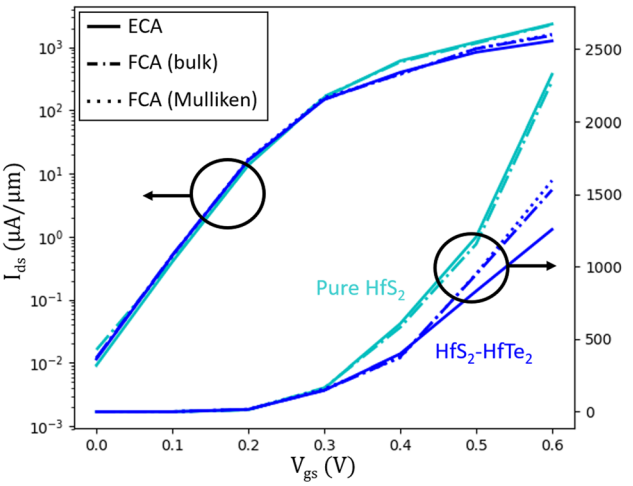

Fig. 9 IV curves of the nMOS HfS_2 transistor with $HfTe_2$ lateral contacts and pure HfS_2 references, simulated with both the ECA and FCA. The IV curves are plotted both on a linear scale (right) and a logarithmic scale (left)

The potential profiles along the transport direction for the different choices of model and n_{cores} are shown in Fig. 10. In the lateral contact configuration, there are no dangling bonds, and hence, there is no need for surface passivation. The interface charge in the FCA is therefore expected not to be unphysically large and hence give an accurate representation of the actual interface charge. This is consistent with the fact that there are no large peaks in the potential profile of the lateral contact configuration, compared to the top contact configuration. Note that the ECA gives rise to an effective Schottky barrier height at the source of ~ 100 mV,~ 40 mV higher than the FCA, which is consistent with the lower ON-current for the ECA. The FCA shows little dependency on the choice of n_{cores} at the source side, consistent with the minimal difference in ON-current between the two choices. At the drain side, the choice of n_{cores} significantly influences the potential profile and corresponding Schottky barrier. However, as this barrier is below the injection energy at the source, there is little effect on the ON-current. Scattering can be assumed to be simulated correctly for similar reasons. Transport is dominated by the source side and happens around the source-side injection energy, which is several $k_B T$ above the reference level, and hence, it can be assumed that $G^{<}_{\text{neq}} \approx G^{<}$ at the energies most relevant for transport.

5 Conclusion

There are typically two approaches to calculate the charge density in the NEGF formalism: the ECA and FCA. The ECA does not account for electronic screening and relies on neutrality levels, for which the validity is unclear at

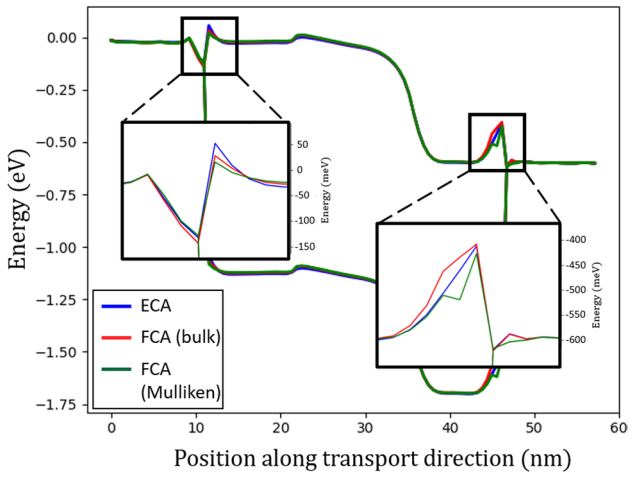

Fig. 10 The potential profile of the nMOS HfS_2 transistor with $HfTe_2$ lateral contacts in ON-state. The potential profile is given as a function of position along the transport direction, for both the ECA and the FCA and for both choices of n_{cores} at the interface. The potential is denoted with the bottom of the conduction band and the top of the valence band in the semiconducting part and with the neutrality level in the metal parts. The insets repeat the potential profiles at the interfaces for the sake of clarity

heterojunctions. The FCA automatically accounts for electronic screening and relies on a choice of reference levels and ionic charges n_{cores}. For single doping type pure semiconducting devices, the two approaches are expected to be equivalent except for a difference in electronic screening. We used an nMOS WS_2 transistor to verify the equivalence of the ECA and FCA and test our implementation of the FCA.

It was found that the explicit screening behavior of the FCA depends on the DFT package used to extract the Hamiltonian, negating this advantage of the FCA. The reason for this discrepancy is presently unclear. However, this dependency can be compensated for by the use of a matched permittivity value in the Poisson solver.

Furthermore, it was found that under certain circumstances, the FCA is compatible with electron–phonon scattering, if adjusted scattering self-energy expressions are used. We also resolved some uncertainty concerning the choice of the reference energy level E_{f_r} used in the FCA and provided new expressions for rescaling of the Green's functions at the contacts.

Finally, we investigated the impact of the model choice for the simulation of metal–semiconductor interfaces, where both models can feature different currents in ON-state for the same device. These differences in ON-current are attributed to differences in the charges at the interface, which can give rise to potential peaks and changes in the Schottky barrier height. For adequate choices of n_{cores}, e.g., the Mulliken charges of the corresponding atoms at the DFT level, and assuming an adequate matched permittivity can be found, the FCA is expected to simulate the interface charge at heterojunctions more accurately. However, the two models give qualitatively similar results for our case of a HfS_2 transistor with $HfTe_2$ side contacts, characterized by a reasonable low Schottky barrier height of ~ 100 mV. For devices containing unpassivated edge atoms, the difference between the two models can be more pronounced. Unrealistic interface states make the FCA susceptible to the surface termination. While introducing correct passivation at the DFT level is a possibility, it makes simulations more cumbersome and is considered outside the scope of this paper.

Author contributions AA developed the theory and simulation code. RD implemented the FCA model, performed the transport simulations, analyzed the data and wrote the manuscript. All authors discussed the results and proofread the manuscript.

Funding This research was funded by the FWO as part of the PhD fellowship 1100321N.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

1. Schwierz, F., Pezoldt, J., Granzner, R.: Two-dimensional materials and their prospects in transistor electronics. Nanoscale **7**(18), 8261–8283 (2015). https://doi.org/10.1039/C5NR01052G
2. Logoteta, D., Zhang, Q., Fiori, G.: What can we really expect from 2d materials for electronic applications? In: 72nd Device Research Conference, pp. 181–182 (2014). https://doi.org/10.1109/drc.2014.6872357
3. Kang, J., Cao, W., Xie, X., Sarkar, D., Liu, W., Banerjee, K.: Graphene and beyond-graphene 2d crystals for next-generation green electronics. In: Micro-and Nanotechnology Sensors, Systems, and Applications VI, International Society for Optics and Photonics, vol. 9083, p. 908305 (2014). https://doi.org/10.1117/12.2051198
4. Chhowalla, M., Jena, D., Zhang, H.: Two-dimensional semiconductors for transistors. Nat. Rev. Mater. **1**(11), 16052 (2016). https://doi.org/10.1038/natrevmats.2016.52

5. Afzalian, A.: Ab initio perspective of ultra-scaled CMOS from 2d-material fundamentals to dynamically doped transistors. npj 2D Mater. Appl. (2021). https://doi.org/10.1038/s41699-020-00181-1
6. Zhao, Y., Xu, K., Pan, F., Zhou, C., Zhou, F., Chai, Y.: Doping, contact and interface engineering of two-dimensional layered transition metal dichalcogenides transistors. Adv. Funct. Mater. **27**(19), 1603484 (2016). https://doi.org/10.1002/adfm.201603484
7. Chu, Y., Sarangapani, P., Charles, J., Klimeck, G., Kubis, T.: Explicit screening full band quantum transport model for semiconductor nanodevices. J. Appl. Phys. **123**(24), 244501 (2018). https://doi.org/10.1063/1.5031461
8. Lake, R.K., Pandey, R.R.: Non-equilibrium green functions in electronic device modeling. arXiv preprint cond-mat/0607219 (2006)
9. Sancho, M.P.L., Sancho, J.M.L., Sancho, J.M.L., Rubio, J.: Highly convergent schemes for the calculation of bulk and surface green functions. J. Phys. F Metal Phys. **15**(4), 851–858 (1985). https://doi.org/10.1088/0305-4608/15/4/009
10. Afzalian, A.: Computationally efficient self-consistent born approximation treatments of phonon scattering for coupled-mode space non-equilibrium green's function. J. Appl. Phys. **110**(9), 094517 (2011). https://doi.org/10.1063/1.3658809
11. Afzalian, A., Akhoundi, E., Gaddemane, G., Duflou, R., Houssa, M.: Advanced DFT-NEGF transport techniques for novel 2-d material and device exploration including HfS_2/WSe_2 van der waals heterojunction TFET and WTe_2/WS_2 metal/semiconductor contact. IEEE Trans. Electron Devices **68**(11), 5372–5379 (2021). https://doi.org/10.1109/ted.2021.3078412
12. Reddy, P., Bryan, I., Bryan, Z., Tweedie, J., Washiyama, S., Kirste, R., Mita, S., Collazo, R., Sitar, Z.: Charge neutrality levels, barrier heights, and band offsets at polar AlGaN. Appl. Phys. Lett. **107**(9), 091603 (2015). https://doi.org/10.1063/1.4930026
13. Papior, N., Lorente, N., Frederiksen, T., García, A., Brandbyge, M.: Improvements on non-equilibrium and transport green function techniques: The next-generation transiesta. Comput. Phys. Commun. **212**, 8–24 (2017). https://doi.org/10.1016/j.cpc.2016.09.022
14. Datta, S.: Quantum Transport: Atom to Transistor. Appendix: advanced formalism, pp. 319–342. Cambridge University Press, Cambridge (2005)
15. Brandbyge, M., Mozos, J.-L., Ordejón, P., Taylor, J., Stokbro, K.: Density-functional method for nonequilibrium electron transport. Phys. Rev. B **65**, 165401 (2002). https://doi.org/10.1103/physrevb.65.165401
16. Neale, M.C., Hunter, M.D., Pritikin, J.N., Zahery, M., Brick, T.R., Kirkpatrick, R.M., Estabrook, R., Bates, T.C., Maes, H.H., Boker, S.M.: OpenMx 2.0: extended structural equation and statistical modeling. Psychometrika **81**(2), 535–549 (2016). https://doi.org/10.1007/s11336-014-9435-8
17. Kühne, T.D., Iannuzzi, M., Ben, M.D., Rybkin, V.V., Seewald, P., Stein, F., Laino, T., Khaliullin, R.Z., Schütt, O., Schiffmann, F., Golze, D., Wilhelm, J., Chulkov, S., Bani-Hashemian, M.H., Weber, V., Borštnik, U., Taillefumier, M., Jakobovits, A.S., Lazzaro, A., Pabst, H., Müller, T., Schade, R., Guidon, M., Andermatt, S., Holmberg, N., Schenter, G.K., Hehn, A., Bussy, A., Belleflamme, F., Tabacchi, G., Glöß, A., Lass, M., Bethune, I., Mundy, C.J., Plessl, C., Watkins, M., VandeVondele, J., Krack, M., Hutter, J.: CP2k: an electronic structure and molecular dynamics software package–quickstep: efficient and accurate electronic structure calculations. J. Chem. Phys. **152**(19), 194103 (2020). https://doi.org/10.1063/5.0007045
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P.: Numerical Recipes in Fortran, The Art of Scientific Computing. Cambridge University Press, Cambridge (1992)
19. Maciejko, J.: An introduction to nonequilibrium many-body theory. Springer (2007)
20. Laturia, A., Put, M.L.V., Vandenberghe, W.G.: Dielectric properties of hexagonal boron nitride and transition metal dichalcogenides: from monolayer to bulk. npj 2D Mater. Appl. **2**(1), 6 (2018). https://doi.org/10.1038/s41699-018-0050-x
21. Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G.L., Cococcioni, M., Dabo, I., Corso, A.D., Gironcoli, S., Fabris, S., Fratesi, G., Gebauer, R., Gerstmann, U., Gougoussis, C., Kokalj, A., Lazzeri, M., Martin-Samos, L., Marzari, N., Mauri, F., Mazzarello, R., Paolini, S., Pasquarello, A., Paulatto, L., Sbraccia, C., Scandolo, S., Sclauzero, G., Seitsonen, A.P., Smogunov, A., Umari, P., Wentzcovitch, R.M.: QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. J. Phys. Condensed Matter. **21**(39), 395502 (2009). https://doi.org/10.1088/0953-8984/21/39/395502
22. Marzari, N., Mostofi, A.A., Yates, J.R., Souza, I., Vanderbilt, D.: Maximally localized wannier functions: theory and applications. Rev. Mod. Phys. **84**(4), 1419–1475 (2012). https://doi.org/10.1103/revmodphys.84.1419
23. Galitskii, V.M., Migdal, A.B.: Application of quantum field theory methods to the many body problem. Sov. Phys. JETP **7**(96), 18 (1958)
24. Duflou, R., Pourtois, G., Houssa, M., Afzalian, A.: Fundamentals of Low-Resistive 2D-Semiconductor Metal Contacts: An Abi-initio NEGF Study. preprint on webpage at https://doi.org/10.21203/rs.3.rs-2202758/v1 (2022)
25. Lang, N.D., Kohn, W.: Theory of metal surfaces: induced surface charge and image potential. Phys. Rev. B **7**, 3541–3550 (1973). https://doi.org/10.1103/PhysRevB.7.3541
26. Srikantaiah, J.G., DasGupta, A.: Quantum mechanical effects in bulk mosfets from a compact modeling perspective: A review. IETE Tech. Rev. **29**(1), 3–28 (2012). https://doi.org/10.4103/0256-4602.93119
27. Ozaki, T.: Variationally optimized atomic orbitals for large-scale electronic structures. Phys. Rev. B **67**, 155108 (2003). https://doi.org/10.1103/physrevb.67.155108
28. Ozaki, T., Kino, H.: Numerical atomic basis orbitals from h to kr. Phys. Rev. B **69**, 195113 (2004). https://doi.org/10.1103/physrevb.69.195113