Unit cell restricted Bloch functions basis for first-principle transport models: Theory and application

M. G. Pala,[1] P. Giannozzi,[3,4] and D. Esseni[2]

[1]*Université Paris-Saclay, CNRS, Centre de Nanosciences et de Nanotechnologies, 10 Boulevard Thomas Gobert, 91120 Palaiseau, France*
[2]*DPIA, University of Udine, Via delle Scienze 206, 33100 Udine, Italy*
[3]*DMIF, University of Udine, Via delle Scienze 206, 33100 Udine, Italy*
[4]*CNR-IOM, Istituto dell'Officina dei Materiali, SISSA, I-34136 Trieste, Italy*

(Received 13 April 2020; revised 19 June 2020; accepted 19 June 2020; published 8 July 2020)

We present the theory and the application of a first-principle transport model employing a basis set obtained directly from the *ab initio* Bloch functions. We use a plane-wave density functional theory Hamiltonian and show that a judicious choice of the reduced basis set can effectively suppress the potentially thorny problem of the unphysical solutions. Our methodology enables *ab initio* transport simulations with a huge reduction of the size of the problem compared to the original *ab initio* formulation. Moreover, the approach can also be used for local and nonlocal empirical pseudopotential Hamiltonians, thus promising a wide range of possible applications. We report results for *ab initio* simulations of $\mathrm{MoS_2}$ field effect transistors, where the transport and electrostatics equations are solved self-consistently for channel lengths up to about 20 nanometers. The simulation results rapidly converge with the size of the basis set, so that the blocks of the Hamiltonian matrix can be reduced to a size below 100. Our methodology is a viable approach for *ab initio* and semiempirical quantum transport simulations and, in particular, it offers an alternative to the use of maximally localized Wannier functions.

DOI: 10.1103/PhysRevB.102.045410

I. INTRODUCTION

A quantum transport methodology relying on an *ab initio* description of the physical system is, in several respects, the frontier of the transport modeling in nanoscale systems. One reason why such an approach has become necessary to steer the technological developments is that the cross section of many devices has reached truly nanometric dimensions and the transistor length has reached the 10 nm range, where quantum transport effects become important, such as the source-drain tunneling in metal-oxide-semiconductor field-effect transistor (MOSFET) [1–3], or the band-to-band tunneling (BTBT) in tunnel FETs (TFETs) [4,5]. Moreover, a quantum transport formalism based on *ab initio* methods has become indispensable to explore the potentials of new device concepts exploiting the recently discovered atomically thin two-dimensional (2D) materials, and their many possible combinations in terms of vertical or lateral heterojunction options [6–8].

The first-principle electronic-structure calculations are typically based on density functional theory (DFT) and on either a plane-wave basis [9,10] or on the linear combination of atomic orbitals (LCAO) [11]. The former basis may be considered the most natural option for periodic crystals, while the latter is closely related to the chemical bonding picture.

The tight-binding method is the most widely used approximated implementation of the LCAO approach. It is also the most popular method for quantum transport based either on a fixed set of orbitals per atom with empirical coupling parameters (e.g., the $sp^3d^5s^*$ model [12,13]) or on maximally localized Wannier functions extracted as a postprocessing, sometimes quite delicate and computationally demanding, of first-principle calculations [14–18].

Plane waves form a complete set of orthogonal functions. They allow for a good control of accuracy and convergence in electronic-structure calculations through a cutoff of the kinetic energy (see, also, Sec. II). However, frequently they result in a large basis set, particularly for those supercells that include vacuum regions, where a good description of the exponential wave-function decay demands a large plane-wave set. Consequently, a direct use of first-principle calculations based on plane waves is often considered computationally prohibitive for electronic transport in technologically relevant systems. Some contributions have recently appeared for the empirical pseudopotentials method and use either a quantum-transmitting-boundary approach [19–22], or a nonequilibrium Green's function (NEGF) method [23,24]. Until now, DFT-based transport calculations in relatively large systems have been addressed only by using the LCAO basis [25] with the adoption of equivalent transport model techniques [26].

In this work, we present a method for quantum transport in nanoscale devices and physical systems, based on a plane-wave DFT Hamiltonian. This method employs a basis set of Bloch functions of the underlying system to drastically reduce the size of the transport problem. Our approach does not require the solution of any eigenvalue problem besides those addressed by first-principle calculations [27]; in fact, the basis set is obtained directly from the Bloch wave functions determined by the *ab initio* solver. An appropriate choice of Bloch functions allows us to effectively avoid the problem of unphysical solutions, whose filtering can be theoretically and computationally challenging [25,26]. We found that the size of the basis set for transport simulations is essentially independent upon the cutoff energy used in first-principle calculations, which is extremely beneficial because it allows one to decouple the size of the transport problem from the computational effort necessary to obtain full convergence and high accuracy in first-principle calculations.

Our results demonstrate that the Bloch functions form an extremely effective basis set, which enables band structure and transport calculations using a basis size that is hundreds of times smaller than the plane-wave basis used to calculate the Bloch functions in the *ab initio* solver. The ability of the Bloch functions to retain most of the physics with a small basis set is not surprising in consideration of the results obtained for band structure calculations with empirical pseudopotential models [28–32], and it is a precious asset for future developments of quantum transport methods based on a first-principle Hamiltonian. Recently, we became aware of a recent contribution [33], where a basis of Bloch functions was used for quantum transport simulations based on an empirical pseudopotential Hamiltonian and a quantum-transmitting-boundary approach.

The paper is organized as follows. In Sec. II, we provide the necessary information about the first-principle methods employed in our calculations and, in particular, we clarify the relevant connections to the transport model. In Sec. III, we introduce the reduced basis set used in this work, which consists of unit cell restricted Bloch functions. This section also illustrates several tests and comparisons which validate the basis in terms of the reconstruction of the *ab initio* electronic structure. Section IV presents the transport model based on the NEGF formalism, and the procedure to achieve simulations accounting for a self-consistent description of the electrostatics via the Poisson equation. Then, in Sec. V, we illustrate some examples of complete, self-consistent device simulations for an $\mathrm{MoS_2}$-based nanoscale transistor. Finally, in Sec. VI, we offer some concluding remarks.

II. *AB INITIO* HAMILTONIAN

Electronic-structure methods from first principles are typically based on density functional theory, where one-electron states ("Kohn-Sham orbitals") are obtained by self-consistently solving the Kohn-Sham equations,

$$H_{KS}\,\Psi_n = E_n\,\Psi_n, \quad H_{KS} = T + V_{scf}, \tag{1}$$

where the Kohn-Sham Hamiltonian H_{KS} is the sum of the kinetic energy T and the self-consistent potential V_{scf}. In turn, $V_{scf} = V_{eI} + V_H + V_{xc}$, where V_{eI} is the electron-ion interaction potential, V_H is the Hartree electrostatic potential, and V_{XC} is the "exchange-correlation" potential. The two latter terms depend upon the charge density, which can be written as the sum of the squares of all occupied Kohn-Sham orbitals.

Let us use a plane-wave basis set and pseudopotentials to represent the valence electron-nuclei interactions. The solution of the Kohn-Sham equations reduces to a secular problem, in which the potential is computed self-consistently. Leaving apart the problem of how to compute the charge density and the self-consistent potential, the only difference between Kohn-Sham and empirical-pseudopotential Hamiltonians is the presence of a nonlocal term in the Kohn-Sham Hamiltonian.

Both the Hartree and the exchange-correlation potentials, with the latter in the typical generalized gradient approximation (GGA) form, are *local* functions $v(\mathbf{r})$ of the position $\mathbf{r}$. The nonlocal term stems from atomic norm-conserving pseudopotentials, which contain two types of contributions: (a) a local $v_L(\mathbf{r})$ part with the expected asymptotic $v_L(\mathbf{r}) \sim -Z_v e^2/r$ behavior at large r (where e is the electron charge and Z_v is the number of valence electrons of the atom), and (b) a *nonlocal, short-ranged* $v_{NL}(\mathbf{r}, \mathbf{r}')$ part.

For each atom μ, the nonlocal term v^{μ}_{NL} can be expressed as a sum of N_μ *projectors*, defined via atomic pseudopotential parameters $\beta_n(\mathbf{r})$ and $D_{nn'}$ as follows:

$$v^{\mu}_{NL}(\mathbf{r}, \mathbf{r}') = \sum_{n,n'} \beta^{\mu}_n(\mathbf{r})\, D^{\mu}_{nn'} \left[\beta^{\mu}_{n'}(\mathbf{r}')\right]^*, \tag{2}$$

and, throughout this paper, we use a^* to denote the complex conjugate of a scalar a, and $\mathbf{M}^\dagger$ to denote the adjoint of a matrix or a vector $\mathbf{M}$. Equation (2) corresponds to the "separable" form of pseudopotentials. For the simple norm-conserving pseudopotentials used in this work, the $D_{nn'}$ matrix is diagonal: $D_{nn'} = D_n \delta_{nn'}$. The index n is a combined index, running on angular momentum quantum numbers l and m, up to the highest angular momentum values present in the atomic core. Typically, just a few projectors (<10) per atom need to be taken into account.

The $\beta^{\mu}_n(\mathbf{r})$ functions are short ranged and vanish for $r > r_c$, where r_c is the radius beyond which pseudoatomic and true atomic Kohn-Sham orbitals are the same. For most atoms, $r_c \sim 0.1 \div 0.3$ nm.

The nonlocal term in the Kohn-Sham Hamiltonian of a crystal is thus

$$\begin{aligned} V_{NL}(\mathbf{r}, \mathbf{r}') &= \sum_{\mu,\mathbf{R}} v^{\mu}_{NL}(\mathbf{r} - \mathbf{d}_\mu - \mathbf{R}, \mathbf{r}' - \mathbf{d}_\mu - \mathbf{R}) \\ &= \sum_{\mu,\mathbf{R}} \sum_{nn'} \beta^{\mu}_n(\mathbf{r} - \mathbf{d}_\mu - \mathbf{R})\, D^{\mu}_{nn'} \\ &\quad \times \left[\beta^{\mu}_{n'}(\mathbf{r}' - \mathbf{d}_\mu - \mathbf{R})\right]^*, \end{aligned} \tag{3}$$

where $\mathbf{d}_\mu$ is the position of atom μ in the unit cell, the $\mathbf{R}$'s are lattice vectors, and it is understood that

$$(V_{NL}\,\psi)(\mathbf{r}) = \int V_{NL}(\mathbf{r}, \mathbf{r}')\psi(\mathbf{r}')d\mathbf{r}'. \tag{4}$$

Kohn-Sham orbitals have the Bloch form and can be expanded into plane waves,

$$P_{\mathbf{k}+\mathbf{G}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} e^{i(\mathbf{k}+\mathbf{G})\cdot\mathbf{r}}, \tag{5}$$

where $\mathbf{k}$ is the Bloch vector, the $\mathbf{G}$'s are reciprocal lattice vectors, and Ω is the volume of the crystal. A finite set is obtained by choosing plane waves up to a given kinetic energy value E_w (the "cutoff"): $\frac{\hbar^2}{2m_0}(\mathbf{k}+\mathbf{G})^2 \leqslant E_w$, where m_0 is the electron mass. The Kohn-Sham Hamiltonian can be expanded into plane waves as well:

$$\begin{aligned} \langle \mathbf{k}+\mathbf{G}|H|\mathbf{k}+\mathbf{G}'\rangle \equiv H_{\mathbf{k}}(\mathbf{G}, \mathbf{G}') &= \frac{\hbar^2}{2m}(\mathbf{k}+\mathbf{G})^2 \delta_{\mathbf{G},\mathbf{G}'} \\ &\quad + V_L(\mathbf{G}-\mathbf{G}') + V_{NL}(\mathbf{k}+\mathbf{G}, \mathbf{k}+\mathbf{G}'), \end{aligned} \tag{6}$$

where $V_L(\mathbf{G}-\mathbf{G}')$ is the Fourier transform of the local part of the total potential (local pseudopotential plus Hartree and exchange potential). The nonlocal contribution, coming from the pseudopotential, is

$$V_{NL}(\mathbf{k}+\mathbf{G},\mathbf{k}+\mathbf{G}') = \frac{1}{\Omega}\int V_{NL}(\mathbf{r},\mathbf{r}')e^{-i(\mathbf{k}+\mathbf{G})\mathbf{r}}e^{i(\mathbf{k}+\mathbf{G}')\mathbf{r}'}\,d\mathbf{r}\,d\mathbf{r}'. \tag{7}$$

In principle, the solution of the secular equation for the Kohn-Sham Hamiltonian matrix,

$$\sum_{\mathbf{G}'} H_{\mathbf{k}}(\mathbf{G},\mathbf{G}')\,B(\mathbf{G}') = E\,B(\mathbf{G}), \tag{8}$$

provides the electronic structure $E_n(\mathbf{k})$ and the Bloch functions $\Psi_{n\mathbf{k}}(\mathbf{r})$, which are completely determined by the eigenvectors $B_{n\mathbf{k}}(\mathbf{G})$. In practice, $H_{\mathbf{k}}(\mathbf{G},\mathbf{G}')$ is a very large matrix and thus it is not stored or directly diagonalized, but rather one resorts to iterative techniques and to on-the-fly computation of $H_{\mathbf{k}}\Psi$ products exploiting fast Fourier transform techniques [34]. The potential and the charge density contain plane waves up to a cutoff energy $E_\rho = 4E_w$.

In our calculations, we used an orthorhombic unit cell (see examples in Sec. III B), where the real-space unit vectors can be written as $\mathbf{a_1}{=}(a_x,0,0)$, $\mathbf{a_2}{=}(0,a_y,0)$, $\mathbf{a_3}{=}(0,0,a_z)$, with x being the transport direction. Hence, the unit vectors of the reciprocal lattice are $\mathbf{b_1}{=}(2\pi/a_x,0,0)$, $\mathbf{b_2}{=}(0,2\pi/a_y,0)$, $\mathbf{b_3}{=}(0,0,2\pi/a_z)$, and the reciprocal lattice vectors are $\mathbf{G}{=}n_x\mathbf{b_1}{+}n_y\mathbf{b_2}{+}n_z\mathbf{b_3}$ (with n_x, n_y, $n_z = 0, \pm1, \pm2, \ldots$). The Brillouin zone can be taken as the parallelepiped defined by the conditions $-\pi/a_s < k_s \leqslant \pi/a_s$ (with $s = x, y, z$), that has a volume $\Omega_{RZ} = (2\pi)^3/\Omega_{\text{cell}}$, with $\Omega_{\text{cell}} = a_xa_ya_z$ being the volume of the unit cell. It is understood that for a 2D crystal in the (x, y) plane, for example, the unit cell includes a relatively large vacuum region in the z direction that makes the extension of the reduced zone along z practically negligible, thus resulting in a 2D electron gas. Ultrathin films or nanowires consisting of an underlying 3D crystal can be similarly described as a 2D or 1D system by inserting vacuum regions in the unit cell.

III. REDUCED BASIS OF UNIT CELL RESTRICTED BLOCH FUNCTIONS

In our methodology for transport simulations, the expansion volume for $\mathbf{G}$ vectors is given by the cube inscribed to the sphere used in *ab initio* calculations, namely, the cube set by the condition $\frac{\hbar^2}{2m_0}G_s^2 \leqslant E_\rho/3$ with $s = x, y, z$. The grid of $\mathbf{G}$ vectors naturally defines a corresponding grid of points in real space and, if we denote by N_{Gs} the number of $\mathbf{G}_s$ vectors (with $\mathbf{G}_x$, $\mathbf{G}_y$, and $\mathbf{G}_z$ lying, respectively, along the x, y and z axis), the spacing of the grid in real space is $d_s = a_s/N_{ds}$, where $N_{ds} = N_{Gs}$ is the number of grid points along s inside the unit cell.

In the remainder of the paper, we will often refer to the Hamiltonian matrix in different basis sets. Matrices are denoted by using capital letters in square brackets and, whenever necessary, a subscript indicates the basis. For example, $[\mathbf{H}]_{\mathbf{K}}$, $[\mathbf{H}]_{x\mathbf{K}_{yz}}$, and $[\mathbf{H}]_{\Phi}$ denote the Hamiltonian matrix, respectively, in the plane-wave basis used in Sec. II, in the hybrid $x\mathbf{K}_{yz}$ basis described in Sec. III A, and in the reduced basis of Bloch functions discussed in Sec. III B. The subscript may be omitted to lighten the notation when there is no ambiguity about the basis set. When we refer to the elements of the matrices, instead, we drop the square brackets and the subscript because the symbols used for the elements identify the basis: for example, we write $\mathbf{H}(x\mathbf{K}_{yz}, x'\mathbf{K}'_{yz})$ to denote the elements of the Hamiltonian matrix $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in the $x\mathbf{K}_{yz}$ basis. Finally, we use curly brackets to denote column or row vectors: for example, $\{\Psi\}$ will be used for the column vectors representing the wave functions.

A. Hybrid $x\mathbf{K}_{yz}$ basis

Let us now consider a system having N_{cx} unit cells in the transport direction x and subject to periodic boundary conditions. The Hamiltonian matrices in the plane-wave basis are given by Eq. (6) and are identified by $\mathbf{k} = (k_x,\mathbf{k}_{yz})$, with $\mathbf{k}_{yz} = (k_y,k_z)$. The corresponding Hamiltonian matrix in the hybrid basis, $x\mathbf{K}_{yz}$, consisting of real-space grid points along x and plane waves in the (y, z) directions, can be obtained by using the unitary transformation described below. Before discussing the transformation, however, here we notice that the short-range nature of the nonlocal pseudopotential defined in Eqs. (2) and (3) implies that the nonlocal terms practically vanish on distances much smaller than a_x, as already mentioned in Sec. II. In particular, we numerically verified that the Hamiltonian in the $x\mathbf{K}_{yz}$ basis can be accurately expressed by the block tridiagonal form,

$$[\mathbf{H}]_{x\mathbf{K}_{yz}} = \begin{bmatrix} \mathbf{H}_{0,0} & \mathbf{H}_{0,1} & 0 & 0 & \cdots & \mathbf{H}_{0,1}^\dagger \\ \mathbf{H}_{0,1}^\dagger & \mathbf{H}_{0,0} & \mathbf{H}_{0,1} & 0 & \cdots & 0 \\ \cdots & & \cdots & & \ddots & \vdots \\ \mathbf{H}_{0,1} & 0 & \cdots & 0 & \mathbf{H}_{0,1}^\dagger & \mathbf{H}_{0,0} \end{bmatrix}, \tag{9}$$

where each block describes an a_x long region consisting of N_{dx} discretization points $x_j{=}0, d_x, 2d_x \cdots (a_x{-}d_x)$, so that blocks $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$ have a rank $N_G = N_{dx}N_{Gy}N_{Gz}$ (with $N_{dx} = N_{Gx}$).

The $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (9) corresponds to a given $\mathbf{k}_{yz}$ and it has N_{cx}^2 blocks. The matrix is sorted so that for each discretization point x_j, we have all the $N_{Gy}N_{Gz}$ entries corresponding to the spectral components $\mathbf{K}_{yz} = \mathbf{k}_{yz}{+}\mathbf{G}_{yz}$ [with $\mathbf{G}_{yz} = (G_y,G_z)$].

According to Eq. (9), the knowledge of $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ coincides with the knowledge of $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$. In Appendix A, we discuss in more detail the structure of $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (9). In particular, we argue that the short-range nonlocal pseudopotentials make the off-diagonal blocks other than $\mathbf{H}_{0,1}(i,j)$ negligible, and result in an $\mathbf{H}_{0,1}$ block that is a lower triangular matrix, namely, we have $\mathbf{H}_{0,1}(i,j) \simeq 0$ for $j \geqslant i$.

For a system having N_{cx} unit cells along x, the $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (9) can be obtained by transforming from K_x to x the Hamiltonian matrices $\mathbf{H}_{\mathbf{k}}$ in Eq. (6) for the corresponding N_{cx} wave vectors $\mathbf{k}{=}(k_x,\ \mathbf{k}_{yz})$, with k_x values spaced by $2\pi/(N_{cx}a_x)$. However, we argue that $N_{cx}{=}\ 2$ is sufficient to determine $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$. In fact, for $N_{cx}{=}\ 2$, we can rewrite Eq. (9) as

$$[\mathbf{H}]_{x\mathbf{K}_{yz}} = \begin{bmatrix} \mathbf{H}_{0,0} & \mathbf{H}_{0,1} + \mathbf{H}_{0,1}^\dagger \\ \mathbf{H}_{0,1}^\dagger + \mathbf{H}_{0,1} & \mathbf{H}_{0,0} \end{bmatrix}, \tag{10}$$

and Eq. (10) allows us to unambiguously determine $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$ in virtue of the above-mentioned property $\mathbf{H}_{0,1}(i,j) \simeq$

0 for $j \geqslant i$. The $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (10) can be obtained by transforming the two matrices $\mathbf{H}_{\mathbf{k}_0}$, $\mathbf{H}_{\mathbf{k}_1}$ given by Eq. (6) for $\mathbf{k}_0{=}(k_{x0},\, \mathbf{k}_{yz})$ and $\mathbf{k}_1{=}(k_{x1},\, \mathbf{k}_{yz})$, with $k_{x0}{=}0$ and $k_{x1}{=}\pi/a_x$.

For any $\mathbf{k}{=}(k_x,\, \mathbf{k}_{yz})$, the transformation from K_x to the home unit cell given by $x_j = 0, d_x, \cdots (a_x{-}d_x)$ is governed by the $N_G \times N_G$ unitary matrix [35]

$$[\mathbf{U}^0_{k_x}] = \frac{1}{\sqrt{N_{dx}}}\begin{bmatrix} \mathbf{I} & e^{i(k_x+G_{x,1})d_x}\mathbf{I} & \cdots & e^{i(k_x+G_{x,1})(a_x-d_x)}\mathbf{I} \\ \mathbf{I} & e^{i(k_x+G_{x,2})d_x}\mathbf{I} & \cdots & e^{i(k_x+G_{x,2})(a_x-d_x)}\mathbf{I} \\ \cdots & \cdots & \ddots & \vdots \\ \mathbf{I} & e^{i(k_x+G_{x,N_{Gx}})d_x}\mathbf{I} & \cdots & e^{i(k_x+G_{x,N_{Gx}})(a_x-d_x)}\mathbf{I} \end{bmatrix}, \tag{11}$$

where $\mathbf{I}$ is an identity matrix with rank $N_{Gy}N_{Gz} = N_G/N_{dx}$ and $N_{dx} = N_{Gx}$. Hence, the corresponding transformation to the unit cell p extending from $x_j = pa_x$ to $x_j = [(p+1)a_x{-}d_x]$ (with $p = 1, 2, \dots$) is governed by the matrix $[\mathbf{U}^p_{k_x}]{=}[\mathbf{U}^0_{k_x}]\, e^{ik_x\, p\, a_x}$. Consequently, the transformation from K_x to the two unit cells necessary to calculate $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (10) can be written as

$$[\mathbf{U}^{(2a_x)}_{k_x}] = \frac{1}{\sqrt{2}}[[\mathbf{U}^0_{k_x}], [\mathbf{U}^0_{k_x}]e^{ik_x\, a_x}]. \tag{12}$$

By using Eq. (12), we can express the $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (10) in terms of the two plane-wave DFT Hamiltonian matrices $[\mathbf{H}_{\mathbf{k}_0}]$, $[\mathbf{H}_{\mathbf{k}_1}]$ as

$$[\mathbf{H}]_{x\mathbf{K}_{yz}} = \sum_{\mathbf{k}=\mathbf{k}_0,\mathbf{k}_1} [\mathbf{U}^{(2a_x)}_{k_x}]^\dagger\, [\mathbf{H}_\mathbf{k}]\, [\mathbf{U}^{(2a_x)}_{k_x}] = \frac{1}{2}\sum_{\mathbf{k}=\mathbf{k}_0,\mathbf{k}_1} \begin{bmatrix} [\mathbf{U}^0_{k_x}]^\dagger[\mathbf{H}_\mathbf{k}][\mathbf{U}^0_{k_x}] & [\mathbf{U}^0_{k_x}]^\dagger[\mathbf{H}_\mathbf{k}][\mathbf{U}^0_{k_x}]\, e^{ik_x\, a_x} \\ [\mathbf{U}^0_{k_x}]^\dagger\, [\mathbf{H}_\mathbf{k}][\mathbf{U}^0_{k_x}]\, e^{-ik_x\, a_x} & [\mathbf{U}^0_{k_x}]^\dagger[\mathbf{H}_\mathbf{k}][\mathbf{U}^0_{k_x}] \end{bmatrix}, \tag{13}$$

where, for $\mathbf{k}_0{=}(0,\, \mathbf{k}_{yz})$ and $\mathbf{k}_1{=}(\pi/a_x,\, \mathbf{k}_{yz})$, the exponential terms simply evaluate to ±1. Hence, we can finally identify $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$ in the hybrid $x\mathbf{K}_{yz}$ basis as [36]

$$[\mathbf{H}_{0,0}] = \tfrac{1}{2}\big([\mathbf{U}^0_{k_{x0}}]^\dagger[\mathbf{H}_{\mathbf{k}_0}][\mathbf{U}^0_{k_{x0}}] + [\mathbf{U}^0_{k_{x1}}]^\dagger[\mathbf{H}_{\mathbf{k}_1}][\mathbf{U}^0_{k_{x1}}]\big), \tag{14a}$$

$$[\mathbf{H}_{0,1}] + [\mathbf{H}_{0,1}]^\dagger = \tfrac{1}{2}\big([\mathbf{U}^0_{k_{x0}}]^\dagger[\mathbf{H}_{\mathbf{k}_0}][\mathbf{U}^0_{k_{x0}}] - [\mathbf{U}^0_{k_{x1}}]^\dagger[\mathbf{H}_{\mathbf{k}_1}][\mathbf{U}^0_{k_{x1}}]\big). \tag{14b}$$

In Appendix A, we discuss a generalization of Eqs. (12)–(14) that can be used to build the Hamiltonian in the hybrid $x\mathbf{K}_{yz}$ basis by using any number N_{cx} of k_x Bloch vectors.

B. Unit cell restricted Bloch functions

The blocks $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$ in Eq. (9) have an N_G rank so that, for high-precision *ab initio* calculations using a relatively large cutoff energy E_ρ, the direct manipulation of such matrices for transport simulations is practically intractable. A drastic reduction of the size of $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$ can be achieved by moving to an appropriate basis set consisting of Bloch functions restricted to a unit cell. To this purpose, we first argue that thanks to the block tridiagonal form of the Hamiltonian in Eq. (9), the corresponding Bloch functions in the unit cell p [with $p{=}0, 1, \cdots (N_{cx}{-}1)$] take the form [37]

$$\Psi_\mathbf{k}(x_j + p\, a_x, \mathbf{G}_{yz}) = \Psi^0_\mathbf{k}(x_j, \mathbf{G}_{yz})\, e^{i\, k_x p\, a_x}, \tag{15}$$

where the $\{\Psi^0_\mathbf{k}\}$ are Bloch functions restricted to the home unit cell [i.e., for $x_j = 0, d_x, \cdots (a_x{-}d_x)$], which are in turn the solutions of N_{cx} eigenvalue problems,

$$[\ \mathbf{H}^\dagger_{0,1}\, e^{-ik_x a_x} + \mathbf{H}_{0,0} + \mathbf{H}_{0,1}\, e^{ik_x a_x}\ ]\, \{\Psi^0_\mathbf{k}\} = E(\mathbf{k})\, \{\Psi^0_\mathbf{k}\}, \tag{16}$$

with $\mathbf{k}{=}(k_x{,}\mathbf{k}_{yz})$. Equation (16) requires that the $\Psi^0_\mathbf{k}(x_j, \mathbf{G}_{yz})$ functions fulfill the k_x-dependent boundary condition $\Psi^0_\mathbf{k}(a_x, \mathbf{G}_{yz}){=}\Psi^0_\mathbf{k}(0, \mathbf{G}_{yz})\, e^{i\, k_x\, a_x}$.

The reduced basis employed in this paper is identified as a single basis set suitable for all the eigenvalue problems in Eq. (16). In this respect, we recall that the Bloch functions $\{\Psi^0_\mathbf{k}\}$ are in effect known because they are determined by the eigenvectors $B_{n\mathbf{k}}(\mathbf{G})$ of the secular Eq. (8) solved in *ab initio* calculations. We can express the column vectors $\{\Psi^0_{n\mathbf{k}}\}$ in matrix notation as

$$\{\Psi^0_{n\mathbf{k}}\} = [\mathbf{U}^0_{k_x}]^\dagger\{B_{n\mathbf{k}}\}, \tag{17}$$

where $[\mathbf{U}^0_{k_x}]$ is defined in Eq. (11) and $\mathbf{k}{=}(k_x{,}\mathbf{k}_{yz})$. Our reduced basis set consists of a subset of the $\{\Psi^0_{n\mathbf{k}}\}$ corresponding to a few k_x values in the reduced zone $-\pi/a_x < k_x \leqslant \pi/a_x]$ and, for each k_x, to some tens of energies $E_n(\mathbf{k})$. There is substantial flexibility in such a definition of the basis set, which can be used to find a good compromise between the size N_B of the basis and the accuracy in the reconstruction of the electronic structure. In general, the size of the basis can be written as

$$N_B = \sum_{i=1}^{N_{kB}} N_E(k_{x,i}), \tag{18}$$

where N_{kB} denotes the number of k_x values and $N_E(k_{x,i})$ denotes the number of energies at $k_{x,i}$ included in the basis.

Appendix B offers more details about the choice of the basis functions. Most of the calculations and simulations reported in this paper were obtained by using either two k_x values (i.e., $k_x{=}0$, π/a_x) or four k_x values (i.e., $k_x{=}0$, $\pm 0.5\,\pi/a_x$, π/a_x), as exemplified by the results in Fig. 1 discussed below.

Because the $\{\Psi^0_{n\mathbf{k}}\}$ for different k_x are not orthogonal over a_x, we apply an orthonormalization procedure. Here, we remark that the use of an orthonormal basis is not really necessary, but it is, in fact, very convenient in the

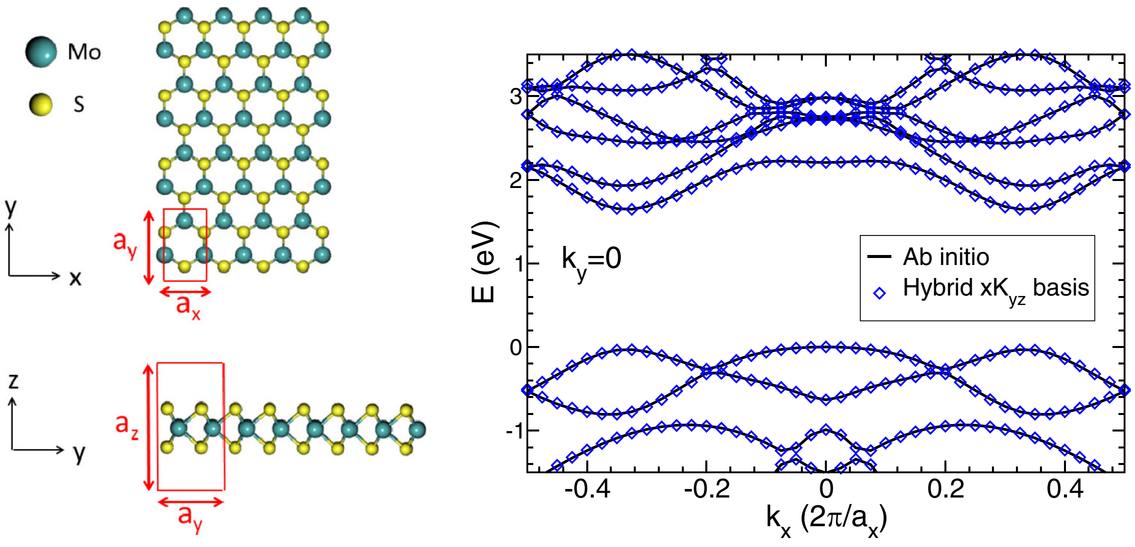

FIG. 1. Left: Cross section and the top view of the unit cell for the single-layer MoS_2 employed in this work. Right: Electronic structure for the single-layer MoS_2 vs k_x and for k_y=0. *Ab initio* calculations (solid lines) are compared to results obtained by using Eq. (16) in the hybrid $x\mathbf{K}_{yz}$ basis (diamonds).

transformations from the reduced basis to real space that are necessary, for example, for the calculation of space charge density in self-consistent simulations (see Sec. IV A). The orthonormalization procedure simply consists of an appropriate linear combination of the Bloch functions $\{\Psi^0_{n\mathbf{k}}\}$, but it is prone to numerical instabilities. We overcame this problem by employing the modified Gram-Schmidt algorithm, which is more robust than the standard approach in dealing with rounding errors [38]. Since the $\{\Psi^0_{n\mathbf{k}}\}$ are column vectors with N_G components and we select only a number N_B of basis functions much smaller than N_G, the orthonormalization procedure can always be successfully completed.

We will denote by $\Phi_m(x_j, \mathbf{G}_{yz})$ the orthonormalized Bloch functions restricted to a unit cell, with $m = 1, 2, \cdots N_B$ and $x_j = 0, d_x, 2d_x \cdots (a_x - d_x)$. The rectangular transformation matrix from the hybrid $x\mathbf{K}_{yz}$ basis to the Φ basis is defined as

$$[\mathbf{U}_\Phi] = [\,\{\Phi_1\}, \{\Phi_2\} \cdots \{\Phi_{N_B}\}\,], \tag{19}$$

and $[\mathbf{U}_\Phi]$ has N_B columns and N_G rows. Each $\{\Phi_m\}$ is a column vector such that for each discretization point x_j=0, $d_x, 2d_x \cdots (a_x - d_x)$, we have all the $N_{Gy}N_{Gz}$ spectral components. Because the $\{\Phi_m\}$ have been orthonormalized, we have $[\mathbf{U}_\Phi]^\dagger\,[\mathbf{U}_\Phi]$=$[\mathbf{I}]_{N_B}$, with $[\mathbf{I}]_{N_B}$ being the identity matrix with rank N_B.

By recalling the expression in Eqs. (14) for the $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$ in the $x\mathbf{K}_{yz}$ basis, we can readily write $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$ in the reduced basis as

$$\begin{aligned}[\mathbf{H}_{0,0}]_\Phi &= \tfrac{1}{2}([\mathbf{W}_{k_{x0}}]^\dagger[\mathbf{H}_{\mathbf{k}_0}][\mathbf{W}_{k_{x0}}]\\ &\quad + [\mathbf{W}_{k_{x1}}]^\dagger[\mathbf{H}_{\mathbf{k}_1}][\mathbf{W}_{k_{x1}}]), \qquad (20a)\\ [\mathbf{H}_{0,1}]_\Phi + [\mathbf{H}_{0,1}]^\dagger_\Phi &= \tfrac{1}{2}([\mathbf{W}_{k_{x0}}]^\dagger[\mathbf{H}_{\mathbf{k}_0}][\mathbf{W}_{k_{x0}}]\\ &\quad - [\mathbf{W}_{k_{x1}}]^\dagger[\mathbf{H}_{\mathbf{k}_1}][\mathbf{W}_{k_{x1}}]), \qquad (20b)\end{aligned}$$

where we have introduced transformation matrices $[\mathbf{W}_{k_{x0}}] = [\mathbf{U}^0_{k_{x0}}][\mathbf{U}_\Phi]$, $[\mathbf{W}_{k_{x1}}] = [\mathbf{U}^0_{k_{x1}}][\mathbf{U}_\Phi]$.

Equations (20) express $[\mathbf{H}_{0,0}]_\Phi$, $[\mathbf{H}_{0,1}]_\Phi$ directly in terms of the two plane-wave DFT Hamiltonian matrices $[\mathbf{H}_{\mathbf{k}_0}]$, $[\mathbf{H}_{\mathbf{k}_1}]$. The expressions for $[\mathbf{H}_{0,0}]_\Phi$, $[\mathbf{H}_{0,1}]_\Phi$, in turn, allow us to transform the Hamiltonian in Eq. (9) and the eigenvalue problems in Eq. (16) to the reduced Φ basis, where the size of such blocks is much smaller than N_G, as exemplified below.

Figure 1(a) illustrates the unit cell of the single-layer MoS_2 that we used in this work as a baseline material for electronic-structure calculations and device simulations; the unit vectors are $\mathbf{a_1}$=$(a_x, 0, 0)$, $\mathbf{a_2}$=$(0, a_y, 0)$, $\mathbf{a_3}$=$(0, 0, a_z)$ with a_x = 3.18818 Å, a_y = 5.52208 Å, and a_z = 20.2 Å, while the vertical distance between the Mo and S atoms is 1.564 Å. This cell was built by expanding the relaxed primitive unit cell of the monolayer MoS_2, whose band structure along the high-symmetry points of the primitive Brillouin zone matches very well with the results reported in Ref. [39] (not shown). Figure 1(b) reports the corresponding band structure obtained from the DFT Hamiltonian in the plane-wave basis (solid line). The DFT calculation was performed by means of the QUANTUM ESPRESSO code [10], using a norm-conserving pseudopotential [40], and the Perdew-Burke-Ernzerhof [41] (PBE) approximation to the exchange-correlation functional. The self-consistent solution was obtained by employing a 15 × 12 × 1 Monkhorst-Pack k-points grid and a cutoff energy of E_w = 90 Ryd, resulting in a number on plane waves N_G = 26 733. Figure 1(b) also shows the band structure obtained with Eq. (16) in the hybrid $x\mathbf{K}_{yz}$ basis (symbols), still having a size N_G = 26 733.

Figure 2 addresses the reconstruction of the electronic structure in the reduced basis Φ, and in particular it reports the absolute energy difference ΔE (x axis) between the energies obtained by using Eq. (16), either in the reduced basis (i.e., for $[\mathbf{H}_{0,0}]_\Phi$, $[\mathbf{H}_{0,1}]_\Phi$ with an N_B size) or in the complete $x\mathbf{K}_{yz}$ basis (i.e., for $[\mathbf{H}_{0,0}]$, $[\mathbf{H}_{0,1}]$ with an N_G size). The results of Fig. 2(a) were obtained by using two k_x values to build the reduced basis (i.e., N_{kB}=2 and k_x=0, π/a_x), while those in Fig. 2(b) correspond to N_{kB}=4 (i.e., k_x=0, $\pm 0.5\pi/a_x$, π/a_x).

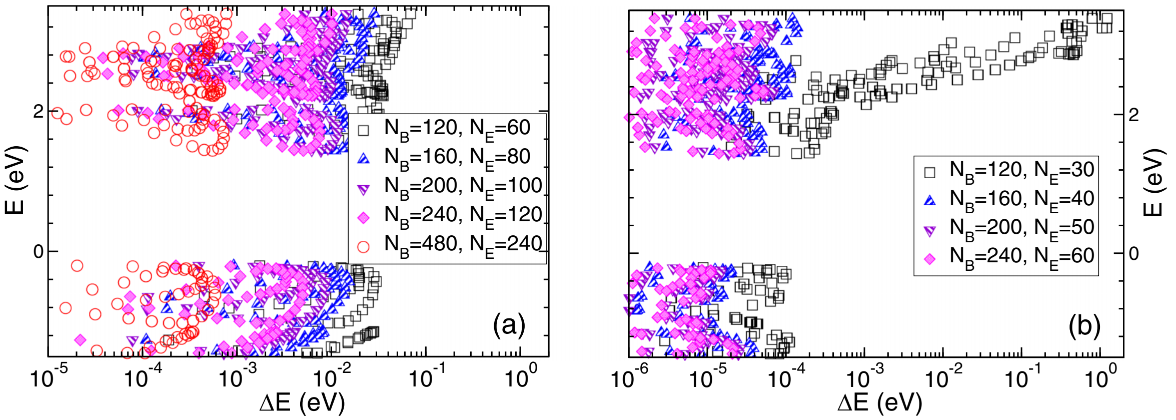

FIG. 2. Absolute energy difference ΔE between the electronic structure calculated either in the hybrid $x\mathbf{K}_{yz}$ basis [i.e., Eq. (16) with $[\mathbf{H}_{0,0}]$, $[\mathbf{H}_{0,1}]$ blocks] or in the reduced basis [i.e., Eq. (16) with $[\mathbf{H}_{0,0}]_\Phi$, $[\mathbf{H}_{0,1}]_\Phi$ blocks]. (a) ΔE for reduced basis calculations obtained with two k_x values (i.e., k_x=0, π/a_x) and different number N_E of basis functions at each k_x. (b) Same as (a), but for reduced basis calculations obtained with four k_x values (i.e., k_x=0, $\pm 0.5\pi/a_x$, π/a_x).

For a given N_{kB}, the accuracy of the reduced basis improves by increasing the number N_E of Bloch functions at each k_x. As it can be seen, the error for $N_B = 160$ is small enough for most applications in both Fig. 2(a) and Fig. 2(b), which enables a drastic reduction of the problem size compared to $N_G = 26\,733$. For N_B=160 and 240, we also see that for a given N_B, the reconstruction of the electronic structure improves by increasing N_{kB}. This behavior is not unexpected because the $\{\Psi^0_{n\mathbf{k}}\}$ are continuous functions of k_x, so that a linear combination of the $\{\Psi^0_{n\mathbf{k}}\}$ for a few k_x can still approximate well the remaining Bloch functions. Figure 2 shows, however, that very few k_x values are sufficient to achieve a close agreement with the reference results.

We finally notice in Fig. 2(b) that for N_B =120, the energy error in the conduction band steeply increases for energies above about 2 eV. This is because in this system we have 26 valence bands, all of which are included in the reduced basis set. Consequently, for N_E =30, the basis set includes the Bloch functions for only the four lowest conduction bands, which results in relatively large errors for higher conduction bands. In this respect, we verified that by including in the basis set the Bloch functions for all the valence bands at each k_x, we can effectively suppress the problem of unphysical solutions [25,26]. Hence, the number of valence bands in the system sets a lower bound for the size N_B of the reduced basis.

As for the unphysical states that have been observed sometimes upon the introduction of a reduced basis, we have extended our analysis by inspecting the transmission across a single-layer MoS_2 in the flat band condition, namely, with neither built-in nor externally applied potential. The transmission calculated by using the reduced basis (not shown) is very steeply, exponentially suppressed in the energy gap. In other words, we do not observe any feature of the transmission indicating the presence of evanescent, spurious states in the energy gap. Moreover, for energies belonging to the valence or the conduction band, the transmission at a given lateral wave vector k_y equals, as expected, the number of available bands. The inspection of the transmission is a good complement to the analysis of the electronic states, and it reinforced our confidence in an effective suppression of unphysical effects in our reduced basis calculations.

IV. NEGF BASED TRANSPORT MODEL

In this section, we describe the procedure based on the NEGF formalism to achieve a self-consistent simulation of a nanodevice or a mesoscopic system subject to external bias conditions. All the relevant physical quantities, such as density of states, carrier concentration, and currents, were computed in terms of retarded and Green's function matrices, which are calculated in the reduced Bloch function basis.

A. Charge, current, and self-consistent calculations

In the Bloch function basis, the retarded (advanced), $[\mathbf{G}^{r(a)}]_\Phi$, and lesser- (greater)-than Green's functions, $[\mathbf{G}^{<(>)}]_\Phi$ at a given energy E are defined as

$$[\mathbf{G}^{r(a)}]_\Phi = [(E + i\eta)[\mathbf{I}]_\Phi - [\mathbf{H}]_\Phi - [\mathbf{\Sigma}^{r(a)}]_\Phi]^{-1} \tag{21}$$

and

$$[\mathbf{G}^{<(>)}]_\Phi = [\mathbf{G}^r]_\Phi[\mathbf{\Sigma}^{<(>)}]_\Phi[\mathbf{G}^a]_\Phi, \tag{22}$$

where η is a positive (negative) infinitesimal, $[\mathbf{\Sigma}^{r(a)}]_\Phi = [\mathbf{\Sigma}^{r(a)}_L]_\Phi + [\mathbf{\Sigma}^{r(a)}_R]_\Phi + [\mathbf{\Sigma}^{r(a)}_{ph}]_\Phi$ and $[\mathbf{\Sigma}^{<(>)}]_\Phi = [\mathbf{\Sigma}^{<(>)}_L]_\Phi + [\mathbf{\Sigma}^{<(>)}_R]_\Phi + [\mathbf{\Sigma}^{<(>)}_{ph}]_\Phi$ are the retarded (advanced) and the lesser- (greater)-than self-energies describing the connection to contacts (i.e., left lead L and right lead R), or possible interaction with photons or phonons [42]. Thanks to the block-tridiagonal structure of the Hamiltonian matrix, the submatrices of the retarded (advanced) and lesser- (greater)-than Green's functions that are needed to calculate carrier concentrations and current density can be efficiently computed with well-known recursive algorithms [43], and by manipulating matrix blocks of rank N_B.

More precisely, in order to calculate the 3D real-space concentration of mobile carriers, we need to compute the diagonal terms of the real-space Green's functions starting from the Green's functions in the Bloch functions basis. To this end, we first transform the Green's functions from the Bloch functions to the $x\mathbf{K}_{yz}$ basis, and then compute the charge in real space. Inside each unit cell, the Green's functions in the $x\mathbf{K}_{yz}$ basis can be concisely written in matrix notation as $[\mathbf{G}^{<(>)}]_{x\mathbf{K}_{yz}}=[U_\Phi]\,[\mathbf{G}^{<(>)}]_\Phi\,[U_\Phi]^\dagger$, with $[U_\Phi]$ given by Eq. (19). An explicit expression is given by

$$\begin{aligned}&\mathbf{G}^{<(>)}(x_j\mathbf{K}_{yz}, x_j\mathbf{K}'_{yz};E)\\ &\quad=\sum_{n,m=1}^{N_B}\mathbf{G}^{<(>)}(n,m;E)\,\Phi_n(x_j,\mathbf{G}_{yz})\,\Phi_m^*(x_j,\mathbf{G}'_{yz}).\end{aligned}\tag{23}$$

Then, we evaluated the free electron concentration on the fine mesh grid with discretization steps (d_x, d_y, d_z) as

$$\begin{aligned}n(x_j,\mathbf{r}_{yz}) &= \frac{-i}{d_xd_yd_z}\int_{E_0(x_j)}^{\infty}\frac{dE}{2\pi N_{Gy}N_{Gz}}\\ &\quad\times\sum_{\mathbf{G}_{yz},\mathbf{G}'_{yz}}\mathbf{G}^<(x_j\mathbf{K}_{yz},x_j\mathbf{K}'_{yz};E)\,e^{i(\mathbf{G}_{yz}-\mathbf{G}'_{yz})\cdot\mathbf{r}_{yz}},\end{aligned}\tag{24}$$

and, similarly for the free hole concentration,

$$\begin{aligned}p(x_j,\mathbf{r}_{yz}) &= \frac{i}{d_xd_yd_z}\int_{-\infty}^{E_0(x_j)}\frac{dE}{2\pi N_{Gy}N_{Gz}}\\ &\quad\times\sum_{\mathbf{G}_{yz},\mathbf{G}'_{yz}}\mathbf{G}^>(x_j\mathbf{K}_{yz},x_j\mathbf{K}'_{yz};E)\,e^{i(\mathbf{G}_{yz}-\mathbf{G}'_{yz})\cdot\mathbf{r}_{yz}},\end{aligned}\tag{25}$$

with $\mathbf{K}_{yz}=(\mathbf{k}_{yz}+\mathbf{G}_{yz})$ and with $E_0(x_j)$ being the neutrality point that we assumed to be at the center of the energy band gap.

It is understood that all equations in this section refer to a given $\mathbf{k}_{yz}$ and that if the system is periodic along either y or z, a sum over $\mathbf{k}_{yz}$ is necessary to calculate all physical quantities.

In order to simulate the transport properties of realistic devices, it is necessary to evaluate the electrostatic potential induced by external biases, ionized dopants, and mobile carriers. Such an electrostatic potential $\phi(\mathbf{r})$ can be accurately described within the Hartree approximation, namely, by self-consistently solving the equations for the Green's functions [that in turn give the carrier concentrations via Eqs. (24) and (25)], with the 3D Poisson equation

$$\nabla\cdot[\epsilon(\mathbf{r})\nabla\phi(\mathbf{r})] = -e[p(\mathbf{r})-n(\mathbf{r})+N_D(\mathbf{r})-N_A(\mathbf{r})],\tag{26}$$

where $\epsilon(\mathbf{r})$ is the material-dependent permittivity, and $n(\mathbf{r})$, $p(\mathbf{r})$, $N_A(\mathbf{r})$, and $N_D(\mathbf{r})$ are the electron, hole, acceptor, and donor concentration, respectively.

Here, we assume that the electrostatic potential $\phi(\mathbf{r})$ varies over a relatively large spatial scale compared to the discretizations (d_x, d_y, d_z) used to calculate the atomistic Green's functions. Consequently, in order to reduce the size of the Poisson equation problem in devices with a technologically relevant size, we transfer the free electron and hole concentrations computed in Eqs. (24) and (25) on a coarser mesh with a discretization $(\Delta_x, \Delta_y, \Delta_z) \sim 0.1 \div 0.2$ nm. The conversion from the finer to the coarser mesh is performed so as to conserve the integral of the carrier concentration. The effect of external biases was imposed by setting Dirichlet boundary conditions at the contacts and Neumann boundary conditions to noncontacted boundary regions.

Finally, once the self-consistent solution has been obtained, we express the electron current as a function of the Green's functions and self-energies computed at the contact $L(R)$ as

$$\begin{aligned}I_{L(R)} = &-i\frac{e}{\hbar}\int dE\ \ \mathrm{tr}\big\{[\mathbf{\Gamma}_{L(R)}]\big[\big([\mathbf{G}^r_{L(R)}]\\ &-[\mathbf{G}^r_{L(R)}]^\dagger\big)f_{L(R)}+[\mathbf{G}^<_{L(R)}]\big]\big\},\end{aligned}\tag{27}$$

where $[\mathbf{\Gamma}_{\mathrm{L(R)}}] = i([\mathbf{\Sigma}_{\mathrm{L(R)}}]-[\mathbf{\Sigma}_{\mathrm{L(R)}}]^\dagger)$, $\mathrm{tr}\{\cdots\}$ is for the trace operation, and $f_{L(R)}$ is the Fermi-Dirac distribution. The calculation of the current can be carried out by using the Green's functions in the reduced Bloch functions basis, where the size of the matrices is the smallest.

B. Implementation and computational burden

Our simulation procedure takes full advantage of the first-principle calculations carried out by the *ab initio* solver [10]. In fact, after a duly converged *ab initio* simulation has been achieved for the unit cell of the physical system, our approach can be summarized in the following steps.

(i) Assuming $N_{kB}=2$, for example, select a subset of the Bloch states for $\mathbf{k}_0=(0,\mathbf{k}_{yz})$ and $\mathbf{k}_1=(\pi/a_x,\mathbf{k}_{yz})$ obtained by DFT calculations and transform them to the $x\mathbf{K}_{yz}$ basis using Eq. (17). Then orthogonalize the basis functions so as to obtain $\{\Phi_{\mathbf{k}_0}\}$, $\{\Phi_{\mathbf{k}_1}\}$ with $n=1, 2, \cdots N_B$ and assemble the U_Φ in Eq. (19).

(ii) Calculate $[\mathbf{H}_{0,0}]_\Phi$, $[\mathbf{H}_{0,1}]_\Phi$ in the reduced Bloch function basis by using Eq. (20).

(iii) Solve for the Green's functions in the reduced basis by using an initial guess of the electrostatic potential $\phi(\mathbf{r})$.

(iv) Calculate carrier concentrations from Eqs. (24) and (25) and then solve the Poisson equation for a new guess of $\phi(\mathbf{r})$. Loop between steps (iii) and (iv) until a specified convergence is reached and, finally, calculate the current using Eq. (27).

Here it should be mentioned that for any physical system or nanoscale device, steps (i) and (ii) above have to be performed only once. In fact, they correspond to material properties, so that the relevant quantities calculated in these steps can be stored and reused in subsequent simulations corresponding to different bias conditions. The number N_B of Bloch basis functions is the most important parameter affecting the computational load of the Green's functions equations.

Several important optimizations are possible and have been introduced in the implementation of our methodology. For example, it is apparent from Eqs. (24) and (25) that the sums over $\mathbf{G}_{yz}$, $\mathbf{G}'_{yz}$ for any couple of basis functions Φ_n, Φ_m can be carried out only once for a given physical system. In other words, we can introduce the new quantity $K_{n,m}(x_j,\mathbf{r}_{yz})$ defined as

$$\begin{aligned}K_{n,m}(x_j,\mathbf{r}_{yz}) &= \frac{1}{N_{Gy}N_{Gz}}\sum_{\mathbf{G}_{yz},\mathbf{G}'_{yz}}\Phi_n(x_j,\mathbf{G}_{yz})\\ &\quad\times\Phi_m^*(x_j,\mathbf{G}'_{yz})\,e^{i(\mathbf{G}_{yz}-\mathbf{G}'_{yz})\cdot\mathbf{r}_{yz}},\end{aligned}\tag{28}$$

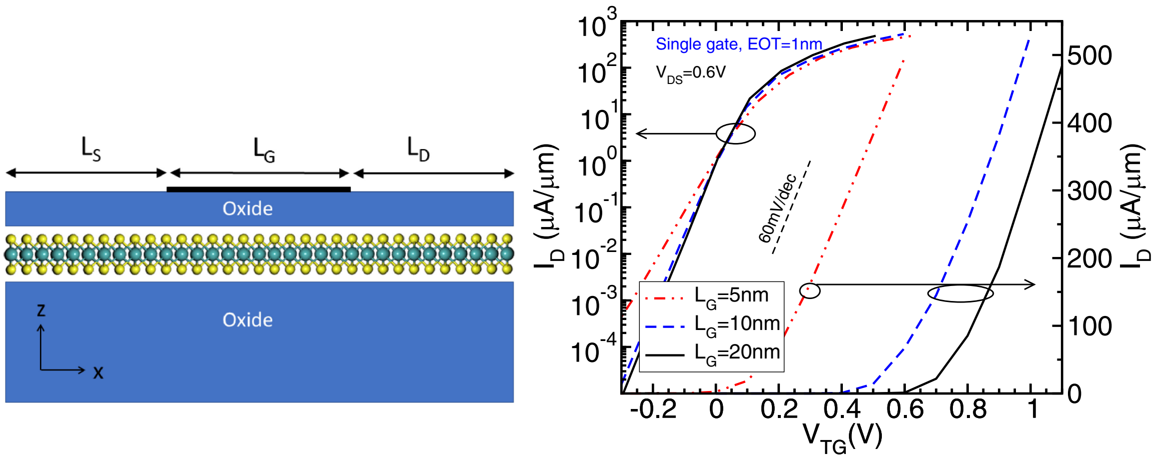

FIG. 3. Left: Sketch of the single-gate MOSFET simulated in this work and consisting of a single-layer MoS$_2$ channel material. The length of the source and drain extensions is $L_S = L_D \simeq 9$ nm and L_G indicates the top gate length. Right: Simulated drain current I$_{DS}$ vs top gate voltage V_{TG} characteristics at V$_{DS}$=0.6 V for n-type MoS$_2$ MOSFETs featuring different gate lengths L_G $\simeq$ 20, 10, and 5 nm. The I$_{DS}$ curves reported in semilogarithmic scales have been V_{TG} shifted so as to have the same I$_{DS}$=1 μA/μm at $V_{TG} = 0$ V for all gate lengths, whereas the curves in the linear scales have not.

and then notice that the free electron concentration can be written in terms of $K_{n,m}(x_j, \mathbf{r}_{yz})$ as

$$n(x_j, \mathbf{r}_{yz}) = \frac{-i}{d_x d_y d_z} \int_{E_0(x_j)}^{\infty} \sum_{n,m=1}^{N_B} \mathbf{G}^<(n, m; E)\, K_{n,m}(x_j, \mathbf{r}_{yz})\, \frac{dE}{2\pi}. \tag{29}$$

A similar expression holds for the free hole concentration $p(x_j, \mathbf{r}_{yz})$.

As it can be seen, once $K_{n,m}(x_j, \mathbf{r}_{yz})$ has been calculated for a given physical system, the carrier concentrations during self-consistent simulations can be obtained by using Eq. (29), that is, by skipping the sums over $\mathbf{G}_{yz}$, $\mathbf{G}'_{yz}$.

V. SIMULATION RESULTS

We present in this section an example of a nanoscale transistor that we could efficiently simulate by using the reduced basis of Bloch functions computed directly from plane-wave DFT calculations.

The device under investigation is composed by the monolayer MoS$_2$ n-MOSFET sketched in Fig. 3(a), where the 2D semiconductor is sitting on a 10-nm-thick back oxide having a dielectric constant $\varepsilon_{ox} = 3.9\varepsilon_0$ (with ε_0 being the vacuum permittivity). Source and drain regions are considered chemically doped with a donor concentration of 10^{20} cm^{-3} and the top gate oxide has an equivalent oxide thickness of about 1 nm. The lateral direction was assumed to be periodic and was described by including a discrete sampling of the wave vector k_y with a constant $\Delta k_y = 0.1\ \pi/a_y$.

Figure 3 shows the drain current I$_{DS}$ versus the top gate voltage V_{TG} that is characteristic for a gate length of $L_G = 64a_x \simeq 20$ nm, $32a_x \simeq$ 10 nm, and $16a_x \simeq$ 5 nm. The large I$_{DS}$ values are due to the fact that neither scattering nor series resistance is included in the simulations. Thanks to the subnanometer thickness of the MoS$_2$ layer, the I$_{DS}$ versus V_{TG} characteristics of the transistor are still well behaved for a channel length of about 5 nm, even if a degradation of the subthreshold swing is observed with respect to longer FETs. The onset of short channel effects also manifests itself in a significant left shift of the I$_{DS}$ versus V_{TG} characteristic in the shortest gate lengths.

The subthreshold swing degradation in the shortest device is mainly due to the onset of a sizable source-to-drain quantum

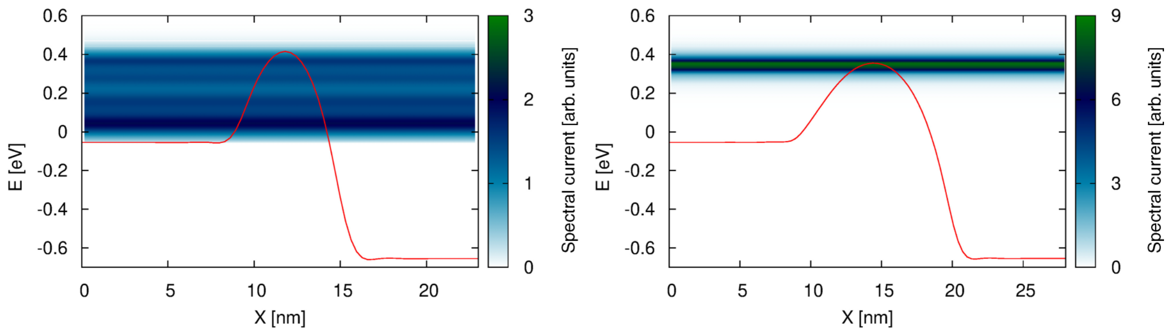

FIG. 4. Color map of the current density spectrum and profile of the lowest conduction band at $k_y = 0$ (red line) for a single-layer MoS$_2$ MOSFET having a gate length of either L_G $\simeq$ 5 nm (left) or 10 nm (right). Both devices have V$_{DS} = 0.6$ V and a gate bias corresponding to approximately the same subthreshold current I$_{DS} \simeq$ 5 nA/μm.

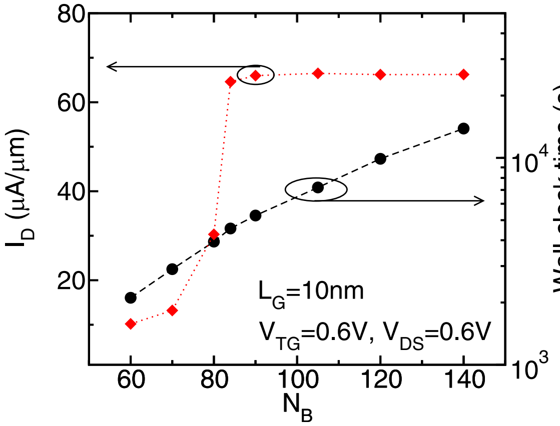

FIG. 5. Self-consistently calculated drain current for the L_G = 10 nm single-layer MoS$_2$ MOSFET sketched in Fig. 3 (left) for the bias point corresponding to V_{TG} = 0.6 V and V$_{DS}$ = 0.6 V, and plotted vs the number N_B of reduced basis functions. The right y axis reports the corresponding wall-clock simulation time. For this specific bias point, 11 iterations were necessary to obtain the self-consistent solution of the Poisson-NEGF equations. Calculations were performed by using a parallelization of the energy points over 10 cores (Intel Xeon Gold 6150 CPU at 2.70 GHz).

tunneling, as illustrated in Fig. 4, reporting the current spectra and the profile of the lowest conduction band along the transport direction for the MOSFET with either L_G = 5 or 10 nm. For the shortest gate length, the spectral current is spread over energies well below the top of the barrier, thus confirming a significant source-to-drain tunneling contribution to the off current in this specific bias condition. In the longer transistor, instead, the off current is dominated by thermionic emission above the top of the barrier.

Figure 5 illustrates, for the device in Fig. 3(a) subject to a specific external bias, the dependence of the self-consistently calculated I$_{DS}$ on the number N_B of Bloch states in the reduced basis. The corresponding CPU time versus N_B is also reported. It can be observed that for the case at study, the self-consistent solution of the Poisson-NEGF equations rapidly converges to a stable value for $N_B \geqslant 90$, which in turn enables the simulation of one bias point with a wall-clock computation time slightly longer than one hour by using only 10 CPU cores.

VI. CONCLUSIONS

We presented theoretical developments and a sound implementation for a first-principle transport model based on the NEGF formalism, and on a basis set obtained directly from the *ab initio* Bloch functions. Differently from previous papers proposing a similar approach for DFT calculations based on an LCAO basis, we used plane-wave *ab initio* calculations and we argue that thanks to an appropriate choice of the basis functions, we could effectively suppress the problem of unphysical solutions, whose treatment is delicate and computationally demanding [25,26].

We found that the unit cell restricted Bloch functions basis enables band structure and transport calculations with hundreds-of-times reductions of the size of the problem compared to the original DFT formulation. Moreover, while here we have reported results only for a homogeneous system consisting of a single-layer MoS$_2$, we envision that our approach can also be used for heterostructures, thus paving the way for a number of technologically and physically important applications, such as contacts between metals and 2D materials, as well as vertical or horizontal heterojunctions between 2D semiconductors.

The methods of this work can also be applied to first-principle calculations based on an LCAO basis; however, we think that the herein reported demonstration for plane-wave DFT Hamiltonians is particularly promising for future developments concerning the electron-phonon interaction. In fact, the NEGF formalism can naturally include electron-phonon scattering [42] and, moreover, the plane-wave DFT approach is especially suitable for the calculation of phonon spectra [44] and electron-phonon coupling coefficients [45].

The benefits of a Bloch function basis are not confined to first-principle methods; on the contrary, they also directly apply to empirical pseudopotential Hamiltonians in both their local and nonlocal formulation [46,47], and promise large computational advantages compared to the methods recently proposed by some of the present authors [23,24].

We believe that the results of this work qualify the methodology based on unit cell restricted Bloch functions as a viable approach for *ab initio* and semiempirical quantum transport simulations and, in particular, as an alternative to maximally localized Wannier functions. In this respect, while the direct use of Bloch functions is attractive in that it circumvents the *a posteriori* determination of the Wannier functions basis, further work is admittedly needed to demonstrate the general applicability of the methods of this paper, and the feasibility of the above-mentioned extensions to heterostructures and dissipative transport.

ACKNOWLEDGMENT

P.G. acknowledges support by the European Union via the H2020-INFRAEDI-2018-1 project MaX “Materials Design at the Exascale. European Centre of Excellence in materials modelling, simulations, and design” (Grant No. 824143).

APPENDIX A: HAMILTONIAN MATRIX IN THE $x\mathbf{K}_{yz}$ BASIS

In Sec. III, the Hamiltonian in the hybrid basis was built by using two DFT Hamiltonian matrices $[\mathbf{H}_{\mathbf{k}_0}]$, $[\mathbf{H}_{\mathbf{k}_1}]$, but the extension of the methodology to more than two $\mathbf{k}$ values is quite natural. To this purpose, we first generalize Eq. (12) and define the transformation matrix

$$\big[\mathbf{U}_{k_x}^{(N_{cx})}\big] = \frac{1}{\sqrt{N_{cx}}}\big[\big[\mathbf{U}_{k_x}^{0}\big], \big[\mathbf{U}_{k_x}^{0}\big]e^{ik_x a_x}, \cdots, \big[\mathbf{U}_{k_x}^{0}\big]e^{ik_x(N_{cx}-1)a_x}\big], \tag{A1}$$

where $[\mathbf{U}_{k_x}^{0}]$ has been defined in Eq. (11). Then we can use Eq. (A1) and reformulate Eq. (13) as

$$[\mathbf{H}]_{x\mathbf{K}_{yz}} = \sum_{\mathbf{k}} \big[\mathbf{U}_{k_x}^{(N_{cx})}\big]^{\dagger}\,[\mathbf{H}_{\mathbf{k}}]\,\big[\mathbf{U}_{k_x}^{(N_{cx})}\big], \tag{A2}$$

where the sum runs over N_{cx} Bloch vectors $\mathbf{k}$=(k_x,$\mathbf{k}_{yz}$), with k_x taking N_{cx} values in the range $-\pi/a_x < k_x \leqslant \pi/a_x$ with a spacing $2\pi/(N_{cx}a_x)$ and including k_x=0. Equation (A2) allows us to calculate $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ for any value of N_{cx} and the matrix $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ has a rank $N_{cx}N_G$.

Equation (A2) is useful in several respects. First, we see that according to the assumption for $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (9), the blocks $[\mathbf{H}_{0,0}]$ and $[\mathbf{H}_{0,1}]$ can be determined from the element (1,1) and (1,2) of the matrix $[\mathbf{H}]_{x\mathbf{K}_{yz}}$, so that Eq. (A2) allows us to write

$$[\mathbf{H}_{0,0}] = \frac{1}{N_{cx}} \sum_{\mathbf{k}} \big[\mathbf{U}^0_{k_x}\big]^\dagger [\mathbf{H}_\mathbf{k}] \big[\mathbf{U}^0_{k_x}\big], \tag{A3a}$$

$$[\mathbf{H}_{0,1}] = \frac{1}{N_{cx}} \sum_{\mathbf{k}} \big[\mathbf{U}^0_{k_x}\big]^\dagger [\mathbf{H}_\mathbf{k}] \big[\mathbf{U}^0_{k_x}\big]\, e^{ik_x a_x}. \tag{A3b}$$

Equation (A3) is a generalization of Eq. (14) for $N_{cx} \geqslant 3$, which immediately leads to the corresponding extension of Eq. (20) for the Hamiltonian blocks in the reduced basis.

By using Eq. (A3), we verified that the reconstruction of $[\mathbf{H}_{0,0}]$, $[\mathbf{H}_{0,1}]$ from the plane-wave DFT Hamiltonian matrices $[\mathbf{H}_\mathbf{k}]$ is practically independent of N_{cx} for $N_{cx} > 2$. Namely, any further increase of N_{cx} has a negligible effect on the electronic structure calculated in the $x\mathbf{K}_{yz}$ or in the Φ basis that has been analyzed, for example, in Figs. 1 and 2.

In the remainder of this Appendix, we discuss in more detail the assumption in Eq. (9) about the structure of the Hamiltonian in the $x\mathbf{K}_{yz}$ basis.

As discussed in Secs. II and III A, the nonlocal pseudopotential $V_{NL}(\mathbf{r}, \mathbf{r}')$ defined in Eq. (3) is the dominant nonlocal term of the Hamiltonian matrix in the $x\mathbf{K}_{yz}$ basis. However, $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ has to be calculated numerically by using Eq. (13) for N_{cx}=2 or Eq. (A2) for larger N_{cx} values. Consequently, while it is expected that the short-range nature of the nonlocal pseudopotential results in a single off-diagonal block of the Hamiltonian matrix as assumed by Eq. (9), we have no analytical expression for $V_{NL}(x_j, \mathbf{G}_{yz})$ ensuring *a priori* that, for example, a second off-diagonal block $\mathbf{H}_{0,2}$ is indeed negligible. Hence, we carried out a numerical analysis of this aspect.

In this respect, we first notice that if we hypothesize that $\mathbf{H}_{0,2}$ is *not* negligible, then, in order to identify $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$, and $\mathbf{H}_{0,2}$, it is necessary to write the $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ across at least four unit cells and, thus, by using four k_x values (i.e., k_x=0, $\pm 0.5\,\pi/a_x$, π/a_x). In fact, for N_{cx}=4, we can extend Eq. (10) and write

$$[\mathbf{H}]_{x\mathbf{K}_{yz}} = \begin{bmatrix} \mathbf{H}_{0,0} & \mathbf{H}_{0,1} & \mathbf{H}_{0,2} + \mathbf{H}^\dagger_{0,2} & \mathbf{H}^\dagger_{0,1} \\ \mathbf{H}^\dagger_{0,1} & \mathbf{H}_{0,0} & \mathbf{H}_{0,1} & \mathbf{H}_{0,2} + \mathbf{H}^\dagger_{0,2} \\ \mathbf{H}^\dagger_{0,2} + \mathbf{H}_{0,2} & \mathbf{H}^\dagger_{0,1} & \mathbf{H}_{0,0} & \mathbf{H}_{0,1} \\ \mathbf{H}_{0,1} & \mathbf{H}^\dagger_{0,2} + \mathbf{H}_{0,2} & \mathbf{H}^\dagger_{0,1} & \mathbf{H}_{0,0} \end{bmatrix}, \tag{A4}$$

which allows us to identify $\mathbf{H}_{0,2}$ (besides $\mathbf{H}_{0,0}$ and $\mathbf{H}_{0,1}$) by assuming that the $\mathbf{H}_{0,2}$ block is a lower triangular matrix [namely, $\mathbf{H}_{0,2}(i, j) \simeq 0$ for $j \geqslant i$]. In the presence of $\mathbf{H}_{0,2}$, the secular Eq. (16) can be rewritten as

$$\big[\mathbf{H}^\dagger_{0,2}\, e^{-ik_x(2a_x)} + \mathbf{H}^\dagger_{0,1}\, e^{-ik_x a_x} + \mathbf{H}_{0,0} + \mathbf{H}_{0,1}\, e^{ik_x a_x} + \mathbf{H}_{0,2}\, e^{ik_x(2a_x)}\big]\big\{\Psi^0_{n\mathbf{k}}\big\} = E_n(\mathbf{k})\big\{\Psi^0_{n\mathbf{k}}\big\}, \tag{A5}$$

and Eq. (A5) allows one to determine the electronic structure in the $x\mathbf{K}_{yz}$ basis and duly accounting for $\mathbf{H}_{0,2}$.

For the single-layer MoS$_2$ studied in this work, we calculated $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ for N_{cx}=4 by using Eq. (A2) and we verified that the electronic structure obtained with Eq. (A5) and accounting for $\mathbf{H}_{0,2}$ is practically identical to the results of Eq. (16) that instead neglect $\mathbf{H}_{0,2}$. This means that $\mathbf{H}_{0,2}$ can be neglected in the case at study, which legitimates the tridiagonal form of $[\mathbf{H}]_{x\mathbf{K}_{yz}}$ in Eq. (9) and implies that the $\mathbf{H}_{0,1}$ block is a lower triangular matrix. This latter observation, in turn, allows us to determine $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$ from Eq. (10).

We reiterate that the negligibility of $\mathbf{H}_{0,2}$ is an expected result from a physical perspective, in virtue of the relatively short-range nature of the nonlocal pseudopotential discussed in Sec. II.

APPENDIX B: SELECTION OF BLOCH FUNCTIONS FOR THE REDUCED BASIS

As already mentioned in Sec. III B, there exists a significant flexibility in the definition of the Bloch functions basis set, so that we add a few conceptual and practical remarks about the choice of the basis. The first is that the basis set cannot be formed by using the $\Psi^0_{n\mathbf{k}}(x_j, \mathbf{G}_{yz})$ corresponding to a single k_x value [with $\mathbf{k}$=(k_x,$\mathbf{k}_{yz}$)] because all such functions fulfill the same boundary condition $\Psi^0_{n\mathbf{k}}(a_x, \mathbf{G}_{yz})$=$\Psi^0_{n\mathbf{k}}(0, \mathbf{G}_{yz})\, e^{i\,k_x\,a_x}$ for their specific k_x value, so that it is impossible to build solutions of Eq. (16) for a different k_x value.

Quite interestingly, the $\Psi^0_{n\mathbf{k}}(x_j, \mathbf{G}_{yz})$ for two different k_x values appear to be sufficient to obtain solutions of Eq. (16) fulfilling the boundary condition $\Psi^0(a_x, \mathbf{G}_{yz})$=$\Psi^0(0, \mathbf{G}_{yz})\, e^{i\,k_x\,a_x}$ for any k_x value. In order to show this, we first recall the standard notation for Bloch functions $\Psi^0_{n\mathbf{k}}(x_j, \mathbf{G}_{yz})$=$u_{n,\mathbf{k}}(x_j, \mathbf{G}_{yz})\, e^{i\,k_x x_j}$ [where $u_{n,\mathbf{k}}(x_j, \mathbf{G}_{yz})$ is the periodic part of $\Psi^0_{n\mathbf{k}}$], and then we expand the unknown Ψ^0 for a generic k_x value in terms of the $\Psi^0_{n\mathbf{k}_0}$ and $\Psi^0_{n\mathbf{k}_1}$ for $\mathbf{k}_0$=(0, $\mathbf{k}_{yz}$) and $\mathbf{k}_1$=(π/a_x, $\mathbf{k}_{yz}$). By recalling Eq. (17), we write

$$\Psi^0(x_j, \mathbf{G}_{yz}) = \sum_{n=1}^{M_0} C_{n,\mathbf{k}_0}\, u_{n,\mathbf{k}_0}(x_j, \mathbf{G}_{yz}) + \sum_{m=1}^{M_1} C_{m,\mathbf{k}_1}\, u_{m,\mathbf{k}_1}(x_j, \mathbf{G}_{yz})\, e^{i\frac{\pi}{a_x}x_j}, \tag{B1}$$

where x_j=0, d_x, $2d_x \cdots (a_x{-}d_x)$ and M_0, M_1 denote the number of, respectively, $\Psi^0_{n\mathbf{k}_0}$ and $\Psi^0_{n\mathbf{k}_1}$ functions included in the expansion. Because $u_{n,\mathbf{k}_0}$ and $u_{n,\mathbf{k}_1}$ are periodic over

a_x, it is straightforward to see that the boundary condition $\Psi^0(a_x, \mathbf{G}_{yz}) = \Psi^0(0, \mathbf{G}_{yz})\, e^{i\,k_x\,a_x}$ becomes

$$(1 - e^{i\,k_x\,a_x}) \sum_{n=1}^{M_0} C_{n,\mathbf{k}_0}\, u_{n,\mathbf{k}_0}(0, \mathbf{G}_{yz}) = (1 + e^{i\,k_x\,a_x}) \sum_{m=1}^{M_1} C_{m,\mathbf{k}_1}\, u_{m,\mathbf{k}_1}(0, \mathbf{G}_{yz}). \tag{B2}$$

For k_x=0, Eq. (B2) can be fulfilled by taking all $C_{m,\mathbf{k}_1} = 0$ so that, as expected, Ψ^0 can be obtained by using only the $\Psi^0_{n\mathbf{k}_0}$ functions. Likewise, for k_x=π/a_x, we can take all $C_{m,\mathbf{k}_0} = 0$ and build the solution by using only the $\Psi^0_{\mathbf{k}_1}$ functions. For any other k_x value, Eq. (B2) provides the relation between the $C_{n,\mathbf{k}_0}$ and $C_{m,\mathbf{k}_1}$ coefficients, ensuring that Ψ^0 fulfills the boundary condition $\Psi^0(a_x, \mathbf{G}_{yz}) = \Psi^0(0, \mathbf{G}_{yz})\, e^{i\,k_x\,a_x}$, and hence it can be a solution of Eq. (16).

While two k_x values appear to be sufficient to build a basis, we found that sampling the reduced zone with more than two k_x values can improve the accuracy in the reconstruction of the electronic structure for a given overall number of basis functions, as exemplified in Fig. 2. Moreover we found that by including in the basis set the Bloch functions for all the valence bands at each k_x, we can effectively suppress the unphysical solutions sometimes observed upon the introduction of a reduced basis [25,26].

[1] R. Kim, U. E. Avci, and I. A. Young, IEEE Trans. Electron Devices **62**, 713 (2015).

[2] D. Esseni, M. Pala, and T. Rollo, IEEE Trans. Electron Devices **62**, 3084 (2015).

[3] C. Grillet, D. Logoteta, A. Cresti, and M. G. Pala, IEEE Trans. Electron Devices **64**, 2425 (2017).

[4] A. Seabaugh and Q. Zhang, Proc. IEEE **98**, 2095 (2010).

[5] D. Esseni, M. Pala, P. Palestri, C. Alper, and T. Rollo, Semicond. Sci. Technol. **32**, 083005 (2017).

[6] L. Mingda, D. Esseni, G. Snider, D. Jena, and H. G. Xing, J. Appl. Phys. **115**, 074508 (2014).

[7] D. Sarkar, X. Xie, W. Liu, W. Cao, J. Kang, Y. Gong, S. Kraemer, P. M. Ajayan, and K. Banerjee, Nature (London) **526**, 91 (2015).

[8] J. Cao, D. Logoteta, S. Özkaya, B. Biel, A. Cresti, M. G. Pala, and D. Esseni, IEEE Trans. Electron Devices **63**, 4388 (2016).

[9] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[10] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, and R. M. Wentzcovitch, J. Phys.: Condens. Matter **21**, 395502 (2009).

[11] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, J. Phys.: Condens. Matter **14**, 2745 (2002).

[12] M. Luisier, A. Schenk, W. Fichtner, and G. Klimeck, Phys. Rev. B **74**, 205323 (2006).

[13] G. Klimeck, S. S. Ahmed, H. Bae, N. Kharche, R. Rahman, S. Clark, B. Haley, S. Lee, M. Naumov, H. Ryu, F. Saied, M. Prada, M. Korkusinski, and T. B. Boykin, IEEE Trans. Electron Devices **54**, 2079 (2007).

[14] N. Marzari and D. Vanderbilt, Phys. Rev. B **56**, 12847 (1997).

[15] I. Souza, N. Marzari, and D. Vanderbilt, Phys. Rev. B **65**, 035109 (2001).

[16] J. Chang, L. F. Register, and S. K. Banerjee, Appl. Phys. Lett. **103**, 223509 (2013).

[17] A. Szabó, R. Rhyner, and M. Luisier, Phys. Rev. B **92**, 035435 (2015).

[18] G. Pizzi, M. Gibertini, E. Dib, N. Marzari, G. Iannaccone, and G. Fiori, Nat. Commun. **7**, 12585 (2016).

[19] X.-W. Jiang, S.-S. Li, J.-B. Xia, and L.-W. Wang, J. Appl. Phys. **109**, 054503 (2011).

[20] A. Garcia-Lekue, M. Vergniory, X. Jiang, and L. Wang, Prog. Sur. Sci. **90**, 292 (2015).

[21] J. Fang, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, J. Appl. Phys. **119**, 035701 (2016).

[22] J. Fang, S. Chen, W. G. Vandenberghe, B. Fu, and M. V. Fischetti, IEEE Trans. Electron Devices **64**, 2758 (2017).

[23] M. G. Pala and D. Esseni, Phys. Rev. B **97**, 125310 (2018).

[24] M. G. Pala and D. Esseni, J. Appl. Phys. **126**, 055703 (2019).

[25] M. Shin, W. J. Jeong, and J. Lee, J. Appl. Phys. **119**, 154505 (2016).

[26] G. Mil'nikov, N. Mori, and Y. Kamakura, Phys. Rev. B **85**, 035317 (2012).

[27] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Kkbenli, M. Lazzeri, M. Marsili, N. Marzari, F. Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L. Paulatto, S. Ponc, D. Rocca, R. Sabatini, B. Santra, M. Schlipf, A. P. Seitsonen, A. Smogunov, I. Timrov, T. Thonhauser, P. Umari, N. Vast, X. Wu, and S. Baroni, J. Phys.: Condens. Matter **29**, 465901 (2017).

[28] L.-W. Wang, A. Franceschetti, and A. Zunger, Phys. Rev. Lett. **78**, 2819 (1997).

[29] L.-W. Wang and A. Zunger, Phys. Rev. B **59**, 15806 (1999).

[30] F. Chirico, A. Di Carlo, and P. Lugli, Phys. Rev. B **64**, 045314 (2001).

[31] D. Esseni and P. Palestri, Phys. Rev. B **72**, 165342 (2005).

[32] J.-L. van der Steen, D. Esseni, P. Palestri, L. Selmi, and R. J. E. Hueting, IEEE Trans. Electron Devices **54**, 1843 (2007).

[33] M. L. Van de Put, M. V. Fischetti, and W. G. Vandenberghe, Comput. Phys. Commun. **244**, 156 (2019).

[34] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).

[35] From an implementation viewpoint, the transformation in Eq. (11) can be equivalently split in a one-by-one transformation of each block identified by a couple of $\mathbf{G}_{yz}$ and $\mathbf{G}'_{yz}$, which results in a large reduction of the block size from N_G to N_{Gx}=N_{dx}.

[36] Even if Eq. (14) allow us to calculate the entire $\mathbf{H}_{0,0}$, $\mathbf{H}_{0,1}$ blocks, for the kinetic energy operator we actually employed the analytic expressions discussed in detail in Ref. [23]. In particular, we used a eighth-order centered difference discretization scheme. In this respect, we also argue that if we let $2p$ denote the discretization order, any p smaller than the number N_{dx} of discretization points inside a_x is consistent with the block tridiagonal structure of the Hamiltonian matrix in Eq. (9), and with $\mathbf{H}_{0,1}$ being a lower triangular matrix (see Eq. (31) in Ref. [23]).

[37] S. Datta, *Quantum Transport - Atom to Transistor* (Cambridge University Press, Cambridge, 2005).

[38] Watkins, David S, *Fundamentals of Matrix Computations* (Wiley, New York, 2004).

[39] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, and N. Marzari, Nat. Nanotechnol. **13**, 246 (2018).

[40] D. R. Hamann, Phys. Rev. B **88**, 085117 (2013).

[41] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[42] G. D. Mahan, *Many-Particle Physics* (Plenum Press, New York, 1988).

[43] M. P. Anantram, M. S. Lundstrom, and D. E. Nikonov, Proc. IEEE **96**, 1511 (2008).

[44] S. Baroni, S. de Gironcoli, A. Dal Corso, and P. Giannozzi, Rev. Mod. Phys. **73**, 515 (2001).

[45] S. Poncé, E. Margine, C. Verdi, and F. Giustino, Comput. Phys. Commun. **209**, 116 (2016).

[46] J. R. Chelikowsky and M. L. Cohen, Phys. Rev. B **10**, 5095 (1974).

[47] M. L. Cohen and J. R. Chelikowsky, *Electron Structure and Optical Properties of Semiconductors*, Springer Series in Solid-State Sciences (Springer-Verlag, Berlin, 1988).