


Hybrid Mode-Space–Real-Space Approximation for First-Principles Quantum Transport Simulation of Inhomogeneous Devices

Fabian Ducry^{✉,*}, Mohammad Hossein Bani-Hashemian[✉], and Mathieu Luisier
Eidgenössische Technische Hochschule Zürich, 8092 Zürich, Switzerland

 (Received 14 December 2019; revised manuscript received 2 March 2020; accepted 1 April 2020; published 27 April 2020)

We propose a robust strategy for transforming the Hamiltonian of a metallic structure expressed in a nonorthogonal density-functional-theory (DFT) basis into a low-dimensional space compatible with electron-transport simulations. This mode-space approximation is applied locally to inhomogeneous material stacks including amorphous phases and interfaces. The contacts and periodic parts of the device are transformed into the subspace created, while the active inhomogeneous regions remain represented in real space. The various regions are connected to each other through mode-space–real-space hybrid blocks of the Hamiltonian and overlap matrices. This approach allows harnessing of the full flexibility of DFT combined with the power of the nonequilibrium Green’s function formalism at a moderate cost. In particular, for a realistic resistive random-access memory cell composed of 3390 atoms, a performance improvement by a factor of 136 compared with a pure real-space treatment is demonstrated, with an error of less than 2%.

DOI: [10.1103/PhysRevApplied.13.044067](https://doi.org/10.1103/PhysRevApplied.13.044067)

I. INTRODUCTION

Because of the continuous decrease in the dimensions of integrated circuits, quantum-mechanical simulation approaches have become essential for predicting the performance of nanoscale components and optimizing their “current versus voltage” characteristics. Moreover, the emergence of alternative technologies to conventional complementary-metal-oxide-semiconductor devices, such as nonvolatile-memory (NVM) cells relying on the relocation of a few atoms in complex material stacks [1,2], makes accurate simulations even more challenging. The often little-explored materials and interfaces encountered in these memory cells and the stochastic nature of atomic relocations call for *ab initio* modeling approaches. The latter are crucial not only for determining observable quantities, but also for shedding light on the underlying physical principles of operation, which are, to date, not fully understood. Density-functional theory (DFT) [3] satisfies the requirements for atomistic simulations without the need for parameterization. Coupling DFT to the nonequilibrium Green’s function (NEGF) [4] formalism gives rise to a powerful tool for performing electrical, thermal, and coupled electrothermal device investigations of any kind of nanostructure [5,6]. Such simulations are, however, typically limited to a small number of atoms, in the range of few hundreds to thousands, because of their heavy computational burden [7]. Particularly, the

presence of electron-phonon scattering may represent an insurmountable computing obstacle [8]. Taking account of these dissipative interactions is essential for evaluating electrothermal effects such as self-heating, which are responsible for possible device failures.

The computational intensity can be decreased by reducing the size of the Hamiltonian matrix describing the system of interest through, for example, the mode-space (MS) approximation [9–11]. This technique decomposes a real-space (RS) domain into two directions, a transport and a transverse one, the two being orthogonal to each other. Only a small set of the transverse modes need to be retained to accurately reproduce the RS results within a limited energy window. The transverse modes represent the band structure of a unit cell, extracted along the plane orthogonal to the transport direction. It is important that the unit cell can be periodically repeated so that its band structure is unambiguously defined. As a consequence, interfaces and amorphous layers cannot be transformed into MS. Such arrangements are, however, present in many realistic applications that are of great interest to industry and thus to the device-modeling community, e.g., tunnel and break junctions [12,13], grain boundaries and interfaces in interconnects [14,15], and emerging NVMs and structures with surface roughness [16]. A prime example is conductive bridging random-access memories (CBRAMs) [2], a type of emerging NVM where a layer of amorphous oxide separates two metallic electrodes. Through the controlled growth and dissolution of a metallic filament between the contacts, the resistance state of the cell

*fabian.ducry@iis.ee.ethz.ch

can be modulated. Computing the electrical or electrothermal properties of CBRAMs at the *ab initio* level requires an enormous amount of computational power, which prevents large-scale investigations of their I - V operational principle.

On close examination of the open boundary conditions employed in the NEGF, it turns out that every quantum transport simulation contains at least two periodic regions, the semi-infinite leads that are implicitly attached to the central device region. Sometimes, parts of the device region are also made of repeated unit cells. To exploit this property, we propose a hybrid scheme in which a MS transformation is applied to the periodic regions of the simulation domain, while the RS description of any nonperiodic features along the transport direction is retained. This enables the application of a MS transformation to the Hamiltonian for any device geometry, while retaining the advantages of the underlying DFT framework. Therefore, even though the interface and amorphous layers of the targeted CBRAM cells are inherently non-periodic, the corresponding Hamiltonian matrix exhibits local periodicity at least in the metal contacts. This feature can be leveraged by transforming only the periodic sub-Hamiltonian composing the contacts into MS. The amorphous layer and its interfaces with the metal electrodes are left untouched and remain represented in RS [17]. This approach offers two key advantages over pure RS simulations. First, the reduction in the size of the Hamiltonian drastically improves the memory footprint required to perform NEGF calculations and offers significant speedups in solving the resulting equations with, for example, the recursive-Green's-function [18] algorithm. Secondly, as the contact regions are periodic, much smaller boundary Hamiltonian blocks are obtained, rendering the computation of the open boundary conditions (OBCs) [19] much more affordable.

Furthermore, we demonstrate the successful application of MS transformations to metallic layers, which have so far not received as much attention as semiconductors. To do so, we develop a scheme to carefully select all parameters needed to transform a RS block into a MS one, thus leading to efficient and reliable transformations without having to manually sweep a large parameter space. The proposed scheme is applied to a CBRAM cell and benchmarked against a pure RS reference in terms of speed and accuracy. Besides reporting significant speedups, we also provide insight into the electrical properties of the metallic filaments present in CBRAM cells based on ballistic quantum transport simulations.

The paper is organized as follows. In Sec. II, the RS-to-MS transformation and the scheme for parameter selection are introduced, before the MS-RS hybrid approach is described and simulation results are shown in Sec. III. The paper is concluded in Sec. IV.

II. MODE-SPACE TRANSFORMATION

Electron-transport calculations can usually be restricted to a small energy window around the Fermi energy, as only the partially filled states in this window contribute to the current. DFT simulations, on the other hand, require the consideration of all bands below the Fermi energy as well as some above it. This results in the presence of many energy states in the Hamiltonian matrix that are irrelevant to transport simulations with, e.g., the NEGF. Thus, the latter simulations become unnecessarily expensive. One remedy for this situation consists of projecting the Hamiltonian onto a reduced basis that encompasses only the states relevant to electron transport.

Such order-reduction schemes are available for periodic structures only, as is the case for the MS approximation: based on the eigenvectors of certain Hamiltonian blocks, a small basis of transverse modes can be constructed that reproduces the band structure within the energy window of interest [9]. Using the resulting transformed matrices, device characteristics can be computed to high accuracy, but with low cost. The speedup can reach multiple orders of magnitude [10,11]. The fact that real devices very often feature inhomogeneities precludes the use of MS. This is the case for CBRAMs, which contain metal-oxide interfaces and an amorphous layer with no periodicity. The contacts, however, can be treated as periodic. A major difficulty with the contacts comes from their metallic nature. Because of this, procedures for constructing a MS basis for metallic contacts have not yet been demonstrated. The CBRAM cell studied in the present work is illustrated in Fig. 1(a). The unit cells delimited by black boxes are perfectly periodic, with a local block structure as shown in Fig. 1(b). Taking advantage of this, a MS basis can be created for the contacts of the device.

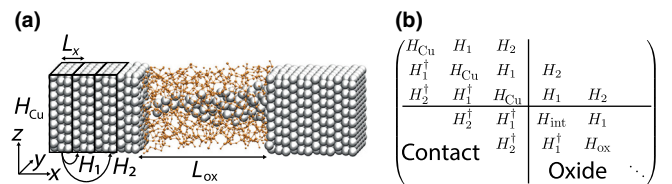


FIG. 1. (a) Atomistic representation of the Cu/a-SiO₂/Cu CBRAM cell structure studied in the present work with the help of the proposed hybrid mode-space–real-space approach. The structure is composed of 4449 atoms, with the large white spheres representing Cu atoms (with 25 orbitals per atom), the orange lines representing bonds between Si and O atoms, and the small orange spheres representing Si and O atoms (both with 13 orbitals per atom). The black rectangles mark the periodic Cu contact blocks. (b) Block pentadiagonal structure of the real-space Hamiltonian matrix H corresponding to the structure in (a). Each block represents a unit cell (three Cu layers). Because of long-range interactions, each unit cell is connected to $N_N = 2$ neighboring cells.

A. Atomic device structure

To explain the RS-to-MS transformation and the hybrid coupling and to demonstrate the applicability of the scheme, a CBRAM cell made of two Cu contacts surrounding a layer of amorphous silicon dioxide ($a - \text{SiO}_2$) is used [20,21]. The $a - \text{SiO}_2$ is generated with the help of a melt-and-quench approach [22], with melting at 4000 K for 800 ps and quenching at a rate of 30 K/ps, using classical molecular dynamics as implemented in the QuantumATK 2017.1 [23,24] package and force-field parameters from Ref. [25]. The final atomic structure of the CBRAM cell is obtained by attaching two Cu electrodes to the slab of $a - \text{SiO}_2$. The metallic filament is inserted by converting all Si and O atoms within a pre-defined cone to Cu. The structure is then annealed with the *ab initio* molecular dynamics solver of the CP2K package [26] for more than 4 ps at 500 °C. The Perdew-Burke-Ernzerhof exchange-correlation functional [27], double zeta-valence polarized Gaussian-type orbital (GTO) basis sets [28], and Goedecker-Teter-Hutter pseudopotentials [29] are employed for this purpose. Two variants, *A* and *B*, of the same material stack are used. The length of the oxide L_{ox} in structure *A* is 3.5 nm, and the size of the cross section is $2.1 \times 2.3 \text{ nm}^2$. The resulting device is schematically illustrated in Fig. 1(a). The contacts are in fact longer than shown in Fig. 1(a) and each feature seven unit cells of Cu. The total number of atoms in the contacts and oxide for structure *A* is 4449. For a direct comparison of RS and hybrid simulations, the smaller structure *B* is required. The number of Cu unit cells in the contacts is reduced to six on either side, and L_{ox} is shortened to 1.6 nm, while the cross section remains the same. This brings the number of atoms in structure *B* down to 3390.

The RS-to-MS transformation is performed on one unit cell of the Cu contact, which is extracted from the CBRAM cell. The unit cell is composed of 240 atoms, with 25 orbitals each. This unit cell corresponds to the region encapsulated in a black box in Fig. 1(a) and is identical in the two variants (*A* and *B*) of the cell.

B. Transformation matrix

Converting a RS Hamiltonian H_{RS} , as produced by CP2K, to its MS equivalent H_{MS} requires one first to construct a transformation matrix U such that $H_{\text{MS}} = U^\dagger H_{\text{RS}} U$. To obtain this transformation matrix, we follow the procedure developed by Shin *et al.* in Ref. [11]. Equations are repeated where needed to facilitate the understanding of our approach, but no formal derivations are given. The procedure depends sensitively on a number of parameters that can be chosen freely. In the next section, we provide guidance for the selection of these parameters, avoiding time-consuming searches for suitable configurations.

The energy-momentum dispersion of a RS Hamiltonian with a periodicity along the x direction can be computed from the following generalized eigenvalue problem:

$$H_{k_x} \Psi_{k_x} = E(k_x) S_{k_x} \Psi_{k_x}, \quad (1)$$

where H_{k_x} and S_{k_x} are the k_x -dependent Hamiltonian and overlap matrices, respectively, while $E(k_x)$ is the energy of the transverse mode Ψ_{k_x} at momentum k_x . The matrix H_{k_x} is given by

$$H_{k_x} = H_{\text{Cu}} + \sum_{n=1}^{N_N} (H_n e^{ink_x} + H_n^\dagger e^{-ink_x}). \quad (2)$$

Here, N_N represents the number of connected neighboring cells. Also, in Eq. (2), H_{Cu} is the on-site RS Hamiltonian block corresponding to a periodic unit cell of width L_x , as depicted in Fig. 1(a). It is connected to its n th nearest-neighboring cell, situated at a distance nL_x from it, through the RS Hamiltonian block H_n . The structure of H_{RS} with a contact with a periodic nature is illustrated in Fig. 1(b). Note that the overlap matrix S_{k_x} , which is equal to the identity matrix in the case of an orthogonal basis set, has the same form as H_{k_x} .

To construct an initial guess U_0 for the transformation matrix U , Eq. (1) is solved for all energies within a pre-defined window with upper and lower limits E_1 and E_2 , and for a set of n_k k_x -points selected according to

$$k_{x_j} = 2\pi j / (n_k - 1) - \pi, \quad j = 0, \dots, n_k - 1. \quad (3)$$

Overall, for each k_x -point, a total of $m(k_x)$ modes $\Psi(k_x)$ is calculated, which satisfy $E_1 \leq E \leq E_2$. A basis $\Psi = [\Psi_{k_{x1},1}(E), \Psi_{k_{x1},2}(E), \dots, \Psi_{k_{x1},m(k_{x1})}(E), \Psi_{k_{x2},1}(E), \dots, \Psi_{k_{xn_k},m(k_{n_k})}(E)]$ is formed by grouping all these modes. It is orthonormalized using singular-value decomposition (SVD). To reduce the size of the MS basis when building U_0 , only those modes corresponding to the largest singular values (SVs) are retained, and all the others are discarded.

After the MS transformation of H_{k_x} with U_0 , the band structure computed with the reduced k_x -dependent Hamiltonian blocks contains unphysical (spurious) states in addition to the real ones. This is illustrated in Fig. 2(a) for a contact of the device in Fig. 1. To eliminate these unphysical energies and obtain the final transformation matrix U , a refinement procedure must be applied. This is achieved through the iterative addition of carefully selected modes to U_0 , which shifts the unphysical energies below E_1 or above E_2 , until no unphysical energies remain between E_1 and E_2 . The cleaned band structure is shown in Fig. 2(b) for the same cell with 240 Cu atoms. Finding these modes is done by minimizing the following functional $\mathcal{F}(C)$ [10,11]

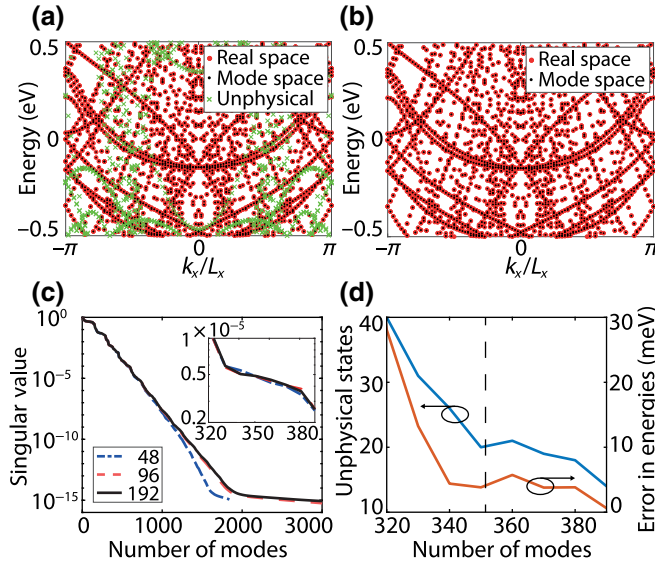


FIG. 2. (a) Band structure of a Cu cell containing 240 atoms as computed in RS (large red dots) and MS (small black dots). The unphysical energies coming from the initial MS transformation are shown as green crosses. (b) As (a), but after applying the basis refinement in Eq. (4). No unphysical energies remain. (c) Distribution of singular values (normalized by their largest entry) resulting from the orthonormalization of the basis Ψ corresponding to a Cu cell made of 240 atoms. Sets with different values of n_k from 48 to 192 are presented. The inset shows the same data enlarged around the range 320–390. (d) Maximum number of unphysical energy levels at a single k -point (left axis) and maximum energy error (right axis) when the RS and MS bands are compared as a function of the number of initial modes N_M composing U_0 .

of a column vector C :

$$\mathcal{F}(C) = \frac{1}{n_z} \sum_{q=1}^{n_q} \sum_{j=1}^{n_z} \frac{C^T A(k_q, z_j) C}{C^T B(k_q, z_j) C} + (C^T C - 1)^2. \quad (4)$$

The procedure starts with a set of n_q trial k -points k_q , where n_q is not necessarily equal to n_k . The z_j are n_z points on a complex contour around the energy $(E_1 + E_2)/2$ [10]. Equation (4) involves matrices A and B , defined as [11]

$$A(k_q, z_j) = (z_j - \epsilon_c) \Xi^\dagger (\Lambda U \tilde{\Lambda}^{-1} U^\dagger S_{k_q} U \tilde{\Lambda}^{-1} U^\dagger \Lambda - S_{k_q}^\dagger U \tilde{\Lambda}^{-1} U^\dagger \Lambda - \Lambda U \tilde{\Lambda}^{-1} U^\dagger S_{k_q} + S_{k_q}) \Xi, \quad (5)$$

$$B(k_q, z_j) = \Xi^\dagger (\Lambda - \Lambda U \tilde{\Lambda}^{-1} U^\dagger \Lambda) \Xi, \quad (6)$$

where ϵ_c , z_j , and the matrices Λ and $\tilde{\Lambda}$ obey the following equations:

$$\epsilon_c = \frac{E_1 + E_2}{2}, \quad (7)$$

$$z_j = \left[\epsilon_c + \left(\frac{E_2 - E_1}{2} \right) \right] e^{2\pi i(j-1/2)/n_z}, \quad (8)$$

$$\Lambda(k_q, z_j) = z_j S_{k_q} - H_{k_q}, \quad (9)$$

$$\tilde{\Lambda}(k_q, z_j) = U^\dagger \Lambda U. \quad (10)$$

Each individual mode added to U is given by

$$\Xi C, \quad (11)$$

where Ξ is a trial basis. As A and B appear only in quadratic form, i.e., $C^T A C$ and $C^T B C$, in Eq. (4), they can be symmetrized without loss of generality. This trick is found to improve the stability of the minimization, and it simplifies the form of the analytical derivative of $\mathcal{F}(C)$ to

$$\frac{\partial \mathcal{F}(C)}{\partial C} = \sum_q \sum_z \left[\frac{C^T A(q, z)}{C^T B(q, z) C} - \frac{C^T A(q, z) C}{(C^T B(q, z) C)^2} C^T B(q, z) \right] + 4(C^T C - 1)C. \quad (12)$$

With this form of $\partial \mathcal{F}(C)/\partial C$ and the use of a derivative-based minimizer such as the Broyden-Fletcher-Goldfarb-Shanno algorithm [30–33], the calculation of the C that minimizes \mathcal{F} can be considerably accelerated. The success of the minimization, however, depends strongly on the quality of the initial guess for C .

C. Selection of a robust set of parameters

Equations (1)–(12) are sufficient to implement the desired MS transformation. Six parameters need to be carefully selected for the process to be successful, namely (1) n_k in Eq. (3) for the calculation of the RS band structure, (2) the number N_M of modes composing U_0 , (3) the trial k -point set k_q used for the refinement procedure, (4) the integration points z_j , (5) the initial guess for C in Eq. (4), and (6) the trial basis Ξ . We now provide guidelines for selecting all these parameters for transforming RS Hamiltonian blocks corresponding to metallic cells to their MS equivalents. The strategy discussed here does not necessarily return the set of minimal parameters, but it returns one that leads reliably to a clean MS band structure without unphysical states, as shown in Fig. 2(b).

First, n_k is chosen based on the SVD of the basis of transverse modes Ψ . This choice should ensure that the SVs are converged in the sense that adding more modes Ψ_{k_x} to Ψ does not change the SVs anymore. This concept is illustrated in Fig. 2(c), where the SVs of Ψ are plotted for

several values of n_k , using a rectangular Cu cell with 240 atoms as shown in Fig. 1(a). For $n_k = 96$ and 192, the SVs hardly change anymore, indicating that $n_k = 96$ is a robust choice, whereas any lower value might cause convergence issues.

To find the number of modes N_M included in U_0 , the MS band structure must be computed as a function of this quantity with

$$[U_0^\dagger H_{k_x}(E) U_0] \Psi_{k_x}(E) = E [U_0^\dagger S_{k_x} U_0] \Psi_{k_x}(E), \quad (13)$$

and compared with the RS reference obtained from Eq. (1). With increasing N_M , the error in the physical energies computed in MS is reduced. Hence, N_M should be large enough that the largest error, measured as the difference between the RS and physical MS energies, decreases to below 5 meV. To find the physical MS energies, each RS energy level is mapped to the closest MS energy. Consequently, all MS energies without a corresponding RS energy are termed unphysical. Increasing N_M beyond the minimum value satisfying the aforementioned conditions improves the accuracy of the final U , but at the cost of its size and of computational time. Following the same trend as the error in the energies, the number of unphysical bands, i.e., the number of spurious bands that exist only in MS, decreases as N_M increases, reducing the number of refinement iterations required. The dependence of the error in the physical energies, i.e., those present in both RS and MS, and the number of unphysical states on N_M is illustrated in Fig. 2(d). The smallest N_M allowing a reliable refinement is marked with a vertical dashed black line. The impact of N_M on the result of the refinement procedure is summarized in Table I. A suitable guess for N_M corresponds to the number of normalized SVs that are larger than 10^{-2} .

Selecting the proper set of trial k -points k_q has a large impact on the refinement procedure, regarding both convergence and computational resources. A large n_q is

TABLE I. Comparison of the efficiency of the real-space-to-mode-space transformation as a function of the size of U_0 for a rectangular Cu cell with 240 atoms. The energy window covers a range of ± 0.5 eV around the Fermi energy. A total of $n_k = 96$ k -points is used to sample the band structure of the Cu cell. The original RS Hamiltonian block size is equal to 6000, made up of 240 atoms times 25 orbitals per atom. Using a larger initial guess leads to a reduction in both the number of refinement iterations and the error in the resulting MS band structure, but it also increases the MS block size.

Size of initial guess (N_M)	350	380	390
Refinement iterations	38	34	26
MS block size	388	414	416
Size reduction (%)	93.5	93.1	93.1
Max. error (meV)	0.89	0.62	0.34

needed for a comprehensive sampling of the k -space when k_q is selected with the help of Eq. (3). To overcome this limitation, Huang *et al.* [34] suggested choosing k_q according to the position of the unphysical energies. To do that, a nonhomogeneous k -point grid is necessary. In the absence of equidistant k -points, however, instabilities can occur during the minimization, which may lead to divergence of the refinement procedure. We opt for an intermediate scheme where the k_q are selected based on Eq. (3) with $n_q = 5$. Such a grid is too coarse to sample all unphysical bands and does not guarantee convergence on its own. This shortcoming is eliminated by augmenting k_q with points based on an analysis of Eq. (13). The $\pm k_x$ that correspond to the positions with the largest number of unphysical energy states are added to k_q . At every iteration of the refinement procedure, the MS band structure must be recomputed with the updated U_0 , and $\pm k_x$ must be adjusted accordingly. This scheme ensures the same stability as that obtained with a regular k -point grid, it converges through sampling of at least one unphysical band at a time, and it keeps n_q small for computational efficiency.

The number of integration points in Eq. (4), n_z , is not a critical parameter. With $n_z = 6$, convergence can be readily achieved. Increasing n_z could avoid the addition of a few modes to the final transformation matrix, but this rapidly increases the computational cost.

Minimizing Eq. (4) is a challenging task, as the functional has many local minima. This is one of the biggest issues encountered in the refinement procedure, as there is no algorithm available that guarantees that the global minimum is identified if the sum in Eq. (4) has more than one term [35]. Having a good initial guess is thus essential for finding a mode that improves the MS band structure. From numerical testing, it appears that the rows of U form a suitable initial guess for the C 's. To enable fast convergence of the refinement with the lowest possible number of iterations, we solve Eq. (4) with multiple initial guesses and always select the mode that produces the smallest value of $\mathcal{F}(C)$. To reduce the computational burden, we perform the minimization with the first 10% of the rows of U .

The last parameter to be selected is the trial basis Ξ . Two different matrices can be used for that purpose,

$$\Xi(U) = (\mathcal{I} - UU^\dagger)[S_{k_1}^{-1}H_{k_1}, S_{k_2}^{-1}H_{k_2}]U \quad (14)$$

or

$$\Xi(U) = (\mathcal{I} - UU^\dagger)[S_{k_1}^{-1}H_{k_1} \oplus S_{k_2}^{-1}H_{k_2}]U, \quad (15)$$

where \mathcal{I} is the identity matrix of appropriate size, $[,]$ is the commutator, and $[\oplus]$ is the concatenation operator. We find that the small variational space of Eq. (14) is sufficient and often outperforms that of Eq. (15) in terms of accuracy, besides offering lower computational cost. It is possible, however, that towards the end of the refinement,

the Ξ obtained from Eq. (14) does not give enough degrees of freedom anymore. Thus, when Eq. (4) stops providing modes satisfying $\mathcal{F}(C) < -1$, we switch to Eq. (15) to determine Ξ . To fully define Ξ , wave vectors k_1 and k_2 must be selected; the first corresponds to the position with the highest number of unphysical states in the initial MS band structure, and the second is fixed at the Γ point.

The computational resources required to obtain the MS transformation matrix and to eliminate spurious states from the band structure depend on the size of the matrices H_{RS} and S_{RS} , on the width of the energy window [11], and on the choice of the parameters discussed in this section. With these, the times needed to acquire the MS transformation matrix for the band structure shown in Fig. 2(b) are listed in Table II. The time per refinement iteration, 65.4 s, is roughly split in half between solving Eqs. (5)–(10) and minimizing Eq. (4), both needing around 30 s. Calculating the MS band structure, on the other hand, hardly contributes to the computational effort. The 38 refinement iterations needed to clean up the band structure add up to 3644 s. Together with the computation of the RS band structure, the entire procedure takes 4740 s.

We verify that the strategy discussed in this section can be applied to metals featuring more complex band structures, e.g., Pt, with its d -like bands around the Fermi energy. Without modification of the approach, the MS bands accurately reproduce the RS ones. The sizes of the transformation matrix and of the MS Hamiltonian, however, are directly related to the density of bands within the energy window of interest. Therefore, metals with densely packed bands around the Fermi energy require one to include far more transverse modes than, e.g., Cu does. A substantial reduction in the matrix size can still be achieved, but the benefit of the MS transformation is limited in these cases.

TABLE II. Timing data for the MS transformation procedure. The profiling is performed on a workstation featuring two Intel Xeon E5-2680 v4 CPUs with 28 cores in total, and the procedure is implemented in MATLAB 2018b. The refinement iteration is broken down into parts: (1) the preparation of the matrices A and B , (2) the global minimization, and (3) the calculation of the band structure. The numbers in the first column correspond to the time needed for the execution of one iteration of a task on one CPU, e.g., one k -point on one CPU. The second column specifies the time needed to perform the entire computation, parallelized over 28 cores.

	Per task and core	Total
RS band structure (s)	167.7	1096.7
Initial guess U_0 (s)	155.7	19.4
Refinement iteration ($N_M = 350$) (s)		65.4
(1) Matrices A and B (s)	56.9	28.3
(2) Minimization of $\mathcal{F}(C)$ (s)	10.4	34.1
(3) MS band structure (s)	0.05	3.0

III. ELECTRON TRANSPORT

For electron-transport simulations, the time-independent Schrödinger equation has to be solved in the presence of OBCs [4]. This equation takes the following form:

$$[ES - H - \Sigma^R(E)]\Phi(E) = I_{nj}(E). \quad (16)$$

In Eq. (16), $\Sigma^R(E)$ is the retarded boundary self-energy and I_{nj} is the injection vector; both of these arise from the OBCs and are defined in Ref. [19]. The Hamiltonian matrix H can be expressed in any RS basis (effective-mass approximation, tight-binding, DFT) or MS. This equation needs to be solved for a set of discrete energies E that belong to the range where electron transport happens, i.e., around the Fermi energy. In the ballistic limit, all energies are independent and Eq. (16) can be solved directly for the wave function Φ . This approach is known as the quantum transmitting-boundary method (QTBM) [36]. In the absence of scattering, the NEGF [4] formalism is equivalent to the QTBM. It relies on the solution of the following system of equations:

$$\begin{aligned} [ES - H - \Sigma^R(E)]G^R(E) &= I, \\ G^<(E) &= G^R(E)\Sigma^<(E)G^A(E). \end{aligned} \quad (17)$$

The Green's functions here can be the retarded (G^R), advanced ($G^>$), or lesser ($G^<$) Green's function. The same holds true for the self-energies $\Sigma(E)$. An advantage of the NEGF over the QTBM is its ability to describe electron transport in the presence of scattering, which can be included naturally via self-energies. Here, we restrict ourselves to ballistic transport, where the QTBM offers greater computational efficiency. Once Eq. (16) or (17) is solved, the current and electron density can be extracted from the Green's function or the wave function [37,38]. Using the QTBM, computing $\Sigma^{<,R}(E)$ and solving Eq. (16) are performed in parallel on CPUs and graphics processing units (GPUs), respectively. In a postprocessing step, the two solutions are combined. The resulting transmission function $T(E)$ and the density of states (DOS) can then be calculated [39].

In general, H is a function of the electron density ρ , which can be extracted from $G^<(E)$ or $\Phi(E)$ and changes under the application of an external bias. This creates an interdependence between $H(\rho)$ and $G^<(E)$ that must be resolved self-consistently [5]. Previous studies have shown that this effect can be treated properly in MS calculations by transforming the charge density from MS back into RS to solve Poisson's equation in RS [11]. The same procedure can be applied to the MS part of the MS-RS hybrid scheme introduced in the next section. However, in the present work all simulations are performed in the low-bias regime, where the charge is close to its equilibrium value, so that H can be assumed to be constant, thus avoiding the need for a self-consistent solution.

A. Hybrid mode-space–real-space device hamiltonian

The Hamiltonian and overlap matrices H and S in Eqs. (16) and (17) are first computed with DFT, as explained in Sec. II A, using a localized GTO [28] basis set. Such a basis has the advantage of producing a banded H_{RS} that is directly suitable for Eqs. (16) and (17) without further processing [5,6]. To reduce the size of the matrices H and S , the MS transformation introduced in Sec. II B is employed. Because these matrices are not homogeneous throughout the device structure, the transformation is applied locally. In the contact extensions, the diagonal and off-diagonal blocks of H_{MS} are given by

$$\mathcal{H}_n = U^\dagger H_n U. \quad (18)$$

The H_n in Eq. (18) correspond to an on-site term ($n = 0, H_0 = H_{Cu}$) and to the coupling of one Cu unit cell, with 240 atoms, to its n th nearest neighbor ($n > 0$). The Cu blocks at the SiO_2 interface and the oxide regions remain in RS. At the transition between the periodic area of the device, which is transformed into MS, and the RS active region, the off-diagonal blocks must be treated carefully, as they exhibit a hybrid character. They undergo a one-sided transformation

$$\mathcal{H}_{RM,n} = U^\dagger H_n, \quad \mathcal{H}_{MR,n} = H_n U. \quad (19)$$

The subscript RM indicates that $\mathcal{H}_{RM,n}$ connects a RS to a MS region. The MS transformation is not optimized for the surface blocks of the Cu contacts, which are in direct contact with the SiO_2 layer and whose atomic structure is different from that of the other Cu cells due to the additional relaxation imposed by the neighboring oxide. Still, the MS transformation can theoretically be applied to them using the same matrix U as for the other Cu cells because they are made of the same number of atoms. In the next section, we present two strategies, one with all metal blocks transformed into MS and one with one metal block at the oxide interface remaining in RS. The sparsity patterns of the RS and the two hybrid Hamiltonian matrices are depicted in Fig. 3. The large Cu blocks at the interface between the RS and MS domains disappear and become rectangular. Only the active region, through which a metallic filament grows and then dissolves, remains in RS. As this region is not affected by the MS transformation, the matrix U does not need to be recomputed when structural changes occur within the oxide, e.g., atomic relocations. This feature is especially appealing when one is considering components whose atomic geometry evolves with time, such as CBRAM cells.

An alternative more general approach is to treat the entire simulation domain in RS and compute only the OBCs in MS. The coupling between MS and RS is achieved by transforming the boundary self-energies from MS to RS with $\Sigma_{RS}^R(E) = U \Sigma_{MS}^R(E) U^\dagger$ or by constructing

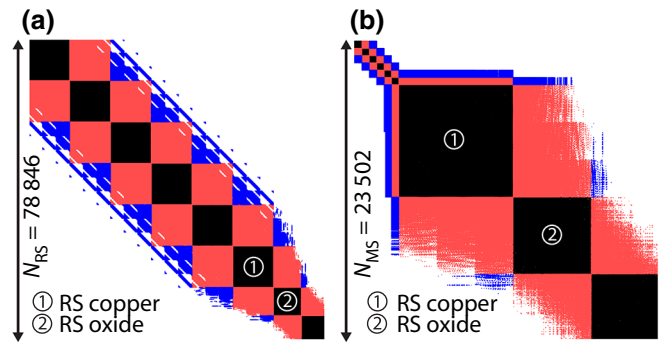


FIG. 3. Sparsity patterns of (a) the RS and (b) the hybrid MS-RS Hamiltonian matrices corresponding to the left half of structure B defined in Sec. II A. The presence of second-nearest-neighbor block connections translates into a pentadiagonal matrix structure, with on-site blocks (black), first-nearest-neighbor blocks (red), and second-nearest-neighbor blocks (blue). The contact of the hybrid Hamiltonian is composed of six MS blocks and one RS block. The overall matrix size is reduced from 78 846 to 23 502. Although the contact region is transformed into MS, the active device area remains in RS, which is recognizable from the much larger block sizes. The hybrid off-diagonal blocks connecting the RS and MS take the form of long thin rectangular blocks.

hybrid blocks connecting the MS boundary self-energy to the RS central part of the device. The benefit from the evaluation of the OBCs in MS is equal to that offered by the first approach. Solving the transport equations [Eq. (16) or (17)], on the other hand, does not benefit from the MS transformation. While providing a lower gain in computational efficiency, this method is generally applicable to any quantum transport simulation, regardless of the length of the contacts. As such, the proposed MS-RS scheme finds applications beyond the CBRAM cells considered.

It should be noted that the two contacts are independent of each other. While the device illustrated in Fig. 1(a) features two identical electrodes, the proposed hybrid scheme does not enforce such a constraint. The MS transformation matrix can be computed and applied to the two contacts independently by transforming the respective blocks obtained from the Hamiltonian and overlap matrices with Eqs. (18) and (19).

B. Ballistic-simulation benchmark

In the present section, we benchmark the MS-RS hybrid scheme and compare two different versions of it with a RS reference, in terms of both accuracy and computational efficiency. The first hybrid version, MS-RS₁, has all contact blocks transformed into MS using Eqs. (18) and (19), except for the surface Cu blocks, which form the interface with the insulating layer. In the second version, MS-RS₂, all metal blocks, including those in contact with SiO_2 , are

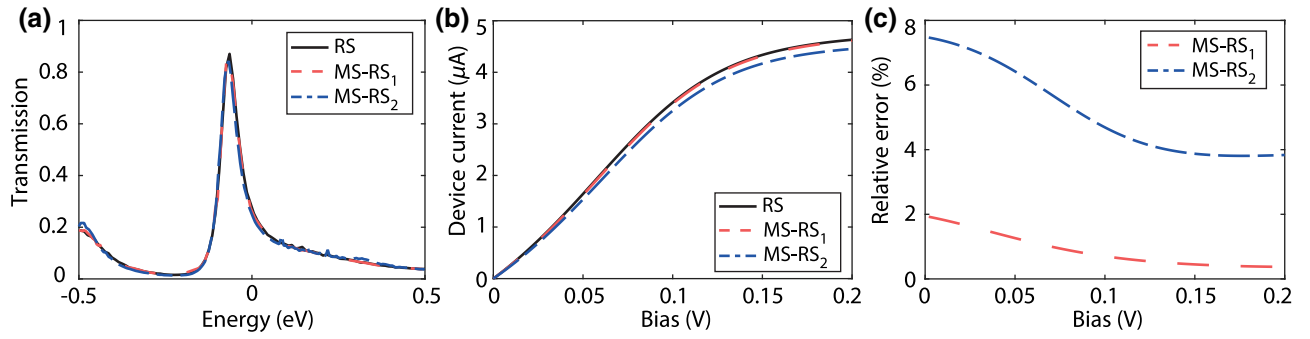


FIG. 4. (a) Ballistic transmission through the device structure shown schematically in Fig. 1 with different representations of the Hamiltonian. The RS reference curve is plotted in black, while the red and blue dashed curves correspond to the hybrid models MS-RS₁ and MS-RS₂, respectively. (b) Low-field current versus applied voltage between the Cu electrodes for the same structure and Hamiltonian matrices as in (a) (with the same plotting conventions). (c) Relative error between the hybrid models 1 (dashed red curve) and 2 (blue curve with small dashes) and the reference RS results.

transformed into MS. A single transformation matrix U is used in all cases.

To reduce the size of the RS simulations and fit the matrices into memory, structure B , the shorter version of the CBRAM cell introduced in Sec. II A, is used as a test bed. The length of the oxide layer is 1.6 nm, and the Cu contacts include six blocks, resulting in a total of 3318 atoms. The number of Cu atoms per contact cell is still equal to 240, so that the same transformation matrix U as before can be used.

The quantities of interest are the ballistic transmission function and the current flowing through the device. The transmission function is shown in Fig. 4(a) within the energy window where the MS transformation is valid. The MS-RS₁ model agrees almost perfectly with the RS reference; MS-RS₂, on the other hand, shows some discrepancies. The latter are attributed to the fact that projecting the relaxed surface block onto the same MS basis as for the perfectly crystalline Cu blocks does not work fully. The same behavior is reflected in the ballistic current

in Fig. 4(b), where it can be seen that MS-RS₁ follows almost exactly the RS reference, while MS-RS₂ slightly underestimates the current. The observed error is quantified in Fig. 4(c), where the relative differences between the currents obtained from the hybrid and the RS simulation are reported. For MS-RS₁, this error does not exceed 2% at very low bias, and vanishes at larger voltages. MS-RS₂ displays deviations of up to 8%.

All calculations are performed on CPU-GPU hybrid nodes with 64 GB of memory each, using 20 CPUs and 2 GPUs to treat each energy point. These computational resources are distributed over as few nodes as possible, while accommodating memory constraints. Detailed timings are recorded in Table III. The size of the matrices is reduced by factors of 3 and 7 for MS-RS₁ and MS-RS₂, respectively. The number of nonzero elements is cut down by even more, by factors of 6 and 25, reducing the memory requirement accordingly. The times required to calculate the OBCs are equal for the two hybrid schemes, as they depend only on the size of the boundary blocks, which is

TABLE III. Summary of computing times for the RS and the different RS-MS hybrid approaches. The calculation of the OBCs and the solution of the resulting linear system of equations are done in parallel, the former on CPUs, the latter on GPUs. The total time includes the aforementioned components plus postprocessing, such as the extraction of the density of states and the transmission function. The numbers are averaged over the calculation of 20 energy points. The number of nodes required to run the simulation is directly related to the memory needed to store and process the data. It shrinks with the number of nonzero elements in the Hamiltonian matrix. The total cost of the simulation, given as run time \times number of nodes, is reduced by 2 and almost 3 orders of magnitude with the help of the MS-RS₁ and MS-RS₂ schemes, respectively.

	RS	MS-RS ₁	Gain vs RS	MS-RS ₂	Gain vs RS
Total matrix size	78 846	22 726	3.47	11 502	6.86
Nonzero elements	673.8×10^6	104.9×10^6	6.43	27.5×10^5	24.51
Time for OBCs (s)	308.8	1.61	191.5	1.84	167.5
Time for linear system (s)	273.9	5.72	47.9	1.44	190.7
Time for $T(E)$ and DOS (s)	27.2	6.58	4.14	1.66	16.35
Total time (s)	336.0	12.3	27.3	3.52	95.5
Nodes	20	4	5	2	10
Cost (time \times nodes)	6720	49.3	136.4	7.19	954.9

the same in both cases. A speedup of more than 150 is achieved there. Solving the linear system resulting from Eq. (16) is shortened by factors of 48 and 191. The gain in overall computational time, on the other hand, is less significant than could be expected from the timings for the OBCs and the linear system, with speedups of 27 and 95 for MS-RS₁ and MS-RS₂, respectively. In effect, in the current implementation of the algorithm, the postprocessing stage, where $T(E)$ and the DOS are derived from the wave function, does not offer a substantial speedup in MS, as that depends mainly on the number of states injected into the simulation domain. This number is the same in RS and MS. The postprocessing accounts for less than 10% of the run time in the RS simulation, but it increases to about 50% of the time in both of the MS approaches. Addressing this bottleneck and optimizing the underlying algorithms for the hybrid scheme could lead to an even larger speedup. This is, however, outside the scope of this paper.

Other than the timing, the memory footprint is a large cost factor. While the RS simulations need to be performed on 20 nodes, the MS-RS simulations can run on 4 or 2 nodes, depending on the model used (1 or 2), adding to the cost reduction. Combining these two aspects into a cost function “run time \times number of nodes,” we report a reduction factor of 136 and 955 for MS-RS₁ and MS-RS₂, respectively.

C. Ballistic transport characteristics of the filament

In this section, we leverage the efficiency of the hybrid MS-RS approach to investigate how current flows through a filament formed in a CBRAM cell and how the surrounding oxide affects the transport properties. In accordance with the results in the previous section, the MS-RS₁ model is employed, with one RS metal block at the interface between the MS region and the amorphous oxide.

Structure *A*, introduced in Sec. II A, is used for this analysis. The oxide thickness is set to 3.5 nm, and the cross section of the overall structure to $2.1 \times 2.3 \text{ nm}^2$; the thin filament embedded in the SiO₂ matrix has a conical shape and is composed of 96 Cu atoms. The filament extends through the entire oxide layer and creates a conductive path between the two metallic electrodes, as shown in Fig. 1(a). A rendering of the lines of the current field in RS is plotted in Fig. 5(a). As the oxide layer and the immediately surrounding Cu cells are still in RS, the extraction of the RS current is straightforward; no scheme to transform the results back from MS to RS is needed. It can be observed that the current is confined within the filament. The field lines are focused onto the tip of the filament, where the highest current density is found. These lines also reveal that a substantial part of the current flows through the oxide around the tip, so that the narrowest current distribution happens slightly before the filament tip. The current

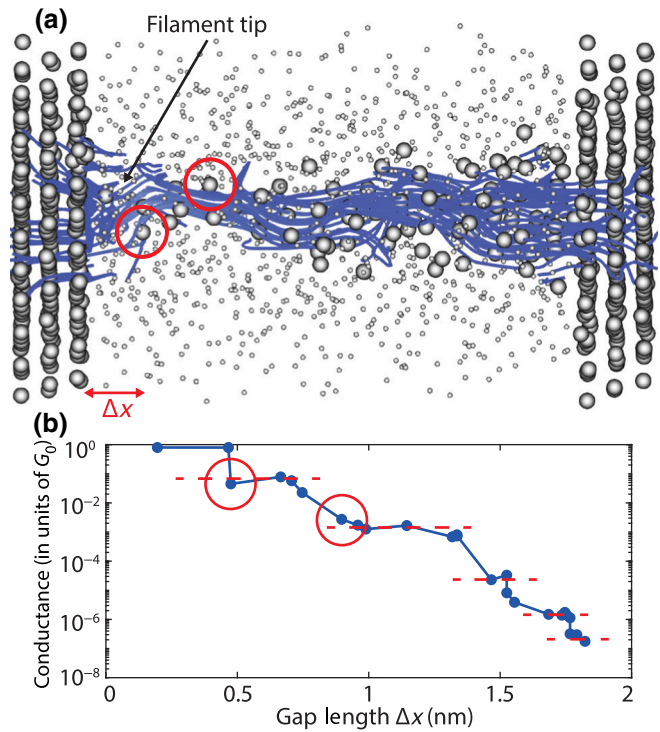


FIG. 5. (a) Rendering of the current field lines for the device shown in Fig. 1(a) under an applied bias of 1 mV. The large gray spheres represent Cu atoms, and the small spheres Si and O atoms. The current follows a “curly” path through the filament and is focused at the tip before spreading again through the right contact. Not all Cu atoms carry current; in particular, surface atoms and certain atoms at the tip are avoided, such as the ones circled in red. (b) Conductance versus Δx , the gap created at the filament tip when atoms are removed. Each point corresponds to a different filament configuration. The first point on the left represents the conductance of the device in part (a), and the second point the same structure but with one atom less at the tip of the filament. This procedure is repeated until a gap length $\Delta x = 1.8 \text{ nm}$ is obtained.

is rather “curly,” and not all atoms appear to contribute to the current.

The hysteretic I - V characteristics of CBRAM cells arise due to the relocation of atoms, a phenomenon that is computationally expensive to account for in *ab initio* quantum transport (QT) simulations [40], where atoms are typically considered to be static. The simplest method for implementing the dissolution of metallic filaments directly within a QT framework (QTBM or NEGF) consists of removing the electronic representation of the affected atoms from the Hamiltonian matrix. This trick enables one to monitor the electronic current during an *on*→*off* switching process at the atomic level. The filament is shortened by removing atoms from its tip, one by one, and the conductance of the remaining structure is recorded. Results are reported in Fig. 5(b) as a function of the length of the gap that is formed between the filament tip and the closest

metallic electrode. The conductance decreases exponentially as the gap increases, showing the same dependence as the current produced by the tunneling of an electron through a potential barrier. The decrease is, however, not uniform. It resembles a steplike process, as indicated by the presence of multiple plateaus in Fig. 5(b). The removal of certain atoms does not reduce the conductance, thus leading to the observed behavior. The atoms whose absence does not affect the conductance are those that do not carry any current in Fig. 5(a), i.e., those that are not crossed by field lines and that are circled in red.

Close investigations of experimental *on*→*off* switching characteristics in Ag/a-SiO₂/Pt CBRAM cells [41] show the same qualitative behavior. The conductance in a voltage-sweep experiment does not decrease continuously, but instead shows a steplike descent, just as observed in the simulation. In the experiment, this characteristic was attributed to discrete relocations of filament atoms that change the electrical properties of the filament. In the simulation, the deletion of a filament atom corresponds to the relocation of a Cu atom out of the simulation domain, thus mimicking the experimental diffusion.

IV. CONCLUSION AND OUTLOOK

We develop a robust scheme to select parameters for successful MS transformations of metallic layers. Based on a nonorthogonal DFT RS Hamiltonian, a size reduction of over 90% can be achieved. We also introduce a scheme to locally transform the Hamiltonian matrix of a device into MS. This hybrid MS-RS approach extends the use of MS to structures with inhomogeneities such as interfaces along the electron-transport direction. Numerical benchmarks against a RS reference reveal the accuracy of the proposed method, with relative errors of the electrical current remaining below 2%. The gain in computational efficiency together with the reduction in the memory footprint results in performance improvements by 2–3 orders of magnitude.

More generally, if the region transformed to MS and the energy window of interest remain the same, as in most resistive switching memory cells, the same transformation matrix can be reused each time the active layer undergoes a structural change. This feature renders our approach very powerful for investigating dynamically evolving devices. Here, this is demonstrated with a simplified model for the switching of a CBRAM cell. We find a steplike noncontinuous decrease in the conductance as the central metallic filament dissolves. We attribute this behavior to the presence of conductive and nonconductive atoms within the filament.

Apart from CBRAMs, the hybrid MS-RS approach is especially appealing for investigating devices with metallic contacts, where the treatment of these contacts dominates the overall computational time. This could allow,

for instance, the characterization of the contact resistance for a multitude of materials. While this resistance is often ignored in quantum transport simulations, it can severely limit device performance in reality.

To further improve the hybrid MS-RS approach, a scheme for including scattering through self-energies should be introduced. Mil'nikov *et al.* [10] demonstrated the feasibility of including scattering in pure mode-space simulations, but the proposed technique needs careful adjustments and verifications in the hybrid approach. Moreover, the Hamiltonian is not treated self-consistently in the present work. Shin *et al.* [11] performed self-consistent mode-space-only calculations. Errors in the charge density arising from the partial mode-space treatment of the simulation domain, even though small, could affect the electrostatic potential. Fully self-consistent hybrid simulations should be performed and compared with real-space simulations to verify the correctness of the approach. Solving the Poisson equation would enable the simulation of larger bias voltages than those considered here.

ACKNOWLEDGMENTS

This work was supported by the Werner Siemens Stiftung, by ETH Research Grant No. ETH-35 15-2, and by a grant from the Swiss National Supercomputing Centre (CSCS) under Projects s714 and s971.

-
- [1] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, The missing memristor found, *Nature* **453**, 80 (2008).
 - [2] I. Valov, R. Waser, J. R. Jameson, and M. N. Kozicki, Electrochemical metallization memories – fundamentals, applications, prospects, *Nanotechnology* **22**, 254003 (2011).
 - [3] W. Kohn and L. J. Sham, Self-consistent equations including exchange and correlation effects, *Phys. Rev.* **140**, A1133 (1965).
 - [4] L. P. Kadanoff and G. A. Baym, *Quantum Statistical Mechanics* (Benjamin, New York, 1962).
 - [5] J. Taylor, H. Guo, and J. Wang, Ab initio modeling of quantum transport properties of molecular electronic devices, *Phys. Rev. B: Condens. Matter Mater. Phys.* **63**, 245407 (2001).
 - [6] M. Brandbyge, J. L. Mozos, P. Ordejón, J. Taylor, and K. Stokbro, Density-functional method for nonequilibrium electron transport, *Phys. Rev. B: Condens. Matter Mater. Phys.* **65**, 1654011 (2002).
 - [7] J. Maassen, M. Harb, V. Michaud-Rioux, Y. Zhu, and H. Guo, Quantum transport modeling from first principles, *Proc. IEEE* **101**, 518 (2013).
 - [8] R. Lake, G. Klimeck, R. C. Bowen, and D. Jovanovic, Single and multiband modeling of quantum electron transport through layered semiconductor devices, *J. Appl. Phys.* **81**, 7845 (1997).

- [9] R. Venugopal, Z. Ren, S. Datta, M. S. Lundstrom, and D. Jovanovic, Simulating quantum transport in nanoscale transistors: Real versus mode-space approaches, *J. Appl. Phys.* **92**, 3730 (2002).
- [10] G. Mil'nikov, N. Mori, and Y. Kamakura, Equivalent transport models in atomistic quantum wires, *Phys. Rev. B: Condens. Matter Mater. Phys.* **85**, 035317 (2012).
- [11] M. Shin, W. J. Jeong, and J. Lee, Density functional theory based simulations of silicon nanowire field effect transistors, *J. Appl. Phys.* **119**, 154505 (2016).
- [12] M. Qiu, S. Ye, W. Wang, J. He, S. Chang, H. Wang, and Q. Huang, Spin transport properties of magnetic tunnel junction based on zinc blende CrS, *Superlattices Microstruct.* **133**, 106199 (2019).
- [13] M. Li and M. Smeu, Atomistic simulation of the structural and conductance evolution of Au break junctions, *Comput. Mater. Sci.* **164**, 147 (2019).
- [14] M. Ye, X. Jiang, S.-S. Li, and L.-W. Wang, in *2019 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, California, 2019), p. 24.6.1.
- [15] N. A. Lanzillo, B. D. Briggs, R. R. Robison, T. Standaert, and C. Lavoie, Electron transport across Cu/Ta(O)/Ru(O)/Cu interfaces in advanced vertical interconnects, *Comput. Mater. Sci.* **158**, 398 (2019).
- [16] S. Kim, M. Luisier, A. Paul, T. B. Boykin, and G. Klimeck, Full three-dimensional quantum transport simulation of atomistic interface roughness in silicon nanowire FETs, *IEEE Trans. Electron. Devices* **58**, 1371 (2011).
- [17] F. Ducry, M. Bani-Hashemian, and M. Luisier, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)* (IEEE, Udine, Italy, 2019), p. 1.
- [18] R. Haydock, The recursive solution of the Schroedinger equation, *Comput. Phys. Commun.* **20**, 11 (1980).
- [19] A. Svizhenko, M. P. Anantram, T. R. Govindan, B. Biegel, and R. Venugopal, Two-dimensional quantum mechanical modeling of nanotransistors, *J. Appl. Phys.* **91**, 2343 (2002).
- [20] F. Ducry, A. Emboras, S. Andermatt, M. H. Bani-Hashemian, B. Cheng, J. Leuthold, and M. Luisier, in *2017 IEEE International Electron Devices Meeting (IEDM)* (IEEE, San Francisco, California, 2017), p. 4.2.1.
- [21] B. Cheng, A. Emboras, Y. Salamin, F. Ducry, P. Ma, Y. Fedoryshyn, S. Andermatt, M. Luisier, and J. Leuthold, Ultra compact electrochemical metallization cells offering reproducible atomic scale memristive switching, *Commun. Phys.* **2**, 1 (2019).
- [22] J. Sarnthein, A. Pasquarello, and R. Car, Structural and Electronic Properties of Liquid and Amorphous SiO₂: An Ab Initio Molecular Dynamics Study, *Phys. Rev. Lett.* **74**, 4682 (1995).
- [23] QuantumATK version 2017.1, <https://www.synopsys.com/silicon/quantumatk.html>.
- [24] S. Smidstrup *et al.*, QuantumATK: An integrated platform of electronic and atomic-scale modelling tools, *J. Phys. Condens. Matter* **32**, 015901 (2020).
- [25] N. Onofrio, D. Guzman, and A. Strachan, Atomic origin of ultrafast resistance switching in nanoscale electrometallization cells, *Nat. Mater.* **14**, 440 (2015).
- [26] J. Hutter, M. Iannuzzi, F. Schiffrin, and J. VandeVondele, CP2K: Atomistic simulations of condensed matter systems, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **4**, 15 (2014).
- [27] J. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [28] J. VandeVondele and J. J. Hutter, Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases, *J. Chem. Phys.* **127**, 114105 (2007).
- [29] S. Goedecker, M. Teter, and J. Hutter, Separable dual-space Gaussian pseudopotentials, *Phys. Rev. B* **54**, 1703 (1996).
- [30] C. G. Broyden, The convergence of a class of double-rank minimization algorithms 1. General considerations, *IMA J. Appl. Math. (Inst. Math. Appl.)* **6**, 76 (1970).
- [31] R. Fletcher, A new approach to variable metric algorithms, *Comput. J.* **13**, 317 (1970).
- [32] D. Goldfarb, A family of variable-metric methods derived by variational means, *Math. Comput.* **24**, 23 (1970).
- [33] D. F. Shanno, Conditioning of quasi-Newton methods for function minimization, *Math. Comput.* **24**, 647 (1970).
- [34] J. Z. Huang, H. Hatikhameh, M. Povolotskiy, and G. Klimeck, Robust mode space approach for atomistic modeling of realistically large nanowire transistors, *J. Appl. Phys.* **123**, 044303 (2018).
- [35] L.-F. Wang and Y. Xia, A linear-time algorithm for globally maximizing the sum of a generalized rayleigh quotient and a quadratic form on the unit sphere, *SIAM J. Optim.* **29**, 1844 (2019).
- [36] C. S. Lent and D. J. Kirkner, The quantum transmitting boundary method, *J. Appl. Phys.* **67**, 6353 (1990).
- [37] Y. Meir and N. S. Wingreen, Landauer Formula for the Current Through an Interacting Electron Region, *Phys. Rev. Lett.* **68**, 2512 (1992).
- [38] M. Luisier and A. Schenk, Atomistic simulation of nanowire transistors, *J. Comput. Theor. Nanosci.* **5**, 1031 (2008).
- [39] M. Calderara, S. Brück, A. Pedersen, M. H. Bani-Hashemian, J. VandeVondele, and M. Luisier, in *SC'15: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (ACM, Austin, Texas, 2015), p. 1.
- [40] D. Stradi, U. G. Vej-Hansen, P. A. Khomyakov, M. E. Lee, G. Penazzi, A. Blom, J. Wellendorff, S. Smidstrup, and K. Stokbro, in *International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)* (IEEE, Udine, Italy, 2019).
- [41] A. Emboras, A. Alabastri, F. Ducry, B. Cheng, Y. Salamin, P. Ma, S. Andermatt, B. Baeuerle, A. Josten, C. Hafner, M. Luisier, P. Nordlander, and J. Leuthold, Atomic scale photodetection enabled by a memristive junction, *ACS Nano* **12**, 6706 (2018).