

Efficient Simulation of Gate Leakage Currents in Ultra-scaled Transistors from First Principles

Mauro Dossena, Nicolas Vetsch, Alexander Maeder, Anders Winka, Vincent Maillou,
Denghui Lu, Alexandros Nikolaos Ziogas, Jiang Cao and Mathieu Luisier
Integrated Systems Laboratory (IIS), ETH Zurich, Zurich, 8092, Switzerland
e-mail: mdossena@iis.ee.ethz.ch

Abstract—To ensure high electrostatic integrity in ultra-scaled field-effect transistors (FETs), thin gate dielectric layers are typically needed, potentially giving rise to (strong) leakage currents. Being able to model this effect is therefore crucial. Here, we show how gate leakage currents can be computed at the *ab initio* level in WS₂-hBN FETs by extending the well-established quantum transmitting boundary method (QTBM). Two main challenges are tackled: (i) the calculation of the open boundary conditions (OBCs) for the gate contacts and (ii) the solution of the resulting sparse linear system of equations $Ax = b$ where the matrix A possesses an unstructured sparsity pattern. We demonstrate that the proposed method outperforms the state of the art and can be readily applied to simulate transistors with gate lengths exceeding 25 nm.

I. INTRODUCTION

As a consequence of the continuous miniaturization of the transistor dimensions and of their equivalent oxide thickness (EOT), gate leakage currents could become a major limiting factor in future technology nodes. This issue might specifically affect logic switches with channels made of two-dimensional transition metal dichalcogenides (TMDCs), e.g., WS₂, for which no intrinsic oxides exist. To shed light on the importance of gate leakage in ultra-scaled devices, an *ab initio* quantum transport (QT) approach is crucial. Although such QT solvers have been used to investigate semiconductor-dielectric interfaces [1], [2], to the best of our knowledge, they have not yet been used to study gate leakage. While conventional *ab initio* QT tools typically support carrier injections at two contacts (source and drain) only, accounting for gate leakage currents requires introducing a third injection path at the gate location, which is placed on top of the oxide and extends across the entire channel region. In standard QT approaches, the presence of such a large spatial extent significantly degrades the computational performance because of (i) the calculation of the required open boundary conditions (OBCs) for the gate contact and (ii) the change in sparsity pattern of the resulting linear system of equations. In this work, we address these shortcomings by extending the quantum transmitting boundary method (QTBM) [3]. The developed approach is applied to the computation of the source-to-gate and drain-

to-gate transmission functions of a monolayer WS₂ field-effect transistor (FET) with a 1.8 nm thick hBN dielectric layer and a Pt gate metal contact as a representative example (Fig. 1(a)). The Hamiltonian and overlap matrices of the three different material regions defining this geometry, namely, the contacts consisting only of WS₂, the buffer area including both WS₂ and hBN, and the central area consisting of WS₂, hBN, and Pt, are determined separately at the density functional theory (DFT) level with CP2K [4], whose basis relies on a linear combination of atomic orbitals (LCAO). Subsequently, the three regions are combined to form the complete device Hamiltonian, H , and the corresponding orbital overlap matrix, S , following the procedure outlined in Ref. [1]. These device matrices are passed to our open-source QT solver, *quatrex* [5], which implements the QTBM to compute the transport properties of multi-terminal nanoelectronic devices.

II. METHODS

A. Quantum Transmitting Boundary Method

The QTBM is a computationally efficient framework to evaluate phase-coherent electron transport through nanostructures. It has been widely used, in combination with the effective mass approximation, tight-binding methods, and *ab initio* models. It allows for the incorporation of the effects of semi-infinite leads into the Schrödinger equation, giving rise to an equation of the form [6]

$$[ES - H - \Sigma^B(E)] \Phi(E) = \text{Inj}(E). \quad (1)$$

Here, E is the electron energy at which the system is solved, Σ^B is the open boundary self-energy, capturing the interaction of the system with all connected semi-infinite leads, and Inj is the injection matrix describing the electron modes entering the device from the semi-infinite leads. The unknown quantity, Φ , represents the electron wavefunctions generated inside the device. In Eq. (1), Σ^B is a sparse matrix whose non-zero entries correspond to orbitals directly coupling the central device region to its contact extensions. It can be partitioned as $\Sigma^B = \Sigma_S^B + \Sigma_D^B + \Sigma_G^B$, where each individual term refers to a contribution from a different contact, i.e., the source, drain, and gate, respectively. The same decomposition applies to the injection matrix $\text{Inj} = \text{Inj}_S + \text{Inj}_D + \text{Inj}_G$. The computation of all terms in Σ^B and Inj involves the solution of non-linear eigenvalue problems (NEVPs) [6], whose time complexity scales with $\mathcal{O}(N^3)$, N being the number of orbitals directly

This work was supported by the Swiss National Science Foundation (SNSF) under grant 205602 (NCCR MARVEL) and grant 209358 (QuaTrEx), and by the Platform for Advanced Scientific Computing in Switzerland (BoostQT). We acknowledge support from the Swiss National Supercomputing Center (CSCS) under projects c33, lp16, and lp82.

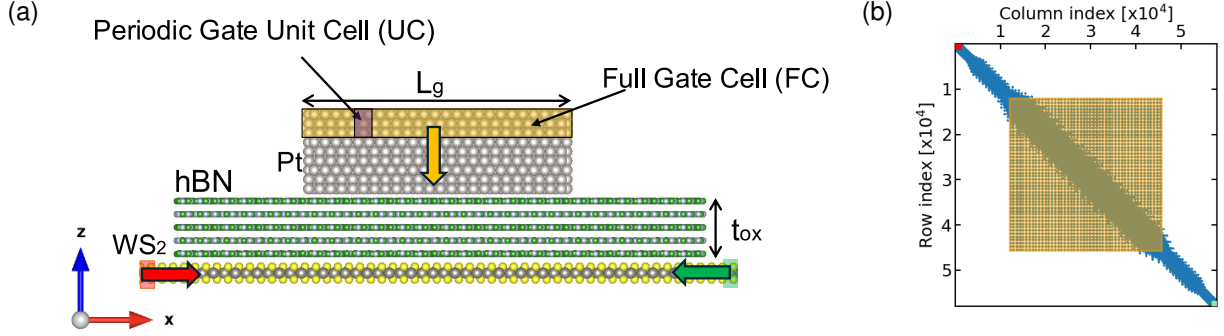


Fig. 1. (a) Overview of the WS₂-hBN-Pt transistor structure considered in this work. It comprises a semiconducting WS₂ channel, a 1.8 nm thick hBN insulator layer, and a Pt gate of varying length. Periodic boundary conditions are assumed along the y direction, and transport occurs along the x and z axes. The source, drain, and gate contact regions are highlighted in red, green, and yellow, respectively, while the arrows represent electron injection. The gate contact can either be treated as a single entity (Full Gate Cell, FC) or decomposed into smaller, periodically repeated unit cells depicted in purple (Periodic Gate Unit Cell, UC). (b) Sparsity pattern of the $[ES - H - \Sigma^B]$ matrix corresponding to the device geometry in (a) ($L_g \approx 6$ nm) and entering our QTBM solver. Specifically, the entries of Σ^B are marked in red, green, and yellow for the source, drain, and gate contacts.

connected to the contacts. A first problem becomes apparent here: Computing Σ_G^B becomes increasingly difficult for long gate electrodes due to the cubic scaling of the NEVP. The sparsity pattern of the system matrix $[ES - H - \Sigma^B]$ for a device with $L_g \approx 6$ nm is shown in Fig. 1(b), where the contribution of each contact is highlighted with a different color. This illustrates the second challenge: The Σ_G^B term leads to an unstructured sparsity pattern, which drastically increases the fill-in occurring during a direct solution of Eq. (1) via LU factorization. In addition, standard block Gaussian elimination techniques, such as the recursive Green's function algorithm (RGF), do not perform well here due to the large size of the central block induced by the gate contact.

B. Bloch Expansion of Open Boundary Conditions

To overcome the first limitation, we divide the full cell (FC) gate contact (yellow-shaded region in Fig. 1) into smaller, repeatable unit cells (UC) that correspond to the purple-shaded region in Fig. 1. In the case of a periodic contact along two directions, e.g., x and y , the UC can be repeated α times along x and β times along y . Following the Bloch expansion procedure of Refs. [7], [8], we introduce a set of $N_k = \alpha \times \beta$ auxiliary wave vectors (k_a, k_b) sampled over a uniform Monkhorst-Pack grid and compute a UC boundary self-energy $\sigma_G^B(k_a, k_b)$ and UC injection matrix $\text{inj}_G(k_a, k_b)$ for each (k_a, k_b) combination. The FC Σ_G^B and Inj_G are reconstructed as

$$\Sigma_G^B = \frac{1}{N_k} \sum_{k_a, k_b} \begin{bmatrix} 1 & e^{-ik_a} & \dots & e^{-i(\alpha-1)k_a} \\ e^{ik_a} & 1 & \dots & e^{-i(\alpha-2)k_a} \\ \vdots & \vdots & \ddots & \vdots \\ e^{i(\alpha-1)k_a} & e^{i(\alpha-2)k_a} & \dots & 1 \end{bmatrix} \otimes \begin{bmatrix} 1 & e^{-ik_b} & \dots & e^{-i(\beta-1)k_b} \\ e^{ik_b} & 1 & \dots & e^{-i(\beta-2)k_b} \\ \vdots & \vdots & \ddots & \vdots \\ e^{i(\beta-1)k_b} & e^{i(\beta-2)k_b} & \dots & 1 \end{bmatrix} \otimes \sigma_G^B(k_a, k_b) \quad (2)$$

$$\text{Inj}_G = \frac{1}{\sqrt{N_k}} [\gamma(k_0, k_0) | \gamma(k_0, k_1) | \dots] \quad (3)$$

$$\gamma(k_a, k_b) = \begin{bmatrix} 1 \\ e^{ik_a} \\ \vdots \\ e^{i(\alpha-1)k_a} \end{bmatrix} \otimes \begin{bmatrix} 1 \\ e^{ik_b} \\ \vdots \\ e^{i(\beta-1)k_b} \end{bmatrix} \otimes \text{inj}_G(k_a, k_b), \quad (4)$$

where $\otimes$ denotes the Kronecker product.

Solving the NEVP on a periodic UC grid reduces the computational complexity of the OBC calculation from $\mathcal{O}(N^3)$ down to $\mathcal{O}(N_k(N/N_k)^3)$. However, it should be noted that this approach is exact only in the case of a truly periodic contact cell extending infinitely in all transverse directions. In our device configuration, since the gate has finite dimensions along the x direction, this method must be considered an approximation and requires careful validation (see below).

C. Efficient Solution of the System of Equations

The second issue, i.e., the unstructured sparsity pattern of the system matrix, is tackled with a special low-rank treatment of Σ_G^B . The $N_G \times N_G$ open boundary self-energy matrix Σ_G^B lends itself to a low-rank decomposition [6]

$$\Sigma_G^B \approx \text{Ref}_G \Lambda_G^{-1} \bar{\varphi}_G, \quad (5)$$

where Ref_G is an $N_G \times n_G$ matrix representing the decaying reflected modes at the contact, Λ_G^{-1} is an $n_G \times n_G$ diagonal matrix containing the phase factor of each mode, and $\bar{\varphi}_G$ is an $n_G \times N_G$ matrix representing the dual (inverse) of every decaying mode. Notably, $n_G \ll N_G$ (Fig. 2). Analogous decompositions apply to Σ_S^B and Σ_D^B , and can thus be extended to the entire Σ^B matrix. By combining the low-rank approximation with the Woodbury identity [9], a modified QTBM equation can be derived

$$(ES - H) [\Phi', \Phi''] = [\text{Inj}, \text{Ref}], \quad (6)$$

where Σ^B no longer explicitly enters the system matrix, but is replaced by additional columns on the right-hand side of

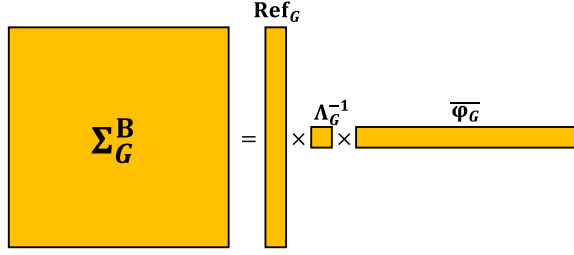


Fig. 2. Overview of the low-rank approximation employed to decompose the gate boundary self-energy Σ_G^B into the product of three matrices, modifying the QTBM numerical problem.

the equation. The solution of the original QTBM problem can be recovered via matrix-vector multiplications:

$$\Phi = \Phi' + \Phi'' (\Lambda - \bar{\varphi} \Phi'')^{-1} \bar{\varphi} \Phi'. \quad (7)$$

The system of equations in the modified formulation of Eq. (6) now exhibits a structured sparsity pattern that is more amenable to LU factorization and can potentially be solved via block Gaussian elimination algorithms. Another favorable property of the modified system is that it is Hermitian, requiring only roughly half of the memory and operations compared to the solution of the non-Hermitian system in Eq. (1). As with the full-rank boundary treatment, the low-rank approximation of Σ^B allows for a Bloch expansion. In this case, the three matrices $\mathbf{Ref}$, Λ^{-1} , and $\bar{\varphi}$ can be upscaled from their UC counterparts, following an approach similar to the upscaling of Eq. (4). Combining the two approximations enables the simulation of devices with L_g exceeding 25 nm.

III. RESULTS

A. Validation of Approximations

The two approximations introduced in the previous section are examined here. By selecting a device with $L_g = 6.4$ nm, we can compute a baseline solution without approximations where the NEVP of the full gate cell is solved and the full-rank Σ_G^B in Eq. (1) is used. This baseline is then compared to a solution employing only the low-rank technique and another one exploiting both the Bloch expansion and low-rank OBC approximation. To make the simulation conditions as realistic as possible, an external electrostatic potential bending the channel's conduction band is added to the Hamiltonian, $\mathbf{H}$. It is shown in the insets of Fig. 3. The resulting drain-to-gate transmission functions for the three aforementioned cases are displayed in Fig. 3(a). All curves agree excellently, demonstrating the validity of the low-rank and Bloch expansion approximations, even at $L_g = 6.4$ nm.

Importantly, this method enables the simulation of gate leakage in devices with realistically-sized gate contacts. To illustrate our model's capabilities, the source-to-gate and drain-to-gate transmission functions are computed for a structure with $L_g = 25.6$ nm, as depicted in Fig. 3(b). As can be expected from the steeper potential drop on the drain side of the FET, the drain-to-gate transmission function is larger than its source-to-gate counterpart. Both gate leakage transmission

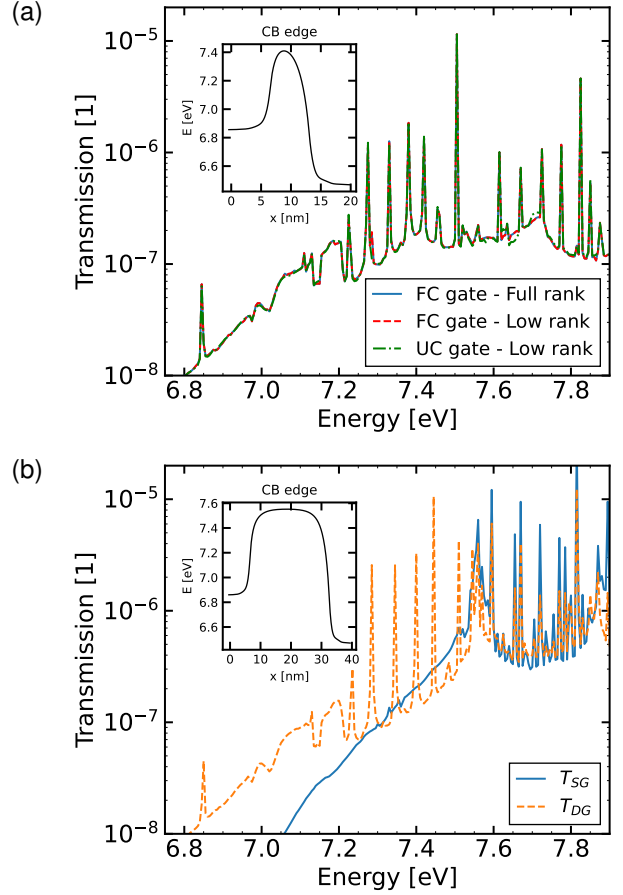


Fig. 3. (a) Comparison of the drain-to-gate transmission function in the WS₂-hBN-Pt transistor from Fig. 1(a) with a gate length $L_g = 6.4$ nm, as calculated with the original QTBM approach (blue line), with the low-rank approximation (dashed red line), and with the additional division of the gate electrode into small, repeatable unit cells (dashed-dotted green line). The conduction band (CB) edge of the device is shown in the inset. (b) Source-to-gate (blue) and drain-to-gate (orange) transmission functions for the same WS₂-hBN-Pt structure as in (a), but with $L_g = 25.6$ nm.

functions exhibit distinct peaks around 7.55 eV, which corresponds to conduction band states in the WS₂ channel. Electrons injected into the channel can tunnel into the gate along its entire extent, increasing the overall leakage probability at this specific energy.

B. Numerical Performance

To complete our analysis, in Fig. 4 we provide performance measurements to illustrate the computational benefits of our approach, in terms of runtime and memory. All benchmarks were run on a single NVIDIA GH200 chip with 96 GB of GPU memory. Figure 4(a) demonstrates how the Bloch expansion method decreases the simulation time and memory footprint, assuming the low-rank approximation is employed to solve the QTBM system. In particular, explicitly treating the full gate contacts limits simulations to $L_g \leq 12.8$ nm. Conversely, applying the Bloch expansion approximation, we find more favorable time and memory scaling (up to $5\times$ savings for $L_g = 12.8$ nm), enabling OBCs computation in structures with $L_g =$

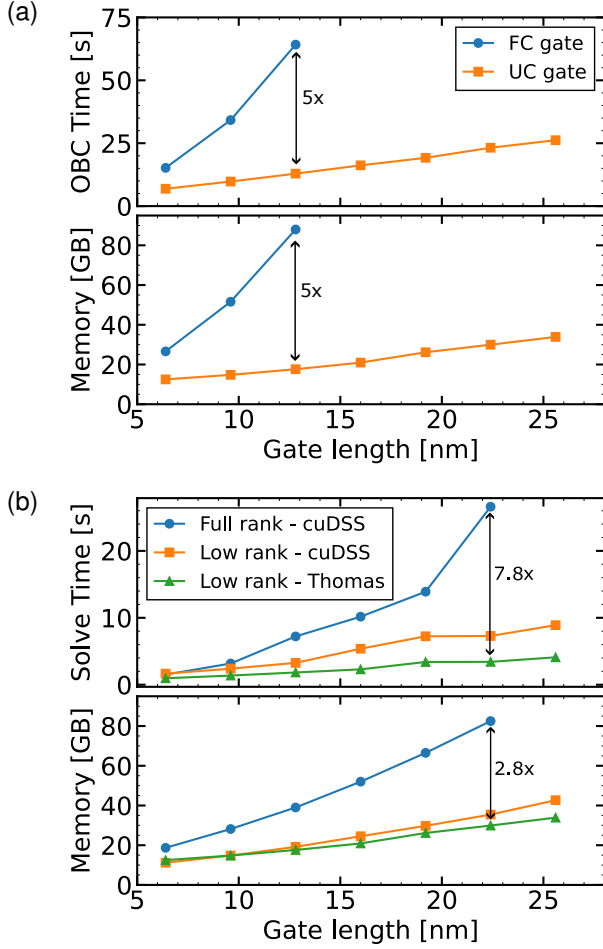


Fig. 4. (a) Runtime (top) and memory footprint (bottom) on an NVIDIA GH200 GPU during the calculation of the gate boundary self-energy Σ_G^B as a function of the gate length for the WS₂-hBN-Pt transistor from Fig. 1(a). The blue lines with circles refer to the case where the entire gate contact is treated as a single block, while the orange lines with squares correspond to the unit cell Bloch expansion method. (b) Same as (a), but for the solution of the QTBM problem. Three solvers are compared: NVIDIA cuDSS using the full-rank Σ^B (blue lines with circles), cuDSS when employing the low-rank approximation (orange lines with squares), and a custom implementation of the block Thomas algorithm with low-rank approximation (green lines with triangles).

25.6 nm. In Fig. 4(b), the performance impact of the low-rank approximation is reported under the assumption that the Bloch expansion approximation is used when solving the OBCs. We performed simulations with and without the low-rank approximation using NVIDIA's cuDSS sparse direct solver. In addition, with the low-rank approximation, we can employ a custom implementation of the block Thomas algorithm [9], which is a block Gaussian elimination algorithm, similar to RGF. Due to the large size of the central matrix block, the solution is not possible in the full-rank QTBM formulation. The low-rank approximation provides further computational benefits, namely runtime and memory savings of 7.8 $\times$ and 2.8 $\times$, respectively, at $L_g = 22.4$ nm. Table I summarizes the performance results.

TABLE I
SUMMARY OF NUMERICAL PERFORMANCE RESULTS

Gate Length (L_g [nm])	6.4	9.6	12.8	16.0	19.2	22.4	25.6
Time (s)							
Full Cell Gate	15.2	34.2	64.2	—	—	—	—
OBCs Unit Cell Gate	6.9	9.8	12.9	16.2	19.2	23.2	26.2
Speedup	2.2 $\times$	3.5 $\times$	5.0 $\times$	—	—	—	—
Full Rank - cuDSS	1.5	3.2	7.2	10.2	13.9	26.6	—
Solve Low Rank - Thomas	1.0	1.4	1.8	2.3	3.4	3.4	4.1
Speedup	1.5 $\times$	2.3 $\times$	4.0 $\times$	4.4 $\times$	4.1 $\times$	7.8 $\times$	—
GPU Memory (GB)[†]							
Full Cell Gate	26.6	51.5	88.0	—	—	—	—
OBCs Unit Cell Gate	12.5	14.8	17.6	21.0	26.2	30.0	33.9
Memory Saving	2.1 $\times$	3.5 $\times$	5.0 $\times$	—	—	—	—
Full Rank - cuDSS	18.7	28.2	39.0	52.0	66.5	82.5	—
Solve Low Rank - Thomas	12.5	14.8	17.6	21.0	26.2	30.0	33.9
Memory Saving	1.5 $\times$	1.9 $\times$	2.2 $\times$	2.5 $\times$	2.5 $\times$	2.8 $\times$	—

[†] Memory values are the peak of the full simulation (OBCs + solve).

IV. CONCLUSION

In this paper, we presented a QTBM formulation that combines a Bloch expansion of the OBCs and a low-rank approximation. We demonstrated that our approach allows us to compute the gate leakage transmission functions of realistic nanoelectronic devices with a 2D channel and gate lengths $L_g > 25$ nm. This achievement paves the way for investigating the influence of different dielectric layers on the leakage properties of ultra-scaled FETs.

REFERENCES

- [1] M. Dossena et al., "Mobility calculation in disordered WS₂-Al₂O₃ stacks from first principles," npj 2D Mater. Appl., vol. 9, no. 1, p. 67, Aug. 2025, doi: 10.1038/s41699-025-00587-9.
- [2] F. Ducry, B. Van Troeye, M. Dossena, M. Luisier, G. Pourtois, and A. Afzalian, "Oxide induced degradation in MoS₂ field-effect transistors," npj 2D Mater. Appl., vol. 10, no. 1, p. 49, Mar. 2026, doi: 10.1038/s41699-026-00677-2.
- [3] C. S. Lent and D. J. Kirkner, "The quantum transmitting boundary method," J. Appl. Phys., vol. 67, no. 10, pp. 6353–6359, May 1990, doi: 10.1063/1.345156.
- [4] T. D. Kühne et al., "CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations," J. Chem. Phys., vol. 152, no. 19, p. 194103, May 2020, doi: 10.1063/5.0007045.
- [5] quatrex/quatrex. Accessed: Jun. 22, 2026. [Online]. Available: <https://github.com/quatrex/quatrex>
- [6] S. Brück, M. Calderara, M. H. Bani-Hashemian, J. VandeVondele, and M. Luisier, "Efficient algorithms for large-scale quantum transport calculations," J. Chem. Phys., vol. 147, no. 7, p. 074116, Aug. 2017, doi: 10.1063/1.4998421.
- [7] G. Gandus, Y. Lee, D. Passerone, and M. Luisier, "Efficient partitioning of surface Green's function: toward ab initio contact resistance study," in 2020 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD), Sep. 2020, pp. 177–180. doi: 10.23919/SISPAD49475.2020.9241649.
- [8] N. Papior, N. Lorente, T. Frederiksen, A. García, and M. Brandbyge, "Improvements on non-equilibrium and transport Green function techniques: The next-generation transiesta," Comput. Phys. Commun., vol. 212, pp. 8–24, Mar. 2017, doi: 10.1016/j.cpc.2016.09.022.
- [9] A. Quarteroni, R. Sacco, and F. Saleri, Numerical Mathematics, in Texts in Applied Mathematics, vol. 37. New York, NY: Springer, 2007. doi: 10.1007/b98885.