Accelerated Modelling of Interfaces for Electronic Devices using Graph Neural Networks

Pratik Brahma[1*], **Krishnakumar Bhattaram**[1*], **Sayeef Salahuddin**[1,2]

[1] Department of Electrical Engineering and Computer Sciences
University of California Berkeley
[2] Materials Science Division, Lawrence Berkeley National Laboratory
{pratik_brahma, krishnabhattaram, sayeef}@berkeley.edu

Abstract

Modern microelectronic devices are composed of interfaces between a large number of materials, many of which are in amorphous or polycrystalline phases. Modeling such non-crystalline materials using first-principles methods such as density functional theory is often numerically intractable. Recently, graph neural networks (GNNs) have shown potential to achieve linear complexity with accuracies comparable to ab-initio methods. Here, we demonstrate the applicability of GNNs to accelerate the atomistic computational pipeline for predicting macroscopic transistor transport characteristics via learning microscopic physical properties. We generate amorphous heterostructures, specifically the HfO_2-SiO_2-Si semiconductor-dielectric transistor gate stack, via GNN predicted atomic forces, and show excellent accuracy in predicting transport characteristics including injection velocity for nanoslab silicon channels. This work paves the way for faster and more scalable methods to model modern advanced electronic devices via GNNs.

1   Introduction

The modern Si transistor gate stack comprises of multiple material interfaces whose electronic interactions significantly affect electron transport and ultimately transistor performance. In particular, the heterogeneous semiconductor-dielectric gate stack introduces many atomic-scale modeling challenges, which, if addressed, can help design higher-performance gate stacks for next-generation transistors. Fundamentally, the starting point for modeling any transport process of an atomistic electronic device stems the continuity equation $\frac{\partial Q}{\partial t} = -\bar{\nabla} \cdot \bar{J}$, where Q is the non quasi-static charge and J is current, which in turn is a function of Q and injection velocity (v_{inj}) [10, 8]. In practical devices, a significant contribution to Q comes from parasitic sources, which traditional Poisson solvers[15] capture well. The main challenge in transport calculations is calculating the intrinsic Q and v_{inj}, which depend on the specific combination of materials interfaces and confinement effects. When amorphous/polycrystalline phases are involved, one cannot directly leverage the E-k diagram (eigenvalues of the quantum Hamiltonian of a crystalline material). Therefore, the Density of States (DOS) is then used to calculate all relevant parameters, including Q and v_{inj}, which are later used in electrostatic solvers and transport models to directly estimate the fast behavior of nanoscale devices. Fig.1 summarises this atomistic computational pipeline for calculating transistor characteristics from the macroscopic transistor dimensions. This pipeline has two bottlenecks for scalability: (i) Molecular Dynamics (MD), which generates atomistic transistor gate stacks with different structural phases, and (ii) Electronic Structure Calculations, which generate atomistic properties like DOS from the quantum Hamiltonian. These bottlenecks arise as the current state-of-the-art atomistic simulations
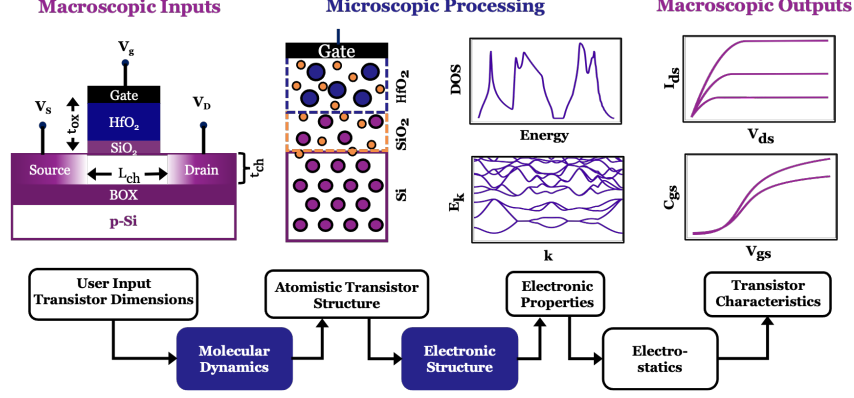
---

* These authors contributed equally to this work

Figure 1: **Atomistic Computational Pipeline**: Procedure for ab initio accurate predictions of advanced transistor characteristics containing various material interfaces given the macroscopic transistor dimensions. The first figure shows a Fully Depleted Silicon on Insulator(FDSOI) transistor, with a gate oxide stack containing HfO_2 and SiO_2 and a metal gate terminal. BOX is buried silicon oxide that lowers the parasitic capacitance between the source and drain and also efficiently confines the flow of electrons. The blue color boxes represent the current bottlenecks for scalable simulations of large atomistic devices. We propose to substitute these blocks with GNNs for accelerated predictions.

that model interactions from first principles[6], diagonalize the quantum Hamiltonian, an operation that scales cubically with system size. This poses a challenge for fast and accurate simulations of practically large material systems containing thousands of atoms and varied crystalline states.

Following the success of graph neural networks (GNNs)[11, 18] in learning ab initio molecular forces[9] and predicting atomistic properties [4, 3], we propose to overcome the scaling challenge by learning the functional relationship between atomic local chemical environments and macroscopic transistor properties. As GNN inference (summarized in Fig.2) scales linearly with system size, orders of magnitude speedup can be realized. Other existing neural network algorithms for transistor characterization prioritize speed but sacrifice generalizability to unseen geometries and complex material interfaces [17] by inferring on macroscopic scales and ignoring microscopic properties. Our work focuses on learning atomistic contributions to macroscopic transistor characteristics, which we demonstrate yields accurate and generalizable predictions even on unseen transistor geometries.

2   Methods

Neural Network Architecture: Our GNN architecture is composed of the Schnet[11] and SpookyNet[18] architectures. The forward pass of the GNN (Fig.2) is divided into three phases:

Atom Embedding: The atomistic transistor structure is modeled as a graph where each atom is a node, and each neighboring atom interaction within a given cutoff r_c is a weighted edge. Every node is initialized with a random vector (x_v^t) according to its atomic number, and every distance vector $(\vec{r})$ to its neighboring atoms is projected in radial (R) and spherical harmonic basis (Y_l^m)[18]. This phase emulates a system of isolated atoms unaware of its local chemical environment.

Message Passing: This phase starts with an isolated system of atoms and allows interactions with atomic neighbors to influence the atomic chemical state. Continuous convolutional filter-generated messages (m_v^t) are sent over the edges as given in Fig.2b. The state vector (x_v^t) of each node is updated by summing up the incoming messages (m_{vj}^t) along the edges j as $x_v^{t+1} = x_v^t + \sum_{j \in \mathcal{N}(v)} m_{vj}^t$.

Readout: This phase calculates the local atomic contributions of the desired global property from the state vector of all nodes. We focus on two sets of properties: (i) Energy and Atomic Forces: These properties generate the atomistic transistor gate stack using MD. A dense linear layer predicts the local atomic energy using the final atomic state of each node. Summing up the local atomistic energy predictions give total energy, and its derivative gives the atomic forces. (ii) Injection Velocity: This property characterizes the drain current through small channel transistors[10, 8]. It relates to the average velocity of all electrons over the source-channel potential barrier (U). In the ballistic limit,
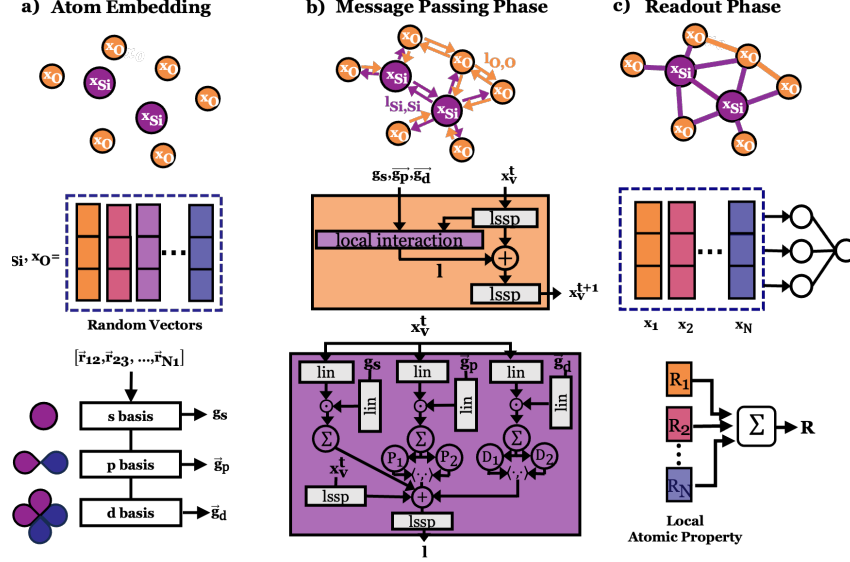
Figure 2: **Graph Neural Network Architecture**: The forward pass of a GNN is divided into three phases: a) Atom Embedding, b) Message Passing and c) Readout.

drain current (I_D) through a transistor is related to injection velocity as $I_D = qN_{inv}v_{inj}$, where:

$$v_{inj} = \frac{\int dE v_x(E)D(E)f(E+U-E_f)}{N_{inv}}, \quad N_{inv} = \int dE D(E)f(E+U-E_f)$$

N_{inv} is the inversion electron density present in the silicon channel, $v_x(E)$ is the bandstructure velocity of the electron at energy E and $D(E)$ is the DOS in the silicon channel. A dense linear layer with multiple outputs predicts $D(E)$ and $J_x(E) = v_x(E)D(E)$ from the final node state vectors. We perform PCA on the dataset to reduce the number of output nodes[1]. v_{inj} and N_{inv} are subsequently calculated using the predicted properties at a given fermi level (E_f).

Datasets: Two datasets are generated related to the transistor gate stack and silicon channel.

Gate Stack: The transistor gate stack, a heterostructure of crystalline silicon, amorphous silica, and hafnia, is generated via a high-temperature quench using the LAMMPS MD package[16]. Forces between atoms are estimated using the Charge-Optimized Many Body (COMB) potential[13, 12]. The crystalline forms β-cristobalite and orthorhombic HfO_2 are melted using constant number, pressure, temperature (NPT) dynamics at 2700K and 3500K, respectively. Subsequently, we reshape the melts to match the silicon substrate area using non-equilibrium constant number, volume, temperature (NVT) dynamics and quenched via a damped force minimizer. The generated amorphous silica, hafnia, and crystalline silicon are stacked on each other, and the material interfaces are allowed to relax using the COMB potential. This procedure generates a dataset of ∼200k molecular structures ranging from 25-96 atoms.

Silicon channel: We consider the silicon channel of the transistor as a nanoslab passivated by hydrogen atoms. The empirical $sp^3d^5s^*$ tight-binding model in Quantum ATK[14, 7] generates the electronic properties of the silicon channel, primarily $D(E)$ and $J_x(E)$. Around 1k structures are generated by varying the strain (0.900-1.035) and the nanoslab silicon channel thickness (0.6-2.4 nm).

3   Results

Gate Stack Generation: We simultaneously train on atomic forces and energy using a 90-10 weighted sum of mean squared error (MSE) losses, which yields a final mean absolute error of 3.0e-2 eV/Å (0.26% error) for the MD dataset on force predictions. The trained model is then used to generate gate stacks of around 200 atoms, a factor of 2 larger than the structures provided during training. The validity of the GNN-generated heterostructure (Fig.3a) is confirmed by the
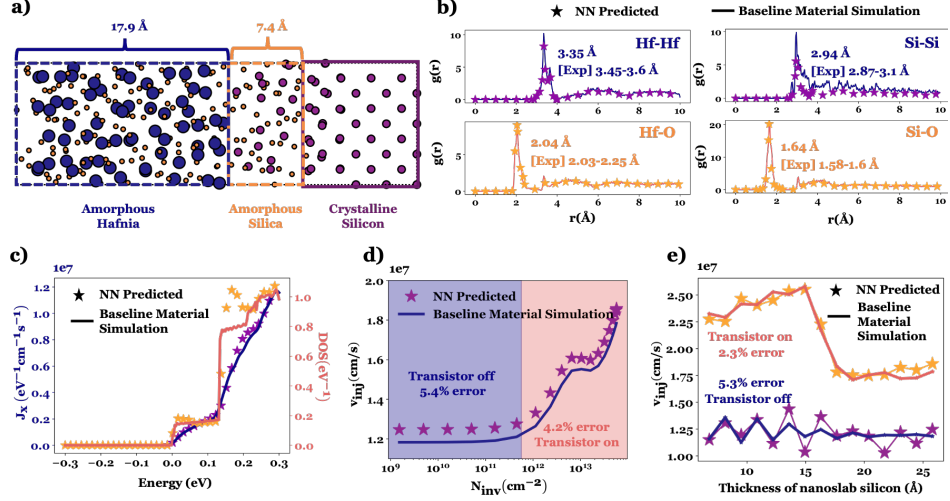
Figure 3: **Gate Stack Generation and Injection Velocity Prediction**: (a) The gate stack heterostructure generated via the GNN predicted atomic forces. (b) Excellent agreement between the $g(r)$ of the heterostructure generated by GNN and the baseline material simulation. [Exp] refers to the experimental bond length values[2, 5]. (c), (d) Excellent agreement of the predicted $D(E)$, $J_x(E)$, and v_{inj} between the GNN and the baseline model for a 2.6nm thick unstrained nanoslab silicon. (e) GNN correctly predicts the transistor v_{inj} trend as a function of the unstrained nanoslab silicon channel thickness.+

excellent match of the pair distribution functions $g(r)$ of amorphous silica and hafnia to the baseline simulation model (Fig.3b). Predicted Si-Si, Hf-Hf, Si-O, and Hf-O bond lengths are within 3% of their experimental values Fig.3b)[2, 5] This demonstrates that our approach of learning local chemical environments can be generalized to predicting atomic forces and energy in transistor geometries unseen in the initial training set.

Injection Velocity: We simultaneously train on $D(E)$ and $J_x(E)$ using an equally weighted sum of MSE losses, which yields a final mean absolute error of 9.0e-4 /eV (0.18% error) for $D(E)$ and 4.9e4 cm/s-eV (0.82% error) for $J_x(E)$. The trained task-specific GNN predicts PCA coefficients for $D(E)$ and $J_x(E)$ over 200 and 15 basis functions respectively for the crystalline silicon nanoslab. We found high model performance in reproducing $D(E)$ and $J_x(E)$ for a 2.6nm thick unstrained silicon nanoslab, a structure larger than in the training set, as shown in Fig.3c. From these predictions, we derived v_{inj} for a range of chemical potentials E_f, finding errors within 5.4% for both the on and off states of the transistor (Fig. 3d) Furthermore, we evaluate our neural network on unstrained silicon nanoslabs with thicknesses ranging from 0.67 to 2.4 nm. We reproduce v_{inj} as a function of thickness with high fidelity (within 5.3%) (Fig.3e), demonstrating the ability of our model to successfully predict macroscopic dynamics of unseen geometries of silicon channel. The runtime for predicting injection velocity by the GNN is around 20ms, while the baseline simulation takes 430s, demonstrating four orders of speed improvement.

4 Conclusion

In this study, we demonstrate the efficacy of GNNs to accelerate an end-to-end simulation framework for predicting the material properties of complex material interfaces. Starting from macroscopic transistor dimensions, we use our neural network to predict forces for generating the modern transistor gate stack (HfO$_2$-SiO$_2$-Si) with bond lengths within 3% of the experimental values. We furthur reproduce global electronic ($D(E)$, $J_x(E)$) and transport properties (v_{inj}) of crystalline silicon channels. We show agreement within 5.4% for injection velocity on unseen geometries and demonstrate high performance on a structure size outside the training domain. The scalability and accuracy of our predictions over a wide range of material and transport properties demonstrate the viability of our approach for the modeling of advanced heterogeneous devices, paving the way for rapid design and modeling of next-generation transistors.

Acknowledgements

References

[1] K. Bang, B. C. Yeo, D. Kim, S. S. Han, and H. M. Lee. Accelerated mapping of electronic density of states patterns of metallic nanoparticles via machine-learning. *Scientific Reports*, 11, 12 2021.

[2] L. C. Gallington, Y. Ghadar, L. B. Skinner, J. K. R. Weber, S. V. Ushakov, A. Navrotsky, A. Vazquez-Mayagoitia, J. C. Neuefeind, M. Stan, J. J. Low, and C. J. Benmore. The structure of liquid and amorphous hafnia. *Materials*, 10(11), 2017.

[3] R. Gurunathan, K. Choudhary, and F. Tavazza. Rapid prediction of phonon structure and properties using the atomistic line graph neural network (alignn). *Physical Review Materials*, 7, 2 2023.

[4] P. R. Kaundinya, K. Choudhary, and S. R. Kalidindi. Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (alignn). *JOM*, 2022.

[5] L. Khouchaf, K. Boulahya, P. P. Das, S. Nicolopoulos, V. K. Kis, and J. L. Lábár. Study of the microstructure of amorphous silica nanostructures using high-resolution electron microscopy, electron energy loss spectroscopy, x-ray powder diffraction, and electron pair distribution function. *Materials*, 13(19), 2020.

[6] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996.

[7] Y. Liu, N. Neophytou, T. Low, G. Klimeck, and M. S. Lundstrom. A tight-binding study of the ballistic injection velocity for ultrathin-body soi mosfets. *IEEE Transactions on Electron Devices*, 55:866–871, 3 2008.

[8] K. Natori. Ballistic metal-oxide-semiconductor field effect transistor. *Journal of Applied Physics*, 76: 4879–4890, 1994.

[9] C. W. Park, M. Kornbluth, J. Vandermause, C. Wolverton, B. Kozinsky, and J. P. Mailoa. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Computational Materials*, 7, 12 2021.

[10] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom. Theory of ballistic nanotransistors. *IEEE Transactions on Electron Devices*, 50:1853–1864, 9 2003.

[11] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, and K. R. Müller. Schnet - a deep learning architecture for molecules and materials. *Journal of Chemical Physics*, 148, 6 2018.

[12] T. R. Shan, B. D. Devine, J. M. Hawkins, A. Asthagiri, S. R. Phillpot, and S. B. Sinnott. Second-generation charge-optimized many-body potential for si/ sio 2 and amorphous silica. *Physical Review B - Condensed Matter and Materials Physics*, 82, 12 2010.

[13] T. R. Shan, B. D. Devine, T. W. Kemper, S. B. Sinnott, and S. R. Phillpot. Charge-optimized many-body potential for the hafnium/hafnium oxide system. *Physical Review B - Condensed Matter and Materials Physics*, 81, 3 2010.

[14] S. Smidstrup, K. Stokbro, A. Blom, T. Markussen, J. Wellendorff, J. Schneider, T. Gunst, B. V. v Petr A Khomyakov, U. G. Vej-Hansen, M. Brandbyge, et al. Quantumatk: An integrated platform of electronic and atomic-scale modelling tools. *J. Phys: Condens. Matter (APS)*, 32:015901, 2020. doi: 10.1088/1361-648X/ab4007.

[15] Synopsys. *Sentaurus Device User Guide, Version O-2018.06*. 2018.

[16] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.*, 271:108171, 2022.

[17] C. T. Tung, M. Y. Kao, and C. Hu. Neural network-based and modeling with high accuracy and potential model speed. *IEEE Transactions on Electron Devices*, 69:6476–6479, 11 2022.

[18] O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, and K. R. Müller. Spookynet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature Communications*, 12, 12 2021.