

A Machine-Learning Accelerated Atom-to-Transistor Simulation Framework

Pratik Brahma*

Department of Engineering
University of California, Berkeley
Berkeley, United States
pratik_brahma@berkeley.edu

Krishna Bhattaram*

Department of Engineering
University of California, Berkeley
Berkeley, United States
krishnabhattaram@berkeley.edu

Sayeef Salahuddin

Department of Engineering
University of California, Berkeley
Berkeley, United States
sayeef@berkeley.edu

Abstract—Technology Computer-Aided Design (TCAD) has long proved invaluable in developing semiconductor devices. This is becoming especially true as transistor design enters the nanoscale, where interfacial and quantum effects lead to nontrivial deviation from bulk properties. However, modern TCAD approaches to capturing these effects are often inadequate, relying on coarse approximations dependent on fitting parameters. Full atomistic simulations using first-principles methods such as density functional theory (DFT) can be more accurate but are numerically intractable for realistic devices. In this work, we introduce FASTCAD, a graph neural network designed to accelerate atomistic simulation of semiconductor devices. Starting from machine-learning force-field generated atomic structures, we employ equivariant graph networks to infer density of states, dielectric profiles, and ballistic transport properties to derive current-voltage characteristics. Our approach shows current errors within 6.5% and gate capacitance errors within 2.5% of first-principles methods with 100-10,000x speedup, paving the path to designing advanced semiconductor devices from the atomic level.

Index Terms—Atomistic Transistor Simulation, Graph Neural Networks, TCAD

I. INTRODUCTION

Modern transistors feature multiple interfacing materials with sub-nanometer dimensions, where quantum and atomic-scale effects strongly influence transport and electrostatics [1], [2]. Conventional continuum TCAD methods are often used to simulate such systems, but their reliance on coarse band-structure models and numerous fitting parameters often leads to inaccuracy, limited generalizability, and inability to capture interfacial and atomic-scale variation [3].

First-principles simulations such as density functional theory (DFT) offer generalizability and atomistic fidelity, but become prohibitively expensive at realistic device scales and for high-throughput applications [4], [5]. Recently, machine-learning methods have shown promise in accelerating device and material characterization tasks through compact-model surrogates, physics-informed solvers, and molecular and bulk material property prediction [6]–[8]. However, a clear method for propagating the impact of atomic-level material variation to electronic device performance is lacking.

In this work, we introduce FASTCAD, an equivariant graph neural network designed to accelerate atomistic transistor sim-

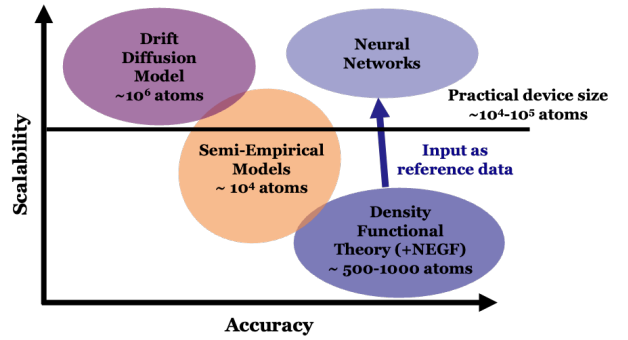


Fig. 1. Accuracy and scalability trade-off in device simulation. Neural networks trained on DFT data can extend near-DFT fidelity toward practical device sizes.

ulation to enable prediction of transistor characteristics directly from the atomistic gate stack. FASTCAD learns the mapping from atomic structure to material properties, which are used in conjunction with electrostatics to obtain device-level characteristics including charge- and current-voltage relations. Using our pipeline, we achieve current errors within 6.5% and gate-capacitance errors within 2.5% of first-principles methods, while delivering a 10^2 – 10^4 times speedup. Our framework enables fast, atomically-informed device characterization, unlocking rapid design-space exploration and technology co-optimization of advanced transistor technologies.

II. ATOMISTIC SIMULATION PIPELINE

We study silicon-on-insulator (SOI) transistors with crystalline Si channels and amorphous SiO_2 or SiO_2 - HfO_2 gate stacks, representative of modern high- κ gate-dielectric technologies [9]. Amorphous oxide layers are generated from crystalline SiO_2 and HfO_2 by melt-quench molecular dynamics accelerated with machine-learned moment tensor potentials [10]. Interfaces are then formed by stacking crystalline Si, amorphous SiO_2 , and, when included, amorphous HfO_2 , followed by localized anneal-quench steps and DFT relaxation using an LCAO basis with PBE exchange-correlation functional [11]. Finally, dangling bonds on terminal surfaces are passivated with hydrogen, with vacuum regions and metallic electrodes

*These authors contributed equally

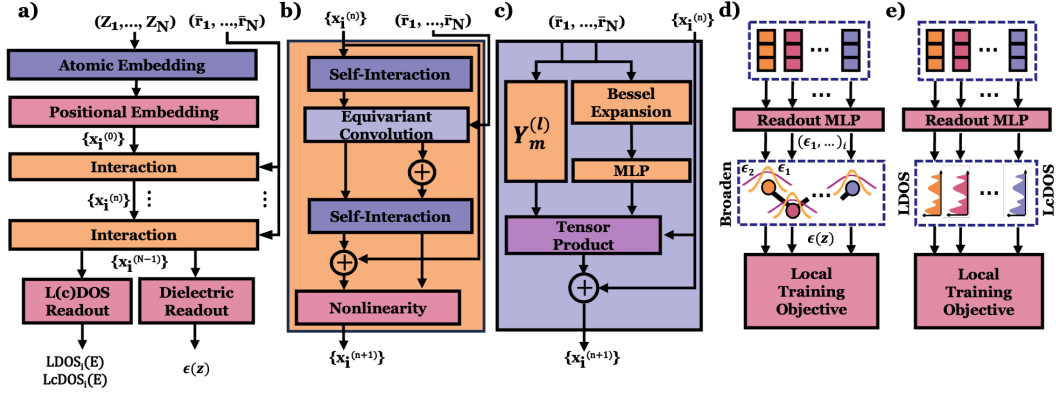


Fig. 2. FASTCAD architecture for LDOS/LcDOS and dielectric-profile prediction. The shared equivariant message-passing trunk is followed by readout heads for atom-local spectra and Gaussian-broadened dielectric fields.

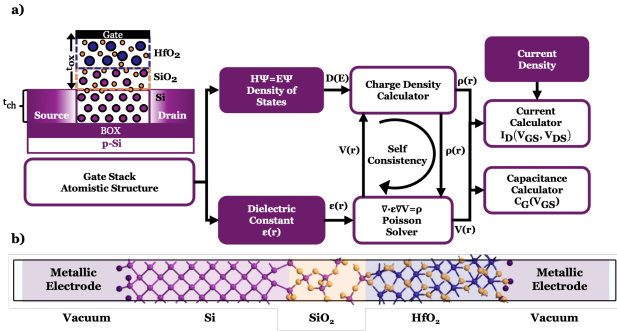


Fig. 3. Simulation pipeline for computing macroscopic transistor behavior from an atomistic SOI gate stack. FASTCAD accelerates the colored property-prediction blocks while preserving coupling to charge, current, capacitance, and Poisson calculations.

added for dielectric profile evaluation. An example structure is visualized in Fig. 3b.

A. Transport Model and Electrostatics

With modern transistors operating at $\sim 40\text{--}50\%$ of the ballistic limit [12], the ballistic transport model provides a simple yet physically meaningful framework for analyzing current flow in ultra-scaled devices [13]. In this model, the drain current I_D and inversion electron density N_{inv} are evaluated at the source-channel barrier:

$$I_D = q \int J_x(E) [f_s(E + U_b) - f_d(E + U_b)] dE \quad (1)$$

$$N_{\text{inv}} = \int D(E) [f_s(E + U_b) + f_d(E + U_b)] dE \quad (2)$$

where $D(E)$ is the density of states (DOS) in the channel, $f_s(E)$ and $f_d(E)$ are the source and drain Fermi functions, and U_b is the self-consistent barrier potential. The product of the group velocity and DOS defines the current density of states $J_x(E) := v_x(E)D(E)$ (cDOS). We compute the atom-projected DOS using DFT in an LCAO basis and obtain the real-space local DOS (LDOS) $D(E, r)$ by Gaussian smearing

of $D(E, a)$. The atomic and real-space projections of the cDOS (LcDOS) are obtained analogously.

The spatially-resolved dielectric profile is extracted by applying a small voltage $V_{\text{bias}} = 0.1 \text{ V}$ between metallic electrodes and comparing the biased and unbiased Hartree potentials. Leveraging the difference potential $\Delta V_H(z) = V_H(z; V_{\text{bias}}) - V_H(z; 0)$, the spatial relative permittivity is estimated as [14]:

$$\frac{1}{\epsilon_r(z)} = -\frac{1}{\epsilon_{\text{ref}}} \frac{\partial \Delta V_H(z) / \partial z}{E_{\text{ext}}} \quad (3)$$

where $E_{\text{ext}} = V_{\text{bias}}/L$, and ϵ_{ref} is a reference bulk dielectric constant that accounts for ionic contributions in each dielectric region (~ 11.9 for Si, ~ 3.9 for SiO_2 , and ~ 25 for HfO_2).

Finally, using the LDOS of the channel, the spatial inversion charge density is computed as

$$\rho_{\text{inv}}(\mathbf{r}) = -q \int D(E, \mathbf{r}) f(E + U(\mathbf{r}) - E_F) dE \quad (4)$$

Right- and left-moving electrons are equilibrated with the source or drain Fermi levels E_F , respectively. Charge density is solved self-consistently with the electrostatic potential using the dielectric profile $\epsilon_r(z)$. Once converged, the current is computed with the LcDOS and potential at the source-channel barrier [13]. The gate capacitance is obtained by differentiating integrated channel charge with respect to gate voltage.

Following this simulation pipeline, we generate 3,006 Si-SiO₂ and Si-SiO₂-HfO₂ heterostructures for LDOS/LcDOS prediction, and 2,098 for dielectric prediction, each with varied cell surface areas, random point defects, and Si, SiO₂, and HfO₂ thicknesses ranging from 8–30 Å.

III. FASTCAD ARCHITECTURE

The computational bottleneck in this pipeline is repeated diagonalization of quantum Hamiltonians to compute forces, L(c)DOS, and dielectric properties, which scales cubically with the number of atoms. While the former is mitigated through machine-learned MTPs for molecular dynamics, FASTCAD is used to replace the latter two calculations with

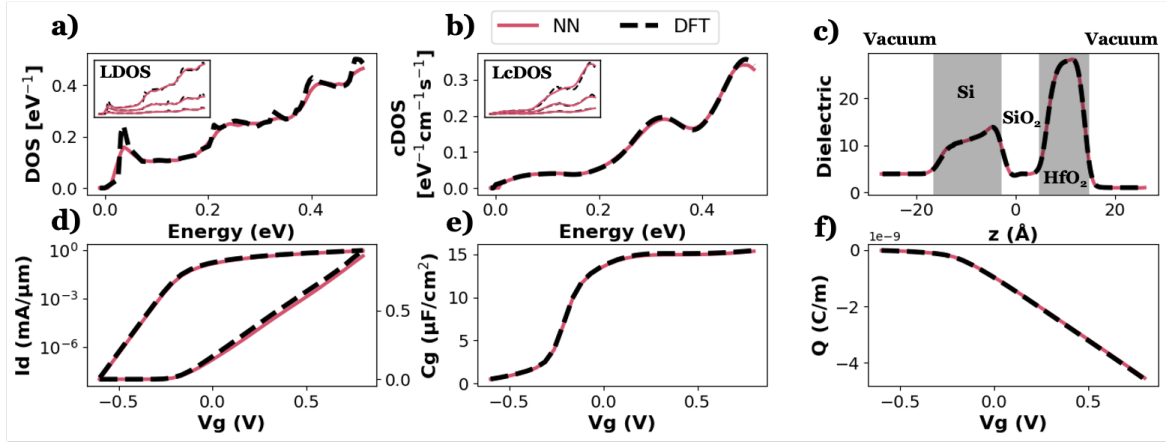


Fig. 4. Comparison of DFT and FASTCAD predictions for a representative 16 Å Si, 8 Å SiO₂, 8 Å HfO₂ gate stack. FASTCAD reproduces DOS/cDOS spectra, dielectric response, drain current, gate capacitance, and charge-voltage behavior.

a locality-bounded graph neural network, preserving linear scaling and providing two to four orders-of-magnitude speed-up. The runtime for electronic property calculation can thereby be reduced from hours or days to seconds on a single GPU (Fig. 6).

FASTCAD takes an atomistic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as input, where node i carries atomic number Z_i and position $\mathbf{r}_i$, and edges connect atoms within a 5 Å cutoff. Each atom is assigned a vector $\mathbf{x}_i$ initialized with a trainable embedding, with a vertical positional encoding added to encode the atom position along the gate stack:

$$\mathbf{x}_i^0 = \text{Embedding}(Z_i) + \sum_{k=1}^{K_{\text{pos}}} \sin^2\left(\frac{k\pi z_i}{t}\right) \quad (5)$$

where t denotes the channel thickness for LDOS/LcDOS prediction and the total gate-stack thickness for dielectric prediction.

Message passing uses E(3)-equivariant interaction layers implemented with spherical harmonics, learnable radial Bessel expansions, and tensor products as in e3nn [15]. At interaction step l , messages depend on neighboring descriptors and the interatomic direction:

$$\mathbf{m}_{ij}^l = \phi_B(\|\mathbf{r}_{ij}\|)Y_m^\ell(\hat{\mathbf{r}}_{ij}), \quad (6)$$

$$\mathbf{x}_i^{l+1} = \sigma\left(\mathbf{W}\mathbf{x}_i^l + \sum_{j \in \mathcal{N}(i)} \text{TP}(\mathbf{x}_i^l, \mathbf{m}_{ij}^l)\right) \quad (7)$$

Y_m^ℓ here denotes a real spherical-harmonic basis expansion of the interatomic displacement vector $\mathbf{r}_{ij}$, ϕ_B is a multilayer perceptron (MLP), TP is an e3nn tensor product, and $\mathbf{W}$ is a learnable matrix.

Finally, task-specific readout MLPs map the atom descriptors to the desired property. For electronic spectra,

$$\text{LDOS}_i(E) = \phi_R^D(\mathbf{x}_i^T), \quad \text{LcDOS}_i(E) = \phi_R^J(\mathbf{x}_i^T), \quad (8)$$

with total DOS and cDOS obtained by summing over atoms. For dielectric prediction, amplitudes $\{s_k\}_i = \phi_R^\varepsilon(\mathbf{x}_i^T)$ are

TABLE I
ERROR METRICS FOR FASTCAD. TOP ROWS EVALUATE MATERIAL-PROPERTY PREDICTIONS; BOTTOM ROWS EVALUATE DOWNSTREAM DEVICE QUANTITIES OVER 11 TEST GATE STACKS.

Metric	MAE	RMSE	R ²	WMAPE
LDOS	4.01×10^{-3}	18.5×10^{-3}	0.959	9.60%
LcDOS	3.82×10^{-3}	14.8×10^{-3}	0.958	11.6%
Dielectric	1.97×10^{-1}	8.12×10^{-1}	0.966	4.41%
I_d	2.24×10^{-2}	4.09×10^{-2}	0.991	6.02%
Q	1.75×10^{-11}	3.08×10^{-11}	0.999	1.81%
C_g	1.49×10^{-1}	2.15×10^{-1}	0.998	2.34%

derived from the atom-embeddings and broadened over a basis of k fixed-width Gaussians to the real grid:

$$\varepsilon(z) = \sum_i \sum_k \frac{s_{k,i}}{\sqrt{2\pi}\sigma_k} \exp\left[-\frac{(z - z_i)^2}{2\sigma_k^2}\right] \quad (9)$$

FASTCAD is designed to preserve linear scaling in system size. In addition, our approach maintains input permutation symmetry, translation invariance, and rotational equivariance of intermediate features, which improves model generalization and reduces the amount of training data required for accurate inference [16].

IV. RESULTS

We train two separate networks, one for joint LDOS/LcDOS prediction and one for dielectric profiles, each with four message-passing layers, 64-dimensional invariant embeddings, and vector features up to order $\ell = 1$. The electronic spectra network is trained with a 10 to 1 LDOS/LcDOS weighted sum of atom-local mean-squared-error (MSE) spectral losses, while the dielectric network is trained on MSE over a uniform spatial grid. Final dataset errors are summarized in Table I.

FASTCAD is evaluated on held-out structures not used during training. Table I summarizes the material- and device-level errors. For electronic spectra, we achieve R^2 values of 0.959 and 0.958 for LDOS and LcDOS, with 9.6 and 11.6% weighted mean absolute percent error (WMAPE, the mean

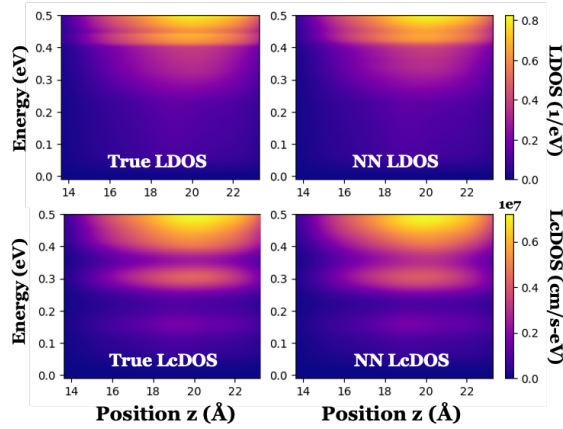


Fig. 5. Ground-truth and FASTCAD-predicted energy-resolved spatial maps of LDOS and LcDOS in the silicon channel.

absolute error over the mean). The dielectric model obtains an $R^2 = 0.966$, with WMAPE below 5%.

Fig. 4 shows representative spectra, dielectric profiles, and downstream device curves. FASTCAD predictions reproduce DFT-derived DOS/cDOS features, spatial dielectric variation across Si, SiO₂, and HfO₂, and downstream I_D - V_G , C_G - V_G , and Q - V_G characteristics. Across 11 held-out gate stacks in the intersection of the spectral and dielectric test sets, the framework reaches 6.02% WMAPE for drain current and 2.34% WMAPE for gate capacitance.

Finally, Fig. 5 visualizes two-dimensional energy-position maps of both the true and predicted LDOS and LcDOS in the silicon channel region. We observe strong qualitative agreement and faithful reproduction of key spatial features such as the decay of LDOS and LcDOS near the silicon edges, a signature of quantum confinement in ultrathin channels.

V. CONCLUSION

In this work, we presented FASTCAD, an equivariant graph neural network-accelerated framework capable of rapidly and accurately capturing the impact of atomic-scale variation on the electronic properties of transistors. By replacing traditional DFT-based simulation blocks with fast, accurate neural surrogates, we demonstrate acceleration of the transistor simulation pipeline while retaining atomistic fidelity. Our framework captures the influence of local disorder and material interfaces on device behavior and enables scalable exploration of atomic variation effects on performance. Compared to DFT, our approach achieves two to four orders of magnitude speedup and linear rather than cubic asymptotic scaling (Fig. 6) while preserving less than 6.5% error in downstream device properties, paving a path toward large-scale and high-throughput atomistic simulations of device structures for design and optimization of next-generation transistors.

REFERENCES

[1] K. Das *et al.*, “Ferroelectricity in atomic-scale titanium dioxide dielectric films,” *Science*, vol. 392, pp. 280–284, 2026, doi: 10.1126/science.aec9417.

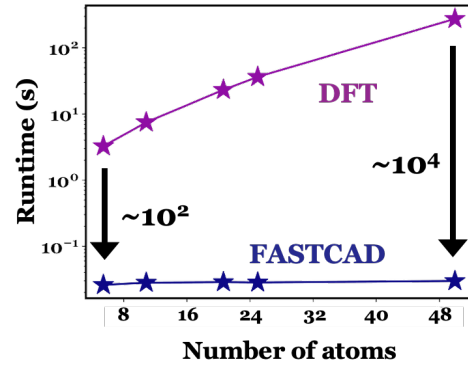


Fig. 6. Runtime comparison for local electronic-property prediction. FASTCAD gives 10^2 - $10^4 \times$ acceleration and linear scaling.

- [2] “The interface is still the device,” *Nature Materials*, vol. 11, no. 2, pp. 91–91, Feb. 2012, doi: 10.1038/nmat3244.
- [3] Synopsys, Inc., *Sentaurus Device Monte Carlo User Guide*, Version N-2017.09. Mountain View, CA, USA: Synopsys, Inc., 2017.
- [4] N. Vetsch, A. Maeder, V. Maillou, A. Winka, J. Cao, *et al.*, “Ab-initio quantum transport with the GW approximation, 42,240 atoms, and sustained exascale performance,” in *SC ’25*, New York, NY, USA: Association for Computing Machinery, 2025, pp. 1–13, doi: 10.1145/3712285.3771784.
- [5] P. Brahma, K. Bhattaram, and S. Salahuddin, “Accelerated modelling of interfaces for electronic devices using graph neural networks,” arXiv preprint arXiv:2310.06995, 2023, doi: 10.48550/arXiv.2310.06995.
- [6] K. Mehta and H.-Y. Wong, “Prediction of FinFET current-voltage and capacitance-voltage curves using machine learning with autoencoder,” *IEEE Electron Device Letters*, vol. 42, no. 2, pp. 136–139, 2021, doi: 10.1109/LED.2020.3045064.
- [7] B. Cao, Y.-D. Wang, N.-E. Zhang, Y.-Z. Liang, and W.-Y. Yin, “A physics-informed neural networks algorithm for simulating semiconductor devices,” in *2023 ACES-China*, 2023, pp. 1–3, doi: 10.23919/ACES-China60289.2023.10249274.
- [8] S. Batzner, *et al.*, “E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials,” *Nature Communications*, vol. 13, Art. no. 2453, 2022, doi: 10.1038/s41467-022-29939-5.
- [9] T. Ando, “Ultimate scaling of high- κ gate dielectrics: Higher- κ or interfacial layer scavenging?,” *Materials*, vol. 5, no. 3, pp. 478–500, 2012, doi: 10.3390/ma5030478.
- [10] A. V. Shapeev, “Moment tensor potentials: A class of systematically improvable interatomic potentials,” *Multiscale Modeling & Simulation*, vol. 14, no. 3, pp. 1153–1173, 2016, doi: 10.1137/15M1054183.
- [11] S. Smidstrup *et al.*, “QuantumATK: An integrated platform of electronic and atomic-scale modelling tools,” *Journal of Physics: Condensed Matter*, vol. 32, no. 1, Art. no. 015901, Oct. 2020, doi: 10.1088/1361-648X/ab4007.
- [12] D. Antoniadis, I. Djomehri, and A. Lochtefeld, “Electron velocity in sub-50-nm channel MOSFETs,” Jan. 2001, doi: 10.1007/978-3-7091-6244-6_35.
- [13] A. Rahman, J. Guo, S. Datta, and M. S. Lundstrom, “Theory of ballistic nanotransistors,” *IEEE Transactions on Electron Devices*, vol. 50, no. 9, pp. 1853–1864, 2003, doi: 10.1109/TED.2003.815366.
- [14] D. Stradi, *et al.*, “Atomistic modeling of nanoscale ferroelectric capacitors using a density functional theory and non-equilibrium Green’s-function method,” in *SISPAD*, 2019, pp. 1–4, doi: 10.1109/SISPAD.2019.8870397.
- [15] M. Geiger and T. Smidt, “e3nn: Euclidean neural networks,” arXiv:2207.09453, 2022.
- [16] C. Zhang, S. Bengio, M. Hardt, M. C. Mozer, and Y. Singer, “Identity crisis: Memorization and generalization under extreme overparameterization,” in *ICLR*, 2020. [Online]. Available: <https://openreview.net/forum?id=B116y0VFPr>