Machine-Learned Hamiltonians for Quantum Transport Simulation of Valence Change Memories

1st Chen Hao Xia
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
chexia@iis.ee.ethz.ch

2nd Manasa Kaniselvan
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
mkaniselvan@iis.ee.ethz.ch

3rd Marko Mladenović
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
mmladenovic@iis.ee.ethz.ch

4th Mathieu Luisier
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland
mluisier@iis.ee.ethz.ch

Abstract—**The construction of the Hamiltonian matrix H is an essential, yet computationally expensive step in *ab-initio* device simulations based on density-functional theory (DFT). In homogeneous structures, the fact that a unit cell repeats itself along at least one direction can be leveraged to minimize the number of atoms considered and the calculation time. However, such an approach does not lend itself to amorphous or defective materials for which no periodicity exists. In these cases, (much) larger domains containing thousands of atoms might be needed to accurately describe the physics at play, pushing DFT tools to their limit. Here we address this issue by learning and directly predicting the Hamiltonian matrix of large structures through equivariant graph neural networks and so-called augmented partitioning training. We demonstrate the strength of our approach by modeling valence change memory (VCM) cells, achieving a Mean Absolute Error (MAE) of 3.39 to 3.58 meV, as compared to DFT, when predicting the Hamiltonian matrix entries of systems made of ∼5,000 atoms. We then replace the DFT-computed Hamiltonian of these VCMs with the predicted one to compute their energy-resolved transmission function with a quantum transport tool. A qualitatively good agreement between both sets of curves is obtained. Our work provides a path forward to overcome the memory and computational limits of DFT, thus enabling the study of large-scale devices beyond current *ab-initio* capabilities.**

Index Terms—**Machine Learning, Memory, Amorphous, DFT.**

I. Introduction

As the dimensions of modern-day electronic devices keep decreasing, *ab-initio* calculations are gaining momentum to capture atomistic details and their influence on key performance metrics. Density-functional theory (DFT) lends itself optimally to this task. Besides the energy and wavefunction of atomic systems, it can also return their Hamiltonian matrix **H**. This quantity can then be passed to a quantum transport (QT) solver to compute the "current vs. voltage" characteristics of the targeted device. In structures with atomic disorder such as the amorphous oxides used at the switching layer of resistive random access memories (ReRAM), the construction of **H** can be tedious. Large unit cells containing thousands of atoms are required to accurately represent ReRAM geometries. As DFT scales with $O(N^3)$, N being the number of atoms, it becomes prohibitively expensive at these scales, limiting the size of the devices that can be investigated at the *ab-initio* level.

In this work, we present an approach that overcomes this bottleneck by replacing DFT-computed Hamiltonians with machine-learned (ML) ones. We demonstrate this approach by simulating a valence change memory (VCM) cell, a ReRAM type with applications in neuromorphic computing [1]. The VCM of interest is made of a TiN- HfO_2- Ti/TiN stack and 5,268 atoms. Its conductance can be modulated by applying an external voltage, which leads to the generation/recombination of oxygen vacancies at the Ti-HfO_2 interface and the formation/dissolution of conductive filaments. From a modeling perspective, the underlying structural evolution can be tracked with, for example, kinetic Monte Carlo (KMC) [2] or molecular dynamics (MD) [3] simulations. To compute the corresponding "I-V" characteristics, the Hamiltonian matrix of samples extracted at regular time intervals should be produced with DFT and passed as input to a QT solver. Bypassing the DFT step with ML requires a model that can learn complex features, e.g., non-regular lattices, the influence of oxygen vacancies, and the interaction between atoms of different types. For that purpose, we adopt an equivariant graph neural network (EGNN) [4] and show that, if it is trained on very few VCM configurations, it can accurately predict the Hamiltonian of large structures with unseen vacancy distributions and filament morphologies. The electrical current calculated with the ML Hamiltonian matrices are in good qualitative agreement with DFT results.

II. Method

A. Network Architecture

One of the key challenges in machine learning electronic structure predictions (MLESP) is the rotational covariance of the Hamiltonian matrix involved. In a localized spherical harmonic basis, a geometric rotation of the input geometry leads to a change in the Hamiltonian matrix. Equivariant graph neural networks have been proposed as a solution to drastically reduce the data required to learn such transformation [4]. The organization of our EGNN is presented in Fig. 2. The equivariant convolutions embed rotational covariance as a

We acknowledge funding from the ALMOND project (SNSF Sinergia grant no. 198612) and from the Swiss State Secretariat for Education, Research, and Innovation (SERI) through the SwissChips research project. This work was also supported by the Swiss National Supercomputing Center (CSCS) under project lp16 and lp82.

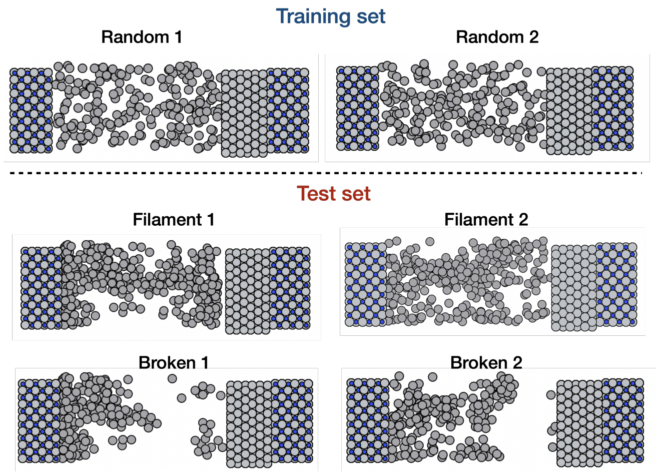

Fig. 1. Training (left) and testing (right) TiN- HfO_2- Ti/TiN VCM structures with the Hf and O atoms omitted for better visualization. A total of 20 slices with 1 Å thickness was extracted from the Random 1 and 2 devices with arbitrary vacancy distributions to train the ML model. Another unseen 1-Å slice from Random 2 was used for validation. The trained model was then tested on unseen samples with either two fully-formed (Filament 1 and 2) or two broken (Broken 1 and 2) filament configurations.

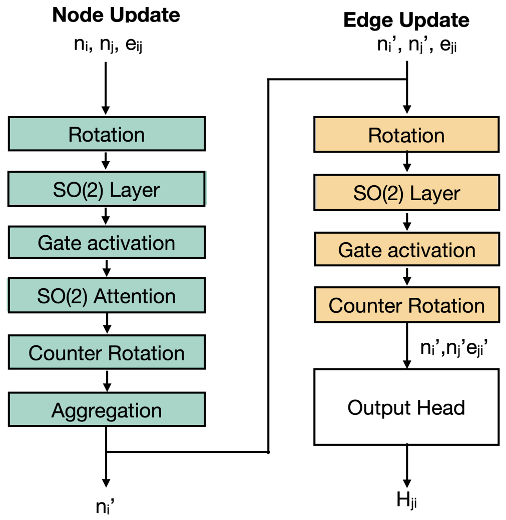

Fig. 2. Higher level overview of the equivariant graph neural network architecture used in this work. It is a strictly local network consisting of a single message passing layer, with node (edge) embeddings n_i (e_{ji}) representing atoms (interactions between atoms). The nodes aggregate messages from each other, weighted by an attention layer, with the output embeddings being used to update the edges. The updated embeddings n_i' and e_{ji}' are then fed into an output head to reconstruct the Hamiltonian sub-blocks.

physical constraint within the network [5]. The nodes (edges) of our graph represent onsite (offsite) Hamiltonian blocks, and the outputs of the nodes are used to update the edges. To enable large-scale $\mathbf{H}$ predictions, strict locality is enforced [6], while multi-headed attention is included [7] to better distinguish between complex atomic environments. Finally, the node and edge embeddings are fed into an output head to reconstruct the Hamiltonian blocks [8].

B. Experiment Setup

Our task is to predict the Hamiltonian matrix of TiN- HfO_2-Ti/TiN valence change memory cells with different vacancy configurations, as illustrated in Fig. 1. The displayed snapshots were created through kinetic Monte Carlo simulations when running a "current vs. voltage" sweep of the devices under consideration [2]. They consist of two broken and two formed filament configurations with a wide range of electrical conductivity. This quantity is highly sensitive to the distribution of oxygen vacancies within the central HfO_2 switching layer.

We first train a single model on the Hamiltonian matrix of two device structures with uniformly distributed vacancies. Both $\mathbf{H}$ were produced with the CP2K DFT package [9], which relies on a localized basis set of Gaussian-type orbitals. Note that the training examples were not generated via KMC, but by randomly inserting oxygen vacancies into device structures with a stoichiometric oxide layer. This makes them significantly different from the clustered, physically meaningful vacancy configurations that are part of the test set. The large discrepancies between the training and test samples is essential to rigorously assess the model's ability to learn a generalizable, useful function and apply it to unseen instances.

C. Simulation Workflow

Our workflow is illustrated in Fig. 3. During training, augmented partitioning [10] is applied to allow for large structures to be divided into multiple slices and fit into the memory of a single GPU. Importantly, the atomic connectivity to neighbor partitions is fully accounted for to maintain high prediction accuracy. Also, partitioning occurs longitudinally, in the $x - y$ plane in Fig. 3, to capture the interface between the different materials composing the VCM stack (TiN, Ti, HfO_2).

The trained EGNN is then tested on the full structure of the selected device examples by constructing/predicting their Hamiltonian matrix $\mathbf{H}$. The latter are finally inserted into our in-house quantum transport simulator [11] that returns the corresponding energy-resolved transmission function $T(E)$ and electrical current I_d with the non-equilibrium Green's function (NEGF) formalism. In all cases, a voltage of 1 V is applied between both TiN electrodes of the VCM cells. The obtained results for the entries of $\mathbf{H}$, $T(E)$, and I_d are compared to reference DFT calculations.

III. Results

A. Hamiltonian Prediction

The predicted node (ϵ_n) and edge (ϵ_e) errors of the predicted Hamiltonian matrices with respect to DFT are summarized in Table I and their distributions are visualized in Fig. 4 in the form of violin plots for the four TiN- HfO_2- Ti/TiN VCM configurations from Fig. 1. The model performs consistently across different examples, with prediction errors per entry lying within a small range (1.54 - 1.82 mE_H for nodes and ~0.12 mE_H for edges), with minimal outliers in all cases. The total errors, averaged over all Hamiltonian entries, are between

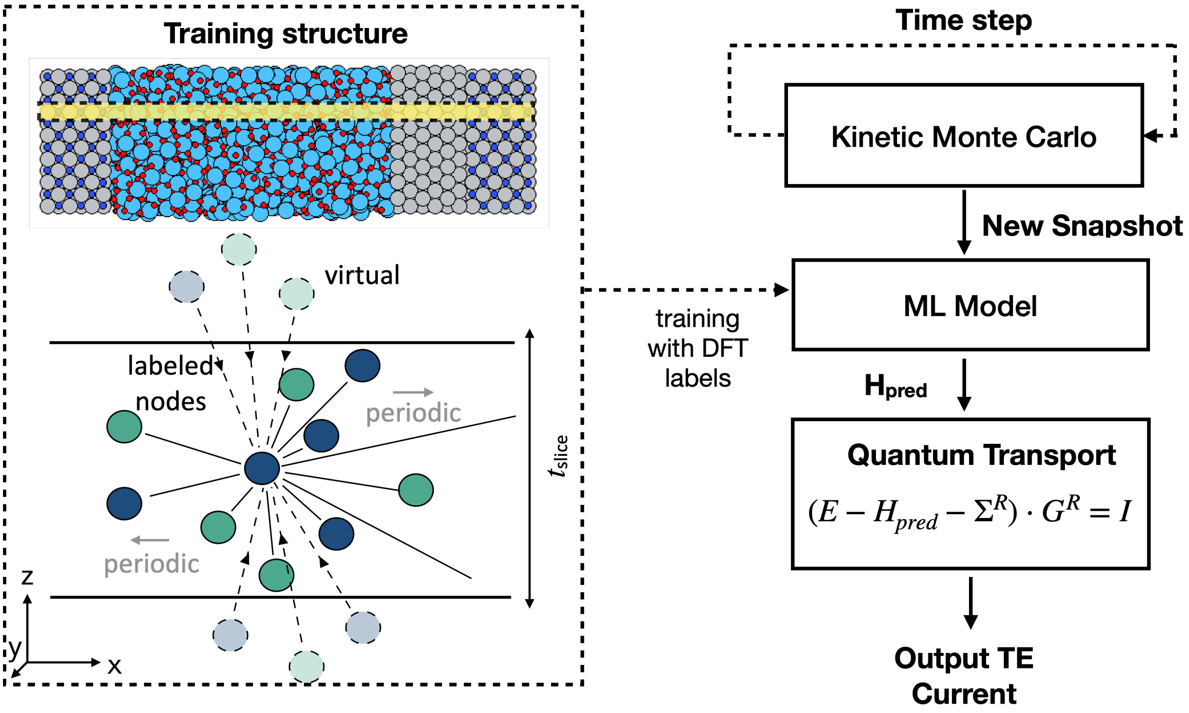

Fig. 3. General overview of the workflow to train and test our machine learning model. The VCM cells considered (top left) consist of a HfO_2 layer (blue: Hf atoms; red: O atoms) with TiN and Ti/TiN electrodes (gray: Ti; dark blue: N). The device dimensions are set 93.4×26.2×26.3 Å^3 along the x, y, and z directions. Slices in the x-z plane (yellow) are used to train our EGNN (middle). A so-called augmented partitioning method (bottom left) [10] is leveraged to ensure that the atomic connectivity is preserved at the slice boundaries. The trained model is then used to directly infer the Hamiltonian matrix H_{pred} of test structures generated with KMC at different times of a full "I-V" sweep. Finally, $\mathbf{H}_{pred}$ serves as input to quantum transport calculations. The transmission function $T(E)$ and electrical current I_d are the final outcomes.

3.39 and 3.58 meV. As such, they are very close to the state-of-the-art (2.2 meV) [12], but for much larger structures (5,268 vs. $\leq$150 atoms).

TABLE I

SUMMARY OF NODE (ϵ_n) AND EDGE (ϵ_e) PREDICTION ERRORS FOR THE DIFFERENT TEST STRUCTURES CONSIDERED IN THIS WORK. THE MEAN ABSOLUTE ERROR (MAE) IS REPORTED IN ALL CASES. THE ELECTRICAL CURRENT VALUES, AS COMPUTED WITH A DFT (I_{ref}) AND MACHINE-LEARNED (I_{pred}) HAMILTONIAN ARE LISTED AS WELL.

Device	$\epsilon_\mathbf{n} [mE_h]$	$\epsilon_\mathbf{e} [mE_h]$	I_{ref} [A]	I_{pred} [A]
Broken 1	1.82	0.12	3.38×10^{-11}	4.35×10^{-10}
Broken 1	1.65	0.12	8.00×10^{-9}	4.66×10^{-9}
Filament 1	1.62	0.12	1.48×10^{-5}	1.28×10^{-5}
Filament 2	1.54	0.12	6.99×10^{-6}	4.59×10^{-6}

B. Transmission and Current

The predicted energy-resolved transmission functions $T(E)$ are plotted in Fig. 4. Although the underlying Hamiltonian entries do not differ by more than a few meV from their DFT reference, the corresponding $T(E)$ only qualitatively agree. The trends are the same in both cases, the band edges of the HfO_2 switching layers are well reproduced, but several features, especially transmission peaks, are not accurately captured. On the other hand, the predicted electrical currents, as computed from the transmission functions through the Landauer-Büttiker formula [13], are close enough to their DFT counterparts to assess the conductance state of the device under test (see Table I). Generally, it can be observed that our ML model is more accurate for configurations with fully formed filaments than for samples with broken filaments where the transmission is much smaller and therefore more sensitive to errors in the predicted Hamiltonian matrices.

Since direct ML inferences are much faster than DFT (2 seconds for the forward pass vs. 3.94 node hours for DFT), our results indicate that electronic structure predictions could facilitate the investigation of devices with evolving morphologies. By replacing DFT with ML, we can rapidly construct the Hamiltonian matrices of hundreds of intermediate samples along the high-to-low resistance transition of VCM cells, thus fully amortizing the training cost (<40 node hours). This will, however, require further enhancement in the accuracy of the ML models.

IV. CONCLUSION

We presented an ML-based approach that can be integrated into a large-scale device simulation platform capable of computing from first-principles the "I-V" characteristics of atomic structures changing with time. Our model not only bypasses computationally intensive DFT calculations, it

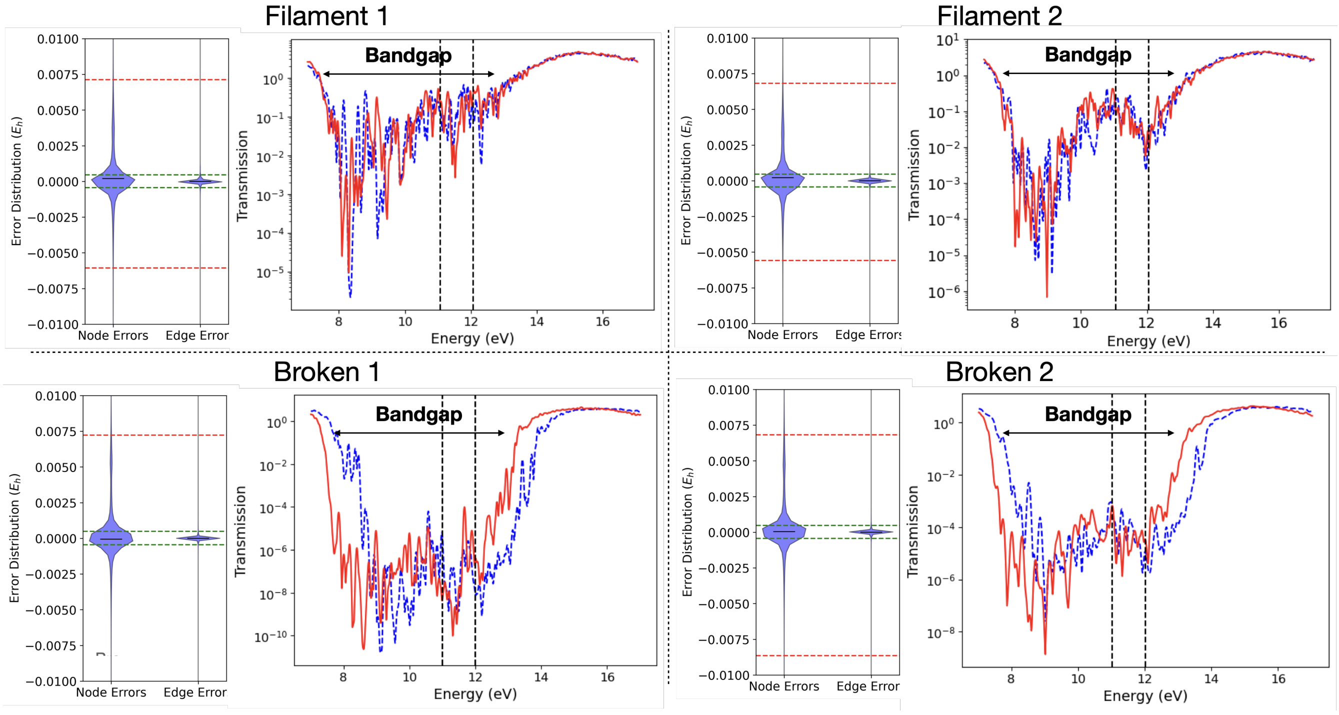

Fig. 4. Prediction results for the four TiN- HfO_2- Ti/TiN VCM test structures from Fig. 1. Each of them contains two sets of data. Left: Violin plot of the error distribution for the predicted Hamiltonian matrix $\mathbf{H}_{pred}$. The 5th and 95th percentile values of ϵ_n (node error) and ϵ_e (edge error) are indicated using red and green dashed lines, respectively. Note that for the ”Broken 1” case, the 5th percentile lies outside of the plotted range. Right: Comparison between the energy-resolved transmission function $T(E)$ as obtained with a predicted Hamiltonian ($\mathbf{H}_{pred}$, solid red curves) with a reference DFT Hamiltonian ($\mathbf{H}_{DFT}$, dashed blue curves) for the Filament 1, Filament 2, Broken 1, and Broken 2 TiN- HfO_2- Ti/TiN VCM configurations from Fig. 1. A bias of 1 V is applied between both metallic electrodes of the devices. The corresponding Fermi window at room temperature is delimited by the black dashed lines. The apparent HfO_2 band gap is indicated by the double arrows.

can potentially also be used to construct the Hamiltonian matrix of systems with sizes beyond current DFT capabilities. Future work includes improving the prediction accuracy of our framework by incorporating more training data and adopting more expressive network architectures. Other applications such as phase-change memories can be envisioned. They undergo gradual amorphous-to-crystalline transitions whose electronic properties could be predicted with ML instead of being computed with DFT [14].

References

[1] Yike Xiao, Cheng Gao, Juncheng Jin, Weiling Sun, Bowen Wang, Yukun Bao, Chen Liu, Wei Huang, Hui Zeng, and Yefeng Yu. Recent progress in neuromorphic computing from memristive devices to neuromorphic chips. *Advanced Devices & Instrumentation*, 5:0044, 2024.

[2] Manasa Kaniselvan, Mathieu Luisier, and Marko Mladenović. An atomistic model of field-induced resistive switching in valence change memory. *ACS Nano*, 17(9):8281–8292, March 2023.

[3] M. Laura Urquiza, Md Mahbubul Islam, Adri C. T. van Duin, Xavier Cartoixà, and Alejandro Strachan. Atomistic insights on the full operation cycle of a hfo2-based resistive random access memory cell from molecular dynamics. *ACS Nano*, 15(8):12945–12954, July 2021.

[4] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018.

[5] Saro Passaro and C. Lawrence Zitnick. Reducing so(3) convolutions to so(2) for efficient equivariant gnns, 2023.

[6] Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, 2023.

[7] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, 2023.

[8] Xiaoxun Gong, He Li, Nianlong Zou, Runzhang Xu, Wenhui Duan, and Yong Xu. General framework for e(3)-equivariant neural network representation of density functional theory hamiltonian. *Nature Communications*, 14(1), May 2023.

[9] Thomas D. Kühne *et al.* CP2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.*, 152(19):194103, May 2020.

[10] Chen Hao Xia, Manasa Kaniselvan, Alexandros Nikolaos Ziogas, Marko Mladenović, Rayen Mahjoub, Alexander Maeder, and Mathieu Luisier. Learning the electronic hamiltonian of large atomic structures, 2025.

[11] Mathieu Luisier and Gerhard Klimeck. Omen an atomistic and full-band quantum transport simulator for post-cmos nanodevices. In *2008 8th IEEE Conference on Nanotechnology*, pages 354–357. IEEE, 2008.

[12] Yuxiang Wang, He Li, Zechen Tang, Honggeng Tao, Yanzhen Wang, Zilong Yuan, Zezhou Chen, Wenhui Duan, and Yong Xu. Deeph-2: Enhancing deep-learning electronic structure via an equivariant local-coordinate transformer, 2024.

[13] Mads Brandbyge, José-Luis Mozos, Pablo Ordejón, Jeremy Taylor, and Kurt Stokbro. Density-functional method for nonequilibrium electron transport. *Physical Review B*, 65(16), March 2002.

[14] En Ma & Volker L. Deringer Yuxing Zhou, Wei Zhang. Device-scale atomistic modelling ofphase-change memory materials. *Nature Electronics*, 6(10):746–754, 2023.