Learning the Hamiltonian of Large, Disordered Atomic Systems

Chen Hao Xia*[a], Manasa Kaniselvan*[a], Alexandros Nikolaos Ziogas[a], Rayen Mahjoub[a], Marko Mladenovic[a], Alexander Maeder[a], Mathieu Luisier[a]

[a] *Integrated Systems Laboratory (IIS), ETH Zurich, Gloriastrasse 35, 8092 Zürich* chexia@iis.ee.ethz.ch

* Presenting author

1. Introduction

The Hamiltonian matrix (**H**) contains rich information about the electronic properties of materials, in particular their bandstructure and electron distribution. It can be typically computed with density functional theory (DFT), an *ab initio* method that consists in solving the Kohn-Sham [1] and Poisson equations self-consistently until convergence is reached. While small unit cells are sufficient to capture the behavior of ideal materials with DFT, realistic components including defects, disorder, or interfaces require larger unit cells (> 10 Å) with hundreds to thousands of atoms to be accurately described (Fig. 1) [2]. As DFT scales with $O(N^3)$, the calculation of Hamiltonian matrices for large-scale materials is prohibitively expensive.

This computational bottleneck can be bypassed using graph neural networks (GNNs), which have proven successful at predicting the **H** of small molecules and periodic structures. However, due to computational limitations, they have yet to be demonstrated on large disordered samples. In this work, we introduce a strictly local network combined with an augmented partitioning approach to break down large graphs for training, obtaining a prediction MAE of 0.99-5.16 meV for amorphous structures with 1,000-3,000 atoms

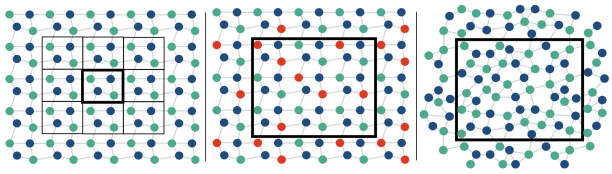

Fig. 1: Illustration of the differences between ideal periodic (left) and compositionally (middle) or structurally (right) disordered materials, Circles (lines) correspond to atoms (bonds).

2. Background

2.1 Related Work

Existing GNNs for Hamiltonian predictions integrate the rotational covariance of the (**H**) elements through tensor product operations that maintain SO(3) equivariance. These equivariant networks [3] achieve state-of-the-art accuracy on small molecule [4] and crystalline [5] datasets. The subsequent reformulation of tensor products from SO(3) to SO(2) significantly improved the scaling from $\mathcal{O}(l_{max}^6)$ down to $\mathcal{O}(l_{max}^3)$, where l_{max} is the maximum degree

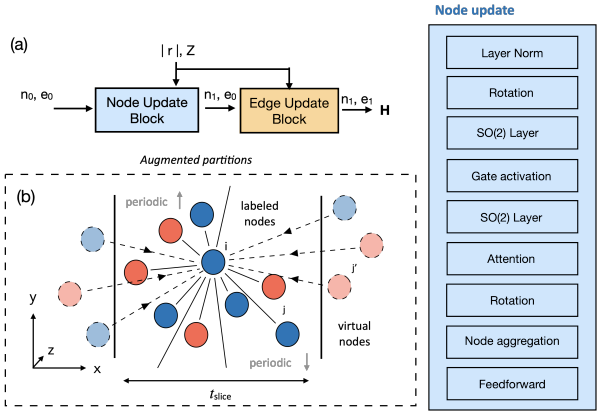

Fig. 2: (a) General organization of the network architecture, with a detailed description of the node update block (blue). (b) Illustration of the augmented partitioning approach. Virtual edges ($e_{j' \to i}$) connect virtual nodes (j') to the labeled ones. Solid vertical lines indicate partition boundaries. Different colors represent different atomic species.

of the angular momentum considered [6]. The introduction of attention [7] further improved the performance of GNNs on **H** datasets [8]. However, the size of the corresponding graphs explodes for disordered materials and large unit cells, resulting in slow/unfeasible training. Being able to predict **H** in such cases remains an open issue that we aim to address in this work. Our dataset includes industry-relevant materials like amorphous HfO_2, PtGe and GeSbTe (a-HfO_2, a-PtGe and a-GST).

3. General Approach

3.1 Network Architecture

The Hamiltonian matrix can be first decomposed into sub-matrices $H_{i,j}$ of size $(N_{orb}^i \times N_{orb}^j)$. Each of them describes the interactions between the $(N_{orb}^i$ basis elements (orbitals) on atoms i and the N_{orb}^j ones on atom j. The diagonal ($H_{i,i}$) and off-diagonal ($H_{i,j}$) blocks of **H** are represented by the nodes and edges of the GNN, respectively. The inputs to the nodes (edges) are the atomic numbers (distances between atoms). Due to the nearsightedness principle of **H** in the local basis, atoms outside of a fixed cutoff radius can be neglected [9]. The network architecture is illustrated in Fig. 2. We adopt efficient eSCN convolutions in our network [6], along with equivariant attention [7], to distinguish and learn complex atomic environments.

3.2 Augmented Partitioning Approach

Large, densely connected graphs representing realistic materials with disorder consume large amounts of memory, often exceeding GPU memory limits. They cannot be straightforwardly partitioned through the removal of edges. Doing so misinforms the network, as it aggregates information through an incorrect graph structure. To be able to train large structures while maintaining connectivity and accuracy, we introduce a so-called *augmented partitioning* approach. The graph is first partitioned into sub-graphs where the data can fit into the memory of a single GPU. Atoms located outside of a given partition, but connected to those inside of it, are represented by virtual nodes (**Fig. 2(c)** - dashed circles). They are connected to the partition through virtual edges (**Fig. 2(c)** - dashed lines). These virtual nodes/edges are initialized similarly to their labeled counterparts with input atomic numbers and distances. However, their outputs are not used. Their only role is to inform each partition of its full connectivity so that the network can then learn an accurate and generalizable aggregation function during message passing. As the virtual nodes and edges represent only the 1-hop neighborhood, the network is strictly local. As demonstrated by previous strictly local architectures, e.g., Allegro [10], information from the local environment is sufficient to achieve state-of-the-art prediction accuracy when locality holds. Additional local layers can also be included to capture many body interactions [11].

4. Results

In all our experiments, we train the model using the proposed augmented partitioning approach, and test it on a full, unseen structure. The Mean Absolute Error (MAE) is reported for nodes (ϵ_n) and edges (ϵ_e). Details on datasets are found in **Appendix A**.

	$\epsilon_{\mathbf{n}}[mE_h]$	$\epsilon_{\mathbf{e}}[mE_h]$
$n' \quad n$	5.18	1.66
$n' \rightarrow n$	2.29	0.20

Table 1: Ablation study on the impact of virtual nodes on the prediction accuracy of a-HfO_2

Compared to training with raw partitions, the addition of virtual nodes and edges reduces the node and edge prediction error by over 50% and 88%, respectively (Table 1). Such an improvement is expected, as raw partitions are characterized by a large proportion of missing edges and thus incorrect atomic neighborhoods. We then investigate the effect of augmented partitioning on accuracy (Table 2). Despite the different divisions ranging from 5 (t_{slice} $\simeq$12 Å) to 27 (t_{slice} $\simeq$2 Å) slices, $\epsilon_{\mathbf{n}}$ and $\epsilon_{\mathbf{e}}$ remain very close to the values obtained by training with the full graph (t_{slice} = 52.346 Å). The prediction error is thus insensitive to the partition size. It also performs consistently across a diverse range of test datasets (Table 3), which consists of large, disordered materials. The total errors range from 0.99 meV to 5.16 meV, for test structures containing 1,000 to 3,000 atoms and 200,000+ to 2,000,000+ edges. These values are comparable to what a previous study obtained (2.2 meV) using equivariant GNNs for smaller structures with $\leq$150 atoms per unit cell [12]. For the case of PtGe, the use of small slices also enables a larger r_{cut} of 16 Å within GPU memory limitations, allowing for further improvements in prediction accuracy.

t_{slice} [Å]	N_t	N_e	Epochs	$\epsilon_{\mathbf{n}}[mE_h]$	$\epsilon_{\mathbf{e}}[mE_h]$
$\sim$2	27	95,398	15,790	2.35	0.20
$\sim$3	18	141,512	18,990	2.15	0.18
$\sim$4	14	184,730	17,098	2.32	0.19
$\sim$8	7	320,324	20,599	2.37	0.17
$\sim$12	5	381,504	19,351	2.69	0.18
$\sim$52	1	533,364	23,396	2.46	0.16

Table 2: Prediction accuracy for a-HfO_2 when the network is trained on differently-sized partitions of the same graph.

Material	r_{cut}[Å]	$\epsilon_{\mathbf{n}}[mE_h]$	$\epsilon_{\mathbf{e}}[mE_h]$	$\epsilon_{\mathbf{tot}}[meV]$
a-HfO_2	8	2.15	0.18	5.16
a-PtGe	8	0.78	0.08	2.39
a-PtGe	16	0.82	0.04	0.99
a-GST	12	0.97	0.10	2.77

Table 3: Summary of the model performance when trained and tested on a-HfO_2, a-PtGe, and a-GST.

Stoichiometry (x)		$\epsilon_{\mathbf{n}}[mE_h]$	$\epsilon_{\mathbf{e}}[mE_h]$
Train set	Test set		
1.8	1.9	2.48	0.18
1.8	1.8	2.50	0.17
1.8	1.7	2.60	0.18

Table 4: Model trained on a-$HfO_{x=1.8}$ and tested on full unseen HfO_x structures

A model trained on one stoichiometry (a-$HfO_{x=1.8}$ with oxygen vacancies), can also generalize well to unseen HfO_x structures with different stoichiometry (Table 4), with ϵ_n and ϵ_e lying within a small range (2.48-2.60 mE_h and 0.17-0.18 mE_h, respectively).

Finally, after reassembling the predicted H, we compute its eigenvalues and compare them to those of the DFT reference. For HfO_2 with 3,000 atoms, we achieve an L1 error of 0.55%, sufficient for practical applications. Compared to full graph training, our method with just 8 augmented slices results in a 6.5$\times$ speedup per epoch (0.38 vs. 2.5 s) and a 7.2$\times$ decrease in memory consumption per rank (8.59 vs. 61.68 GiB) without affecting accuracy (**Appendix B**). Our approach can thus be applied to train and predict the **H** of materials with arbitrarily large structures for a wide range of applications, including phase change compounds, functional oxides, or semiconductor-dielectric interfaces [13, 14].

Acknowledgments

We would like to thank Paul Uriarte Vicandi and Luiz Felipe Aguinsky for providing the HfO_2 and PtGe structures, respectively. This work was supported by the SNSF ALMOND project (grant no. 198612) and by CSCS under project lp16.

References

[1] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, November 1965.

[2] Gil M. Repa and Lisa A. Fredin. Predicting electronic structure of realistic amorphous surfaces. *Advanced Theory and Simulations*, 6(11), June 2023.

[3] 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.

[4] Haiyang Yu, Zhao Xu, Xiaofeng Qian, Xiaoning Qian, and Shuiwang Ji. Efficient and equivariant graph networks for predicting quantum hamiltonian, 2023.

[5] 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.

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

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

[8] 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.

[9] He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, June 2022.

[10] 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.

[11] Zhanghao Zhouyin, Zixi Gan, Shishir Kumar Pandey, Linfeng Zhang, and Qiangqiang Gu. Learning local equivariant representations for quantum operators, 2024.

[12] Yuxiang Wang, Yang Li, Zechen Tang, He Li, Zilong Yuan, Honggeng Tao, Nianlong Zou, Ting Bao, Xinghao Liang, Zezhou Chen, Shanghua Xu, Ce Bian, Zhiming Xu, Chong Wang, Chen Si, Wenhui Duan, and Yong Xu. Universal materials model of deep-learning density functional theory hamiltonian. *Science Bulletin*, 69(16):2514–2521, 2024.

[13] M.Y. Chan, T. Zhang, V. Ho, and P.S. Lee. Resistive switching effects of hfo2 high-k dielectric. *Microelectronic Engineering*, 85(12):2420–2424, December 2008.

[14] A. Pirovano, A.L. Lacaita, A. Benvenuti, F. Pellizzer, and R. Bez. Electronic switching in phase-change memories. *IEEE Transactions on Electron Devices*, 51(3):452–459, March 2004.

[15] 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.

[16] Yong Youn, Youngho Kang, and Seungwu Han. An efficient method to generate amorphous structures based on local geometry. *Computational Materials Science*, 95:256–262, December 2014.

[17] De Nyago Tafen and D. A. Drabold. Realistic models of binary glasses from models of tetrahedral amorphous semiconductors. *Physical Review B*, 68(16), October 2003.

[18] 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.

[19] Thomas D. Kühne, Marcella Iannuzzi, Mauro Del Ben, Vladimir V. Rybkin, Patrick Seewald, Frederick Stein, Teodoro Laino, Rustam Z. Khaliullin, Ole Schütt, Florian Schiffmann, Dorothea Golze, Jan Wilhelm, Sergey Chulkov, Mohammad Hossein Bani-Hashemian, Valéry Weber, Urban Borštnik, Mathieu Taillefumier, Alice Shoshana Jakobovits, Alfio Lazzaro, Hans Pabst, and *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.

[20] Anders Blom Troels Markussen Jess Wellendorff Julian Schneider Tue Gunst Brecht Verstichel Petr A Khomyakov Ulrik G Vej-Hansen Mads Brandbyge Søren Smidstrup, Kurt Stokbro et al. Quantumatk: An integrated platform of electronic and atomic-scale modelling tools. *J. Phys: Condens. Matter (APS)*, 32:015901, 2020.

[21] Joost VandeVondele and Jürg Hutter. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.*, 127(11):114105, 09 2007.

[22] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple [phys. rev. lett. 77, 3865 (1996)]. *Phys. Rev. Lett.*, 78:1396–1396, Feb 1997.

[23] Vladimir I. Anisimov, Jan Zaanen, and Ole K. Andersen. Band theory and mott insulators: Hubbard u instead of stoner i. *Physical Review B*, 44:943–954, Jul 1991.

[24] M. L. Senent and S. Wilson. Intramolecular basis set superposition errors. *International Journal of Quantum Chemistry*, 82(6):282–292, 2001.

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

[26] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 07 2013.

[27] 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.

[28] Gábor Csányi, Steven Winfield, J R Kermode, A De Vita, Alessio Comisso, Noam Bernstein, and Michael C Payne. Expressive programming for computational physics in fortran 95+. *IoP Comput. Phys. Newsletter*, page Spring 2007, 2007.

[29] Frank H. Stillinger and Thomas A. Weber. Computer simulation of local order in condensed phases of silicon. *Phys. Rev. B*, 31:5262–5271, Apr 1985.

[30] K. Nordlund, M. Ghaly, R. S. Averback, M. Caturla, T. Diaz de la Rubia, and J. Tarus. Defect production in collision cascades in elemental semiconductors and fcc metals. *Phys. Rev. B*, 57:7556–7570, Apr 1998.

[31] Z. Q. Wang and D. Stroud. Monte carlo studies of liquid semiconductor surfaces: Si and ge. *Phys. Rev. B*, 38:1384–1391, Jul 1988.

[32] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Damania, and Soumith Chintala. Pytorch distributed: Experiences on accelerating data parallel training, 2020.

Appendix A. Datasets

The structures generated and used for training, validation and testing during the experiments are shown in A1. Atomic structures corresponding to materials in the amorphous phase are not straightforward to generate since they must accurately capture the structural motifs underlying this phase and a realistic range of atomic coordination. To accurately reproduce long-range structural disorder, the structures used must also be large enough to avoid the creation of wavefunctions that repeat over periodic boundaries. Existing methods to do so include melt-quench [15], seed-and-coordinate [16], or 'decorate and relax' [17] approaches.

In this work, we use melt-quench processes with molecular dynamics (MD) to evolve each of the three materials considered from their crystalline phases, following a similar procedure as the ones described in [18] and [15]. We then perform a structural relaxation with CP2K code [19] to correct for any discrepancies between the relaxed bond lengths attained with the force field used for MD and those obtained with DFT. Due to the large cell sizes of the a-HfO_2 structures, all necessary information is contained within the Γ point (where the wavevectors k_x = k_y = k_z = 0). The energies at this location can be computed by directly diagonalizing **H**.

Further details specific to each material are provided in the sections below.

1.1 a-HfO_2

We generate 3 independent structures of a-HfO_2 using the QuantumATK toolkit [20]. As the first step, we run an MD NVT simulation at 3000K for 50 ps with a step size of 1 fs. We use the MTP-HfO_2-2022 potential provided by the software. Next, we run an NPT simulation for 300 ps (and the same 1 fs step size), with an initial reservoir temperature of 3000K and a final temperature of 300K, for a cooling rate of 9K/ps. Finally, we anneal the structure at 300K for 50 ps.

Due to the computational cost of using a more complete Double-ζ Valence Polarized (DZVP) basis set [21], we use a simpler Single-ζ Valence (SZV) basis [21], which uses 4 basis functions per Oxygen atom and 10 basis functions per Hafnium atom. The plane-wave cutoff is set to 500 Ry, while a cutoff of 60 Ry is used for mapping the Gaussian-type orbitals onto the grid. We use the PBE functional for the exchange-correlation energy [22]. To accurately capture the band gap of a-HfO_2, we apply the Hubbard correction [23] of U = 7 eV to the 3d orbital of Ti and the Hubbard correction of U = 10 eV to the 2p orbital of O.

1.2 Substoichiometric HfO_x

We create a dataset for sub-stoichiometric HfO_x structures by introducing randomly distributed oxygen vacancies into the original, pristine HfO_2 structures. The sub-stoichiometric structures are gener-

Material	Structure	Purpose	r_{cut} [Å]	# atoms	# orbitals	# edges	x [Å]	y [Å]	z [Å]
a-HfO_2	1	validate	8	3,000	18,000	527,348	52.876	26.308	26.242
a-HfO_2	2	train	8	3,000	18,000	533,364	52.346	26.237	26.293
a-HfO_2	3	test	8	3,000	18,000	530,920	52.722	26.267	26.191
a-GST	1	train/validate (6:1 split)	12	1,008	13,104	230,848	29.541	25.583	41.777
a-GST	2	test	12	1,008	13,104	226,406	25.857	29.857	41.691
a-PtGe	1	train/validate (10:1 split)	8	2,688	16,128	319,262	82.283	23.171	25.031
a-PtGe	2	test	8	2,688	16,128	319,306	82.283	23.171	25.031
a-PtGe	2	test	16	2,688	16,128	2,154,506	82.283	23.171	25.031

Table A1: Attributes of the generated dataset for three materials, each with its own training, validation, and test set: The $[x, y, z]$ triplet defines the periodic unit cell size. nnz_H is the number of non-zero elements in the Hamiltonian, encompassing all orbital interactions. Edges were defined according to an interaction distance r_{cut}.

ated for x = 1.9, 1.8, and 1.7 (corresponding to vacancy concentrations of 5%, 10%, and 15 %, respectively). Vacancies are treated as ghost atoms (atoms with no orbitals, but with a basis set defined at their locations), to mitigate the basis set superposition error [24], a known problem related to localized basis sets. More precisely, by treating vacancies as ghost atoms, one prevents the excessive borrowing of the basis sets from neighboring atoms by the vacancy, which improves the accuracy of the predicted electronic properties. These ghost atoms are assigned an atomic number of 0.

1.3 a-GST

Two amorphous GST-124 ($Ge(SbTe_2)_2$) structures containing 1008 atoms have been used for training and validation. The first structure is extracted from a crystallization trajectory provided by [25] (Supplementary Material). It is contained in a 25 Å × 30 Å × 40 Å orthorhombic bounding box. The second structure is the result of passing a fully crystalline GST-124 structure (the unit cell of which was from the Materials Project [26] and duplicated to fill a 25 Å × 30 Å × 40 Å monoclinic bounding box) through a standard melt-quench procedure (Randomization at 3,000 K (20 ps), cooling to the melting point of 600K at a rate of $10^{14}K.s^{\text{˘}1}$, holding for 30 ps, quenching to 300 K at $2.5 \times 10^{13}K.s^{\text{˘}1}$ then holding again for another 50 ps). Both structures are relaxed via MD simulations in LAMMPS [27] equipped with the QUIP library for Gaussian Approximation Potential (GAP) [28]. The corresponding Hamiltonian terms are obtained using CP2K, where we run the calculations with the DZVP basis, the plane-wave cutoff of 300 Ry, the Gaussian-type orbitals mapping cutoff of 50 Ry, and the PBE functional.

1.4 a-PtGe

To generate the PtGe structures, germanium structures are taken from the Materials Project database [26]. This is followed by an NVT melt-quench process using LAMMPS and Stillinger-Weber parameters [29, 30, 31]. The structures are heated to a melting temperature of 5000K at a rate of 0.47 $10^{12}K.s^{\text{˘}1}$, kept at the melting temperature for 20000 ps (structure 1) or 22000 ps (structure 2), quenched at a rate of 4.7 $10^{12}K.s^{\text{˘}1}$, and finally annealed at 300K for 100 ps.1/3 of the Ge atoms are then randomly replaced by Pt atoms. The cell of the alloy is then stretched to match the cell of a $PtGe_2$ structure (taken from the Materials Project and optimized using CP2K). Fixed-volume geometry relaxation is then performed on the PtGe alloy. For the structural optimization, as well as for the **H** and **S** generation, SZV basis set and PBE exchange-correlation functionals are used. We apply a plane-wave cutoff of 1000 Ry and a cutoff for Gaussian-type orbitals mapping of 70 Ry.

Appendix B. Compute environment and runtime comparisons

The training is performed with PyTorch Distributed Data Parallel [32], where the graph partitions (slices) can be distributed between GPUs.

2.1 Memory consumption of full-graph training

During the training of the full graph model, the peak memory consumption observed was 61.68 GiB on a single NVIDIA A100 GPU. Most of the consumption does not stem from the network and the structure but from the additional memory needed for the convolution operations.

2.2 Scalability of augmented partitioning

In **Fig. A1**, we show the decrease in time per epoch and resulting speedup when using the *augmented partitioning* approach introduced in **Section 3.2**. Since the partitions are independent, the only communication involved in every epoch is a collective to inform each GPU/rank of the loss of each other rank. The time per epoch thus decreases uniformly with the number of slices (N_t) used.

Despite the independence of each batch and the minimal communication per epoch, the scaling is not perfectly linear. The deviation from an ideal speedup can be attributed to two factors:

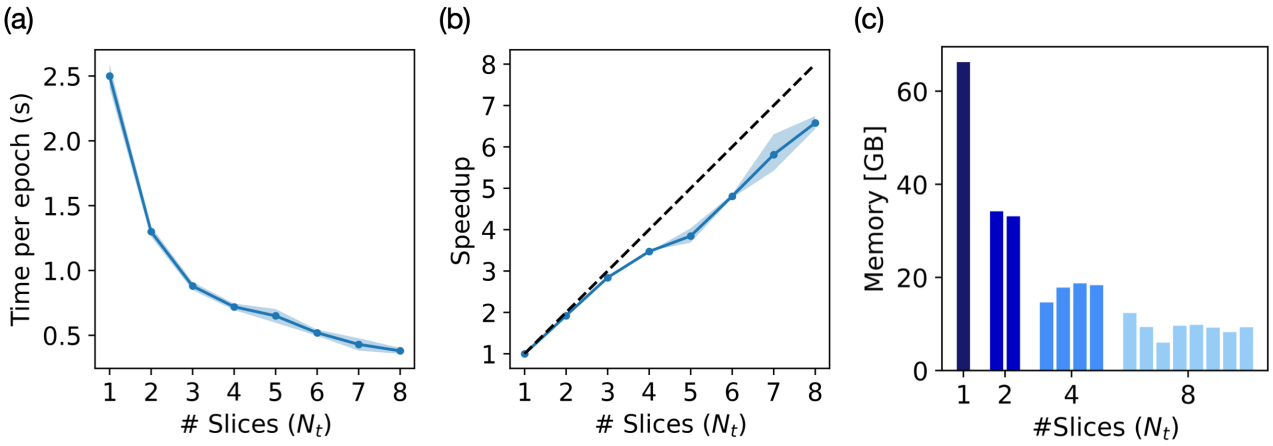

Fig. A1: (a) Time per epoch and (b) speedup resulting from the use of increasing numbers of slices N_t. Median values are shown, while the error bands are one standard deviation. Experiments were run on NVIDIA A100 GPUs with # ranks set to N_t. Measurements are only shown up to 8 slices/8 GPUs due to limitations in available compute resources at the time of submission. The fill-between indicates the range in runtime over the first 30 minutes of training. The dashed black line corresponds to the ideal speedup, in which case the use of N_t slices would enable an $N_t\times$ speedup in the runtime per epoch. (c) Measured peak memory consumption as a function of the number of partitions, where each bar corresponds to a different GPU. Variation in memory consumption between GPUs at each individual value of N_t translates to load imbalance, which correlates with the deviation from ideal scaling shown in (b).

- Load imbalance: The partitioning approach was designed to leverage the periodicity in the y- and z- direction within a straightforward implementation. However, it is not ideal in terms of the number of cuts/number of virtual nodes/edges required, resulting in a slightly different amount of work per rank which leads to an observable load imbalance at higher N_t. This effect can be seen in the allocated memory per partition (**Fig. A1(c)**). We note that the *augmented partitioning* method can be used with any standard graph-partitioning algorithm.
- Computational overhead of the virtual nodes and edges: Individual nodes and edges of the graph can be repeated in labeled and virtual node lists. Treating the replicas introduces additional computational cost while training the network, which increases with N_t. This overhead is maximum with the use of very small slices (large N_t), thus introducing a trade-off between parallelism and time per epoch.

2.3 H_2O vs HfO_2 runtimes

We make a comparison between the computational cost of computing the Hamiltonian for an H_2O molecule and the HfO_2 structure. To approximate the cost of generating them under the same computational conditions, we set up CP2K simulations with a DZVP basis for H_2O. The computation time per H_2O molecule was 7s, when run on 12 nodes with 12-core Intel Xeon E5-2680 CPUs and NVIDIA P100 GPU, resulting in a total of 0.04 node hours. The HfO_2 structures require 3.65 node hours in the same compute environment (but distributed to 27 nodes). The difference, omitting scaling behavior, is $\sim$100$\times$.

Appendix C. Hyperparameters

For all experiments, the embedding size is fixed at 16, and the feedforward dimension size is 64. L_{max} and M_{max} are both set to 4, and 2 attention heads are used.

For training, we use a an Adam optimizer, and a ReduceLRonPlaeau scheduler, with an initial rate of 1×10^{-4}. The scheduler decreases the learning rate by a decay factor of 0.5 when it does not detect a further decrease in validation loss within the decay patience of 500. Once the minimum learning rate of 1×10^{-5} is reached, the training stops. Mean Squared Error (MSE) is used as the loss function.