EVALUATING UNIVERSAL INTERATOMIC POTENTIALS FOR MOLECULAR DYNAMICS OF REAL-WORLD MINERALS

Sajid Mannan
Indian Institute of Technology Delhi
Hauz-Khas, Delhi, India
sajid.mannan@civil.iitd.ac.in

Carmelo Gonzales
Machine Learning Engineer
Intel Labs
carmelo.gonzales@intel.com

Vaibhav Bihani
Indian Institute of Technology Delhi
Hauz-Khas, Delhi, India
vaibhav.bihani525@gmail.com

Kin Long Kelvin Lee
AI Researcher
Intel Labs
kin.long.kelvin.lee@intel.com

Nitya Nand Gosvami
Indian Institute of Technology Delhi
Hauz-Khas, Delhi, India
ngosvami@iitd.ac.in

Santiago Miret
AI Researcher
Intel Labs
santiago.miret@intel.com

N.M. Anoop Krishnan
Indian Institute of Technology Delhi
Hauz-Khas, Delhi, India
krishnan@iitd.ac.in

ABSTRACT

Universal interatomic potentials (UIPs) have emerged as promising models for capturing complex atomic interactions across diverse material families through graph-based representations. Recent UIP architectures, trained on density functional theory (DFT) trajectories spanning the periodic table, have demonstrated accuracy in energy and force predictions for 0 K structures. However, their efficacy for finite temperature molecular dynamics (MD) simulations of experimentally verified materials under physical conditions remains unexplored. We present a comprehensive evaluation of six state-of-the-art UIPs (CHGNET, M3GNET, MACE, MATTERSIM, SEVENNET, ORB) on a curated dataset, namely AMCSD-MD-2.4K, comprising $\sim$2,400 minerals with experimentally validated crystal structures and densities from the American Mineralogist Crystal Structure Database. Our analysis comprises two components: (1) a systematic comparison of model performance across the mineral dataset, and (2) a quantitative assessment of temporal evolution during MD simulations, analyzing structural properties including density and lattice parameters. Our evaluation reveals significant performance variations among UIPs, with ORB and SEVENNET achieving completion rates of 99.96% and 98.75% respectively, while CHGNET completed only 7% of simulations. Furthermore, none of the models achieved the empirically accepted structural variation threshold of $\pm 2.5\%$, with MACE, MATTERSIM, SEVENNET, and ORB showing comparatively better accuracy ($R^2 > 0.8$) in density predictions. This evaluation framework establishes rigorous benchmarks for assessing UIP performance in realistic atomistic simulations of mineral systems.

1 INTRODUCTION

Graph-based machine learning interatomic potentials (MLIPs) have emerged as powerful techniques for large-scale materials simulations, offering unprecedented capabilities in atomic-scale modeling. Recent advances in training graph neural network (GNN) architectures on extensive databases have demonstrated significant potential for accelerating materials property prediction and revolutionizing molecular dynamics (MD) simulations at scale (Duval et al., 2023; Bihani et al., 2024; Batatia et al., 2024; Musaelian et al., 2023; Fu et al., 2023). MD simulations are particularly crucial for understanding materials behavior, as they capture thermal effects, transport properties, structural transitions, and dynamic properties essential for predicting real-world material performance. State-of-the-art models, including CHGNET (Deng et al., 2023), M3GNET (Chen & Ong, 2022), MACE (Batatia et al., 2024), MATTERSIM (Yang et al., 2024), SEVENNET (Park et al., 2024), and ORB (Neumann et al., 2024), have demonstrated remarkable performance in predicting energy and forces across diverse atomic systems. Yet, their performance on simulating the dynamic behavior of finite temperatures, something of utmost importance for understanding material response, remain poorly understood.

The development of Universal Interatomic Potentials (UIPs) has been accelerated by two key factors: the availability of comprehensive computational materials datasets (Deng et al., 2023; Lee et al., 2023; Chanussot et al., 2021; Batzner et al., 2022; Schmidt et al., 2024) and advances in training methodologies. These developments have enhanced the models' ability to generalize across the periodic table while capturing complex atomic interactions with increased precision. However, despite training on millions of structural configurations from simulation trajectories, these models frequently encounter challenges when applied to real-world systems (Bihani et al., 2024; Fu et al., 2023; Gonzales et al., 2024). The absence of comprehensive benchmarks under realistic conditions has made it challenging to assess the true capabilities and limitations of UIPs for practical materials design applications.

To address this gap, we present a systematic study evaluating the numerical stability and accuracy of dynamics simulations using current state-of-the-art UIPs on AMCSD-MD-2.4K, a curated benchmark derived from the American Mineralogist Crystal Structure Database (AMCSD) (Downs & Hall-Wallace, 2003). This database is particularly significant as it contains experimentally validated structures of minerals, which constitute approximately of the entire of Earth's crust and are fundamental to wide range of applications including geological processes, construction, and other industrial applications. The AMCSD-MD-2.4K dataset, comprising *experimentally* determined structures of $\sim$2,400 minerals, enables a rigorous assessment of UIP performance across a diverse range of naturally occurring mineral systems, providing insights into their practical applicability and limitations.

2 AMCSD-MD-2.4K DATASET

Contemporary UIPs derive their training from specialized DFT-generated datasets including MPTRJ (Deng et al., 2023), OC22 (Tran et al., 2023), Alexandria (Schmidt et al., 2024), and related ones. Figure 1 depicts elemental distributions across both MPtrj training data and AMCSD evaluation sets, providing a direct comparison of our benchmark dataset with the elemental distribution in the training data and the compositional coverage of modern UIPs. The visualization employs a dual-color scheme—blue-green gradient for MPtrj and pink for AMCSD frequencies—with logarithmic scaling normalized to the second-highest count. Elements lacking color gradients indicate absence from both datasets. Element-wise frequency analysis in Figure 1a reveals near-complete periodic table coverage between MPtrj and AMCSD, with Americium as the sole exception.

Further, we analyze the structural complexity by considering the number of elements in the unit cell for the both datasets (see Figure 1b). While MPtrj structures exhibit limited compositional diversity, containing maximum 9 unique elements per structure, AMCSD minerals have up to 23 distinct elements suggesting the increased complexity of real-world minerals. Additionally, the Figure 1c shows the distribution of atoms

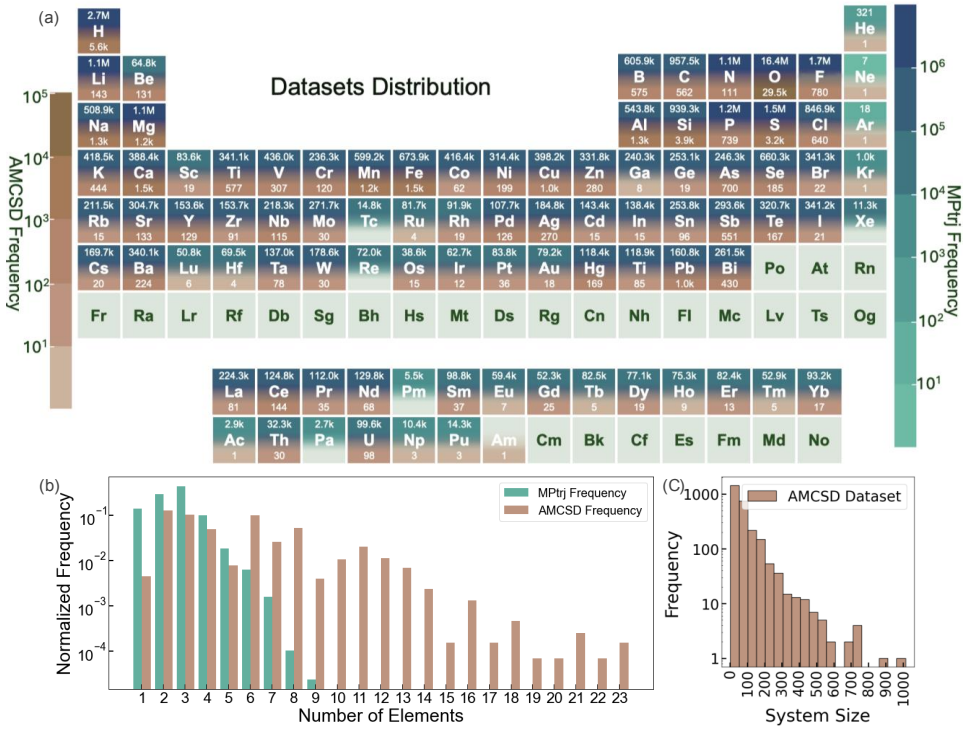

Figure 1: (a) Distribution of atomic elements and their frequency in the MPtrj (blue-green gradient) and AMCSD (pink) datasets respectively. (b) Number of elements present in the unit cells of MPtrj (left abscissa) and AMCSD (right abscissa). (c) Distribution of the number of atoms in the unit cell of AMCSD.

in the AMCSD unit cells, including minerals with partial occupancy. The distribution shows several unit cells with larger number of atoms substantially exceeding typical MPtrj configurations. These disparities in compositional and structural complexity position AMCSD-MD-2.4K as an optimal benchmark for assessing UIP generalization capabilities on real-world mineral systems. All the codes and data is updated in the github repo: `https://anonymous.4open.science/r/Benchmarking_UIPs-55A3/`.

3 MD Simulation Details

AMCSD-MD-2.4K crystallographic information files (CIFs) were manually curated and validated during extraction to ensure appropriate experimental conditions (including temperature and pressure), and physicochemical accuracy. These structures were subsequently preprocessed for compatibility with the Atomic Simulation Environment (ASE) package API (Larsen et al., 2017). The simulation protocol implemented systematic standardization of the simulation supercell: unit cells were replicated to achieve system sizes of 50-100 atoms, with exceptions for structures inherently exceeding this threshold. Spatial replication proceeded sequentially along ascending lattice vectors to optimize toward cubic supercell (similar size in all three directions) while preserving crystallographic integrity and minimizing anisotropic effects. The complete system size distribution is illustrated in Appendix Figure 6.

The computational workflow incorporated a dual-phase equilibration strategy. Initial structural optimization utilized the Fast Inertial Relaxation Engine (FIRE) algorithm (Bitzek et al., 2006) for 200 steps, followed by a 50 ps NPT equilibration phase. Phase-space sampling initiated with Maxwell-Boltzmann velocity distributions at experimentally determined temperatures from CIF metadata, with a canonical temperature of 298 K applied for unspecified cases based on the reference literature. The NPT equilibration implemented the Berendsen thermostat and barostat (Berendsen et al., 1984), maintaining experimentally reported pres-

sures or a standard state pressure of 1 atm. Production MD simulation runs were executed for 50 ps with an integration timestep of 1 fs, capturing trajectory and thermodynamic data at 10-step intervals.

4 RESULTS

A systematic evaluation of pre-trained UIPs through molecular dynamics simulations revealed distinct performance variations across multiple assessment metrics. These are discussed in the following sections.

4.1 PERFORMANCE LANDSCAPE OF UIPS

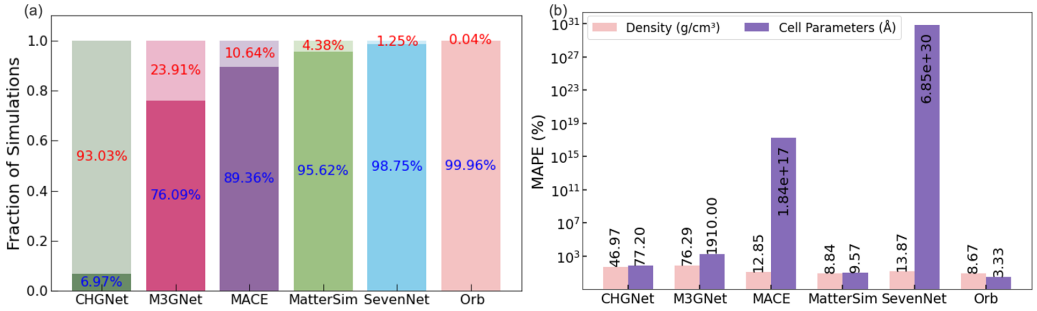

Figure 2: Performance evaluation of Universal Interatomic Potentials on AMCSD-MD-2.4K. (a) Fraction of successfully completed 50 ps MD simulations (dark segments, blue annotations) versus failed simulations (light segments, red annotations) across 2,400 mineral structures. (b) Mean Absolute Percentage Error (MAPE) for density (g/cm^3, pink) and lattice parameters (Å, purple), calculated for predictions with respect to the experimental values.

Figure 2(a) quantifies simulation completion rates across 2,400 mineral structures over 50 ps trajectories. The dark segments with blue annotations indicate successful completions, while light segments with red annotations denote simulation failures. Simulation failures manifested through two primary mechanisms: (1) memory overflow during model forward pass, where structural instabilities generated excessive edge computations in the graph representation, and (2) computationally prohibitive MD timesteps necessitating premature termination. While increased memory allocation delayed failure onset, it proved insufficient for preventing them, leading to the standardized allocation detailed in Appendix A.1.

Quantitative analysis revealed a clear performance hierarchy among UIPs. ORB achieved exceptional stability with 99.96% completion rate, followed by SEVENNET (98.75%) and MATTERSIM (95.6%). MACE maintained reasonable stability at 88.4% completion, while M3GNET completed 74.2% of simulations. CHGNET exhibited significant instability, successfully completing only 7% of simulations. These results suggest that even the best UIPs often exhibit instability and failure, a minimum requirement for performing materials simulations.

Following the simulations, the reasonability of the structures predicted at the end of the MD simulations in comparison to the experimental measurement was evaluated. Structural prediction accuracy, shown in Figure 2(b), evaluated the mean absolute percentage error (MAPE) of the density and lattice parameters of the simulated structures. Only the simulations that were completed successfully were included for this analysis. Structural prediction accuracy, shown in Figure 2(b), was quantified through mean absolute percentage error (MAPE) in density and lattice parameter predictions. ORB achieved superior accuracy with minimal deviations (density: 8.67%, lattice: 3.33%), followed by MATTERSIM (density:8.84%, lattice: 9.57%). SEVENNET maintained comparable precision in density predictions (13.87%) despite higher lattice parameter errors (6.85×10^{30}%) due to exploded simulations. Similar behavior were observed with MACE as well, which exhibited moderate accuracy (density: 12.85%, lattice: 1.84×10^{17}%), while M3GNET and CHGNET showed substantial deviations (density: 76.29%, 46.97%; lattice: 1.91×10^{2}%, 77.20% respectively) confirming their limited capabilities for MD simulations. Complete parity plots for each of the

models for density and lattice parameters along with additional error metrics are provided in the Appendix (see Figure 4 and Figure 5).

These metrics, constrained to predictions on successful simulations, indicate a clear hierarchy in structural property prediction capabilities among current UIPs. Notably, even the best-performing UIPs exceed the experimentally acceptable density variation threshold of $\pm 2.5\%$, highlighting a critical gap between computation

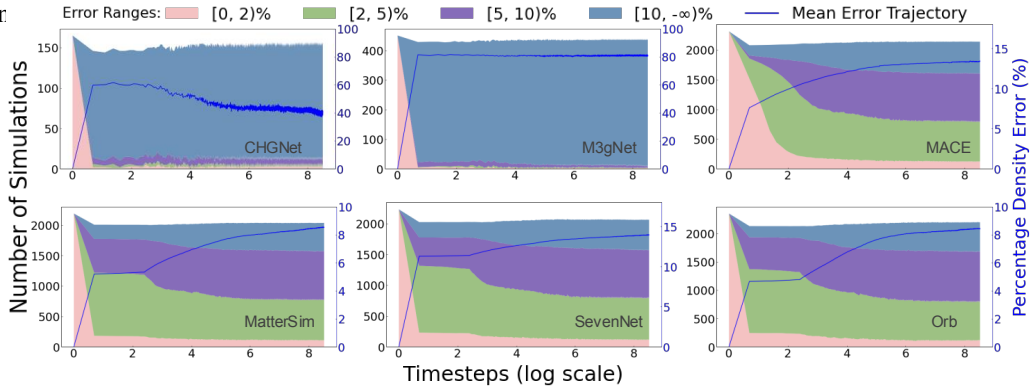

Figure 3: Temporal evolution of density errors during molecular dynamics simulations. The stacked areas represent error distributions across four ranges ([0,2)%, [2,5)%, [5,10)%, [10,∞)%), with the blue line indicating mean error trajectory. Y-axis (left) shows mineral population count; Y-axis (right) displays percentage density error; X-axis represents simulation timesteps on logarithmic scale.

4.2 Temporal evolution of structure

Analysis of density fluctuations during molecular dynamics trajectories provides critical insights into UIP stability and accuracy. Figure 3 presents the temporal evolution of density errors, with simulation timesteps plotted logarithmically against the population distribution of simulated minerals. Each error regime is color-coded ([0,2)%, [2,5)%, [5,10)%, [10,∞)%), while the mean error trajectory (blue line) quantifies aggregate performance. Temporal analysis reveals distinct behavioral patterns among UIPs. CHGNet and M3GNet exhibit substantial instability with mean density deviations of 40% and 80% respectively, indicating fundamental limitations in maintaining structural integrity. MACE, MatterSim, SevenNet, and Orb demonstrate superior stability, with density errors converging below 15%. Notably, MatterSim and Orb achieve exceptional accuracy, maintaining mean density errors below 10% throughout the simulation window. The error distribution reveals that while initial timesteps show broader variance, stable models converge to consistent error ranges, suggesting a reasonable sample of the equilibrium configurations in the energy landscape.

5 Conclusion

This work presents a comprehensive benchmark of six state-of-the-art Universal Interatomic Potentials on a newly proposed AMCSD-MD-2.4K dataset, establishing quantitative metrics for molecular dynamics simulation performance on experimentally validated mineral structures. Our analysis reveals significant variations in model capabilities: Orb and SevenNet exhibited simulation stability for most minerals ($> 98\%$ completion), while CHGNet exhibited critical instabilities (7% completion). Density predictions, even from the best-performing models, exceeded experimentally acceptable thresholds ($\pm$ 2.5%), highlighting fundamental limitations in current architectures to simulate experimentally meaningful structures. Temporal analysis further demonstrated that the properties evolved during the simulations, with even the best performing UIPs, MatterSim and Orb, maintaining structural stability with mean density errors below 10%, suggesting the limitations of UIPs for MD simulations to analyze the materials response.

Limitations and Future Work. Despite the extensive coverage across the periodic table, the current evaluation framework exhibits several key limitations, which will be pursued as part of future efforts. The assessment primarily focuses on structural properties, leaving out the detailed structural analysis including bond length, and angular distributions. MD simulations trajectories obtained could be further analyzed in detail to understand the temporal evolution of the detailed structure of minerals. Additionally, the benchmark lacks evaluations at high temperatures and pressures, which are conditions to which materials are regularly exposed to. Current assessment metrics predominantly address equilibrium properties, overlooking transition state behavior and non-equilibrium phenomena including phase change, and fracture. Finally, evaluation of experimental structures with grain boundaries, interfaces, and surfaces could potentially add to the value of the benchmark.

REFERENCES

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024. URL https://arxiv.org/abs/2401.00096.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1):2453, 2022.

Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

Vaibhav Bihani, Sajid Mannan, Utkarsh Pratiush, Tao Du, Zhimin Chen, Santiago Miret, Matthieu Micoulaut, Morten M Smedskjaer, Sayan Ranu, and NM Anoop Krishnan. Egraffbench: evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery*, 3(4):759–768, 2024.

Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical review letters*, 97(17):170201, 2006.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10):6059–6072, 2021.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022.

Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

Robert T Downs and Michelle Hall-Wallace. The american mineralogist crystal structure database. *American Mineralogist*, 88(1):247–250, 2003.

Alexandre Duval, Simon V Mathis, Chaitanya K Joshi, Victor Schmidt, Santiago Miret, Fragkiskos D Malliaros, Taco Cohen, Pietro Lio, Yoshua Bengio, and Michael Bronstein. A hitchhiker's guide to geometric gnns for 3d atomic systems. *arXiv preprint arXiv:2312.07511*, 2023.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi S Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

Carmelo Gonzales, Eric Fuemmeler, Ellad B. Tadmor, Stefano Martiniani, and Santiago Miret. Benchmarking of universal machine learning interatomic potentials for structural relaxation. In *AI for Accelerated Materials Design - NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=fNyXCCZ0g6.

Tsz Wai Ko, Marcel Nassar, Santiago Miret, Elliott Liu, Ji Qi, and Shyue Ping Ong. Materials Graph Library, June 2021.

Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N Groves, Bjørk Hammer, Cory Hargus, et al. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27): 273002, 2017.

Kin Long Kelvin Lee, Carmelo Gonzales, Marcel Nassar, Matthew Spellings, Mikhail Galkin, and Santiago Miret. Matsciml: A broad, multi-task benchmark for solid-state materials modeling. *arXiv preprint arXiv:2309.05934*, 2023.

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.

Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential, 2024. URL https://arxiv.org/abs/2410.22570.

Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *J. Chem. Theory Comput.*, 20 (11):4857–4868, 2024. doi: 10.1021/acs.jctc.4c00190.

Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures. URL https://github.com/microsoft/mattersim/.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024. URL https://arxiv.org/abs/2405.04967.

A APPENDIX

A.1 EXPERIMENT DETAILS

All experiments were run on an internal cluster, using Kubernetes for orchestration. A single job was launched per material and model in an embarrassingly parallel fashion, using 6 CPUs and 12GB of RAM per run. In simulations that run for 50,000 steps, the computation cost of running a benchmark across materials quickly adds up and becomes a limiting factor given finite computational resources. In addition to computational cost, storage requirements for experiment results and metadata also need to be considered. In total, around 850 GB of data was saved including experiment result logs, metadata, experiment tracking logs, and evaluation data.

Table 1: Model checkpoints.

Model	Checkpoint	Repository/Source
CHGNET	CHGNet-MPtrj-2024.2.13-PES-11M	(Ko et al., 2021)
M3GNET	M3GNet-MP-2021.2.8-PES	(Ko et al., 2021)
MACE	2023-12-10-mace-128-L0_energy_epoch-249	(Batatia et al., 2024)
MATTERSIM	mattersim-v1.0.0-1M	(Yang et al.)
ORB	orb-v2-20241011	(Neumann et al., 2024)
SEVENNET	7net-0_11July2024	(Park et al., 2024)

Table 2: Time taken per MD simulation step per model.

Model	CHGNET	M3GNET	MACE	MATTERSIM	ORB	SEVENNET
time(s)/step	1.452	2.794	1.087	0.780	0.736	2.153

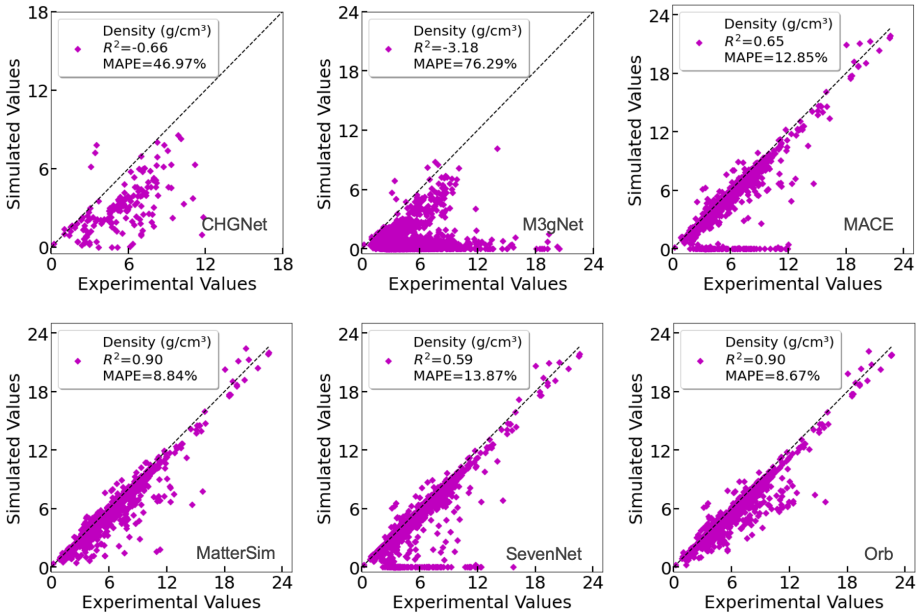

Figure 4: (a) parity plot of density

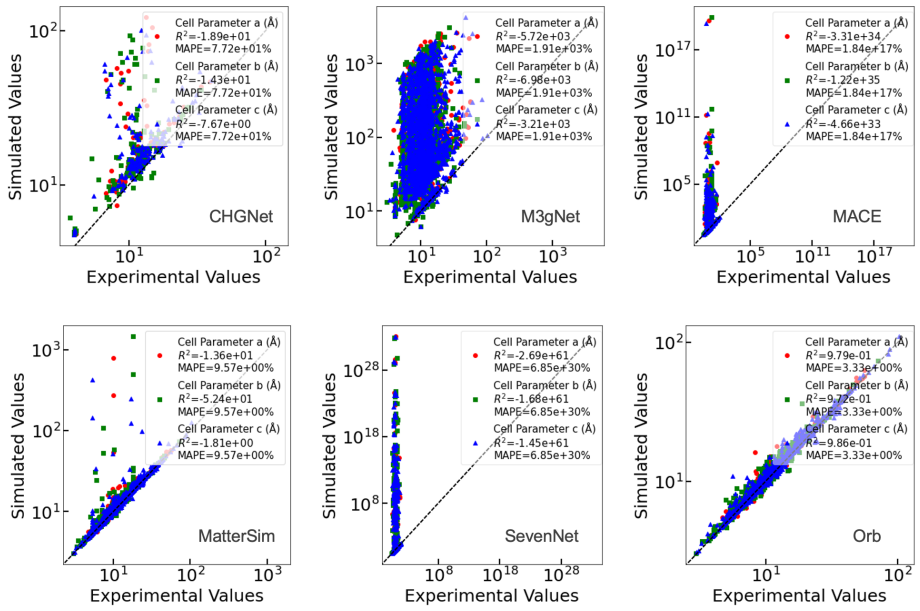

Figure 5: (a) parity plot of lattice parameters

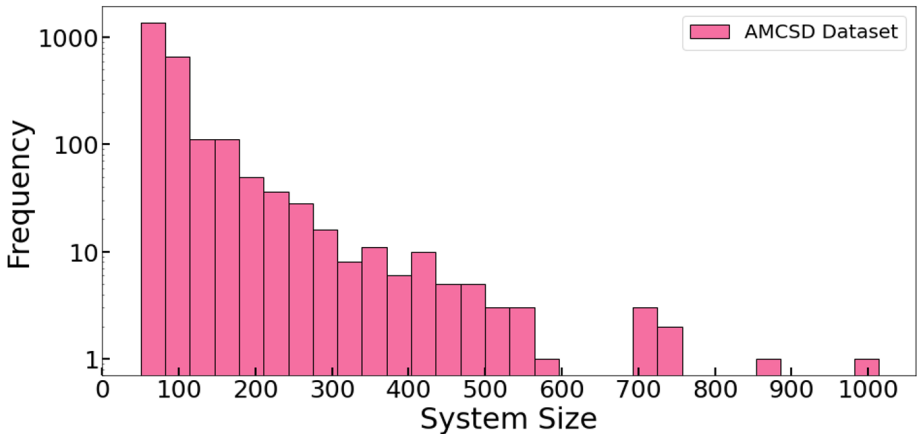

Figure 6: (a) Number of atoms in supercell