MLIP ARENA: ADVANCING FAIRNESS AND TRANSPARENCY IN MACHINE LEARNING INTERATOMIC POTENTIALS VIA AN OPEN, ACCESSIBLE BENCHMARK PLATFORM

Yuan Chiang[1,2] **Tobias Kreiman**[1] **Elizabeth Weaver**[1] **Ishan Amin**[1]
Matthew Kuner[1,2] **Christine Zhang**[1] **Aaron Kaplan**[2] **Daryl Chrzan**[1,2] **Samuel Blau**[2]
Aditi Krishnapriyan[1,2] **Mark Asta**[1,2]

[1]UC Berkeley [2]LBNL

ABSTRACT

Machine learning interatomic potentials (MLIPs) have revolutionized molecular and materials modeling, but existing benchmarks suffer from data leakage, limited transferability, and an overreliance on error-based metrics tied to specific density functional theory (DFT) references. We introduce MLIP Arena, a benchmark platform that evaluates MLIPs based on physics awareness, chemical reactivity, stability under extreme conditions, and predictive capabilities for thermodynamic properties and physical phenomena. Our evaluation challenges previous assumptions about model architectures and performance. MLIP Arena provides a reproducible framework to guide MLIP development toward improved predictive accuracy and runtime efficiency while maintaining physical consistency. The Python package and online leaderboard are available at `huggingface.co/spaces/atomind/mlip-arena`.

1 INTRODUCTION

The accurate prediction of molecular and material properties has driven innovation for decades and remains crucial for addressing challenges in energy technology, climate change, and drug discovery. While first-principles electronic structure methods have long served as the primary workhorse for property prediction, their computational cost remains prohibitive for scaling atomistic modeling beyond hundreds of atoms. Machine learning interatomic potentials (MLIPs), trained on extensive databases comprising millions of density functional theory (DFT) calculations, have emerged as an efficient and accurate alternative. These models have demonstrated remarkably accurate approximations of the DFT potential energy surface (PES) across a wide range of chemical compositions at a fraction of the computational cost of DFT.

However, MLIPs trained on the DFT total energy and interatomic forces do not necessarily capture the correct atomic interactions (Fu et al., 2022), despite excelling in error-based metrics for bulk systems (Riebesell et al., 2023). Analogously, it is well-known (Senftle et al., 2016) that classical force fields fit to describe near-equilibrium radial distribution functions cannot capture the energetics of bond-breaking. This limitation could continue to MLIPs that are primarily trained to near- and on-equilibrium structures. More specifically, error-based benchmarks on near-equilibrium structures can have limitations and might not correlate with utility for downstream scientific applications. We highlight some specific limitations:

First, they are vulnerable to data leakage, failing to accurately assess a model's extrapolation and generalization capabilities. This issue is evident in Matbench Discovery (Riebesell et al., 2023), where non-compliant models rank highly due to energy overfitting at the expense of forces, as this work will show. Additionally, high-ranking models often rely on large datasets, risking test set contamination without proper safeguards.

Second, benchmarks tied to specific datasets or DFT functionals lack flexibility in a rapidly evolving field, where larger, more chemically diverse, or higher-accuracy datasets frequently emerge (Barroso-

Correspondence to `cyrusyc@berkeley.edu`

Table 1: PEC quality of homonuclear diatomics based on physical and geometric measures. **Boldface** and underline represent the **best** and the worst metrics across all MLIPs, respectively. The rankings from all metrics are aggregated to rank the win rate (overall performance) of the MLIPs. Select PECs are shown in Figure S1. Detailed definitions and implementation details are available in Appendix A.2.

Model	Rank	Rank aggregation	Conservation deviation [eV/Å]	Spearman's coefficient E: repulsion	Spearman's coefficient F: descending	Energy jump [eV]	Force flips	Tortuosity
MACE-MP(M)	1	12	0.070	**-0.997**	-0.980	0.038	1.449	1.161
MatterSim	2	16	**0.013**	-0.980	-0.972	**0.008**	2.766	1.021
M3GNet	3	19	0.026	-0.991	-0.947	0.029	3.528	**1.016**
eSCN(OC20)	4	27	2.045	-0.939	-0.984	0.806	0.640	5.335
ORBv2	4	27	9.751	-0.883	**-0.988**	0.991	**0.991**	1.287
CHGNet	6	29	1.066	-0.992	-0.925	0.291	2.255	2.279
SevenNet	7	35	34.005	-0.986	-0.928	0.392	2.112	1.292
ORB	8	36	10.220	-0.881	-0.954	1.019	1.026	1.798
eqV2(OMat)	9	48	15.477	-0.880	-0.976	4.118	3.126	2.515
ALIGNN	10	53	5.164	-0.913	-0.310	9.876	30.669	1.818
EquiformerV2(OC20)	11	64	21.385	-0.680	-0.891	38.282	22.775	8.669
EquiformerV2(OC22)	12	69	27.687	-0.415	-0.855	64.837	21.674	15.880

Luque et al., 2024; Eastman et al., 2024; Kaplan et al., 2025; Schmidt et al., 2024). Static dataset benchmarks quickly become outdated and misleading as newer models trained on larger or proprietary datasets are introduced.

Third, conventional error-based regression metrics often fail to reflect the practical utility and generalizability of MLIPs in real-world applications. Póta et al. (2024) recently demonstrated that while some MLIPs exhibit zero-shot capabilities for lattice thermal conductivity prediction, many top-ranked Matbench Discovery models perform worse due to broken crystal symmetry and rough PES derivatives. This underscores that relying solely on regression metrics while ignoring physical priors can widen the gap between model predictions and experimental observables.

To address these challenges, we introduce **MLIP Arena**, a fair and transparent benchmark platform for foundation MLIPs. This platform evaluates the quality of the learned PES and the physical laws and symmetries critical to chemical modeling. Unlike prior error-based DFT reference benchmarks (Focassio et al., 2024; Riebesell et al., 2023; Wines and Choudhary, 2024; Yu et al., 2024; Zhu et al., 2025), **MLIP Arena focuses on examining physical soundness in order to evaluate the utility of MLIPs for downstream applications**. By moving beyond error-based assessments, it offers more actionable insights for model development and training. Specifically, we examine how well foundation MLIPs capture physics-aware phenomena, their reliability for accurate atomistic modeling, and their readiness for practical scientific research and discovery.

2 MLIP ARENA BENCHMARKS

MLIP Arena assesses the limitations of MLIPs through three primary perspectives. In Section 2.1, we focus on two-body interatomic interactions and propose metrics that enable robust and well-balanced ranking of MLIPs, reducing susceptibility to overfitting on any single metric. Section 2.2 tests MLIPs under extreme conditions using molecular dynamics (MD) simulations, exposing their instabilities and unphysical behaviors. Section 2.3 assesses the predictive capabilities of MLIPs in determining thermodynamic properties and physical phenomena, which requires multiple model passes, higher-order gradients, and more advanced workflows.

2.1 POTENTIAL ENERGY CURVES (PECS) OF DIATOMICS

Pairwise interactions are the most important interactions in atomistic systems. PECs have the benefit of being less vulnerable to data leakage as DFT references for PECs are difficult to calculate due to multiple possible spin configurations, basis set incompleteness in local-orbital DFT codes, and convergence issues in plane-wave DFT codes. In Table 1, we compute six physical and geometric measures to rank the homonuclear PECs of MLIPs in three aspects: conservative field, potential stiffness, and smoothness, as we discuss below and in further detail in Appendix A.2.

Conservative field. Conservative forces are important for energy conserving molecular simulations, and non-conservative forces are known to degrade the stability of thermostats (Bigi et al., 2024). We calculate the *conservation deviation* as the MAE between force and the central difference approximation of the derivative of the energy along all PECs (eq. (S1)). We note that energy conservation is a constraint that can be agnostic of the architecture itself, as the standard way it is enforced is by taking gradients of the predicted potential energy in the loss function.

Potential stiffness. Atoms at close distances should experience strong repulsion. We use *Spearman's coefficients* to measure the monotonicity of PECs at short interatomic distances. Robust MLIPs should have Spearman's coefficients close to -1 for the repulsive region in the energy curve and before the maximum attraction in the force curve.

Smoothness. We quantify the smoothness of the PECs by measuring the *tortuosity* (eq. (S2)), *energy jumps* (eq. (S3)), and *force flips*. Tortuosity measures the arc-chord ratio of PECs projected in the energy dimension. Smooth PECs with a single equilibrium point, like the Lennard-Jones potential, have a tortuosity strictly equal to 1. Energy jump detects the change in the sign of energy gradients and sums up the discontinuity with neighboring points. Force flips count the times force curves flip direction.

The metrics above provide a simplified picture of the quality of the general PES without considering heteronuclear and many-body interactions. A more careful analysis will be needed in those cases, especially when electrostatics become prominent in ionic interactions.

2.2 Stability and Reactivity

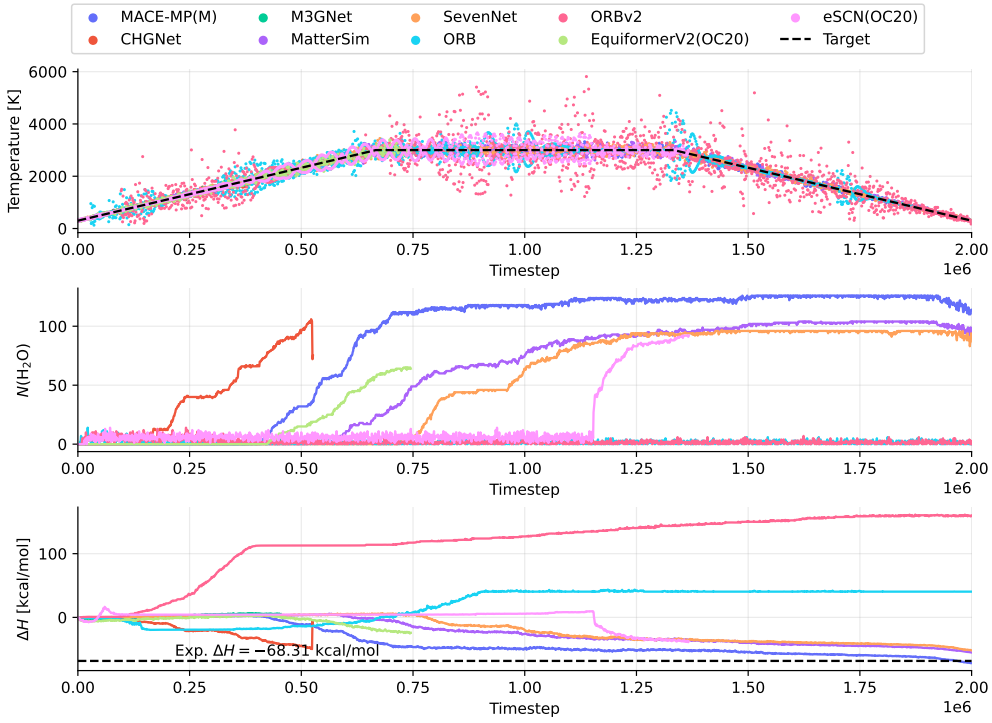

Figure 1: Hydrogen combustion via annealing NVT MD simulation ($128\,H_2 + 64\,O_2 \longrightarrow 128\,H_2O$). Applied temperature schedule is illustrated in the top panel. The experimental reaction enthalpy of $-68.31\,\mathrm{kcal/mol}$ is annotated in the bottom panel (Lide, 2004). CHGNet, EquiformerV2(OC20), eSCN(OC2), and M3GNet could not finish $1\,\mathrm{ns}$ MD trajectories. Experimental adiabatic flame temperature of hydrogen ranges from $2380\,\mathrm{K}$ (air) to $3000\,\mathrm{K}$ (pure O_2) (Hasche et al., 2023). Only MACE-MP(M) and EquiformerV2(OC20) ignite within this region. Runtime performance and center-of-mass drift are available in Figure S2.

Hydrogen combustion (temperature ramp). Hydrogen combustion is a challenging out-of-distribution test since there are multiple bond breaking and formation events that are poorly represented in most of the available MLIP training sets to date (Guan et al., 2023). We evaluate the models on $1\,\mathrm{ns}$ annealing MD simulations (2×10^6 steps with $0.5\,\mathrm{fs}$ timestep) by heating a system of hydrogen and oxygen molecules linearly from $300\,\mathrm{K}$ to $3000\,\mathrm{K}$, holding at $3000\,\mathrm{K}$, and then cooling back to $300\,\mathrm{K}$. Temperature fluctuations, number of water molecules, and enthalpy change ΔH are monitored along MD trajectories (Figure 1).

CHGNet, EquiformerV2(OC20), eSCN(OC20), and M3GNet were not able to finish $1\,\mathrm{ns}$ MD trajectories (see Figure 1). As analyzed in Figure S2, the slow runtime performance of models without built-in equivariance, such as CHGNet and M3GNet, may seem surprising since equivariant models are often more expensive to use. However, we found that molecules condense into droplets at an early stage in CHGNet and M3GNet trajectories, drastically increasing the number of bond and angle edges and therefore slowing down the MD speed.

While ORB and ORBv2 were fastest in terms of MD steps per second (Figure S2), they could not react hydrogen and oxygen at the elevated temperature and keep the number of water molecules close to zero throughout the trajectories; they also have positive reaction enthalpies, contradicting experimental measurements (Lide, 2004). Figure S2 also shows that direct force prediction models (EquiformerV2(OC20), ORB) have large center-of-mass drifts ($> 10^2$ Å) during MD simulations by six orders of magnitude more than gradient-based models. Enforcing net zero forces as implemented by ORBv2 only decreases the drift to (~ 2.4 Å), while other models keep drifts around 10^{-4} Å scales over $1\,\mathrm{ns}$ MD.

2.3 Thermodynamic properties and physical phenomena

Vacancy formation and migration energies. Defects, especially vacancies, play a key role in determining the properties of many functional materials used for photovoltaic, catalytic, thermoelectric, and optoelectronic applications (Choudhary and Sumpter, 2023; Mosquera-Lois et al., 2024). We evaluated six widely used MLIPs capable of predicting stress in elemental face-centered cubic (FCC) and hexagonal close-packed (HCP) crystals, leveraging the vacancy diffusion database by Angsten et al. (2014). The translational symmetry of crystal sites and vacancies requires that the paths and barriers for forward and backward vacancy migration be identical, making this a robust test of a model's ability to respect crystal symmetry.

Climbing image nudged elastic band (CI-NEB) calculations were performed to analyze vacancy migration barriers. The *path asymmetry* (eq. (S4)) and *barrier asymmetry* (eq. (S5)) of the migration energy profiles are analyzed in detail in Appendix A.4. We found the symmetry of NEB profiles has no strong correlation with the underlying model, and in general all models perform worse for HCP crystals.

Second-order dynamical phase transition in perovskite. Perovskites are a versatile class of materials exhibiting diverse properties, including ferroelectricity, magnetoresistance, ionic conductivity, piezoelectricity, and superconductivity. Barium zirconate ($BaZrO_3$, BZO) has been predicted and observed to have a second-order phase transition due to dynamical instability in the cubic polymorph (Fransson et al., 2023; Rosander et al., 2023). In Figure S5, we probe the anharmonic PES of different MLIPs along the octahedral-tilting phonon mode with different unit cell lattice constants. We observe Landau-like second-order phase transition from quartic to quadratic polynomials in MACE-MP(M), MatterSim, CHGNet, and SevenNet. M3GNet remains in quadratic PES across all structures with close degeneracies. ORBv2 has an asymmetrical PES and multiple energy crossings.

3 Discussion and Conclusion

We present MLIP Arena, an open benchmarking platform that avoids simplistic regression metrics susceptible to error cancellation and instead focuses on evaluating physical awareness and practical utility. Our analysis uncovers key insights: gradient-based force predictions may exhibit non-conservative behavior, larger training datasets do not always yield better performance, and equivariant models can sometimes surpass invariant models in runtime efficiency. MLIP Arena serves as

a transparent and reproducible workflow orchestrator, guiding the development of MLIPs with improved adherence to physical principles, runtime performance, and predictive capability.

4 ACKNOWLEDGMENTS

We acknowledge funding through the DOE, Office of Science, Office of Basic Energy Sciences, Materials Sciences and Engineering Division, under Contract No. DE-AC02-05-CH11231 within the Materials Project program (KC23MP). The benchmarks were developed and performed using resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award BES-ERCAP0032604. Yuan Chiang appreciates the support from Taiwan-UC Berkeley Fellowship jointly offered by Ministry of Education in Taiwan and UC Berkeley. TK was supported by the Toyota Research Institute as part of the Synthesis Advanced Research Challenge.

We thank Janosh Riebesell, Philipp Benner, Patrick Huck, Rouxi Yang, Evan Walter Clark Spotte-Smith, and Bowen Deng for early discussions; Hyunsoo Park, Yunsung Lim, Aron Walsh, and Han Yang for valuable exchanges; and Prof. Aron Walsh for his insightful suggestions and comments on the manuscript.

REFERENCES

T. Angsten, T. Mayeshiba, H. Wu, and D. Morgan. Elemental vacancy diffusion database from high-throughput first-principles calculations for fcc and hcp structures. *New Journal of Physics*, 16 (1):015018, 2014.

L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi. A foundation model for atomistic materials chemistry, Mar. 2024. URL http://arxiv.org/abs/2401.00096. arXiv:2401.00096 [cond-mat, physics:physics].

F. Bigi, M. Langer, and M. Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint arXiv:2412.11569*, 2024.

L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, et al. Open catalyst 2020 (oc20) dataset and community challenges. *Acs Catalysis*, 11(10): 6059–6072, 2021.

C. Chen and S. P. Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, Nov. 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00349-3. URL https://www.nature.com/articles/s43588-022-00349-3. Publisher: Nature Publishing Group.

K. Choudhary and B. DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

K. Choudhary and B. G. Sumpter. Can a deep-learning model make fast predictions of vacancy formation in diverse materials? *AIP Advances*, 13(9), 2023.

B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, Sept. 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00716-3. URL https://www.nature.com/articles/s42256-023-00716-3. Publisher: Nature Publishing Group.

B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson, and G. Ceder. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning. *arXiv preprint arXiv:2405.07105*, 2024.

P. Eastman, B. P. Pritchard, J. D. Chodera, and T. E. Markland. Nutmeg and spice: models and data for biomolecular machine learning. *Journal of chemical theory and computation*, 20(19): 8583–8593, 2024.

B. Focassio, L. P. M. Freitas, and G. R. Schleder. Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Applied Materials & Interfaces*, 2024.

E. Fransson, P. Rosander, P. Erhart, and G. Wahnstr"om. Understanding correlations in bazro3: Structure and dynamics on the nanoscale. *Chemistry of Materials*, 36(1):514–523, 2023.

X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, and T. Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237*, 2022.

X. Guan, J. P. Heindel, T. Ko, C. Yang, and T. Head-Gordon. Using machine learning to go beyond potential energy surface benchmarking for chemical reactivity. *Nature Computational Science*, 3 (11):965–974, 2023.

A. Hasche, A. Navid, H. Krause, and S. Eckart. Experimental and numerical assessment of the effects of hydrogen admixtures on premixed methane-oxygen flames. *Fuel*, 352:128964, 2023.

G. Henkelman, B. P. Uberuaga, and H. Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of chemical physics*, 113(22): 9901–9904, 2000.

A. D. Kaplan, R. Liu, J. Qi, T. W. Ko, B. Deng, J. Riebesell, G. Ceder, K. A. Persson, and S. P. Ong. A foundational potential energy surface dataset for materials. *arXiv preprint arXiv:2503.04070*, 2025.

Y.-L. Liao, B. Wood, A. Das, and T. Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

D. R. Lide. *CRC handbook of chemistry and physics*, volume 85. CRC press, 2004.

I. Mosquera-Lois, S. R. Kavanagh, A. M. Ganose, and A. Walsh. Machine-learning structural reconstructions for accelerated point defect calculations. *npj Computational Materials*, 10(1):121, 2024.

M. Neumann, J. Gin, B. Rhodes, S. Bennett, Z. Li, H. Choubisa, A. Hussey, and J. Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

Y. Park, J. Kim, S. Hwang, and S. Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2024.

S. Passaro and C. L. Zitnick. Reducing so (3) convolutions to so (2) for efficient equivariant gnns. In *International Conference on Machine Learning*, pages 27420–27438. PMLR, 2023.

B. Póta, P. Ahlawat, G. Csányi, and M. Simoncelli. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755*, 2024.

E. Qu and A. S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains, 2024. URL https://arxiv.org/abs/2410.24169.

J. Riebesell, R. E. A. Goodall, A. Jain, P. Benner, K. A. Persson, and A. A. Lee. Matbench Discovery – An evaluation framework for machine learning crystal stability prediction, Aug. 2023. URL http://arxiv.org/abs/2308.14920. arXiv:2308.14920 [cond-mat].

P. Rosander, E. Fransson, C. Milesi-Brault, C. Toulouse, F. Bourdarot, A. Piovano, A. Bossak, M. Guennou, and G. Wahnström. Anharmonicity of the antiferrodistortive soft mode in barium zirconate bazro 3. *Physical Review B*, 108(1):014309, 2023.

J. Schmidt, T. F. Cerqueira, A. H. Romero, A. Loew, F. Jäger, H.-C. Wang, S. Botti, and M. A. Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, page 101560, 2024.

T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, and A. C. T. van Duin. The reaxff reactive force-field: development, applications and future directions. *npj Comput. Mater.*, 2(1): 1–14, 2016.

S. Smidstrup, A. Pedersen, K. Stokbro, and H. Jónsson. Improved initial guess for minimum energy path calculations. *The Journal of chemical physics*, 140(21), 2014.

R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, et al. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

D. Wines and K. Choudhary. Chips-ff: Evaluating universal machine learning force fields for material properties. *arXiv preprint arXiv:2412.10516*, 2024.

H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

H. Yu, M. Giantomassi, G. Materzanini, J. Wang, and G.-M. Rignanese. Systematic assessment of various universal machine-learning interatomic potentials. *Materials Genome Engineering Advances*, 2(3):e58, 2024.

S. Zhu, D. Sarıtürk, and R. Arróyave. Accelerating calphad-based phase diagram predictions in complex alloys using universal machine learning potentials: Opportunities and challenges. *Acta Materialia*, page 120747, 2025.

A SUPPLEMENTARY INFORMATION

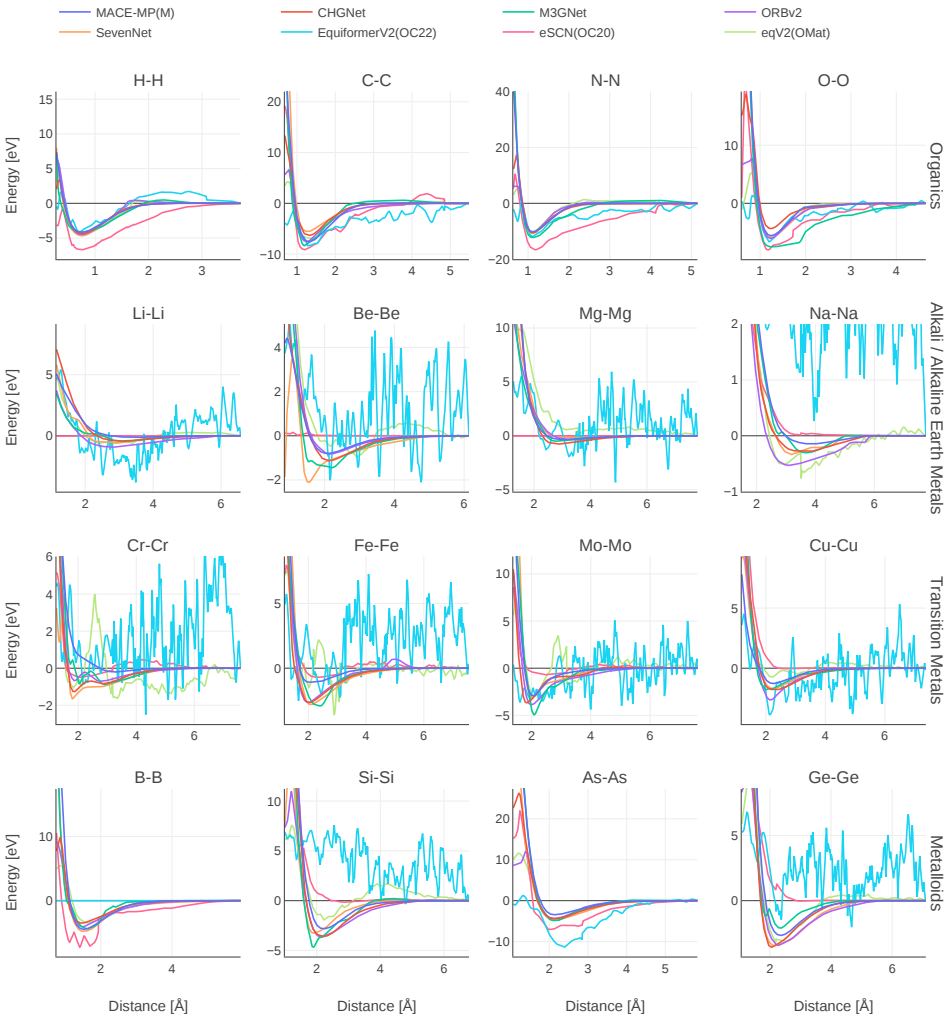

Figure S1: Potential energy curves (PECs) of selected homonuclear diatomic molecules, representing four different chemical characteristics—organics, alkali/alkaline earth metals, transition metals, and metalloids—are presented. The curves from different methods are shifted and aligned to zero at the largest separation distance.

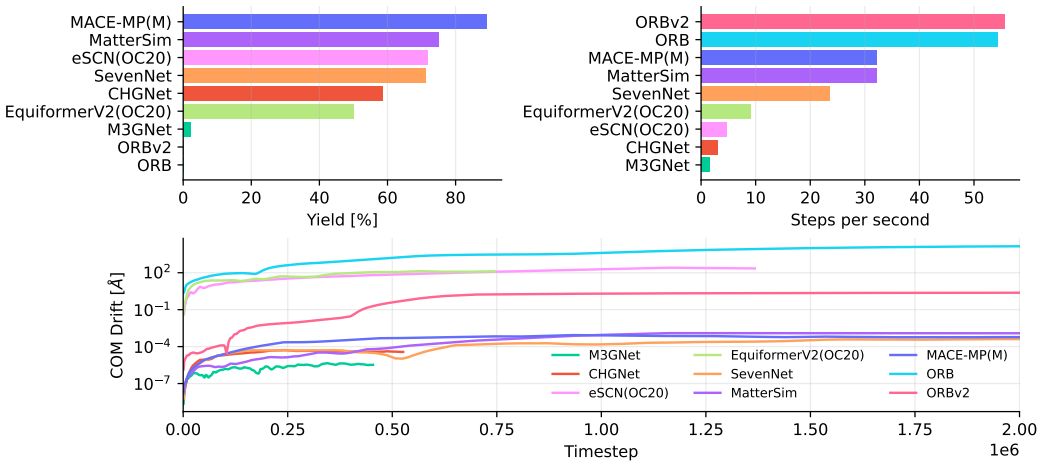

Figure S2: Hydrogen combustion. (Top left) The final reaction yield at the last MD step. (Top right) MD runtime speed measured in steps per second using single NVIDIA A100 GPU. `cuEquivariance` kernel was disabled for MACE-MP(M). (Bottom) The center-of-mass (COM) drift displacement during MD trajectory.

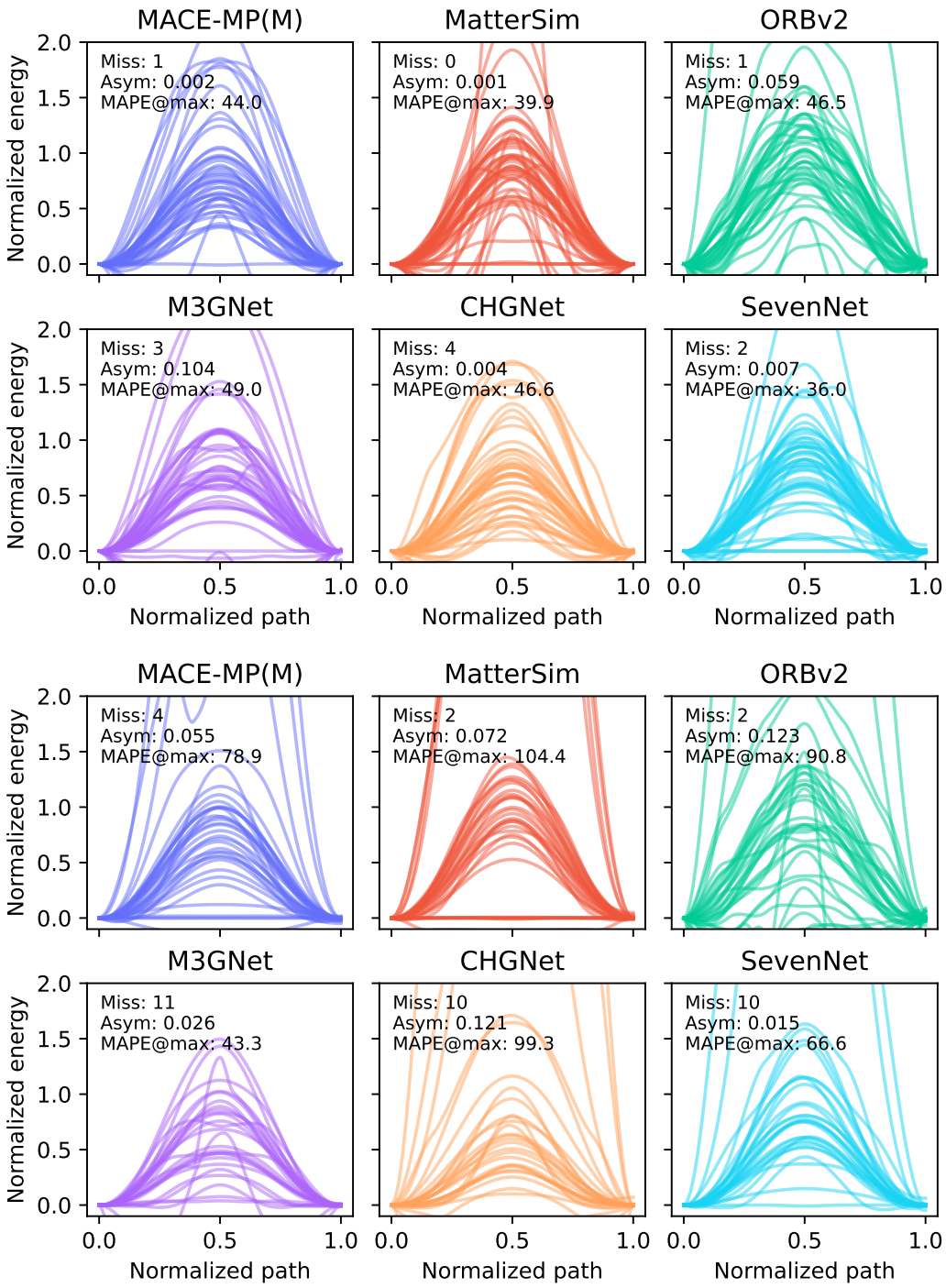

Figure S3: NEB profiles of vacancy migration in FCC (top panel) and HCP (bottom panel) elemental crystals. All path lengths are normalized to 1, and all energies are normalized by PBE vacancy migration energy barrier $E_{\text{vm}}^{\text{PBE}}$ as given in (Angsten et al., 2014). Number of missing predictions, average path asymmetry, and MAPE of maximum energy barrier are annotated on top left.

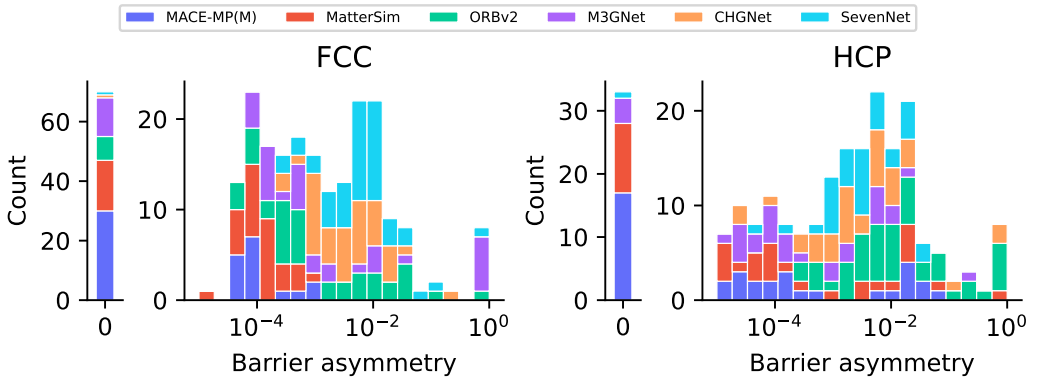

Figure S4: Asymmetry of vacancy migration barrier in FCC and HCP elemental crystals. Compliance to symmetry is not correlated with the (non-)equivariance of the underlying MLIPs. Non-equivariant MLIPs: ORBv2, MatterSim, CHGNet. Equivariant MLIPs: MACE-MP(M), SevenNet.

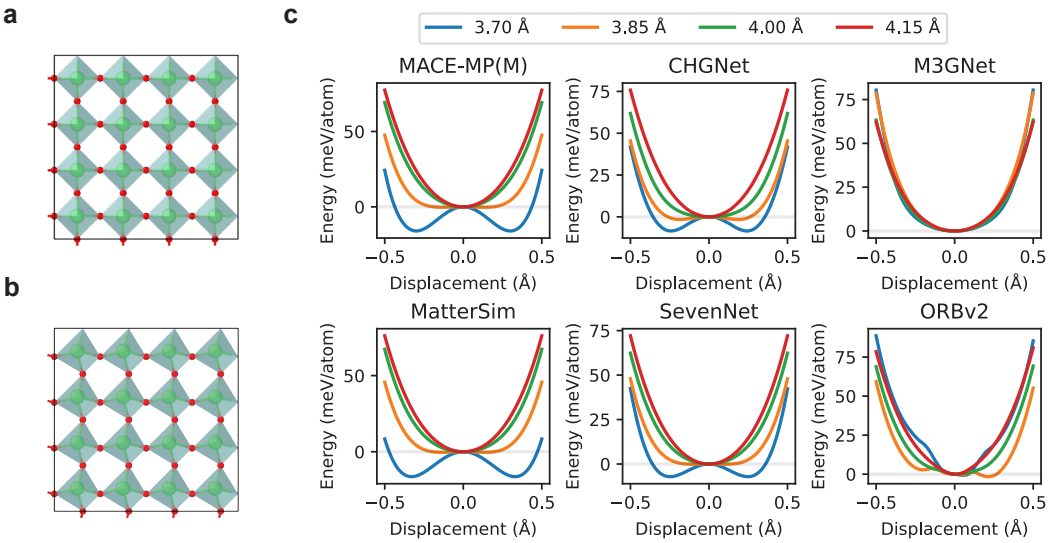

Figure S5: Landau-like second-order phase transition of octahedral-tilting mode in $BaZrO_3$ (BZO). (a) Undeformed $4 \times 4 \times 4$ supercell of BZO with cubic unit cell lattice constant of 4 Å. (b) R-tilt phonon mode with maximum displacement of 0.5 Å. Ba atoms are transparent for better visualization. (c) Transitional behavior from quadratic to quartic Landau-like potential energy landscape as a function of largest modal displacement for different lattice constants from 3.70 Å to 4.15 Å.

A.1 SUPPORTED MODELS

Table S1: List of supported open-source, open-weight models in MLIP Arena. Custom models could be incorporated through convenient class inherited from `ASE Calculator`.

Model	Prediction[1]	NVT	NPT	Training Set[2]	Code	Reference	License	Checkpoint	First Release
MACE-MP(M)	EFS	✓	✓	MPTrj	GitHub	Batatia et al. (2024)	MIT	`2023-12-03-mace-128-L1_epoch-199.model`	2023-12-29
CHGNet	EFSM	✓	✓	MPTrj	GitHub	Deng et al. (2023)	BSD-3-Clause	`v0.3.0`	2023-02-28
M3GNet	EFS	✓	✓	MPF	GitHub	Chen and Ong (2022)	BSD-3-Clause	`M3GNet-MP-2021.2.8-PES`	2022-02-05
MatterSim	EFS	✓	✓	MPTrj, Alex, Proprietary	GitHub	Yang et al. (2024)	MIT	`MatterSim-v1.0.0-5M.pth`	2024-05-10
ORB	EFS	✓	✓	MPTrj, Alex	GitHub	N/A	Apache-2.0	`orbff-v1-20240827.ckpt`	2024-09-03
ORBv2	EFS	✓	✓	MPTrj, Alex	GitHub	Neumann et al. (2024)	Apache-2.0	`orb-v2-20241011.ckpt`	2024-10-15
SevenNet	EFS	✓	✓	MPTrj	GitHub	Park et al. (2024)	GPL-3.0	`7net-0`	2024-07-11
eqV2(OMat)	EFS	✓	✗	OMat, MPTrj, Alex	GitHub	Barroso-Luque et al. (2024)	Apache-2.0*	`eqV2_86M_omat_mp_salex.pt`	2024-10-18
EquiformerV2(OC22)	EF	✓	✗	OC22	GitHub	Liao et al. (2023)	Apache-2.0	`EquiformerV2-lE4-lF100-S2EFS-OC22`	2023-06-21
EquiformerV2(OC20)	EF	✓	✗	OC20	GitHub	Liao et al. (2023)	Apache-2.0	`EquiformerV2-31M-S2EF-OC20-All+MD`	2023-06-21
eSCN(OC20)	EF	✓	✗	OC20	GitHub	Passaro and Zitnick (2023)	Apache-2.0	`eSCN-L6-M3-Lay20-S2EF-OC20-All+MD`	2023-02-07
ALIGNN	EFS	✓	✓	MP22	GitHub	Choudhary and DeCost (2021)	NIST	`2024.5.27`	2021-11-15

[1] E: energy, F: force, S: stress, M: magmom.

[2] MPTrj: Materials Project GGA-PBE relaxation trajectories, Alex: Alexandria GGA-PBE dataset (Schmidt et al., 2024), OMat: Open Materials dataset (Barroso-Luque et al., 2024), MP22: Materials Project 2022, MPF: MPF.2021.2.8: Materials Project snapshot curated to train M3GNet (Chen and Ong, 2022). OC20, OC22: Open Catalyst Project (Chanussot et al., 2021; Tran et al., 2023).

*Modified Apache-2.0 (Meta)

A.2 BENCHMARK DETAILS FOR HOMONUCLEAR DIATOMICS

Two atoms are placed inside a vacuum box and the predictions are made with separation distances ranging from 0.9 covalent radius to 3.1 van der Waals radius or to 6 Å if van der Waals radius is not available. The shortest, equilibrium, and longest separation distances are denoted as r_{min}, r_{eq}, and r_{max} respectively. The energy and force evaluations are performed at 0.01 Å interval.

Conservative field. Conservative forces are important for physical and stable MD simulations, as extra energy will be injected into or extracted from the system through non-conservative forces, degrading the stability of thermostats (Bigi et al., 2024). Some models use direct force prediction (Liao et al., 2023) or apply *post-hoc* correction (Neumann et al., 2024) to achieve better prediction errors or smaller drifts during MD simulations. Despite enhanced speed performance, these non-conservative forces may violate the law of energy conservation, undermining the stability of phonon and MD simulations and the predictive power on finite-temperature thermodynamics quantities (Póta et al., 2024). To quantify the deviation of force prediction from the conservative field, we compute the MAE between force and the central difference of energy along the homonuclear diatomic curves:

$$\text{Conservation deviation} = \left\langle \left| \mathbf{F}(\mathbf{r}) \cdot \frac{\mathbf{r}}{\|\mathbf{r}\|} + \nabla_r E \right| \right\rangle_{r=\|\mathbf{r}\|}. \tag{S1}$$

Note that this definition is only valid for diatomic interaction but a well-defined, manageable alternative for the exploding combinatorics of hetero-nuclear, many-body interactions. Many modern MLIPs have many-body forces, and more careful decomposition of many-body contributions needs to be considered for those cases.

Potential stiffness. Atoms at close distance should experience strong repulsion. Despite the inaccuracies of DFT calculations at short interatomic distances (Appendix A.3), the well-behaved classical FFs and MLIPs should reproduce strong repulsive interactions between atoms at short range distances. In fact, Deng et al. (2024) has indicated prominent softening across MLIPs trained on MPTrj, which consists of crystal relaxation trajectories close to equilibrium. Softened potentials often have early drop in energy and forces at the short range, leading to increased probability of instability. To quantify this behavior, we use *Spearman's coefficients* to evaluate the repulsiveness of energy curves (E: repulsion in Table 1) at the distance range $r \in [r_{\text{min}}, r_{\text{eq}}]$, where $r_{\text{eq}} = \underset{r \in [r_{\text{min}}, r_{\text{max}}]}{\arg\min}\ E(r)$ is taken as the equilibrium internuclear distance. Force curves (F: descending) are evaluated at the distance range between r_{min} and the distance where the largest attractive (negative) force happens.

Smoothness. The smoothness of a PEC can be heuristically estimated by *tortuosity* as the ratio between total variation in energy $\text{TV}_{r_{\text{min}}}^{r_{\text{max}}}(E)$ and the sum of absolute energy differences between shortest separation distance r_{min}, equilibrium distance r_{eq}, and longest separation distance r_{max} . This is essentially the arc-chord ratio projected in the energy dimension:

$$\text{Tortuosity} = \frac{\sum_{r_i \in [r_{\text{min}}, r_{\text{max}}]} |E(r_i) - E(r_{i+1})|}{|E(r_{\text{min}}) - E(r_{\text{eq}})| + |E(r_{\text{eq}}) - E(r_{\text{max}})|} \tag{S2}$$

. The Lennard-Jones potential and any potentials with single repulsion-attraction transition or pure repulsion have tortuosity equal to 1. Note that the true PECs of some elements may have intermediate range energy barriers and thus ideally the elemental average across the periodic table should be slightly above one. For the simplicity of this metric, we rank the models by the absolute difference with 1.

We also identify the sign changes of energy gradients on PECs to extract the *energy jump* on both sides to the neighboring sampled points, which can be written down verbatim:

$$\text{Energy jump} = \sum_{r_i \in [r_{\text{min}}, r_{\text{max}}]} |\text{sign}\left[E(r_{i+1}) - E(r_i)\right] - \text{sign}\left[E(r_i) - E(r_{i-1})\right]| \times \left(|E(r_{i+1}) - E(r_i)| + |E(r_i) - E(r_{i-1})|\right) \tag{S3}$$

. The smoother PEC has lower tortuosity and total energy jump.

A.3 INACCURACIES OF PAW DFT CALCULATIONS AT SHORT INTERATOMIC DISTANCES

Due to the classical treatment of nuclei, frozen core approximation, and smoothed core electron wavefunctions in the projected-augmented wave (PAW) DFT formalism, when two atoms are too close to each other, electron wavefunctions start to overlap and oscillate significantly. In such cases, PAW projectors and plane-wave basis set may not accurately describe core electrons and their interactions with valence electrons, leading to large inaccuracies.

A.4 BENCHMARK DETAILS FOR VACANCY MIGRATION

The benchmarking workflow included geometry optimization of pristine crystals, optimization of defective structure endpoints, and followed by climbing image nudged elastic band (CI-NEB) calculations (Henkelman et al., 2000) to identify transition states and determine vacancy migration barriers. Five intermediate images for NEB calculations were generated using the improved image-dependent pair potential (IDPP) method (Smidstrup et al., 2014).

Figure S3 presents the NEB energy profiles of vacancy migration in FCC and HCP elemental solids. HCP pathways are chosen to be on basal plane to avoid asymmetrical migrations. The *path asymmetry* (eq. (S4)) is measured by mirroring against the midpoint. We found that MACE-MP(M), MatterSim, and ORBv2 generally relax NEB more robustly than M3GNet, CHGNet, and SevenNet. MatterSim, MACE-MP(M), CHGNet, and SevenNet exhibit near-perfect mirror symmetry around the saddle point for most FCC paths, while MACE-MP(M) achieves the best balance between symmetry and robustness for HCP paths.

Figure S4 demonstrates the distribution of *barrier asymmetry* (eq. (S5)) of the vacancy migrations in elemental FCC and HCP crystals. We found that the compliance to symmetry is not strongly correlated with the equivariance and non-equivariance of the underlying MLIPs. MACE-MP(M) and MatterSim produce symmetric pathways. In contrast, ORBv2 and SevenNet tend to have asymmetric migration pathways, possibly due to more corrugated PES with multiple local minima where relaxation trajectories converge to. This might unintentionally lead to more undesirable behaviors and broken symmetries for sophisticated PES and diverse chemistry.

We define *path asymmetry* by calculating the mean difference between the left and right wings of normalized NEB profile $\epsilon(x) = \frac{E^{\text{ML}}(x)}{E^{\text{PBE}}_{\text{vm}}}$ with respect to the middle point $x = 0.5$:

$$\text{path asymmetry} = 2 \int_0^{0.5} |\epsilon(0.5 - x) - \epsilon(0.5 + x)| \, dx \tag{S4}$$

.

Barrier asymmetry is defined as the ratio of reaction energy to forward barrier height:

$$\text{barrier asymmetry} = \frac{\Delta E}{E_{\text{forward}}} = \frac{E_f - E_i}{E_{\text{TS}} - E_i} \tag{S5}$$

, where E_i, E_f are energies of initial and final endpoints, and E_{TS} is the transition state energy.

B ROTATIONAL EQUIVARIANCE EVALUATION

Equivariant MLIPs have been the standard, but recent models (Neumann et al., 2024; Qu and Krishnapriyan, 2024) show competitive performance without explicit encoded equivariance. To evaluate the ability of models to learn symmetries from data, we perform a test to quantify learned rotational equivariance. For rotation matrix $\mathbf{R}$ and atomic positions $\mathbf{r}$, we measure the cosine similarity between rotated force predictions:

$$\text{sim}(\mathbf{F}) = \frac{\mathbf{R}\mathbf{F}(\mathbf{r}) \cdot \mathbf{F}(\mathbf{R}\mathbf{r})}{\|\mathbf{R}\mathbf{F}(\mathbf{r})\|\|\mathbf{F}(\mathbf{R}\mathbf{r})\|} \tag{S6}$$

where $\mathbf{F}(\mathbf{r})$ represents the models' force predictions for atomic positions $\mathbf{r}$. Perfect equivariance would result in a cosine similarity of 1.0 regardless of the rotation angle.

We evaluate models across six rotation angles from 0° to 180°, using two different test sets: SPICEv2 (Eastman et al., 2024) and MPTrj (Deng et al., 2023). SPICEv2 is a dataset consisting of drug-like compounds up to 110 atoms in size. MPTrj consists of inorganic bulk materials. We uniformly sampled 1000 systems from each dataset from the test and validation sets. We then calculate the cosine similarity averaged over the 1000 systems and 100 random rotation axes.

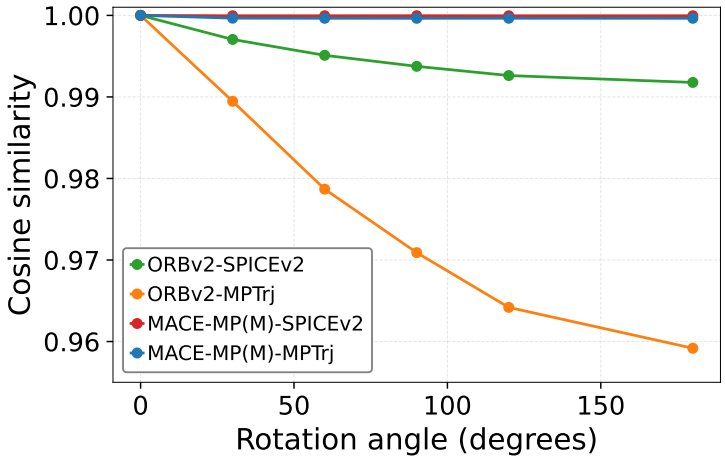

Figure S6: Cosine similarity under rotation for different models and datasets. Perfect equivariance corresponds to a constant value of 1.0.

Figure S6 shows that ORBv2 (Neumann et al., 2024) achieves strong approximate equivariance, maintaining a cosine similarity above 0.99 on SPICEv2 even at a rotation of 180°. However, performance drops to about 0.96 on MPTrj. MACE-MP(M) (Batatia et al., 2024) maintains equivariance on both datasets as expected.

C ADDITIONAL DFT REFERENCE BENCHMARKS

Bulk modulus from equation of state (EOS) calculations. In the vacancy migration task (Appendix A.4), the geometry optimization of each pristine structure is then followed by an EOS fit to compare with GGA-PBE data from Angsten et al. (2014). Figure S7 shows that most of the model can capture the trend up to 400 GPa well, with serious underestimation on a few FCC and several HCP structures.

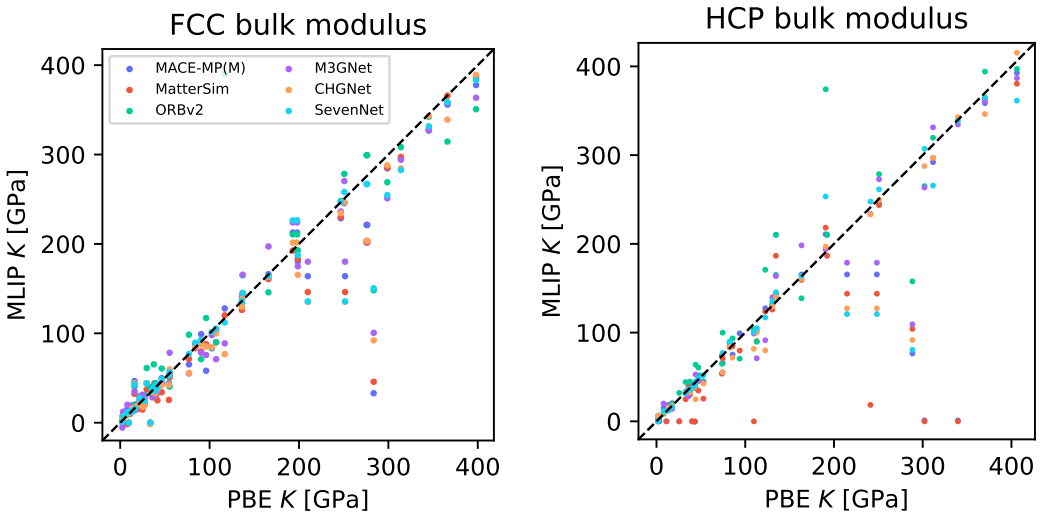

Figure S7: Bulk modulus of FCC and HCP elemental solids compared with GGA-PBE calculations (Angsten et al., 2014).

Table S2: Bulk modulus of FCC elemental crystals. nNA denotes the number of missing predictions out of 57 entries except for noble gases.

model	MAE (GPa)	MAPE	nNA
MACE-MP(M)	18.878	0.287	2
MatterSim	19.142	0.281	1
ORBv2	32.583	0.315	1
M3GNet	21.867	0.370	4
CHGNet	19.815	0.255	6
SevenNet	14.500	0.211	3

Table S3: Bulk modulus of HCP elemental crystals. nNA denotes the number of missing predictions out of 57 entries except for noble gases.

model	MAE (GPa)	MAPE	nNA
MACE-MP(M)	35.969	0.363	5
MatterSim	45.865	0.355	5
ORBv2	41.116	0.364	4
M3GNet	21.321	0.220	16
CHGNet	21.484	0.263	16
SevenNet	21.925	0.170	15