MLIP Arena: Advancing Fairness and Transparency in Machine Learning Interatomic Potentials via an Open, Accessible Benchmark Platform

Yuan Chiang[1,2,*] **Tobias Kreiman**[1] **Christine Zhang**[1] **Matthew C. Kuner**[1,2]
Elizabeth Weaver[1] **Ishan Amin**[1] **Hyunsoo Park**[3] **Yunsung Lim**[4]
Jihan Kim[4] **Daryl Chrzan**[1,2] **Aron Walsh**[3] **Samuel M. Blau**[2]
Mark Asta[1,2] **Aditi S. Krishnapriyan**[1,2,*]

[1]UC Berkeley [2]LBNL [3]Imperial College London [4]KAIST

{cyrusyc,aditik1}@berkeley.edu

Abstract

Machine learning interatomic potentials (MLIPs) have revolutionized molecular and materials modeling, but existing benchmarks suffer from data leakage, limited transferability, and an over-reliance on error-based metrics tied to specific density functional theory (DFT) references. We introduce MLIP Arena, a benchmark platform that evaluates force field performance based on physics awareness, chemical reactivity, stability under extreme conditions, and predictive capabilities for thermodynamic properties and physical phenomena. By moving beyond static DFT references and revealing the important failure modes of current foundation MLIPs in real-world settings, MLIP Arena provides a reproducible framework to guide the next-generation MLIP development toward improved predictive accuracy and runtime efficiency while maintaining physical consistency. The Python package and online leaderboard are available at `https://github.com/atomind-ai/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)—the high-dimensional landscape that maps atomic configurations to their corresponding energies and forces—across a wide range of chemical compositions at a fraction of the computational cost of direct DFT evaluations.

Despite excelling in error-based metrics for bulk systems [1], MLIPs trained on the DFT total energy and interatomic forces do not necessarily capture the correct dynamic interactions of atomistic systems [2]. Analogously, classical force fields [3] fit to describe near-equilibrium radial distribution functions cannot capture the energetics of bond-breaking. These limitations may also extend to MLIPs predominantly trained on near- or on-equilibrium configurations. In particular, energy and force regression metrics based on near-equilibirum structures may not reflect performance in downstream scientific tasks. We highlight some specific limitations below.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

First, energy and force regression metrics are vulnerable to data leakage, failing to accurately assess a model's extrapolation and generalization capabilities. This issue is evident in Matbench Discovery [1], where non-compliant models rank highly for crystal stability metrics due to energy overfitting at the expense of forces and finite-temperature capabilities. This may result in poor generalization to structures more diverse in chemistry and away from the energy convex hull. 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 [4–7]. 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. [8] 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 benchmarking platform for foundation MLIPs. This platform evaluates both the quality of the learned PES and the extent to which models respect the physical laws and symmetries essential to atomistic modeling. Unlike previous error-based DFT reference benchmarks [1, 9–12], **MLIP Arena focuses on assessing physical soundness to better evaluate the utility of MLIPs for downstream applications**. By moving beyond error-centric evaluations, it provides more actionable insights for model development and training. Specifically, we assess how well foundation MLIPs capture physics-informed phenomena, their reliability for accurate atomistic simulations, and their readiness for practical scientific research and discovery. The MLIPs evaluated in this work are listed in Table S4.

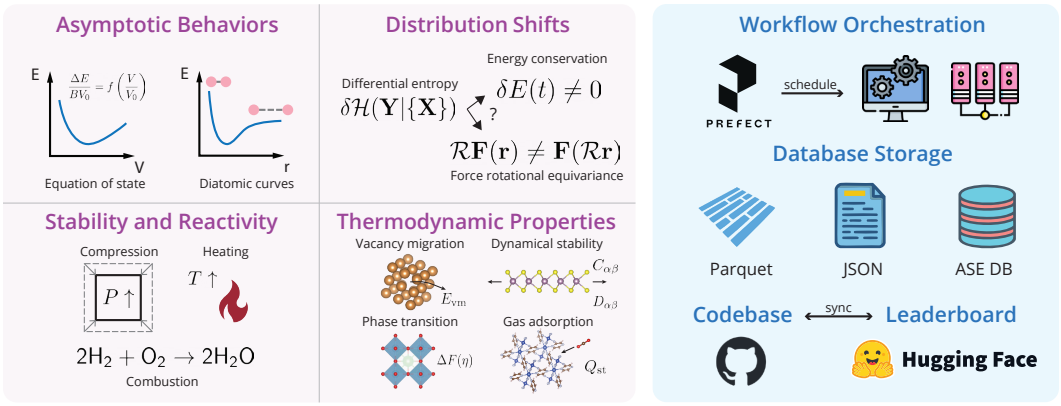


Figure 1: Overview of MLIP Arena. Four benchmark categories beyond error-based regression metrics provide actionable insights agnostic to underlying model architecture and DFT reference. Tasks are defined as Prefect (`https://www.prefect.io/`) workflows to enable advanced task caching, chaining, and parallel/concurrent execution on HPC. Atomic simulation environment (ASE) [13] calculator and database are used. Codebase (`https://github.com/atomind-ai/mlip-arena`) and online leaderboard on Hugging Face Space (`https://huggingface.co/spaces/atomind/mlip-arena`) are available.

2 MLIP Arena benchmarks

MLIP Arena assesses the limitations of MLIPs through four primary perspectives. In Section 2.1, we focus on the asymptotic behaviors of MLIP predictions and propose metrics that enable robust and well-balanced ranking from multi-rank aggregation, reducing susceptibility to overfitting on any single metric. Section 2.2 tests MLIP robustness and reactivity under extreme conditions using molecular dynamics (MD) simulations, exposing their instabilities and unphysical behaviors. Section 2.3 investigate the robustness of MLIP to scenarios with quantified distribution shifts to higher

uncertainty. Section 2.4 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 Asymptotic analyses on off-equilibrium conditions

Asymptotic Behaviors: The benchmarks evaluate the asymptotic behavior of MLIP predictions on the equation of state (EOS) of stable crystals derived from WBM structures [14] and on the potential energy curves (PECs) of homonuclear diatomics across the periodic table. The quality of prediction is quantified using physical and geometric measures of PECs (including derivative flips, tortuosity, and Spearman's coefficient), assessed in terms of deviations from physically correct values agnostic to DFT references.

Robust MLIPs should predict reasonable asymptotic behaviors of an atomic system under extreme conditions and symmetry transformations. We specifically focus on the metrics agnostic to underlying DFT functional the model has been trained on and propose physical and geometric measures to assess the general performance of MLIPs. A new suite of metrics to reflect the important aspects of MLIPs for atomistic modeling beyond regression errors is proposed as follows.

2.1.1 Metrics

Smoothness. The major utility of modern MLIPs is the accurate approximation of DFT PES. High-quality PES should be smooth since DFT, as the ground-state electronic structure theory, predicts smooth PES under the assumption of adiabatic approximation on the lowest-energy Born-Oppenheimer surface. Common training objectives on the energy and force of bulk crystals near equilibrium are subject to many-body error cancellation and do not guarantee smoothness. To quantify this effect, we propose *tortuosity* (eq. (S2)), *energy jumps* (eq. (S3)), and *force/gradient flips* to measure the quality of PES.

Tortuosity measures the arc-chord ratio of potential energy curves (PECs, one dimensional slice of PES) projected in the energy dimension. Smooth PECs with a single equilibrium point, like the Lennard-Jones pairwise PEC, 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. The number of force/gradient flips count the times force/gradient changes sign along the slice.

Short-range repulsion. Atoms at close distances should experience strong repulsion. We use *Spearman's coefficients* to measure the monotonicity of PECs at short interatomic distances or under high compression. Robust MLIPs should have Spearman's coefficients of energy and force close to -1 when approaching the repulsive regime. This metric detects the short-range PES *holes*. The absence of these PES holes and the reasonable repulsion are important for the correct samplings of thermodynamic ensembles essential for correct long-time dynamics and physical property calculations.

Conservative field. Conservative forces are important for energy conserving molecular simulations, and non-conservative forces are known to degrade the stability of thermostats [15]. We calculate the *conservation deviation* as the MAE between force and the central difference approximation of the derivative of the energy along the 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.

2.1.2 Results

The Birch–Murnaghan equation of state (EOS) (eq. (S4)) [16, 17] describes the relationship between the energy and volume of crystalline solids under external pressure and has been computed at scale for materials in the Materials Project [18]. The detailed EOS curves for a set of representative models—each consisting of 21 sampled points evaluated after ionic relaxation at fixed volume (21,000 ionic relaxation trajectories per model)—are visualized in Figure 2. In Table 1, we present the corresponding metrics and their aggregated rankings to assess the quality of the predicted EOS across diverse crystal structures. Both Figure 2 and Table 1 show that most models exhibit the expected

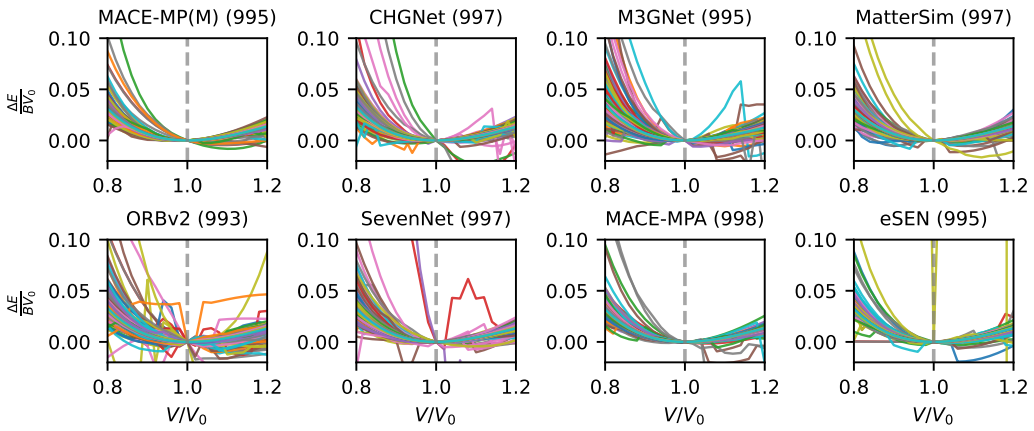


Figure 2: EOS benchmark on 1,000 WBM structures [14]. The reduced relative energy, $\frac{\Delta E}{BV_0}$, is normalized by the bulk modulus B and equilibrium volume V_0 through a rearrangement of the Birch–Murnaghan EOS (eq. (S5)). Color indicates the EOS curve of each crystal structure. The number of valid predictions for each model is shown after the model name.

Table 1: Equation of state (EOS) benchmark on 1,000 WBM structures [14]. **Boldface** and underline indicate the **best** and worst metrics across all MLIPs, respectively. Standard deviations are given in parentheses. Derivative flips are ranked by their absolute deviation from 1. For up-to-date models and aggregated rankings, see the online leaderboard.

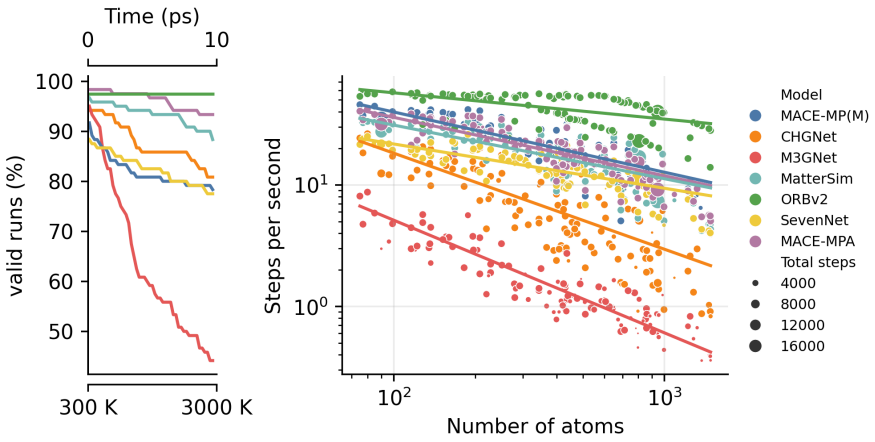
Model	Derivative flips ↓	Tortuosity ↓	Spearman's coefficient			Missing ↓
			E: compression ↓	$\frac{dE}{dV}$: compression ↑	E: tension ↑	
MACE-MPA	**1.037 (0.283)**	**1.005 (0.054)**	**-0.999368 (0.012)**	0.996332 (0.039)	**0.993186 (0.077)**	**2**
eSEN	1.042 (0.314)	1.008 (0.090)	-0.999330 (0.012)	**0.996857 (0.037)**	0.992097 (0.073)	5
MACE-MP(M)	1.042 (0.345)	1.009 (0.129)	-0.999330 (0.011)	0.994116 (0.059)	0.991586 (0.088)	5
MatterSim	1.045 (0.376)	1.006 (0.055)	-0.997350 (0.039)	0.992790 (0.078)	0.988098 (0.115)	3
CHGNet	1.105 (0.540)	1.015 (0.123)	-0.996499 (0.051)	0.992997 (0.052)	0.986642 (0.117)	3
SevenNet	1.109 (0.555)	1.019 (0.275)	-0.998128 (0.026)	0.988912 (0.077)	0.985958 (0.117)	3
M3GNet	1.175 (0.676)	1.018 (0.149)	-0.996321 (0.052)	0.989743 (0.065)	0.980169 (0.133)	5
ORBv2	1.316 (0.870)	1.037 (0.215)	-0.991846 (0.082)	0.970143 (0.132)	0.963746 (0.198)	7

concave-up behavior for the majority of structures, although some models display characteristic failure modes, including short-range holes, shifted energy minima, and spurious spikes.

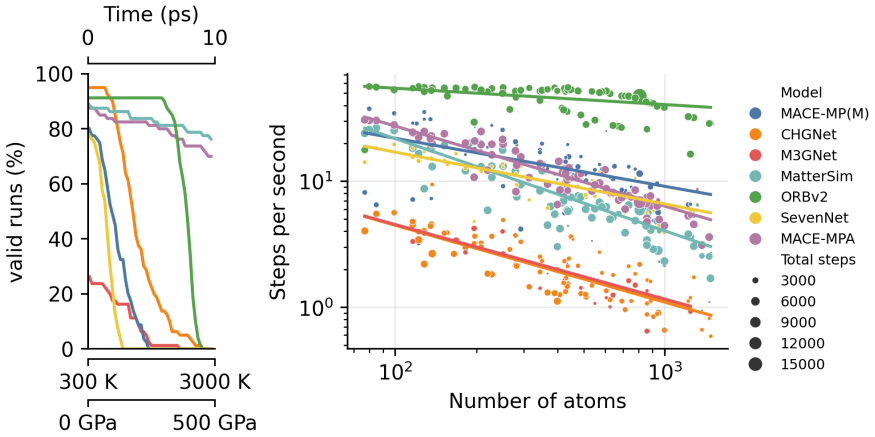
We further perform energy–volume scans under more extreme volumetric strains, ranging from $-49\,\%$ to $75\,\%$, but *without* initial structure or ionic relaxation; that is, the fractional coordinates of ions remain fixed after deformation. See appendix A.4 for further analysis. To evaluate whether the models truly capture the underlying interactions, we also analyze the potential energy curves (PECs) of homonuclear diatomics, with interatomic distances spanning 0.9 times the covalent radius r_{cov} to 3.1 times the van der Waals radius r_{vdw} across the entire periodic table. This range approximately covers the equilibrium covalent bond length ($2r_{\text{cov}}$) and the decay of dispersion interactions (see appendix A.2 for details). Interestingly, many top-ranked models on bulk crystal EOS and Matbench Discovery perform poorly in pairwise interactions, suggesting that apparent benchmark success may result from plausible many-body error cancellation in bulk systems. See Table S1 for rankings and Figure S1 for select PECs in appendix A.2. We encourage reader to visit our interactive leaderboard for complete set of all elements.

2.2 Stability and reactivity from molecular dynamics simulation

Stability and Reactivity: Isochoric–isothermal (NVT) molecular dynamics (MD) simulations with a temperature ramp from $300\,\text{K}$ to $3000\,\text{K}$ over $10\,\text{ps}$, and isobaric–isothermal (NPT) MD simulations from $300\,\text{K}$ at $0\,\text{GPa}$ to $3000\,\text{K}$ at $500\,\text{GPa}$ over $10\,\text{ps}$, are performed on



(a) 120 NVT MD simulations from 300 K to 3000 K.



(b) 80 NPT MD simulations from 300 K to 3000 K and 0 GPa to 500 GPa.

Figure 3: MD stability on RM24 structures. For NVT (a), we perform Nosé-Hoover thermostats with linearly increasing temperature from 300 K to 3000 K. The number of valid trajectories and the scaling of MD steps per second (SPS) with the number of atoms N are shown. For NPT (b), Nosé-Hoover thermostats is performed with an additional pressure ramp from 0 GPa to 500 GPa. The size of each point represents the valid steps along each valid trajectory. The power law SPS $= aN^b$ is used to determine the asymptotic performance of MLIPs (solid line). First 120 structures from RM24 are used for NVT, and first 80 structures are used for NPT. The target length of each trajectory is 10 ps. `cuEquivariance` kernel was disabled for MACE family models.

random mixture structures (RM24). Reactivity is exemplified by annealing MD simulations of hydrogen combustion. Benchmarking metrics include the fraction of valid runs and runtime speed performance (MD steps per second on a single A100 GPU).

Stable and accurate MD simulations are essential for atomistic modeling. As their name suggests, MLIPs should serve as reliable interatomic potentials for running MD simulations. We benchmark MLIPs for stability under extreme temperature and/or pressure conditions and record their runtime performance using the random amorphous mixture structure database RM24 (appendix A.5).

Stability under extreme conditions. We perform MD simulations on RM24 structures with a linear temperature schedule from 300 K to 3000 K for 10 ps using Nosé-Hoover NVT thermostats [19]. The number of valid runs and asymptotic speed scaling with the system size are presented in Figure 3a. Because many MLIPs exhibit short-range holes or require increasingly large neighbor lists under high

pressure, we additionally apply a linearly increasing pressure from $0\,\mathrm{GPa}$ to $500\,\mathrm{GPa}$ over $10\,\mathrm{ps}$ using Nosé–Hoover NPT barostats (fig. 3b). An MD step is considered valid if the atomic structure exists and has finite energy prediction. This is a relatively lenient criterion when treating MLIPs purely as autoregressive samplers, as it does not account for thermodynamic drifts, fluctuations, or potential structural instabilities. Our benchmark challenges the common belief that equivariant models such as MACE and SevenNet are generally slower than non-equivariant models such as CHGNet and M3GNet. In reality, model architecture, engineering optimizations, and checkpoint quality all contribute to overall MD runtime performance, while the speed and stability of MD trajectories also depend on the chemical system. This is illustrated by the fact that MatterSim shares the same architecture as M3GNet but is significantly more stable and performant. ORBv2 is the fastest and has the best scaling exponent among the tested MLIPs in both heating and compression simulations. However, without an explicit short-range core repulsive potential built in, ORBv2 and many earlier MLIPs could not sustain high-pressure conditions up to $500\,\mathrm{GPa}$.

Chemical reactivity. Classical force fields are periled by the inaccurate description of chemical reactions. While a bouquet of reactive force fields [3] has been parametrized to mitigate this limitation, they have shown limited transferability from one system to another. Although MLIPs hold the promise to bypass the limitation, one should not assume the reactivity to be guaranteed from pretraining. As an example to test the reactivity, we perform annealing MD simulation to emulate hydrogen combustion. Hydrogen combustion is a challenging out-of-distribution (OOD) test since there are multiple bond breaking and formation events that are poorly represented in most of the available MLIP training sets to date [20]. We evaluate the select 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 (fig. S3). Our results show that the model reactivity is uncorrelated with the prediction accuracy on bulk crystals. See appendix A.7 for detailed comparison between models.

2.3 Robustness to distribution shifts

Distribution shifts: The benchmarks assess violations of energy conservation in MLIP MD trajectories and of rotational equivariance in static force predictions under input distribution shifts, characterized by the differential entropy of atomic local environments. Energy drifts are monitored over eight MD trajectory windows to evaluate conservation, while rotation-induced force errors are computed and averaged within bins defined by differential entropy.

While model architectures that strictly adhere to known symmetries and physical laws have been the standard, recent models [21–23] have shown competitive performance with non-conservative and non-equivariant force predictions. While models with fewer constraints adhere to symmetries well in-distribution [21, 23], it is important to understand how these models generalize to out-of-distribution systems when considering them for practical use [24]. To this end, we propose an evaluation to measure robustness to symmetries in the face of out-of-distribution structures.

Measuring distribution shifts with differential entropy. To quantify how far a system is from the training distribution, we compute the differential entropy $\delta\mathcal{H}$ for each structure with respect to the training distribution, as the implemented in the QUESTS descriptors proposed by Schwalbe-Koda et al. [25] (see appendix A.9 for details). The differential entropy provides a measure of uncertainty or "surprise" for one to probe how current MLIPs maintain energy conservation and rotational equivariance in the face of distribution shifts.

Energy conservation. We perform $5\,\mathrm{ps}$ NVE simulations with a $1\,\mathrm{fs}$ time step, initializing atomic velocities from a Maxwell-Boltzmann distribution at 1000 K. Simulations are conducted on random subsets of each model's training set. Differential entropy is computed for structures along the simulation trajectories using a sliding window approach. For each $500\,\mathrm{fs}$ window, the differential entropy of the midpoint structure is calculated with respect to MPTrj, and the energy difference between the start and end of the window is recorded.

Figure 4 shows that direct force prediction models such as ORB and Equiformer demonstrate a significant correlation between higher differential entropy and greater energy deviation, indicating

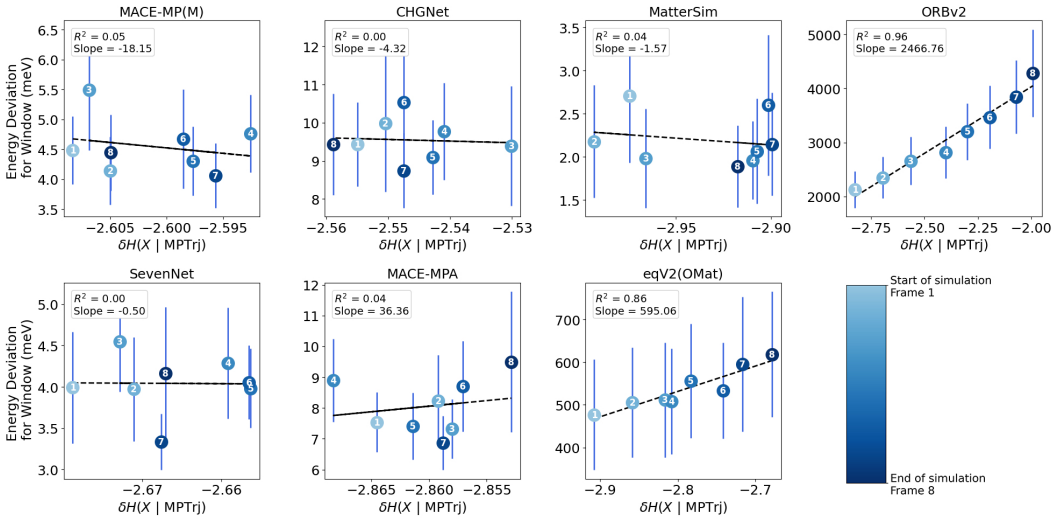


Figure 4: Energy conservation under distribution shift. Energy deviation is calculated for each sliding window during NVE MD simulations for 5 ps. Differential entropy of the structure in the middle of the window is calculated, and the energy deviation from the start to the end of the window is recorded. We report 95% confidence interval error bars and a line of best fit. The order in which windows appear during the simulation is annotated by the number on each point. For direct force prediction models, the simulated trajectories become increasingly surprising over time, as shown by the monotonically increasing numbers from left to right.

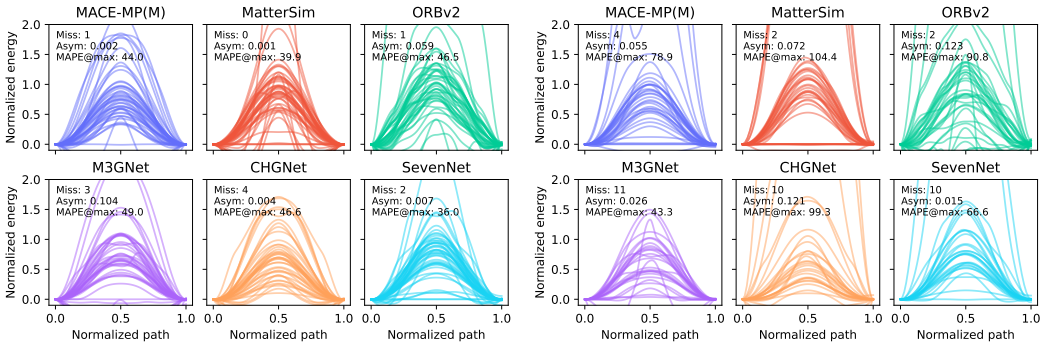
that non-conservative models tend to (increasingly) violate energy conservation on structures that are surprising in their training set. We also find that direct force prediction models reach more surprising regions of phase space over the course of simulation, indicated by the increasing window numbers as the differential entropy increases in Figure 4. However, gradient-based force prediction models show little correlation between differential entropy and energy conservation ability. Unlike non-conservative models, gradient-based models do not show increasing surprise as the simulation progresses.

Force rotational equivariance. To evaluate the ability of models to learn rotational symmetries from data, we perform a test to quantify learned rotational equivariance. For a rotation matrix $\mathbf{R}$ and atomic positions as $\mathbf{r}$, we measure the MAE between rotated force predictions $\text{MAE}(\mathbf{F}) = \frac{1}{3}\sum_{i=1}^{3} |\mathbf{R}\mathbf{F}(\mathbf{r})_i - \mathbf{F}(\mathbf{R}\mathbf{r})_i|$, where $\mathbf{F}(\mathbf{r})$ represents the models' force predictions for atomic positions $\mathbf{r}$. Perfect equivariance would result in a MAE of 0.0 regardless of the rotation angle.

We evaluate models across a random subset of MPTrj [26], a dataset that consists of inorganic bulk materials. We uniformly sample 500 systems and their trajectories from the dataset. We then calculate the force MAE per frame averaged over 10 random rotation axes and 5 angles from 30° to 180°. Figure S6 shows that the non-rotationally equivariant ORB and ORBv2 models [23] exhibit strong correlation between greater differential entropy and higher rotational force MAE. This indicates that while current non-equivariant architectures can adhere to rotational equivariance on in-distribution structures, they may struggle to maintain symmetries for OOD structures with diverse orientations. The rest of the models, which have rotational equivariance built explicitly into the architecture [27–30], achieve perfect rotational equivariance, as expected.

2.4 Thermodynamic properties and phenomenological studies

Thermodynamic Properties: This section provides various benchmarks relevant for downstream property applications and phenomenological studies: vacancy formation and migration from nudged elastic band calculations [31], CO_2 adsorption for metal-organic frameworks



(a) FCC elemental crystals

(b) HCP elemental crystals

Figure 5: NEB profiles of vacancy migration in FCC (a) and HCP (b) 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 [31]. Number of missing predictions, average path asymmetry, and MAPE of maximum energy barrier are annotated on top left.

[32], second-order phase transition in perovskite [33], and dynamical stability screening of 2D materials from C2DB database [34].

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 [35, 36]. 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. [31]. 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. We define *path asymmetry* (eq. (S8)) and *barrier asymmetry* (eq. (S9)) of the migration energy profiles in Appendix A.8. We found the symmetry of NEB profiles has no strong correlation with built-in equivariance or not, and in general all models perform worse for HCP crystals. Figure 5 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. 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.

Extended case studies. In Appendix A.10, we further assess the downstream utility of MLIPs for three extended case studies: CO_2 adsorption in metal-organic frameworks (MOFs) (appendix A.10.1), dynamical stability of 2D materials (appendix A.10.2), and second-order phase transition in perovskite (appendix A.10.3). Each study exposes the certain weakness of modern MLIPs. We found possibly due to poorly described non-bonded interaction between CO_2 molecule and MOFs, many MLIPs deviate at a large degree from experimental adsorption energies and may not be informative enough for MOF virtual screening (appendix A.10.1). Our results on the Landau-like second-order phase transition in $BaZrO_3$ (appendix A.10.3) uncover subtle failure modes (energy degeneracy and asymmetrical PES) of octahedral tilts predicted by M3GNet and ORBv2. These limitations could cause the models unable to reproduce the correct transition behaviors important for exotic functional properties such as superconductivity [33, 37]. Furthermore, despite the recent saturation in the prediction of thermodynamic stability and lattice thermal conductivity of 3D bulk crystals, our results on the dynamical stability of 2D materials (appendix A.10.2) indicate that the discovery rate of stable 2D materials remains poor (highest macro F1 score of 0.420 and 0.412 by MACE family models) and there is still a large performance gap in 2D materials space that the success in 3D materials may not ostensibly translate over.

3 Related Work

Static DFT reference benchmarks. Benchmarking of MLIPs has largely centered around static DFT datasets such as QM9 [38], ANI-1 [39], MD17 [40], SPICE [6, 41], MPTrj [26], GMTKN55 [42], and more. While these have enabled rapid progress, they are tied to specific level of electronic structure theory and the data are generally incompatible with one another. Over time, the community has expanded benchmark domains, but the dependency on static DFT references remains [1, 9–12, 43, 44]. The Matbench test suite introduced by Dunn et al. [43] compiles 13 tasks (e.g., formation energies, band gaps, elastic moduli) drawn largely from DFT-computed data. Matbench Discovery [1] leveraged the WBM database [14] as an extension beyond the MP [45] for crystal stability classification. Some other benchmarks rely on specific DFT reference while comparing models trained on incompatible dataset [12, 44, 46, 47]. While targeting to higher level of theory is desirable, the higher-level of theory however may not be equally transferable (*e.g.* coupled-cluster theory describes metallic solids poorly [48]), leading to misleading, non-cross-comparable assessments.

Many models now saturate these test metrics, yet fail to extrapolate to unseen chemistry, strained configurations, or finite-temperature behavior. MLIP Arena complements these efforts by introducing physically grounded tasks that probe model robustness beyond interpolation to a static reference.

Risk of regression error metrics. Standard metrics like MAE and RMSE are known to poorly reflect real-world MLIP utility. Bigi et al. [15], Fu et al. [49] show that models with low force errors may fail to conserve energy in MD, while Póta et al. [8], Loew et al. [50] demonstrate that good regression error metrics on energy and forces does not ensure accurate phonons or thermal transport. Direct force models often violate energy conservation due to the lack of a consistent potential energy surface [15, 51]. These inconsistencies could be due to a *misalignment* between energy and force pre-training objectives and physical properties of interest [52]. Moreover, average errors can obscure large failures in rare but critical configurations. MLIP Arena addresses these gaps by incorporating task-specific, granular evaluations to provide a more faithful measure of model reliability.

4 Discussion and Conclusion

Limitations. Traditionally, MLIP training and benchmarks rely on DFT references. This is a computationally cheap way to evaluate models since it only requires a few single-point predictions from the ML model, as opposed to autoregressive benchmarks (*i.e.* MD simulations). We acknowledge that moving away from DFT references makes it harder to directly compare models, at least on in-distribution data present in standard test datasets. However, the central promise of foundation MLIPs is to generalize to OOD systems and phases—in which case accuracy with respect to in-distribution DFT data becomes auxiliary rather than primary.

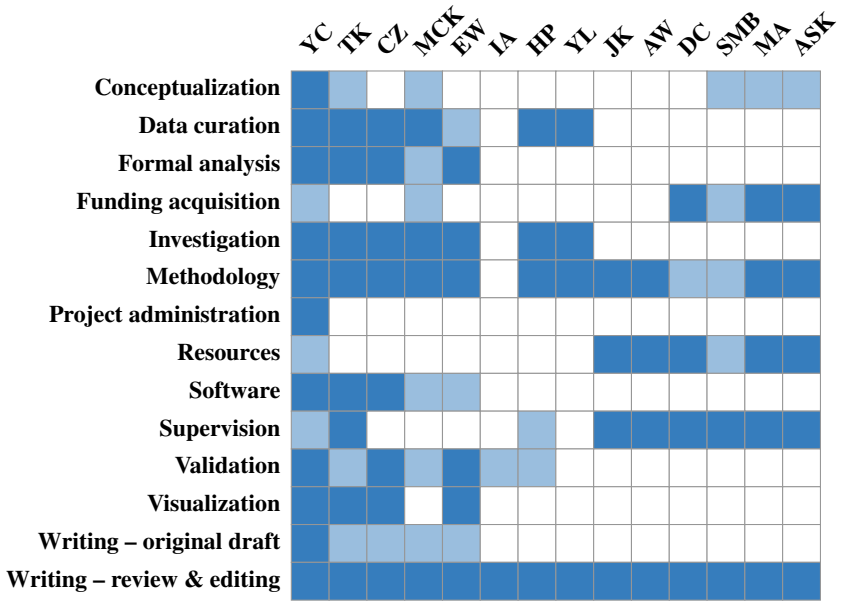
Analogously, many other areas of machine learning have moved beyond standard regression metrics [53]. Large language models, for example, are increasingly evaluated on practical, task-oriented performance rather than raw perplexity on training-like data. In the same spirit, MLIP Arena is our first attempt to prioritize qualities that are essential to atomistic modeling beyond any DFT reference: symmetry, conservation laws, reasonable asymptotic behaviors and thermodynamic properties that any interatomic potential should satisfy for practical utility.

Opportunities. Reference-agnostic benchmarks like MLIP Arena could motivate and guide new directions in model development and training that explicitly tackle generalization and downstream utility, particularly through reinforcement learning [54], implicit differentiation [55], and test-time training [24] approaches. MLIP Arena provides reproducible workflows that can be scaled for high-throughput reward data generation across a broad range of practically relevant OOD tasks, facilitating the exploration of these training paradigms.

In summary, 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 some new insights: gradient-based force predictions may exhibit non-conservative behavior; alignment between training dataset size and better model performance is not always guaranteed but depends on design choice; and current MLIPs have not saturated in reactivity and robustness under distribution shifts. 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.

5 Author Contribution Statement



6 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. YC received 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. MCK was supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-2146752. SMB was supported by the Energy Storage Research Alliance "ESRA" (DE-AC02-06CH11357), an Energy Innovation Hub funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences.

We thank Aaron Kaplan for advice and comments on the manuscript, and Janosh Riebesell, Philipp Benner, Patrick Huck, Rouxi Yang, Evan Walter Clark Spotte-Smith, and Bowen Deng for early discussions; and Jan Janssen, Rhys Goodall, Abhijeet Gangan, and Han Yang for valuable exchanges.

References

[1] Janosh Riebesell, Rhys EA Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A Lee, Anubhav Jain, and Kristin A Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025.

[2] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237*, 2022.

[3] Thomas P Senftle, Sungwook Hong, Md Mahbubul Islam, Sudhir B Kylasa, Yuanxia Zheng, Yun Kyung Shin, Chad Junkermeier, Roman Engel-Herbert, Michael J Janik, Hasan Metin

Aktulga, Toon Verstraelen, Ananth Grama, and Adri C T van Duin. The reaxff reactive force-field: development, applications and future directions. *npj Comput. Mater.*, 2(1):1–14, 2016.

[4] 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*, page 101560, 2024.

[5] Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

[6] Peter Eastman, Benjamin P Pritchard, John D Chodera, and Thomas E Markland. Nutmeg and spice: models and data for biomolecular machine learning. *Journal of chemical theory and computation*, 20(19):8583–8593, 2024.

[7] Aaron D Kaplan, Runze Liu, Ji Qi, Tsz Wai Ko, Bowen Deng, Janosh Riebesell, Gerbrand Ceder, Kristin A Persson, and Shyue Ping Ong. A foundational potential energy surface dataset for materials. *arXiv preprint arXiv:2503.04070*, 2025.

[8] Balázs Póta, Paramvir Ahlawat, Gábor Csányi, and Michele Simoncelli. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755*, 2024.

[9] Haochen Yu, Matteo Giantomassi, Giuliana Materzanini, Junjie Wang, and Gian-Marco Rignanese. Systematic assessment of various universal machine-learning interatomic potentials. *Materials Genome Engineering Advances*, 2(3):e58, 2024.

[10] Bruno Focassio, Luis Paulo M. Freitas, and Gabriel R Schleder. Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Applied Materials & Interfaces*, 2024.

[11] Siya Zhu, Doğuhan Sarıtürk, and Raymundo Arróyave. Accelerating calphad-based phase diagram predictions in complex alloys using universal machine learning potentials: Opportunities and challenges. *Acta Materialia*, page 120747, 2025.

[12] Daniel Wines and Kamal Choudhary. Chips-ff: Evaluating universal machine learning force fields for material properties. *arXiv preprint arXiv:2412.10516*, 2024.

[13] 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.

[14] Hai-Chen Wang, Silvana Botti, and Miguel A. L. Marques. Predicting stable crystalline compounds using chemical similarity. 7(1):1–9. ISSN 2057-3960. doi: 10.1038/s41524-020-00481-6. URL `https://www.nature.com/articles/s41524-020-00481-6`. Publisher: Nature Publishing Group.

[15] Filippo Bigi, Marcel Langer, and Michele Ceriotti. The dark side of the forces: assessing non-conservative force models for atomistic machine learning. *arXiv preprint arXiv:2412.11569*, 2024.

[16] Francis Dominic Murnaghan. The compressibility of media under extreme pressures. *Proceedings of the National Academy of Sciences*, 30(9):244–247, 1944.

[17] Francis Birch. Finite elastic strain of cubic crystals. *Physical review*, 71(11):809, 1947.

[18] Katherine Latimer, Shyam Dwaraknath, Kiran Mathew, Donald Winston, and Kristin A Persson. Evaluation of thermodynamic equations of state across chemistry and structure in the materials project. *npj Computational Materials*, 4(1):40, 2018.

[19] Denis J Evans and Brad Lee Holian. The nose-hoover thermostat. *Journal of Chemical Physics*, 83(8):4069–4074, 1985.

[20] Xingyi Guan, Joseph P Heindel, Taehee Ko, Chao Yang, and Teresa Head-Gordon. Using machine learning to go beyond potential energy surface benchmarking for chemical reactivity. *Nature Computational Science*, 3(11):965–974, 2023.

[21] Eric Qu and Aditi S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Y4mBaZu4vy.

[22] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. *arXiv preprint arXiv:2306.12059*, 2023.

[23] Mark Neumann, James Gin, Benjamin Rhodes, Steven Bennett, Zhiyi Li, Hitarth Choubisa, Arthur Hussey, and Jonathan Godwin. Orb: A fast, scalable neural network potential. *arXiv preprint arXiv:2410.22570*, 2024.

[24] Tobias Kreiman and Aditi S Krishnapriyan. Understanding and mitigating distribution shifts for machine learning force fields. *arXiv preprint arXiv:2503.08674*, 2025.

[25] Daniel Schwalbe-Koda, Sebastien Hamel, Babak Sadigh, Fei Zhou, and Vincenzo Lordi. Model-free quantification of completeness, uncertainties, and outliers in atomistic machine learning using information theory. *arXiv:2404.12367*, 2024. URL https://arxiv.org/abs/2404.12367.

[26] 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, September 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.

[27] 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, March 2024. URL http://arxiv.org/abs/2401.00096. arXiv:2401.00096 [cond-mat, physics:physics].

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

[29] Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

[30] Yutack Park, Jaesun Kim, Seungwoo Hwang, and Seungwu Han. Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 2024.

[31] Thomas Angsten, Tam Mayeshiba, Henry Wu, and Dane Morgan. Elemental vacancy diffusion database from high-throughput first-principles calculations for fcc and hcp structures. *New Journal of Physics*, 16(1):015018, 2014.

[32] Yunsung Lim, Hyunsoo Park, Aron Walsh, and Jihan Kim. Accelerating co2 direct air capture screening for metal-organic frameworks with a transferable machine learning force field. 2024.

[33] Erik Fransson, Petter Rosander, Paul Erhart, and G"oran Wahnstr"om. Understanding correlations in bazro3: Structure and dynamics on the nanoscale. *Chemistry of Materials*, 36(1): 514–523, 2023.

[34] Morten Niklas Gjerding, Alireza Taghizadeh, Asbjørn Rasmussen, Sajid Ali, Fabian Bertoldo, Thorsten Deilmann, Nikolaj Rørbæk Knøsgaard, Mads Kruse, Ask Hjorth Larsen, Simone Manti, et al. Recent progress of the computational 2d materials database (c2db). *2D Materials*, 8(4):044002, 2021.

[35] Irea Mosquera-Lois, Seán R Kavanagh, Alex M Ganose, and Aron Walsh. Machine-learning structural reconstructions for accelerated point defect calculations. *npj Computational Materials*, 10(1):121, 2024.

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

[37] Petter Rosander, Erik Fransson, Cosme Milesi-Brault, Constance Toulouse, Frédéric Bourdarot, Andrea Piovano, Alexei Bossak, Mael Guennou, and Göran Wahnström. Anharmonicity of the antiferrodistortive soft mode in barium zirconate bazro 3. *Physical Review B*, 108(1):014309, 2023.

[38] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.

[39] Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific data*, 4(1):1–8, 2017.

[40] Stefan Chmiela, Alexandre Tkatchenko, Huziel E Sauceda, Igor Poltavsky, Kristof T Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.

[41] Peter Eastman, Pavan Kumar Behara, David L Dotson, Raimondas Galvelis, John E Herr, Josh T Horton, Yuezhi Mao, John D Chodera, Benjamin P Pritchard, Yuanqing Wang, et al. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023.

[42] Lars Goerigk, Andreas Hansen, Christoph Bauer, Stephan Ehrlich, Asim Najibi, and Stefan Grimme. A look at the density functional theory zoo with the advanced gmtkn55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Physical Chemistry Chemical Physics*, 19(48):32184–32215, 2017.

[43] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138, 2020.

[44] Anyang Peng, Chun Cai, Mingyu Guo, Duo Zhang, Chengqian Zhang, Antoine Loew, Linfeng Zhang, and Han Wang. Lambench: A benchmark for large atomic models. *arXiv preprint arXiv:2504.19578*, 2025.

[45] 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, July 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL `https://doi.org/10.1063/1.4812323`.

[46] Rebecca R Brew, Ian A Nelson, Meruyert Binayeva, Amlan S Nayak, Wyatt J Simmons, Joseph J Gair, and Corin C Wagen. Wiggle150: Benchmarking density functionals and neural network potentials on highly strained conformers. *Journal of Chemical Theory and Computation*, 2025.

[47] Ari Wagen. Nnp arena, 2025. URL `https://ariwagen.com/nnp-arena`.

[48] Nikolaos Masios, Andreas Irmler, Tobias Schäfer, and Andreas Grüneis. Averting the infrared catastrophe in the gold standard of quantum chemistry. *Physical Review Letters*, 131(18): 186401, 2023.

[49] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations, August 2023. URL `http://arxiv.org/abs/2210.07237`. arXiv:2210.07237 [physics].

[50] Antoine Loew, Dewen Sun, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Universal machine learning interatomic potentials are ready for phonons. *arXiv preprint arXiv:2412.16551*, 2024.

[51] Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, and C Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. *arXiv preprint arXiv:2502.12147*, 2025.

[52] Yunsheng Liu, Xingfeng He, and Yifei Mo. Discrepancies and error evaluation metrics for machine learning interatomic potentials. *npj Computational Materials*, 9(1):174, 2023.

[53] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL `https://arxiv.org/abs/2403.04132`.

[54] Aditya Koneru, Henry Chan, Sukriti Manna, Troy D Loeffler, Debdas Dhabal, Andressa A Bertolazzo, Valeria Molinero, and Subramanian KRS Sankaranarayanan. Multi-reward reinforcement learning based development of inter-atomic potential models for silica. *npj Computational Materials*, 9(1):125, 2023.

[55] Sanjeev Raja, Ishan Amin, Fabian Pedregosa, and Aditi S Krishnapriyan. Stability-aware training of machine learning force fields with differentiable boltzmann estimators. *arXiv preprint arXiv:2402.13984*, 2024.

[56] Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A Persson, and Gerbrand Ceder. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning. *arXiv preprint arXiv:2405.07105*, 2024.

[57] Leandro Martínez, Ricardo Andrade, Ernesto G Birgin, and José Mario Martínez. Packmol: A package for building initial configurations for molecular dynamics simulations. *Journal of computational chemistry*, 30(13):2157–2164, 2009.

[58] Yuan Chiang. muse, December 2023. URL `https://github.com/chiang-yuan/muse`.

[59] James F Ziegler and Jochen P Biersack. The stopping and range of ions in matter. In *Treatise on heavy-ion science: volume 6: astrophysics, chemistry, and condensed matter*, pages 93–129. Springer, 1985.

[60] David R Lide. *CRC handbook of chemistry and physics*, volume 85. CRC press, 2004.

[61] Anna Hasche, Ali Navid, Hartmut Krause, and Sven Eckart. Experimental and numerical assessment of the effects of hydrogen admixtures on premixed methane-oxygen flames. *Fuel*, 352:128964, 2023.

[62] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[63] Glenn J Martyna, Mark E Tuckerman, Douglas J Tobias, and Michael L Klein. Explicit reversible integrators for extended systems dynamics. *Molecular Physics*, 87(5):1117–1157, 1996.

[64] Graeme Henkelman, Blas P Uberuaga, and Hannes 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.

[65] Søren Smidstrup, Andreas Pedersen, Kurt Stokbro, and Hannes Jónsson. Improved initial guess for minimum energy path calculations. *The Journal of chemical physics*, 140(21), 2014.

[66] Eloy S Sanz-Pérez, Christopher R Murdock, Stephanie A Didas, and Christopher W Jones. Direct capture of co2 from ambient air. *Chemical reviews*, 116(19):11840–11876, 2016.

[67] Osama Shekhah, Youssef Belmabkhout, Zhijie Chen, Vincent Guillerm, Amy Cairns, Karim Adil, and Mohamed Eddaoudi. Made-to-order metal-organic frameworks for trace carbon dioxide removal and air capture. *Nature communications*, 5(1):4228, 2014.

[68] Soumya Mukherjee, Nivedita Sikdar, Daniel O'Nolan, Douglas M Franz, Victoria Gascón, Amrit Kumar, Naveen Kumar, Hayley S Scott, David G Madden, Paul E Kruger, et al. Trace co2 capture by an ultramicroporous physisorbent with low water affinity. *Science advances*, 5 (11):eaax9171, 2019.

[69] Woo Ram Lee, Sang Yeon Hwang, Dae Won Ryu, Kwang Soo Lim, Sang Soo Han, Dohyun Moon, Jungkyu Choi, and Chang Seop Hong. Diamine-functionalized metal–organic framework: exceptionally high co 2 capacities from ambient air and flue gas, ultrafast co 2 uptake rate, and adsorption mechanism. *Energy & Environmental Science*, 7(2):744–751, 2014.

[70] Caitlin E Bien, Kai K Chen, Szu-Chia Chien, Benjamin R Reiner, Li-Chiang Lin, Casey R Wade, and WS Winston Ho. Bioinspired metal–organic framework for trace co2 capture. *Journal of the American Chemical Society*, 140(40):12662–12666, 2018.

[71] Shuvo Jit Datta, Chutharat Khumnoon, Zhen Hao Lee, Won Kyung Moon, Son Docao, Thanh Huu Nguyen, In Chul Hwang, Dohyun Moon, Peter Oleynikov, Osamu Terasaki, et al. Co2 capture from humid flue gases and humid atmosphere using a microporous coppersilicate. *Science*, 350(6258):302–306, 2015.

[72] Jian-Bin Lin, Tai TT Nguyen, Ramanathan Vaidhyanathan, Jake Burner, Jared M Taylor, Hana Durekova, Farid Akhtar, Roger K Mah, Omid Ghaffari-Nik, Stefan Marx, et al. A scalable metal-organic framework as a durable physisorbent for carbon dioxide capture. *Science*, 374 (6574):1464–1469, 2021.

[73] Peter G Boyd, Arunraj Chidambaram, Enrique García-Díez, Christopher P Ireland, Thomas D Daff, Richard Bounds, Andrzej Gładysiak, Pascal Schouwink, Seyed Mohamad Moosavi, M Mercedes Maroto-Valer, et al. Data-driven design of metal–organic frameworks for wet flue gas co2 capture. *Nature*, 576(7786):253–256, 2019.

[74] Alessio Masala, Jenny G Vitillo, Francesca Bonino, Maela Manzoli, Carlos A Grande, and Silvia Bordiga. New insights into utsa-16. *Physical Chemistry Chemical Physics*, 18(1):220–227, 2016.

[75] Omid T Qazvini and Shane G Telfer. Muf-16: A robust metal–organic framework for pre- and post-combustion carbon dioxide capture. *ACS applied materials & interfaces*, 13(10): 12141–12148, 2021.

[76] Bingbing Chen, Dong Fan, Rosana V Pinto, Iurii Dovgaliuk, Shyamapada Nandi, Debanjan Chakraborty, Nuria García-Moncada, Alexandre Vimont, Charles J McMonagle, Marta Bordonhos, et al. A scalable robust microporous al-mof for post-combustion carbon capture. *Advanced Science*, 11(21):2401070, 2024.

[77] Rachel C Rohde, Kurtis M Carsch, Matthew N Dods, Henry ZH Jiang, Alexandra R McIsaac, Ryan A Klein, Hyunchul Kwon, Sarah L Karstens, Yang Wang, Adrian J Huang, et al. High-temperature carbon dioxide capture in a porous material with terminal zinc hydride sites. *Science*, 386(6723):814–819, 2024.

[78] Wendy L Queen, Matthew R Hudson, Eric D Bloch, Jarad A Mason, Miguel I Gonzalez, Jason S Lee, David Gygi, Joshua D Howe, Kyuho Lee, Tamim A Darwish, et al. Comprehensive study of carbon dioxide adsorption in the metal–organic frameworks m 2 (dobdc)(m= mg, mn, fe, co, ni, cu, zn). *Chemical Science*, 5(12):4569–4581, 2014.

[79] Prashant Mishra, Hari Prasad Uppara, Bishnupada Mandal, and Sasidhar Gumma. Adsorption and separation of carbon dioxide using mil-53 (al) metal-organic framework. *Industrial & Engineering Chemistry Research*, 53(51):19747–19753, 2014.

[80] Lukás Grajciar, Andrew D Wiersum, Philip L Llewellyn, Jong-San Chang, and Petr Nachtigall. Understanding co2 adsorption in cubtc mof: comparing combined dft–ab initio calculations with microcalorimetry experiments. *The Journal of Physical Chemistry C*, 115(36):17925–17933, 2011.

[81] Jason M Simmons, Hui Wu, Wei Zhou, and Taner Yildirim. Carbon capture in metal–organic frameworks—a comparative study. *Energy & Environmental Science*, 4(6):2177–2185, 2011.

[82] Hussein Rasool Abid, Huyong Tian, Ha-Ming Ang, Moses O Tade, Craig E Buckley, and Shaobin Wang. Nanosize zr-metal organic framework (uio-66) for hydrogen and carbon dioxide storage. *Chemical Engineering Journal*, 187:415–420, 2012.

[83] Zhijuan Zhang, Shikai Xian, Qibin Xia, Haihui Wang, Zhong Li, and Jing Li. Enhancement of co2 adsorption and co2/n2 selectivity on zif-8 via postsynthetic modification. *AIChE Journal*, 59(6):2195–2206, 2013.

[84] Jarad A Mason, Kenji Sumida, Zoey R Herm, Rajamani Krishna, and Jeffrey R Long. Evaluating metal–organic frameworks for post-combustion carbon dioxide capture via temperature swing adsorption. *Energy & Environmental Science*, 4(8):3030–3040, 2011.

[85] Ben Widom. Some topics in the theory of fluids. *The Journal of Chemical Physics*, 39(11): 2808–2812, 1963.

[86] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry*, 32(7):1456–1465, 2011.

[87] Sten Haastrup, Mikkel Strange, Mohnish Pandey, Thorsten Deilmann, Per S Schmidt, Nicki F Hinsche, Morten N Gjerding, Daniele Torelli, Peter M Larsen, Anders C Riis-Jensen, et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials*, 5(4):042002, 2018.

[88] Jimmy-Xuan Shen and Joel Varley. pymatgen-analysis-defects: A python package for analyzing point defects in crystalline materials. *Journal of Open Source Software*, 9(93):5941, 2024.

[89] Atsushi Togo. First-principles phonon calculations with phonopy and phono3py. *Journal of the Physical Society of Japan*, 92(1):012001, 2023.

[90] Atsushi Togo, Laurent Chaput, Terumasa Tadano, and Isao Tanaka. Implementation strategies in phonopy and phono3py. *Journal of Physics: Condensed Matter*, 35(35):353001, 2023.

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

[92] Duo Zhang, Xinzijian Liu, Xiangyu Zhang, Chengqian Zhang, Chun Cai, Hangrui Bi, Yiming Du, Xuejian Qin, Anyang Peng, Jiameng Huang, et al. Dpa-2: a large atomic model as a multi-task learner. *npj Computational Materials*, 10(1):293, 2024.

[93] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021.

[94] 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.

[95] 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.

A Supplementary Information

A.1 Details on metrics

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 [15]. Some models use direct force prediction [22] or apply *post-hoc* correction [23] 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 [8]. 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:

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

The forces are projected onto the direction of interatomic vectors. Note that this definition is only valid for diatomic interaction but a well-defined, manageable alternative for the exploding combinatorics of hetero-nuclfgear, 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.

Short-range 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 force fields and MLIPs should reproduce strong repulsive interactions between atoms at short range distances. In fact, Deng et al. [56] 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 S1) 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 $\underset{r \in [r_{\text{min}}, r_{\text{max}}]}{\arg\min}\ F(r)$ where the largest attractive (the most 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:

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

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

A.2 Homonuclear 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 S1, we compute six physical and geometric measures to rank the homonuclear PECs of MLIPs in three aspects: conservative field, short-range stiffness, and smoothness, as we discuss in the pervious subsection (appendix A.1).

Two atoms are placed inside a vacuum box and the predictions are made with separation distances ranging from 0.9 covalent radius r_{cov} to 3.1 van der Waals radius r_{vdw} or to 6 Å if van der Waals radius is not available. The range of interatomic distances is chosen heuristically by the fact that the equilibrium bond length is about the sum of covalent radii (for homonuclear diatomics this is $2r_{\text{cov}}$) and the interatomic energy and forces plateau around $2r_{\text{vdw}}$. We increase the distance range by the factor of 50% of $2r_{\text{cov}}$ and $2r_{\text{vdw}}$ and further extend both ends by 10% of radii. The shortest, equilibrium, and longest separation distances are denoted as $r_{\min}$, r_{eq}, and $r_{\max}$ respectively. **Both energy and force curves are performed** at 0.01 Å interval for dense samplings.

Table S1: 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. Select PECs are shown in Figure S1. Detailed definitions and implementation details are available in Appendix A.1.

Model	Conservation deviation [eV/Å]	Spearman's coefficient E: repulsion	Spearman's coefficient F: descending	Energy jump [eV]	Force flips	Tortuosity
MACE-MPA	0.077	**-0.997**	-0.975	0.010	1.371	**1.006**
MACE-MP(M)	0.070	**-0.997**	-0.980	0.038	1.449	1.161
MatterSim	**0.013**	-0.980	-0.972	**0.008**	2.766	1.021
M3GNet	0.026	-0.991	-0.947	0.029	3.528	1.016
ORBv2	9.751	-0.883	**-0.988**	0.991	**0.991**	1.287
eSCN(OC20)	2.045	-0.939	-0.984	0.806	0.640	5.335
CHGNet	1.066	-0.992	-0.925	0.291	2.255	2.279
ORB	10.220	-0.881	-0.954	1.019	1.026	1.798
SevenNet	34.005	-0.986	-0.928	0.392	2.112	1.292
eqV2(OMat)	15.477	-0.880	-0.976	4.118	3.126	2.515
eSEN	1.170	-0.692	-0.919	5.562	4.000	1.838
ALIGNN	5.164	-0.913	-0.310	9.876	30.669	1.818
EquiformerV2(OC20)	21.385	-0.680	-0.891	38.282	22.775	8.669
EquiformerV2(OC22)	27.687	-0.415	-0.855	64.837	21.674	15.880

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 Equation of state and energy-volume scan

Structure selection. Structures were selected from the WBM dataset [14] with a slight bias to adjust for the elemental imbalance in the original paper. That is, each structure was assigned a

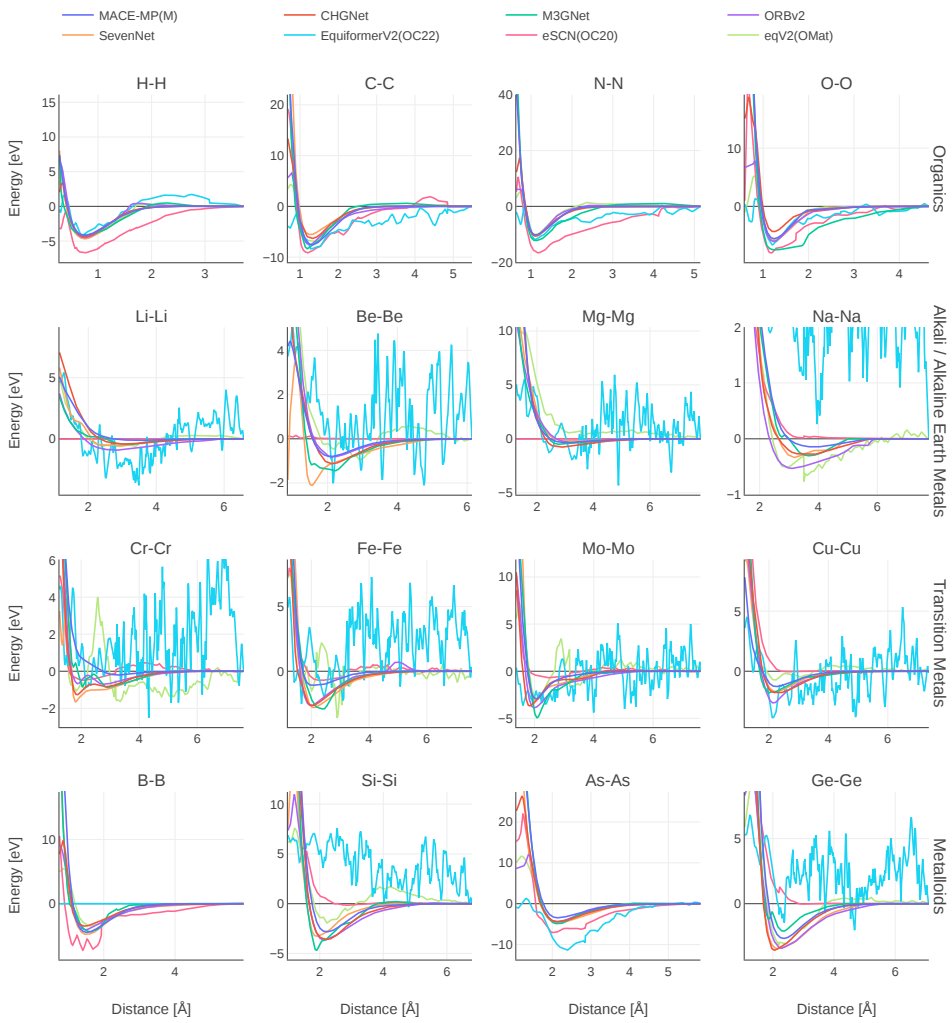


Figure S1: Potential energy curves (PECs) of selected homonuclear diatomic molecules, representing four different chemical characteristics—organics, alkali/alkaline earth metals, transition metals, and metalloids—are presented. The curves from different methods are shifted and aligned to zero at the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

probability for selection based on the prevalence of the elements it contains relative to the overall distribution of elements in the dataset. Elements with lower prevalence in the original dataset were assigned a higher probability of selection; then, 1000 structures from WBM were selected according to these assigned probabilities.

Equation of state. The EOS benchmark protocol in Arena includes first unconstrained structure optimization at 0 K and subsequently multiple energy calculations of isotropic deformations, including ionic relaxation at volumetric strain ranging from $-20\,\%$ to $20\,\%$ of the optimized structure. After ionic relaxation of 21 deformed structures for each crystal, Birch–Murnaghan EOS is fitted with the following equation:

$$E = E_0 + \frac{9BV_0}{16}\left[\left(\eta^2-1\right)^2\left(6+B'\left(\eta^2-1\right)-4\eta^2\right)\right], \quad \eta = \left(\frac{V}{V_0}\right)^{\frac{1}{3}}, \tag{S4}$$

where V_0 is the equilibrium volume after initial structure optimization, and B and B' are the bulk modulus and its pressure derivative from the EOS fit. We calculate the reduced relative energy in

Figure 2 by rearranging Equation (S4) as:

$$\frac{\Delta E}{BV_0} = \frac{E - E_0}{BV_0} = \frac{9}{16}\left[\left(\eta^2 - 1\right)^2 \left(6 + B'\left(\eta^2 - 1\right) - 4\eta^2\right)\right] \tag{S5}$$

Energy-volume scan. MLIPs should provide reasonable predictions at extreme deformations. In this benchmark, we take 1,000 structures selected from the WBM dataset [14] and uniformly deform them by $\pm 20\%$ along each *lattice* vector (i.e. from 0.51 to 1.73 of the initial volume). The energy of each deformed structure is evaluated *without* relaxation—preventing relaxation to another crystal system. Table S2 presents five metrics to evalaute the performance of MLIPs on the energy-volume scan benchmark. Unlike EOS benchmark in Section 2.1.2, all of the select MLIPs have no missing predictions. Our result shows the saturation of all five metrics for top-ranked models.

In comparison with EOS benchmark, energy-volume scan evaluates the orthogonal performance of MLIPs. Birch–Murnaghan EOS is theoretically defined properties based on finite elastic theory under isothermal condition. EOS benchmark evaluates both model's capability to search for energy minima under relaxation protocol, thus mixing the consequences of energy minimum location and relaxation trajectory. Energy-volume scan tests more extreme condition and especially exposes short-range PES holes, penalizing the models with known corrugated short-range PES where DFT however may not converge as well. Energy-volume scan is also based on the assumption that the WBM structures are at local energy minima, but further investigation reveals that this assumption does not hold universally. There are a few structures with shifted energy local minima consistent across different MLIPs.

Expected behavior under compression. When subjected to significant compression, crystalline materials are expected to exhibit strong short-range repulsion. We evaluate this behavior using the Spearman's rank correlation coefficient to quantify the monotonic increase in energy with decreasing volume. Additionally, the energy derivative $\frac{dE}{dV}$ is expected to steepen in the high-compression regime. MLIPs that are physically consistent should yield Spearman's coefficients approaching -1 in the compressive region of the energy–volume curve.

Expected behavior under tension. Under tensile strain, the system's energy should also increase monotonically as atomic bonds are progressively stretched. However, as the crystal approaches dissociation into isolated atoms, the slope of the energy–volume curve ($\frac{dE}{dV}$) is expected to flatten. Thus, we evaluate only the monotonicity of the energy increase under tension using Spearman's coefficients. Reliable MLIPs should produce coefficients close to $+1$ in the tensile regime.

Table S2: Energy-volume scan of 1,000 WBM structures [14]. **Boldface** and underline represent the **best** and the worst metrics across all MLIPs, respectively.

Model	Derivative flips ↓	Tortuosity ↓	Spearman's coefficient		
			E: compression ↓	$\frac{dE}{dV}$: compression ↑	E: tension ↑
eSEN	**1.000000**	**1.000403**	**-0.998339**	**1.000000**	0.999045
MACE-MPA	**1.000000**	1.000676	**-0.998339**	0.999309	0.998718
CHGNet	**1.000000**	1.000629	-0.998279	0.943964	**0.999091**
MatterSim	1.009000	1.000567	-0.998097	0.999709	0.993754
eqV2(OMat)	1.035000	1.000835	-0.998206	0.997224	0.998645
M3GNet	1.002000	1.002001	-0.997588	0.997442	0.996468
ORBv2	1.058000	1.004065	-0.997770	0.970752	0.997600
SevenNet	1.034000	1.010025	-0.995164	0.946558	0.994705
MACE-MP(M)	1.121000	1.080713	-0.943806	0.901188	0.998745
ALIGNN	3.909000	1.375652	-0.889207	0.760271	0.862085

A.5 Random mixture dataset (RM24)

New materials are often found by reacting two stable materials into a single phase. In the conceptually similar procedure, we generate 1,000 random mixture structures at arbitrary ratio of two stable materials from Materials Project (v2024.12.18). As the binary and ternary compounds are already well covered by MP, we aim at higher component systems of up to six elements from the mixture of binary and ternary systems. All stable binary and ternary materials are first retrieved, totaling 24,430

structures. The number of possible 2-combination is >298M. We randomly selected 1,000 pairs from possible combinations and generate the initial structures using Packmol [57] and Muse [58] to consider periodic boundary conditions. Each generated structure is then relaxed via FIRE optimizer and NVT MD simulation at 1500 K for 10 ps with Ziegler-Biersack-Littmark (ZBL) screened nuclear repulsion potential [59]. The final element count distribution of 1,000 structures is presented in Figure S2. The ASE DB file is available at `https://huggingface.co/datasets/atomind/mlip-arena`.

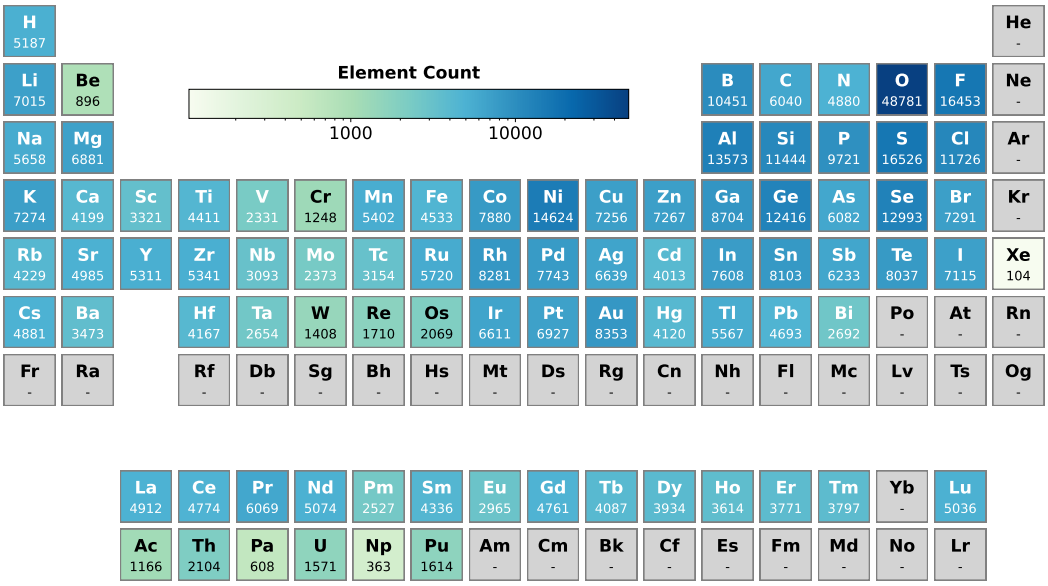


Figure S2: Element counts of random mixture dataset (RM24).

A.6 MD stability

We performed Nosé-Hoover thermostat and barostat on RM24 structures with linear scheduling of temperature from 300 K to 3000 K and/or pressure from 0 GPa to 500 GPa across 10 ps MD. The number of valid runs and asymptotic speed scaling with the system size are presented in Figure 3a. Using Prefect (https://github.com/PrefectHQ/prefect) utility, we ensure each run has access to the same resource of 1 AMD EPYC 7763 (Milan) CPU core and 1 NVIDIA A100 (Ampere) GPU. Each run has two retries, with timeout of 600 elapsed seconds for each retry. In section 2.2, the frames are marked as invalid if the simulation cannot reach the timestep or have non-numerical energy values.

A.7 Hydrogen combustion

CHGNet, EquiformerV2(OC20), eSCN(OC20), and M3GNet were not able to finish 1 ns MD trajectories (see Figure S3). As analyzed in Figure S4, 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 (fig. S4), 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 [60]. Figure S4 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 ns MD.

Here one should note that enforcing net zero forces does not guarantee zero center-of-mass (COM) drift during MD simulations under thermostats. For canonical ensemble like Nosé Hoover thermostats

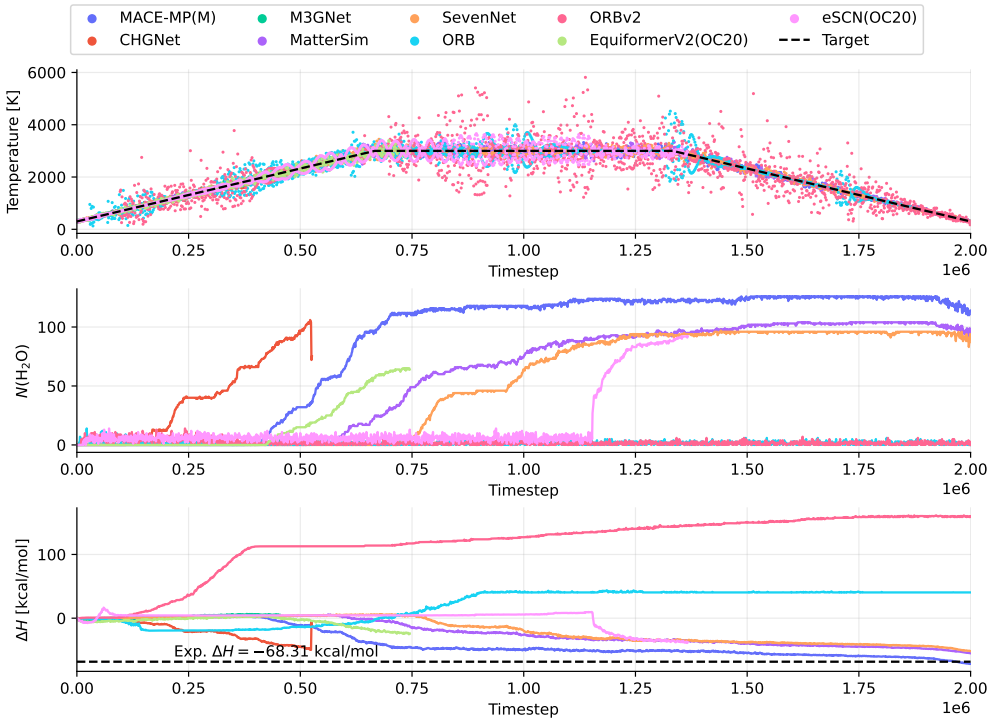


Figure S3: Hydrogen combustion via annealing NVT MD simulation ($128\,H_2 + 64\,O_2 \longrightarrow 128\,H_2O$). Applied temperature schedule is illustrated in the top panel. The experimental reaction enthalpy of $-68.31\,\mathrm{kcal/mol}$ is annotated in the bottom panel [60]. CHGNet, EquiformerV2(OC20), 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) [61]. Only MACE-MP(M) and EquiformerV2(OC20) ignite within this region. Runtime performance and center-of-mass drift are available in Figure S4.

used here [62], the heat bath acts an extra correction on the equations of motion for the system of particles with coordinates $\mathbf{q}_i$, momenta $\mathbf{p}_i$, masses m_i, and interaction potential V

$$\dot{\mathbf{q}}_i = \frac{\mathbf{p}_i}{m_i}, \tag{S6}$$

$$\dot{\mathbf{p}}_i = -\frac{\partial V}{\partial \mathbf{q}_i} - \mathbf{p}_i \frac{p_\xi}{Q}, \tag{S7}$$

where $\mathbf{p}_\xi$ and Q are the artificial momentum and mass of the thermostat particle [13, 63]. *Post-hoc* correction to enforce net zero force from the model prediction will only correct the first term. The total momentum drift is not zero, as can be seen by the following simple proof:

$$\begin{aligned}\sum_i \dot{\mathbf{p}}_i &= \sum_i \left[-\frac{\partial V}{\partial \mathbf{q}_i} - \mathbf{p}_i \frac{p_\xi}{Q} \right] \\ &= -\sum_i \mathbf{p}_i \frac{p_\xi}{Q} \neq \mathbf{0}.\end{aligned}$$

Note that we in our test we have enforced net zero total momentum at the beginning of each MD simulation, but non-conservative forces and slight numerical errors may still accumulate over MD trajectory. This non-zero momentum drift will induce non-zero COM drift over time as the MD simulations progress. Models failed to interact with heat bath correctly may not reproduce correct thermodynamic ensembles and therefore yield COM drift and incorrect partitions of kinetic and potential energies, as kinetic energies might be taken largely by COM velocity.

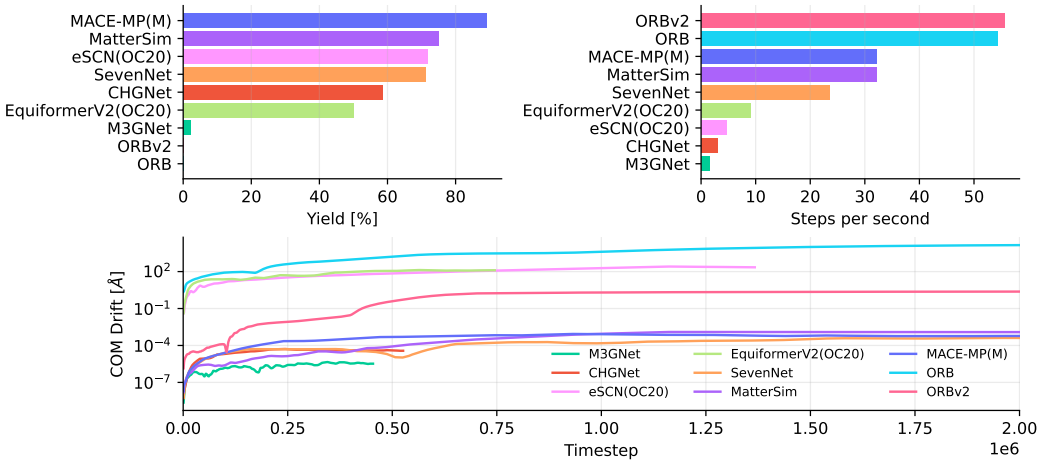


Figure S4: Hydrogen combustion. (Top left) The final reaction yield at the last MD step. (Top right) MD runtime speed measured in steps per second using single NVIDIA A100 GPU. `cuEquivariance` kernel was disabled for MACE-MP(M). (Bottom) The center-of-mass (COM) drift displacement during MD trajectory.

A.8 Vacancy formation and migration in elemental solids

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 [64] 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 [65]. 57 FCC and 57 HCP crystals from Angsten et al. [31] consists of metallic, metalloid, and noble gas elements.

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 $x = 0.5$:

$$\text{path asymmetry} = 2\int_{0}^{0.5} |\epsilon(0.5 - x) - \epsilon(0.5 + x)|\, dx \tag{S8}$$

.

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

$$\text{barrier asymmetry} = \frac{\Delta E}{E_{\text{forward}}} = \frac{E_f - E_i}{E_{\text{TS}} - E_i} \tag{S9}$$

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

Figure S5 demonstrates the distribution of *barrier asymmetry* (eq. (S9)) 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.

A.9 Details on robustness under distribution shifts

Descriptors consist of two concatenated components, X_i^R and X_i^B, that describe each central atom i's radial distances to its k-nearest neighbors and its bond angles, respectively. X_i^R is a vector of length k,

$$\left[\frac{w(r_{i1})}{r_{i1}} \quad \dots \quad \frac{w(r_{ik})}{r_{ik}}\right]^T, \quad r_{ij} \leq r_{i(j+1)}, \tag{S10}$$

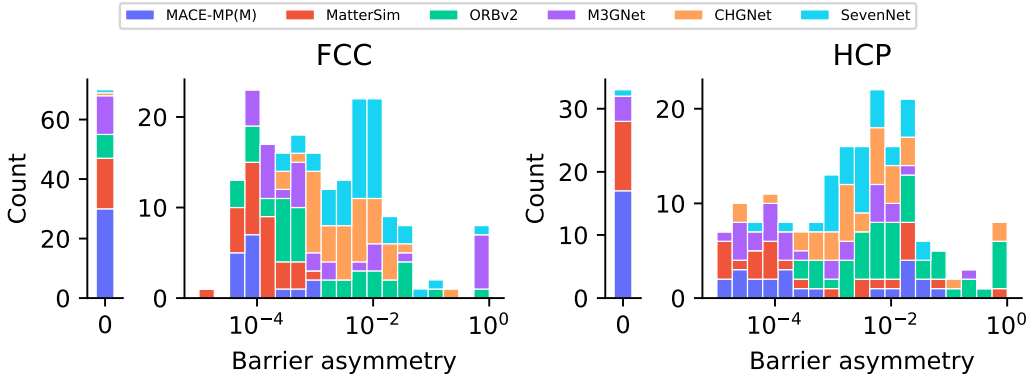


Figure S5: Absolute barrier asymmetry of vacancy migration in FCC and HCP elemental crystals. Compliance to symmetry is not correlated with the (non-)equivariance of the underlying MLIPs. Non-equivariant MLIPs: ORBv2, MatterSim, CHGNet. Equivariant MLIPs: MACE-MP(M), SevenNet.

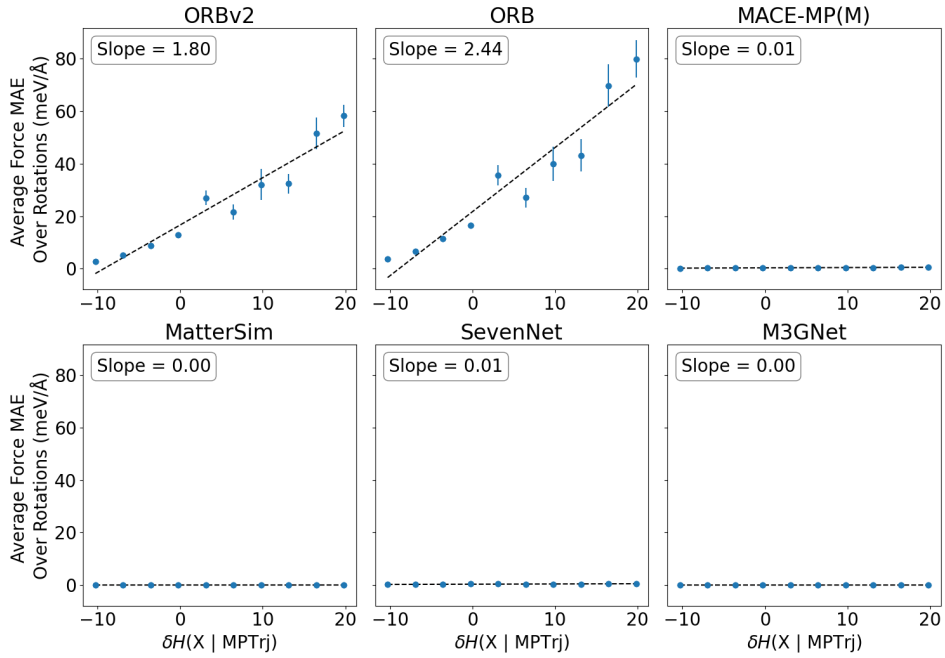


Figure S6: **Rotational equivariance versus differential entropy.** We calculate the mean absolute error (MAE) between each model's predicted forces and the forces predicted for rotated structures after transforming them back to the original reference frame. We compare this to the bin averages of differential entropy and report 95% confidence interval error bars for 10 bins from low to high differential entropy. Perfect rotational equivariance corresponds to a constant MAE of 0.0. Architectures without explicit rotational equivariance struggle to adhere to rotational equivariance with structures farther from the training distribution.

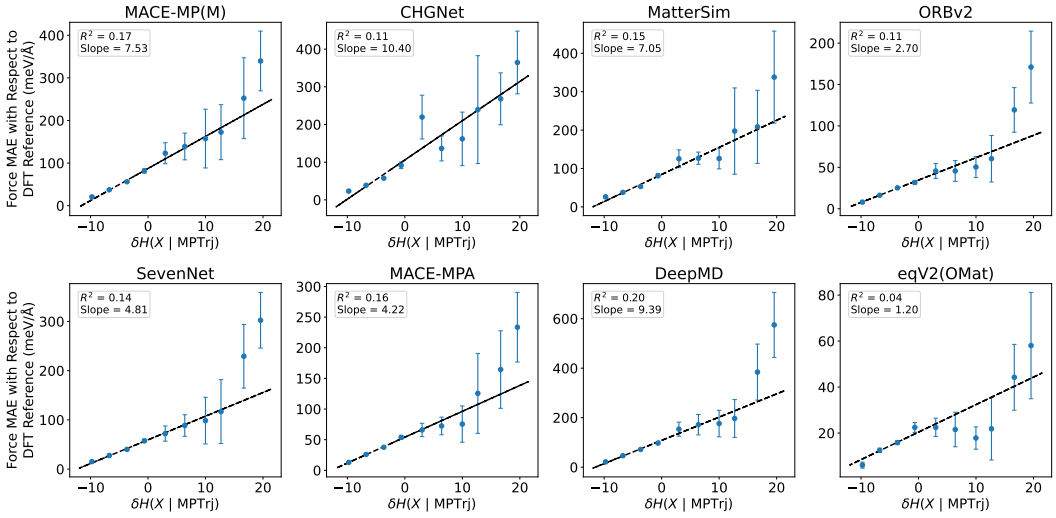


Figure S7: **Force MAE with respect to DFT versus differential entropy.** We compare the force MAE with respect to reference DFT values to the differential entropy of a random subset of 500 MPTrj structures. We report bin averages and 95% confidence interval error bars. Lines of best fit are provided. All models tend to predict forces less accurately when structures are more surprising.

where $1 \leq j \leq k$ due to the k-nearest neighbors approach, r_{ij} is the distance between i and another atom j, and $w(r)$ is a smooth cutoff function:

$$w(r) = \begin{cases} \left[1 - \left(\frac{r}{r_c}\right)^2\right]^2, & 0 \leq r \leq r_c \\ 0, & r > r_c. \end{cases} \tag{S11}$$

To retain information not only about radial distances but also about bond angles, we use X^B_{ijl} given by

$$\mathbf{X}^B_{ijl} = \frac{\sqrt{w(r_{ij})w(r_{il})}}{r_{jl}} \tag{S12}$$

which describes each neighbor l of atom j in the neighborhood of i. We represent the per-neighbor basis as

$$\mathbf{X}^B_{ij} = \left(X^B_{ij1}, \ldots, X^B_{ijk}\right), X^B_{ij1} \geq \cdots \geq X^B_{ij(k-1)}. \tag{S13}$$

For each atom, the bond angle descriptor then becomes

$$\mathbf{X}^B_i = \frac{1}{k}\sum_j \mathbf{X}^B_{ij}. \tag{S14}$$

To represent the training data, we compute these descriptors for every structure in each dataset. We hold out a subset of 50 structures in each dataset and estimate the Shannon information entropy of the remaining data using a kernel density estimate,

$$\mathcal{H}(\{\mathbf{X}\}) = -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{n}\sum_{j=1}^{n} K_h(\mathbf{X}_i, \mathbf{X}_j)\right], \tag{S15}$$

where we use a Gaussian kernel:

$$K_h(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{\|\mathbf{X}_i - \mathbf{X}_j\|^2}{2h^2}\right). \tag{S16}$$

The bandwidth h was selected according to the default provided by QUESTS [25], which was chosen to rescale the metric space of $\mathbf{X}$ according to the distance between two FCC environments with a 1% strain. To quantify the surprise of a data point $\mathbf{Y}$ compared to the existing observations $\{\mathbf{X_i}\}$, we define the differential entropy $\delta\mathcal{H}$ as

$$\delta\mathcal{H}(\mathbf{Y}|\{\mathbf{X}\}) = -\log\left[\frac{1}{n}\sum_{j=1}^{n} K_h(\mathbf{X}_i, \mathbf{X}_j)\right]. \tag{S17}$$

Figure S7 shows a correlation between force MAE and differential entropy for each model trained on MPTrj, indicating that the differential entropy is a reasonable measure of distribution shifts for MLIPs. Although these models perform well on in-distribution data, their error on force predictions increases as structures become more surprising, indicating a potential weakness in ability to generalize to out-of-distribution structures.

A.10 Extended case studies

A.10.1 CO_2 adsorption in metal-organic frameworks (MOFs)

Direct air capture (DAC) targets the removal of CO_2 directly from ambient air (about 400 ppm), and is increasingly recognized as indispensable for achieving net-negative greenhouse-gas emissions[66]. In practical, DAC technologies mainly rely on aqueous KOH slurries or amine-based absorbents whose chemisorptive binding affords the requisite affinity but imposes large thermal regeneration cost and chemical degradation. MOFs offer a promising physisorptive alternative for CO_2 capture. MOFs possess exceptionally high porosity and tunable structures that allow precise incorporation of functional groups such as open metal sites, diamines, thereby enhancing the affinity for CO_2 within their pore to levels suitable for DAC applications. Furthermore, the combination of framework rigidity, high surface area, and chemical stability positions MOFs as highly attractive candidates for durable, high-performance sorbents capable of operating under under relatively mild regeneration conditions.

In this case study, we curated 20 MOFs with experimentally reported Q_{st} values spanning three technologically relevant adsorption regimes and evaluated how accurately MLIPs classify them into correct categories: (1) General ($Q_{\text{st}} < 35\,\text{kJ/mol}$), (2) post-combustion flue gas ($35\,\text{kJ/mol} < Q_{\text{st}} < 50\,\text{kJ/mol}$), and (3) DAC ($50\,\text{kJ/mol} < Q_{\text{st}} < 100\,\text{kJ/mol}$).

The general adsorption class includes seven representative MOFs such as MOF-5, HKUST-1, UiO-66, ZIF-8, MIL-177, MIL-53-Al, and MOF-74-Fe. They exhibit relatively low CO_2 affinities and have not been prominently reported for CO_2 capture applications. The post-combustion flue gas class corresponds to MOFs reported for capturing CO_2 from power-plant exhaust, where the partial pressure of CO_2 is much higher than in ambient air, thus requiring moderate adsorption strengths. This category includes CALF-20, Al-PyMOF, UTSA-16, MUF-16, and ZnH-MFU-4l. Finally, the DAC class comprises MOFs capable of capturing CO_2 at extremely low partial pressures, demanding high binding affinities. DAC-relevant materials considered here include SIFSIX-3-Cu, NbOFFIVE-1-Ni, TIFSIX-3-Ni, SIFSIX-18-Ni-beta, en-Mg2(dobpdc), and CFA-1-OH-Zn, and SGU-29.

Table S3: Experimentally reported CO_2 Q_{st} values used in MLIP-arena, including three categories: DAC, post-combustion flue gas, normal MOFs.

Common name	**CO_2 Q_{st} (kJ/mol)**	**Category**	**Reference**
SIFSIX-3-Cu	54	DAC	[67]
NbOFFIVE-1-Ni	54	DAC	[68]
TIFSIX-3-Ni	49	DAC	[68]
SIFSIX-18-Ni-β	52	DAC	[68]
en-Mg_2(dobpdc)	50	DAC	[69]
CFA-1-OH-Zn	42 (71 in max)	DAC	[70]
SGU-29	51.3	DAC	[71]
CALF-20	39	Post-combustion flue gas	[72]
Al-PyrMOF	28	Post-combustion flue gas	[73]
UTSA-16	39.7	Post-combustion flue gas	[74]
MUF-16	32.3	Post-combustion flue gas	[75]
MIL-120-Al-AP	41	Post-combustion flue gas	[76]
ZnH-MFU-4l	20 (93 in high T)	Post-combustion flue gas	[77]
Fe-MOF74	33.2	General	[78]
MIL-53-Al	26.3	General	[79]
HKUST-1	29	General	[80]
MOF-5	15	General	[81]
UiO-66	28.6	General	[82]
ZIF-8	27	General	[83]
MIL-177	14	General	[84]

The heat of adsorption is calculated from the statistical average of interaction energies E_{int} using Widom insertion method [32, 85]. The interaction energy E_{int} is determined by energy difference between gas-inserted MOF and individual gas and MOF system:

$$E_{\text{int}} = E_{\text{MOF+gas}} - E_{\text{MOF}} - E_{\text{gas}}. \tag{S18}$$

The heat of adsorption is then determined from ensemble average:

$$Q_{\text{st}} = -\frac{\langle E_{\text{int}} e^{-\beta E_{\text{int}}} \rangle}{\langle e^{-\beta E_{\text{int}}} \rangle} + k_B T, \tag{S19}$$

where $\beta = (k_B T)^{-1}$, k_B is Boltzmann constant, and T is temperature.

All MLIP models are used in combination with D3 Becke-Johnson dispersion correction [86] with cutoff of 40 Bohr radius. Initial MOF structures were first alternately optimized with fixed and relaxed cell protocols until the final maximum atomic force is smaller than $0.05\,\text{eV/Å}$. The Widom insertion of CO_2 was then performed at $300\,\text{K}$ for three rounds for each MOF, with 5,000 insertion trials in each round. The grid spacing between gas insertion points was set at $0.15\,\text{Å}$.

Figure S8 shows distribution of predicted heat of CO_2 adsorption and average misclassification margin across 20 MOFs. The misclasification margin is defined as the distance between misclassified point to the closest decision boundary. Our results show that MatterSim is the strong MOF classifiers with misclassification margin of $11.30\,\text{kJ/mol}$ and misclassification count only 4, while MACE-MP(M) and SevenNet have severe overestimation, possibly due to short-range PES holes.

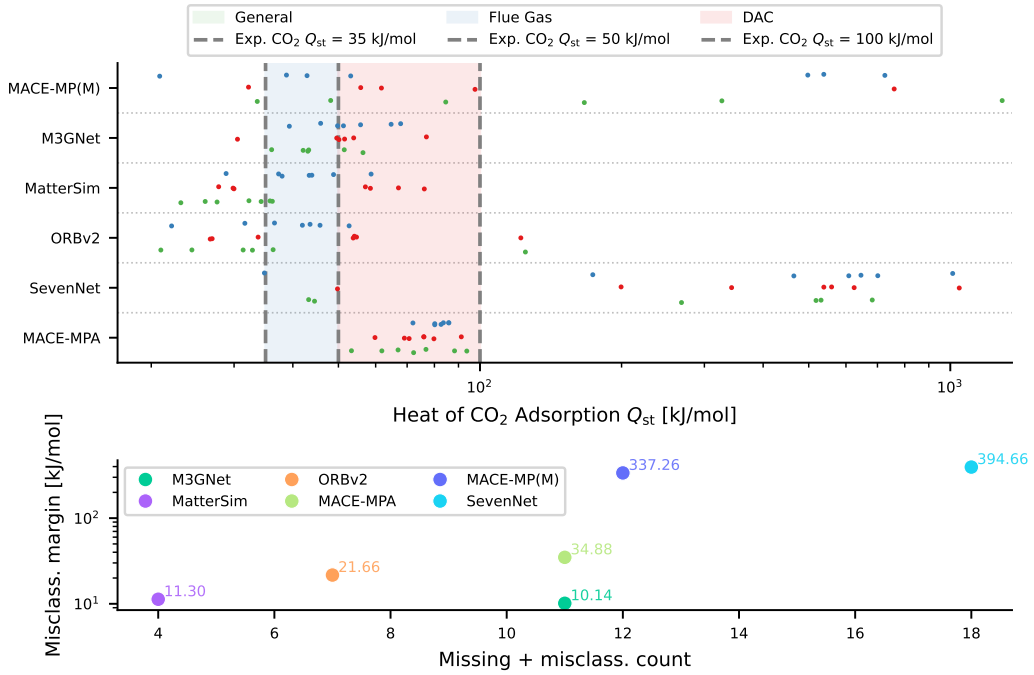


Figure S8: **Classification of MOFs based on predicted heat of CO_2 adsorption.** (Top) Three classes of MOFs based on experimental CO_2 Q_{st} measurements: (1) general (green area and points), (2) flue gas (blue area and blue points), and (3) DAC (red area and points). The perfect classifiers should predict Q_{st} of CO_2 in the corresponding regions. (Bottom) Mean misclassification margin and count of misclassified and missing MOFs.

A.10.2 Dynamical stability of 2D materials

Two-dimensional materials are vital to emerging technologies due to their exceptional physical properties and chemical tunability. To evaluate the ability of MLIPs to predict dynamical stability, we randomly selected 505 monolayers from the C2DB database [34, 87] and computed elastic tensors and phonon band structures using Pymatgen [88] and Phonopy [89, 90]. Following the C2DB protocol, a material is labeled dynamically stable if both the elastic tensor eigenvalues and lowest phonon frequencies are non-negative (specifically, both values should be greater than -10^{-7} to be labeled as stable).

F1 scores (fig. S9) indicate that MACE-MP(M), MACE-MPA, and MatterSim perform best, with macro F1 scores of 0.420, 0.412, and 0.411, respectively. In contrast, CHGNet, ORBv2, and ALIGNN perform significantly worse, with macro F1 scores below 0.30 and stable F1 scores of 0. All models show higher F1 scores for the unstable class—e.g., 0.596 for MACE-MP(M) vs. 0.245 for stable—highlighting a bias toward detecting instability. Confusion matrices (fig. S10) confirm this trend: most models heavily misclassify stable materials as unstable, with CHGNet, ORBv2, and ALIGNN failing to identify any stable structures. These findings reflect current limitations in MLIP generalization to vibrational stability and emphasize the need for improved training strategies that target phonon-related properties.

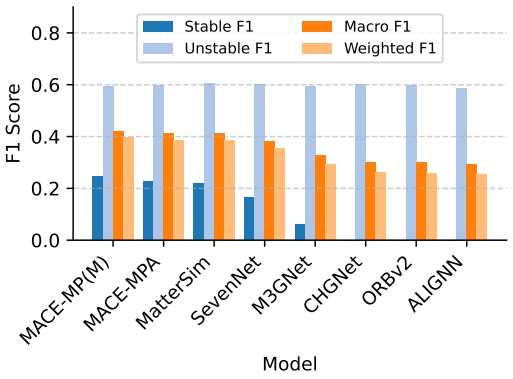


Figure S9: F1 scores of dynamical stability classification for 2D materials from C2DB database.

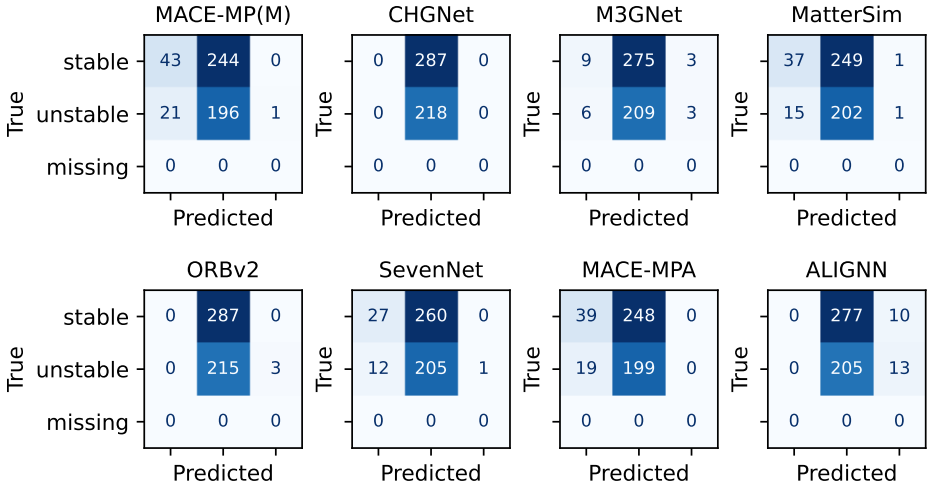


Figure S10: Confusion matrices of dynamical stability classification for 2D materials from C2DB database.

A.10.3 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 [33, 37]. In Figure S11, we probe the anharmonic PES of different MLIPs along the octahedral-tilting phonon mode with different unit cell lattice constants. Energy differences are calculated with respect to the undeformed structures at the respective 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.

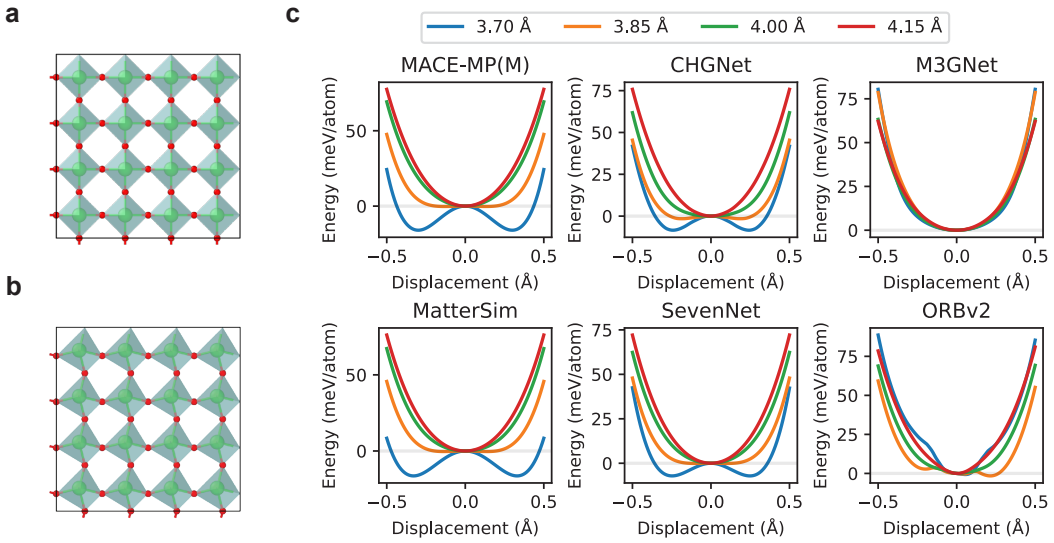


Figure S11: Landau-like second-order phase transition of octahedral-tilting mode in $BaZrO_3$ (BZO). (a) Undeformed $4 \times 4 \times 4$ supercell of BZO with cubic unit cell lattice constant of 4 Å. (b) R-tilt phonon mode with maximum displacement of 0.5 Å. Ba atoms are transparent for better visualization. (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.

B Supported models

Table S4: List of first-class 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. [27]	MIT	`2023-12-03-mace-128-L1_epoch-199.model`	2023-12-29
CHGNet	EFSM	✓	✓	MPTrj	GitHub	Deng et al. [26]	BSD-3-Clause	`v0.3.0`	2023-02-28
M3GNet	EFS	✓	✓	MPF	GitHub	Chen and Ong [28]	BSD-3-Clause	`M3GNet-MP-2021.2.8-PES`	2022-02-05
MatterSim	EFS	✓	✓	MPTrj, Alex, Proprietary	GitHub	Yang et al. [29]	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. [23]	Apache-2.0	`orb-v2-20241011.ckpt`	2024-10-15
SevenNet	EFS	✓	✓	MPTrj	GitHub	Park et al. [30]	GPL-3.0	`7net-0`	2024-07-11
eqV2(OMat)	EFS	✓	✗	OMat, MPTrj, Alex	GitHub	Barroso-Luque et al. [5]	Apache-2.0*	`eqV2_86M_omat_mp_salex.pt`	2024-10-18
eSEN	EFS	✓	✓	OMat, MPTrj, Alex	GitHub	Fu et al. [51]	Apache-2.0*	`esen_30m_oam.pt`	2025-04-14
EquiformerV2(OC22)	EF	✓	✗	OC22	GitHub	Liao et al. [22]	Apache-2.0	`EquiformerV2-lE4-lF100-S2EFS-OC22`	2023-06-21
EquiformerV2(OC20)	EF	✓	✗	OC20	GitHub	Liao et al. [22]	Apache-2.0	`EquiformerV2-31M-S2EF-OC20-All+MD`	2023-06-21
eSCN(OC20)	EF	✓	✗	OC20	GitHub	Passaro and Zitnick [91]	Apache-2.0	`eSCN-L6-M3-Lay20-S2EF-OC20-All+MD`	2023-02-07
DeepMD	EFS	✓	✓	MPTrj	GitHub	Zhang et al. [92]	GNU LGPLv3.0	`dp0808c_v024mixu.pth`	2024-10-09
ALIGNN	EFS	✓	✓	MP22	GitHub	Choudhary and DeCost [93]	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 [4], OMat: Open Materials dataset [5], MP22: Materials Project 2022, MPF: MPF.2021.2.8: Materials Project snapshot curated to train M3GNet [28]. OC20, OC22: Open Catalyst Project [94, 95].
*Modified Apache-2.0 (Meta)

C Additional DFT reference benchmarks

Bulk modulus from equation of state (EOS) calculations. In the vacancy migration task (appendix A.8), the geometry optimization of each pristine structure is then followed by an EOS fit to compare with GGA-PBE data from Angsten et al. [31]. Figure S12 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.

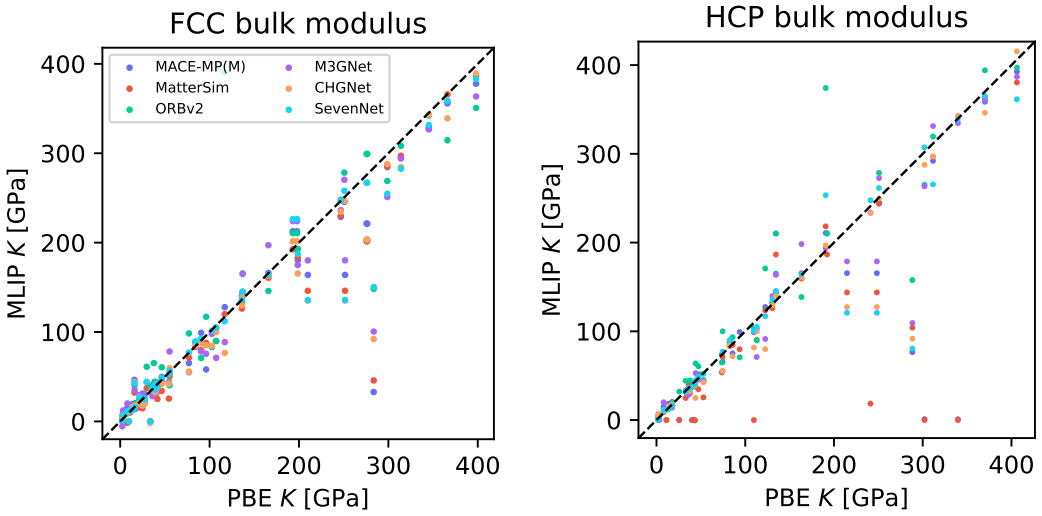


Figure S12: Bulk modulus of FCC and HCP elemental solids compared with GGA-PBE calculations [31].

Table S5: 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	28.7	2
MatterSim	19.142	28.1	1
ORBv2	32.583	31.5	1
M3GNet	21.867	37.0	4
CHGNet	19.815	25.5	6
SevenNet	14.500	21.1	3

Table S6: 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	36.3	5
MatterSim	45.865	35.5	5
ORBv2	41.116	36.4	4
M3GNet	21.321	22.0	16
CHGNet	21.484	26.3	16
SevenNet	21.925	17.0	15

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We discuss the limitations of existing benchmarks in the introduction and highlight the importance of moving beyond regression errors to better evaluate the performance of MLIPs.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitation in section 4 to compare our work with static DFT reference benchmarks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not propose new theory.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the open-source code and relevant data to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We host our code on github (`https://github.com/atomind-ai/mlip-arena`) and huggingface space `https://huggingface.co/spaces/atomind/mlip-arena`. The data is open sourced and hosted on `https://huggingface.co/datasets/atomind/mlip-arena`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We describe our detailed methods in main text and supplementary information.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For our major results in main text, we collect sufficiently large data points and provide statistical measures to fairly benchmark the model performances.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes we provide the compute details in the supplementary materials, in particular for the speed test of MD simulations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and confirmed no violations.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: We are aware that MLIPs and computational chemistry methods can have profound societal impacts in terms of accelerating the discovery of new materials and drugs. The goal of this work is to advance the evaluation of MLIPs. There are many consequences of our work, none which we feel must be specifically highlighted here.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no risks for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes we have cited the models and their associated license information.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We follow the guidelines to host our dataset on hugging face and provide documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM is used only for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.