MS25: Materials Science Focused Benchmark Dataset for Machine Learning Interatomic Potentials

Tristan Maxson[†], Ademola Soyemi[†], Xinglong Zhang[‡#], Benjamin W. J. Chen[‡], Tibor Szilvási[†]*

[†]Department of Chemical and Biological Engineering, University of Alabama, Tuscaloosa, AL 35487, United States

*[‡]Institute of High Performance Computing (IHPC), Agency for Science, Technology, and Research (A*STAR), 1 Fusionopolis Way, #16–16 Connexis, Singapore 138632, Singapore*

[#]Department of Chemistry, The Chinese University of Hong Kong, New Territories, Shatin, Hong Kong 999077, P. R. China

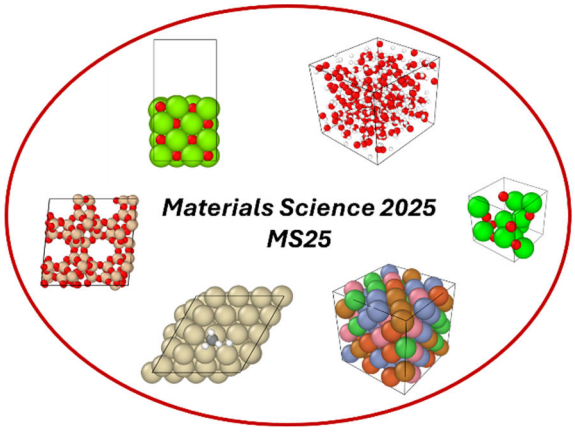
[*]Corresponding author. Email: tibor.szilvasi@ua.edu

ABSTRACT

We present MS25, a benchmark dataset for evaluating machine learning interatomic potentials (MLIPs) across diverse materials-relevant systems including MgO surfaces, liquid water, zeolites, a catalytic Pt surface reaction, high-entropy alloys (HEAs), and disordered Zr-oxides. Five MLIP architectures (MACE, NequIP, Allegro, MTP, and Torch-ANI) are trained and tested, focusing not only on traditional metrics (energies, forces, and stresses) but also explicitly validating derived physical observables such as lattice constants, volumes, and reaction barriers. We find that most models reach comparable accuracy on standard error metrics across the simple systems, although equivariant MLIPs offer 1.5-2X improvements over non-equivariant MLIPs in energy and force error for structurally complex or compositionally disordered environments such as HEAs and Zr–O systems. Our analysis highlights that low errors in energy and force predictions do not guarantee reliable observables, emphasizing the necessity of explicit validation. We demonstrate limitations in cross-framework transferability, as models trained on one zeolite framework (CHA) fail to reliably generalize to predictions of structurally distinct frameworks (e.g., MFI). Size-extensive tests show some dependence on system size for MgO resulting from forced periodicity. The HEA and Zr-O datasets are identified as challenging tests for future benchmarks and MLIP model architecture developments as they show significant differentiation in error between MLIP architectures and are still relatively difficult at 1000 training images. Moving forward, we recommend benchmarking efforts shift their focus from marginal accuracy improvements in energy and force errors toward identifying and understanding model failure modes, assessing transferability rigorously, and how their errors affect observable predictions. For researchers

looking to choose an MLIP architecture, we suggest selecting equivariant MLIP architectures if the complexity of the system is a challenge. For simple materials problems, auxiliary features such as integration with molecular dynamics engines, trade-offs between computational dataset generation cost vs. MLIP inference speed, and framework integration may play a more important decision factor than small differences in error metrics that are unlikely to matter for production level research.

Graphical Table of Contents



INTRODUCTION

Understanding materials properties allows for the development of more performant solutions for applications including, but not limited to catalysis,[1-4] sensors,[5-7] or solar cells[8-10]. Such understanding often requires an accurate representation of the potential energy surface (PES)[11-14] that encodes information about static energy-based quantities (e.g., thermodynamic stability,[15, 16] chemical barriers,[17-19] and binding energies[20-22]) or is used to derive quantities from the dynamics of the system (e.g., vibrational entropy,[23] density,[24] and diffusion rates[25]). Density functional theory (DFT) has been the workhorse of computational research in recent decades and an accurate PES may be constructed, but at a considerable computational cost that limits what problems can be studied within reasonable timescales.[26-29]

Machine learning interatomic potentials (MLIPs) have recently become important as a method of extending the length and time scale of electronic structure theory-based simulations[30-35] while achieving near-DFT accuracy at a fraction of the computational cost. MLIPs circumvent the need for expensive quantum chemical calculations by fitting the PES using DFT reference data and thus open up new avenues for more realistic modeling of complex materials and materials properties, which typically require larger time- and length-scales.

Choosing an MLIP model among the rapidly increasing number of published models is a surprisingly formidable task given the wide variety of available MLIPs. Several software packages are freely available,[36-47] and it is also becoming increasingly difficult even for experienced practitioners to assess performance and follow all the technical details of MLIP development. Thus, various reviews, perspectives, and tutorials have been published to survey state-of-the-art MLIPs.[48-54] Benchmark studies[34, 35, 55-60] have aimed to compare different MLIP architectures for similar model systems;[61-67] however studies are often limited to specific applications and are currently dominated by molecular systems[68-70].

We envision that materials modeling can benefit even more from the advent of MLIPs than small molecule chemistry, as materials are more compositionally complex, and systems investigated are larger. However, most benchmarks published have been focused on either the predictions of electronic properties of small molecules or molecules for which the potential energy surface can be fully explored *ab initio*. On the other hand, focused benchmarks for specific materials has been

limited as most benchmarks relevant to materials science are designed for foundational model development across the entire periodic table. The development of the MPTrj dataset,[71] which contains structures from the Materials Project[72], is a great initiative but suffers from having an overly diverse set of structures. As such, the MPTrj is an ideal dataset for foundational models[73, 74] but does not allow for testing the accuracy of MLIPs trained from scratch on more specific systems as it does not contain enough configurations relevant to applications such as HEAs or disordered oxides. While foundational model datasets are becoming more diverse (higher composition, surfaces, non-equilibrium structures, etc.) they cannot replace specialized datasets that explore the entire potential energy surface to be investigated. Therefore, new benchmark datasets need to be developed[75] that will allow method developers to showcase the performance of their new methods and for practitioners to test their capabilities and be informed on new developments in a comparative manner.

In this work, we provide MS25, a benchmark dataset consisting of a variety of materials-related test systems representative of biomaterials, porous materials, condensed solid and liquid phases, reactive surfaces, metallic alloys, and oxide materials as shown in Figure 1. We start with a model of an oxide surface using MgO(100)[76] as an example to analyze how well clean surfaces can be simulated at high and low temperature[77]. A liquid phase water system is then modeled as water is a material relevant to various problems. To analyze a porous material, we simulate multiple zeolite frameworks such as CHA, MFI, LTA, and BEA[78]. Catalysts represent a large class of materials and as such we simulate a reaction on the surface of Pt(111) as a test of material properties / reactivity, rather than focusing on the simulation of the material structure itself. We also study a FeNiCrCoCu alloy[79] as high entropy alloys have become more commonly modeled computationally[80-82]. Finally, we look at a previously published Zr-O dataset[83] which we found challenging due to the generally disordered, amorphous structure and the highly diverse chemical environment around Zr. We benchmark the NequIP[39], Allegro[84], MACE[85], Torch-ANI[40], and MTP[45] MLIPs. Beyond commonly studied metrics such as energy, force, and stress errors, we also evaluate errors in lattice constants, bulk modulus, energy decomposition analysis, and transition state energies. The more advanced properties provide insights into error cancellation and systematic errors[86, 87] which simple metrics cannot provide.[88-90] Finally, we compare the inference speeds of the error-optimized MLIP models.

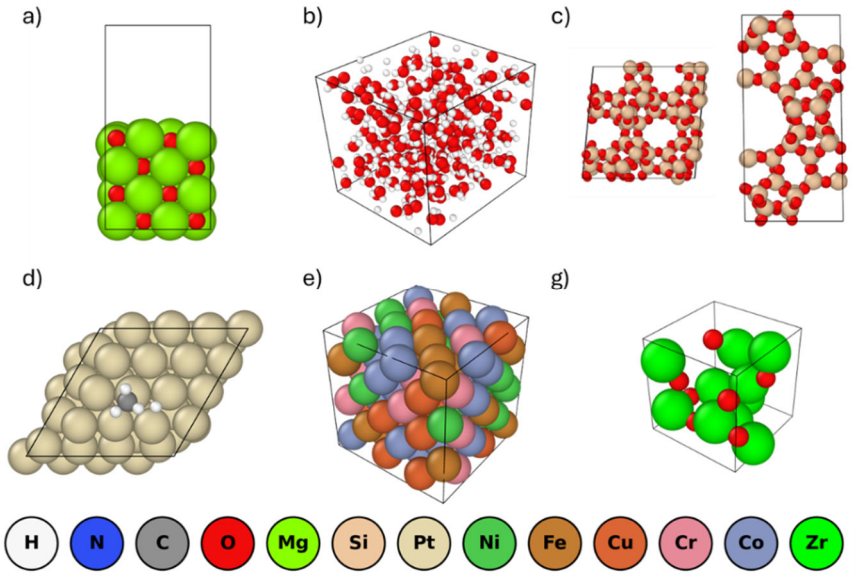


Figure 1. Example structures for (a) MgO(100), (b) Water, (c) CHA and similar zeolites, (d) Chemical Reaction, (e) High entropy alloy (HEA), and (f) amorphous Zr-oxide structures in accordance with the datasets presented in this work. The colors are shown at the bottom of the figure as a legend.

COMPUTATIONAL METHODS

Here we describe the methodology behind the presented benchmark. We begin with a description of the training data generation (in this work or as reported in previous literature as appropriate) for each presented system. We then describe the dataset processing and MLIP training procedures. Finally, we provide additional information on our analysis methods given the trained MLIPs and the errors regarding the respective datasets or observable metrics.

Oxide Surface Training Data

An oxide surface dataset is created by modeling an MgO(100) surface at a size of 2x2 and 4x4 for size extensivity tests. VASP(version 6.3.2)[91-94] calculations are performed at the PBE-D3[95, 96] level of theory with the developer-recommended VASP PAW potentials (version 54)[92] for Mg and O. An energy cutoff of 400 eV, gaussian smearing with 0.1 eV width, and dipole corrections in the z-direction are utilized.

For our MD simulations, the Nose-Hoover thermostat(chain length of 1, characteristic time of 100 fs for the thermostat)[97] is used to partially equilibrate the system at temperatures between 300 and 1200 K with a 2.5 fs timestep in the NVT ensemble. We choose to use 2x2x4 and 4x4x4 MgO slabs, with the bottom two layers fixed to approximate a bulk structure. Monkhorst-Pack[98] gamma-centered K-point grids are chosen to be (3, 3, 1) and (1, 1, 1) for the 2x2x4 and 4x4x4 supercells respectively. Training data is sampled at a frequency no shorter than 25 fs to ensure data is uncorrelated. The 2x2 MgO(100) dataset will be herein referred to as just “MgO with “MgO 2x2” and “MgO 4x4” being referred to for the 2x2 and 4x4 datasets respectively, when both datasets are being discussed.

Water Training Data

A water dataset is created by modeling bulk water in an orthogonal cell with 64 and 192 water molecules for size extensivity tests. CP2K(version 2024.1)[99] calculations are performed at the RPBE-D3(0)[100] level of theory with the GTH-PBE[101] potentials and the DZVP-MOLOPT-SR-GTH[102] basis sets for hydrogen and oxygen. An energy cutoff of 1200 Ry is applied to ensure energy, forces, and stresses are well converged as we have previously observed issues with training on the typical lower energy cutoffs. The orbital transform (OT) method[103] is applied with an energy gap of 0.05 eV and a step size of 0.1. No occupational smearing is applied since it is not supported in the OT mode. Only the gamma k-point is sampled as the system sizes are large enough to not require additional k-points.

MD simulations are run using the Langevin thermostat[104] (0.05 fs^{-1} gamma) with a time step of 0.5 fs in the NVT ensemble at a temperature of 300 K at a distribution of volumes to ensure stresses are sampled properly. To reduce the correlation of the structures, we train an Allegro MLIP on DFT data coming from MACE-MP-0(Medium) generated[105] structures to create an approximate MLIP which is then used to sample structures for the training set sampled at a frequency no shorter than 30 ps apart. DFT single-point calculations are then performed on sampled structures and form the final dataset reported in this benchmark. The 64-water dataset will be herein referred to as just "Water" with "Water-64" and "Water-192" being referred to for the 64 and 192 molecule datasets respectively, when both datasets are being discussed.

Zeolite Training Data

A zeolite dataset is created by modeling a CHA zeolite framework and an extended dataset is created by modeling FAU, LTA, and MFI frameworks as well. CP2K(version 2024.1)[99] calculations are performed at the RPBE-D3[100] level of theory with the GTH-PBE[101] potentials and the DZVP-MOLOPT-SR-GTH[102] basis sets for silicon and oxygen. An energy cutoff of 1000 Ry is applied to ensure energy, forces, and stresses are well converged as we have previously observed issues with training on the typical lower energy cutoffs. The orbital transform (OT) method[103] is applied with an energy gap of 0.05 eV and a step size of 0.1. No occupational smearing is applied since it is not supported in the OT mode. Only the gamma k-point is sampled as the system sizes are large enough to not require additional k-points.

MD simulations are run under Langevin[104] dynamics (0.05 fs^{-1} gamma) with a time step of 2 fs in the NVT ensemble at a temperature of 300 K at a distribution of volumes to ensure stresses are sampled properly. We first obtained the experimental CHA structure from the IZA Structure Commission database of zeolite structures and doubled the primitive cell along each axis to form a supercell. To reduce the correlation of the structures, we train an Allegro MLIP on DFT data with rattled displacements to create an approximate MLIP which is then used to sample structures for the training set sampled at a frequency no shorter than 50 ps apart. DFT single-point calculations are then performed on sampled structures and form the final dataset reported in this benchmark. To assess the transferability of the models, we also provide a similarly generated extended dataset of FAU, LTA, and MFI structures to evaluate the ability of MLIPs to predict other frameworks. The main and extended datasets will be herein referred to by their framework names (CHA, FAU, LTA, MFI) in the following text and figures.

Chemical Reaction Training Data

A chemical reaction dataset is created by modeling a reaction of methane C-H bond cleavage on a Pt(111) surface without explicitly sampling the exact transition state. VASP(version 6.3.2)[91-94] calculations are performed at the PBE-D4[95, 106] level of theory with the developer-recommended VASP PAW potentials (version 54)[92] for Pt, C, and H. An energy cutoff of 400 eV, gaussian smearing with 0.2 eV width, and dipole corrections in the perpendicular (Z) direction are utilized in MD simulations. The Nose-Hoover thermostat(chain length of 1, characteristic time of 20 fs for the thermostat)[97] is used to partially equilibrate the system at temperatures between 300 and 1000 K with a 0.5 fs timestep in the NVT ensemble. The 4x4x4 Pt slab is used as a substrate for the binding of methane, with the bottom two layers of Pt fixed to approximate a bulk structure. The Monkhorst-Pack[98] K-point grid is chosen to be (2, 2, 1).

As we wish to get images representative of the reaction coordinate without directly exploring the transition state, we perform Blue Moon-based[17] MD simulations. In the Blue Moon method, the bond distance is constrained along a reaction coordinate and we choose to break the C-H bond of methane. The bond length constraint is enforced by the SHAKE[107] algorithm as implemented in VASP. Training data is sampled at a frequency no shorter than 10 fs to ensure data is uncorrelated. The C-H bond breaking dataset will be herein referred to as "Reaction" in the following text and figures.

High Entropy Alloy (HEA) Training Data

A HEA dataset is created by forming a diverse composition of 5 transition metal elements whereby the dataset difficulty comes from the diverse chemical environments within the alloy material. Embedded Atom Model[108, 109] calculations are performed in LAMMPS using the FeNiCrCoCu-with-ZBL potential[79] which considers interactions up to 5.804 Å for Fe, Ni, Cr, Co, and Cu. Configurations are generated with a distribution of compositions with each element representing at least 10% of the system, with the remaining composition being chosen randomly and uniformly.

A Nose-Hoover chain[110] (chain length of 3, characteristic time of 100 ps for the thermostat and 200 ps for the barostat) with a time step of 1.5 fs is used to equilibrate the system under NPT conditions between temperatures of 600 and 1200 K and at a pressure of 1 atm. Training data is sampled from configurations at least 5 ps apart to eliminate correlation between datapoints and across an even distribution of temperatures. The HEA dataset will be herein referred to as "HEA" in the following text and figures.

Zr-O Training Data

A dataset of the Zr-O[65] chemical space was generated in a previously published work to evaluate the effectiveness of data generation methods such as molecular dynamics, rattling, and contour exploration for amorphous Zr-containing oxides[83]. VASP(version 5.4.4) calculations were performed at the PBE level of theory with the developer-recommended VASP PAW potentials (version 52)[92] for Zr (Zr_sv) and O. An energy of cutoff of 500 eV is utilized in the data generation, which utilizes random structure generation, MD based structure generation, and dimer search structure generation. The original dataset is then filtered (from 120,068 structures to 10,000) to remove all structures containing only oxygen and bias towards lower force structures as the original dataset explores structures with much higher force than is typically required in standard MLIP work. The dataset samples the entire space of Zr-O compositions at varying cell volumes. The amorphous Zr-O dataset will be herein referred to as "Zr-O" in the following text and figures.

Training Data Post-Processing

10,000 images are selected for MLIP data for all systems. The data is then split into 1000 training images, 80 validation images, and 8920 test images to form a split of data to be used for MLIP training. Additionally, the training data is then split into 50, 100, 200, 400, 600, 800, and 1000 sets of images to determine the effect of training data size on model accuracy. As the number of training images increases, new data is added to the previous training set to simulate the growth of the dataset. By generating the data in an additive manner, we ensure any changes in model performance (as a function of training set size) should be based on the addition of new data and not the removal or substitution of specific data points. The data splitting procedure is then repeated twice to allow for 3 replicates of training, validation, and test set data which are referred to as "Set 1", "Set 2", and "Set 3". In the main text we report the results solely from Set 1, which provides a realistic example of the run-to-run variance, whereas we report Set 2 and Set 3 in the supplementary information (SI). The radial cutoffs for each training data set are presented in Table 1, which were chosen as the minimum radial cutoff that gave reasonable errors via initial testing. Radial cutoff must be kept consistent between architectures as it controls the amount of information available to the MLIP for predictions, but more accurate MLIPs may be possible with larger cutoffs or faster MLIPs with smaller cutoffs.

Table 1. The radial cutoffs for each dataset which are supplied to the MLIPs and kept constant.

Parameter	MgO	Water	CHA	Reaction	HEA	Zr-O
Radial Cutoff (Å)	6	6.5	5	6	5.5	6

NequIP Training Procedure

The NequIP MLIP software package is used to train MLIP from atomistic data, to infer them for error analysis, and to perform MD simulations within LAMMPS (pair_nequip). NequIP models are trained with system-dependent L_{max}, batch size, neural network widths, number of layers, polynomial cutoff, and radial cutoff as seen in Table 2. Models with NequIP are executed using LAMMPS (12 December 2023 – Development, CUDA 12.0, PyTorch 2.0.1) for performance predictions on a 40GB A100 GPU. The training procedure is kept fixed: 1) The first phase uses a learning rate of 0.01 with an annealing learning rate schedule[111] and force loss is set to 10X higher than energy loss, 2) The second phase uses a learning rate of 0.001 with a plateau scheduler to converge errors with no change in loss definitions, 3) A final phase uses a learning rate of 0.0002 with a plateau scheduler while changing the energy loss to be 15x the force loss. The 3-phase training procedure has been derived to allow for the best model to be trained within 24 hours on a single GPU. The 24-hour limitation is applied to other models to help normalize computational efficiency in training, but the time limitation was chosen arbitrarily as a breakpoint where we generally observe models are converged across all codes. When stresses were available, the stresses were also trained, and the stress loss was set to 3x higher than the original energy loss coefficient. Hyperparameter scans are performed by random search, which is refined in stages to find ideal parameters based on the error of the test set of Set 1 as trained on 1000 training points. Loss functions are optimized in an empirical manner to balance error between energy, force, and stress for NequIP and the same approach is applied towards MLIPs of other architectures described later. Details of the training procedure and hardware specs may also be viewed in the included inputs within the SI and Zenodo repository referenced at the end of this manuscript.

Table 2. The NequIP hyperparameters are optimized for the systems of interest. Hyperparameters have been optimized by a multi-stage random search except for radial cutoff, which was predetermined before training. The hyperparameters of the Water (64 vs 192) and MgO (2x2 vs 4x4) MLIPs are not adjusted for training as such they are not distinguished in the table.

Parameter	MgO	Water	CHA	Reaction	HEA	Zr-O
L_{max}	2	2	2	*2*	1	2
Interaction DNN Layers	2	4	3	3	2	3
Interaction DNN Width	64	48	48	48	48	48
Radial DNN Width	48	96	32	32	64	80
Radial DNN Layers	3	2	3	1	1	1
Input Features	64	48	16	48	48	48

Polynomial Cutoff	8	32	8	8	40	24
Stress Included	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE
Batch Size	2	2	1	2	2	1

Allegro Training Procedure

The Allegro module based on NequIP is utilized for training MLIPs but the same training procedure is used as when training standard NequIP models. System-dependent parameters include Lmax, tensor layers, interaction layers, latent layers, neural network width, polynomial cutoff, radial cutoff, and several input features as found in Table 3. Models with Allegro are executed using LAMMPS (12 December 2023 – Development, CUDA 12.0, PyTorch 2.0.1) for performance predictions on a 40GB A100 GPU. When stresses were available, the stresses were also trained and the stress loss was set to 100X higher than the original energy loss coefficient. Hyperparameter scans are performed by random search, which is refined in stages to find ideal parameters based on the error of the test set of Set 1 as trained on 1000 training points. Details of the training procedure and hardware specs may also be viewed via included inputs within the SI and Zenodo repository referenced at the end of this manuscript.

Table 3. The Allegro hyperparameters are optimized for the systems of interest. Hyperparameters have been optimized by a multi-stage random search except for radial cutoff, which was predetermined before training. The hyperparameters of the Water (64 vs 192) and MgO (2x2 vs 4x4) MLIPs are not adjusted for training as such they are not distinguished in the table.

Parameter	**MgO**	**Water**	**CHA**	**Reaction**	**HEA**	**Zr-O**
L_{max}	1	2	2	3	2	2
Interaction DNN Layers	2	2	3	2	2	4
Interaction DNN Width	48	128	128	128	96	64
Tensor DNN Layers	3	3	4	3	2	2
Latent DNN Layers	3	5	4	4	3	4
Input Features	64	48	32	96	32	16
Polynomial Cutoff	16	16	32	8	8	8

Stress Included	FALSE	TRUE	TRUE	FALSE	TRUE	TRUE
Batch Size	1	1	1	1	4	1

MACE Training Procedure

MACE[85] v0.3.7 was utilized for training MLIPs to infer and evaluate test errors, with the MACE LAMMPS (28 Mar 2023 – Development, CUDA 12.2, PyTorch 2.1.0) interface employed for speed benchmarking through molecular dynamics (MD) simulations on a 40GB A100 GPU. The training process was divided into two distinct phases:

Phase 1: Energy and force loss coefficients were set to 1 and 20, respectively. The initial learning rate was fixed at 0.01 and controlled using an on-plateau scheduler based on validation loss, with a patience of 50 epochs and a reduction factor of 0.8.

Phase 2: Activated via the `--swa` option, this phase utilized a larger energy loss coefficient compared to Phase 1. Force and stress loss coefficients were maintained at 1 and 10, respectively, using the "universal" loss function with a Huber delta value of 0.01.

Grid searches were conducted on the 600 training points split of Set 1 to identify optimal hyperparameters, which were subsequently applied across all splits of all datasets (Table 4). The hyperparameters evaluated included: (1) l_{max} (0,1,2), (2) the number of channels (64, 128, 256), (3) batch size (2, 8), and (4) the second phase energy coefficient (100, 10000). For datasets with stresses, a separate grid search over (1) the first phase stress weight (0,1,10,100), and (2) the second phase stress weight (0,1,10,100) was also conducted building on the best parameters obtained from the first grid search. To balance computational efficiency, each hyperparameter trial was limited to 500 epochs. The optimal parameters were identified based on the lowest loss on the validation set.

To manage varying time constraints (from 1.2 hours for the 50-datapoint split to 24 hours for the 1000-datapoint split), we scaled the number of training epochs based on our best hyperparameter configuration. We fixed the ratio of stage 1 to stage 2 epochs as 0.8:0.2. Training was stopped when the number of epochs was reached.

Details of the training procedure and hardware specs may also be viewed via included inputs within the SI and Zenodo repository referenced at the end of this manuscript.

Table 4. MACE hyperparameters optimized for the systems of interest. See main text for details of hyperparameter tuning.

Parameter	**MgO (2×2)**	**MgO (4×4)**	**Water (64)**	**Water (192)**	**CHA**	**Reaction**	**HEA**	**Zr-O**
l_{max}	1	1	1	1	1	1	2	2
Number of channels	128	128	64	32	64	256	64	256

Batch size	2	2	2	2	2	2	8	8
Phase 1 stress weight	10	10	0	0	0	-	10	0
Phase 2 energy weight	10000	10000	10000	10000	10000	10000	1	1
Phase 2 stress weight	1	1	100	100	1	-	10	100
Number of epochs	2200	650	1100	500	2100	1700	2600	3100
E0s	No	No	Yes	Yes	No	No	Yes	Yes

ANI Training Procedure

The TorchANI[40] package was used to train ANI neural networks based on the modified Behler and Parrinello symmetry functions (BPSFs)[112], for inference and error analysis. The LammpsANI[113] package was used to provide an interface for LAMMPS (29 Oct 2020 – Development, CUDA 12.2, PyTorch 2.1.0) for performing MD simulations on a 40GB A100 GPU. Training was performed with the ADAM[114] optimizer, where the initial learning rate was set to 0.001 and controlled by an on-plateau scheduler based on the validation loss with a patience of 100 and a decay factor of 0.5.

Grid searches were conducted on the 600 training points split of Set 1 to identify optimal hyperparameters, which were subsequently applied across all splits of all datasets (Table 5). The hyperparameters evaluated included: (1) NN architecture, where we tuned a varying number of hidden layers (3–6) with different number of neurons per layer (Table 6), (2) batch sizes (2, 8), and (3) force coefficients (0.5, 1.0, 2.0). To balance computational efficiency, each hyperparameter trial was limited to 500 epochs. The optimal parameters were identified based on the lowest loss on the validation set. While ANI is capable of predicting atomic stresses, we omitted stress training because the default training script lacks built-in functionality for this feature.

To manage varying time constraints (from 1.2 hours for the 50-datapoint split to 24 hours for the 1000-datapoint split), we scaled the number of training epochs based on our best hyperparameter configuration. Training was stopped when either (1) the learning rate becomes smaller than the early stopping learning rate (1E-5) or (2) when the number of epochs allocated for training was reached.

Details of the training procedure and hardware specs may also be viewed via included inputs within the SI and Zenodo repository referenced at the end of this manuscript.

Table 5. The optimized ANI hyperparameters used for the systems of interest. Hyperparameters have been optimized using the procedure outlined in the text.

Parameter	**MgO (2×2)**	**MgO (4×4)**	**Water (64)**	**Water (192)**	**CHA**	**Reaction**	**HEA**	**Zr-O**

NN architecture	160-128-96-64-32-16	160-128-96-64-32-16	160-128-96	160-128-96	192-96-32	160-128-96-64-32-16	160-128-96	192-160-128
Batch size	2	2	8	8	2	2	2	8
Force coefficient	2	2	2	2	0.5	0.5	2	2
Number of epochs	2700	700	700	180	800	3800	800	10000

Table 6. ANI architectures tested for hyperparameter tuning.

Architecture	Number of Hidden Layers
160-128-96	3
192-160-128	3
192-96-32	3
192-160-128-96	4
160-128-96-32	4
160-128-96-64-32	5
160-128-96-64-32-16	6

MTP Training Procedure

The MTP[45] MLIP v2 software package was used to train MTPs for inference. To perform speed benchmarking with MD simulations, the LAMMPS (2 Aug 2023 - Update 1, CUDA 12.2, PyTorch 2.1.0) MLIP interface was employed on an AMD EPYC 7713 64-Core Processor. MTPs were trained in a single stage with a fixed energy and force loss coefficients of 1 and 0.02/N, respectively, where N is the average number of atoms in each image of the dataset.

Tuning of additional hyperparameters was conducted via grid search using Set 1 of each dataset and shown in Table 7. To avoid an exponential increase in possible combinations, parameters were optimized in a sequential manner across multiple stages, keeping the from previous stages fixed. The stages involved tuning of: (1) lev_{max} (18, 20, 22, 24, 26) and (2) radial basis size (8, 12). For datasets with stresses, a separate grid search over the stress coefficients (1E-1, 1E-2, 1E-3, 1E-4, 1E-5, 0) was also conducted, building on the best parameters obtained from the first grid search. To balance computational efficiency, each hyperparameter trial was limited to 1000 epochs. The optimal parameters were identified based on the lowest loss on the validation set.

To manage varying time constraints (from 1.2 hours for the 50-datapoint split to 24 hours for the 1000-datapoint split), we scaled the number of training epochs based on our best hyperparameter configuration. For the smallest 50-datapoint split, we halved the epoch count, as we observed non-linear training behavior where smaller datasets required disproportionately longer per-epoch processing time per image. Training was terminated when either: (1) the change in the loss fell below the threshold of 0.001, or (2) the number of epochs allocated for training was reached.

Details of the training procedure and hardware specs may also be viewed via included inputs within the SI and Zenodo repository referenced at the end of this manuscript.

Table 7. MTP hyperparameters optimized for the systems of interest. See main text for details of hyperparameter tuning.

Parameter	MgO 2×2	MgO 4×4	Water 64	Water 192	CHA	Reaction	HEA	Zr-O
lev_{max}	26	26	20	20	26	26	26	26
Radial Basis Size	8	8	12	12	8	12	8	8
Stress Coefficient	0	0	1E-5	1E-5	0	-	1E-4	1E-5
Number of epochs	3000	550	1800	500	1800	3000	900	5000

Error Equations

For error analysis we define the Mean Absolute Error (MAE) and Residual Square Mean Error (RMSE) by equations 1 and 2. n represents the number of values to compare, y_i represents the predicted quantity, and $\hat{y}_i$ represents the actual quantity. y_i and $\hat{y}_i$ may be vectors in the case of forces or stresses, in which case they are flattened to multiple scalars prior to computing errors such that the error is computed component-wise.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

RESULTS AND DISCUSSION

We have computed errors for all MLIPs architectures studied (NequIP, Allegro, MACE, MTP, and ANI) for MgO, Water, CHA, Reaction, HEA, and Zr-O datasets and present our results for analysis. We start with an analysis of general energy, force, and stress errors across all datasets. To highlight training efficiency regarding dataset size, we present what we call a "data training rate" based on a general observation that accuracy tends to scale in a log-log fashion in literature[115-117] and use the slope to directly compare different MLIP architectures. We then focus on more specific tests of size extensivity by MLIPs, observables, and transferability to other frameworks. Inference speeds follow to provide additional insight into how the different architectures perform in terms of computational efficiency. For clarity, all results provided are taken from "Set 1", but additional plots can be found in the supplementary information for "Set 2" and "Set 3".

Energy, Force, and Stress Errors

To evaluate the accuracy of the MLIPs, all datasets are evaluated regarding their predictions of energy, force, and stress in the test set of data that has not been seen during the training of the model. We note that all predictions of force and stress are inferred conservatively in the MLIPs by computing the derivative of energy with respect to position or the cell respectively. We also distinguish precision (noise) error from accuracy error which latter is due to the choice of the level of theory. Precision is concerned only with the chosen level of theory's ability to produce systematic (but not necessarily correct) predictions of energy, force, and stress. We note that energy, force, and stress errors are limited by the precision (noise) of DFT simulations; however, precision is largely controlled by simulation convergence with respect to energy. Typical DFT convergence (as seen in this work) is performed to the sub-meV scale, whereas MLIPs provide total energy errors that are often above 10 meV given a 100-atom system and 0.1 meV/atom error. Thus, we think there is significant room for improvement in the current predictions of MLIPs and data precision (noise) is not the limiting factor in any of the datasets. Here we use the MAE as our desired metric for the energy, force, and stress as the error metric is used for the loss function in training. Figure 2 shows the error heatmaps across the datasets in this work for 100 and 1000 training points to demonstrate the improvements in error due to an order-of-magnitude change in data.

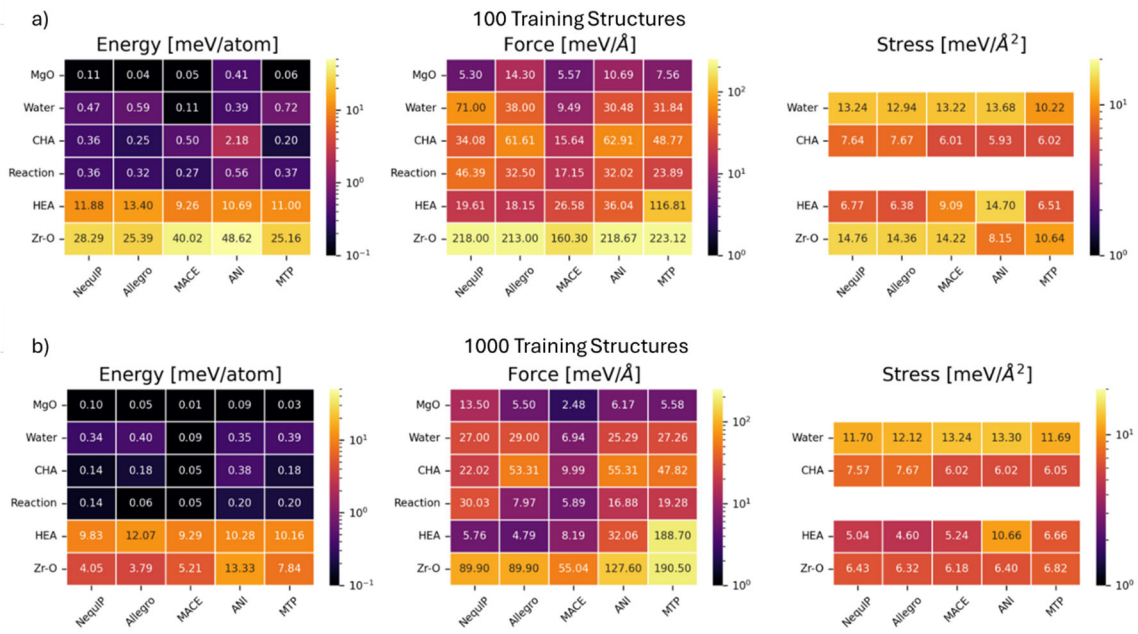


Figure 2. Training errors for 100 (a) and 1000 (b) training structures. Errors are shown in units of meV/atom, meV/Å, meV/Å^2 for energy, force, and stress, respectively, and colored with the inferno matplotlib mapping on a log scale for ease of comparison.

Figure 2 demonstrates very generally that all architectures tend to perform within less than an order of magnitude of each other for a given dataset, with some datasets being generally more difficult and a couple outliers that we address in the following text. The MgO, Water, CHA, and Reaction MLIPs show errors on their respective datasets below the desired thresholds for typical DFT studies (2.5 meV/atom and 35 meV/Å)[118, 119], while HEA and Zr-O are on the high end of error

thresholds we consider acceptable for MLIPs based on literature (5 meV/atom and 50 meV/Å)[90, 120, 121].

The HEA dataset is a classical simulation with a large compositional space, indicating that MLIPs are unable to learn the mathematical formulation of the EAM potential which underlies it. The EAM potential is simple relative to more typical DFT simulations and as such it tests the MLIP's ability to learn simple functional forms in diverse local chemical environments. It is notable that EAM potentials are very easily learned by MLIPs with just one or two elements, such that they are used as training examples in documentation[122]. While the forces are predicted within 35 meV/Å, except for MTP (Force MAE: 189 meV/Å), energies are generally all 4-5X larger than the mentioned threshold of 2.5 meV/atom. The MTP models are particularly interesting to highlight here as the model trained on just 100 points predicts a force error (Force MAE: 117 meV/Å) which is just 61% of the larger training set. The failure of MTP to improve indicates MTP based MLIPs are more sensitive to high elemental composition spaces than other MLIPs and that the interpolation between training points is less stable. For the other architectures, the choice of MLIP architecture appears mostly inconsequential. The difficulty of the HEA dataset was surprising to us as we expected the simple form of the EAM potential to be learnable, but easy training and low errors were not observed using tested MLIPs. It is possible that the EAM potential energy surface is simple enough that a deep learning potential is unneeded and we suggest that simpler kernel based models are considered in future work as they may perform better for this task.

Similarly, Zr-O is a hard dataset for all architectures (Energy MAE: 4-13 meV/atom and Force MAE: 55-190 meV/Å) to train on but the equivariant MLIPs are seen to perform best. The difficulty of the Zr-O dataset comes from the highly disordered structures coming from contour exploration[83] and high temperature MD for a starting structure which was randomly generated. We attribute the Zr-O improvements by equivariance, particularly for MACE (Energy MAE: 5.21 meV/atom and Force MAE: 55 meV/Å), to the improved descriptors of equivariant MLIPs which may capture the disordered structures appearing in the dataset better than the non-equivariant descriptors. The HEA and Zr-O datasets as such represent particular difficulties in MLIPs which can be explored by architecture developers. For example, the HEA dataset failure at the classical level of theory indicates that a DFT based dataset would likely benefit from augmentation by a physical model such as EAM as a physics based underlying potential would greatly simplify the interactions the MLIP must learn. The EAM potential would supply the majority of the repulsive or attractive interactions, with the MLIP only needing to learn deviations from the classical potential similar to previous literature on MLIP augmentation[123].

In the remaining datasets, we note all models perform within our expectation for energy errors not just at 1000 training images as previously described, but even at 100 training set structures as seen in Figure 2a. The observation that 100 training images is often sufficient indicates that a large number of systems relevant to materials science are likely well described by modern MLIP architectures, which aligns well with our observation that MLIPs are becoming routine for use across materials science[32, 49, 50, 87, 124] and the development of foundational models in all types of architectures[105, 125-129].

Data Training Rates

While the models are similar in terms of error, we further check if the accuracy of models as a function of the number of training points (50, 100, 200, 400, 600, 800, 1000) shows any trends

with the architectures explored. Plots of the error with regard to training points on a log-log scale are presented in the SI similar to previous literature[117, 130]. An approximately linear relationship is observed in general and a linear fit with the slope of the line serves as a "data training rate" describing the rate of improvement. The data training rates are presented in Table 8. An example of how a model trains as a function of training set size is presented in Figure 3. Due to the nature of log-log plots, the slope of the fit line is dimensionless and as such the absolute value should not be overinterpreted when comparing between systems or error metrics.

Table 8. The data training rates of each MLIP for energy, force, and stress. Energy, force, and stress is labeled in columns E, F, and S respectively. We note that the fit line is dimensionless.

	MgO		Water			CHA		Reaction		HEA			Zr-O	
	E	F	E	F	S	E	F	E	F	E	F	S	E	F
MACE	0.695	0.391	0.086	0.110	0.020	1.495	0.188	0.838	0.492	0.050	0.519	0.322	0.875	0.541
MTP	0.451	0.268	0.154	0.053	0.025	0.098	0.036	0.317	0.136	0.040	0.219	0.026	0.489	0.639
NequIP	0.063	0.059	0.103	0.004	0.160	0.167	0.161	0.167	0.188	0.066	0.126	0.048	0.876	0.370
Allegro	0.093	0.096	0.118	0.014	0.133	0.127	0.052	0.818	0.629	0.051	0.662	0.065	0.893	0.362
ANI	0.139	0.270	0.105	0.069	0.002	0.091	0.038	0.493	0.313	0.010	0.046	0.021	0.639	0.242

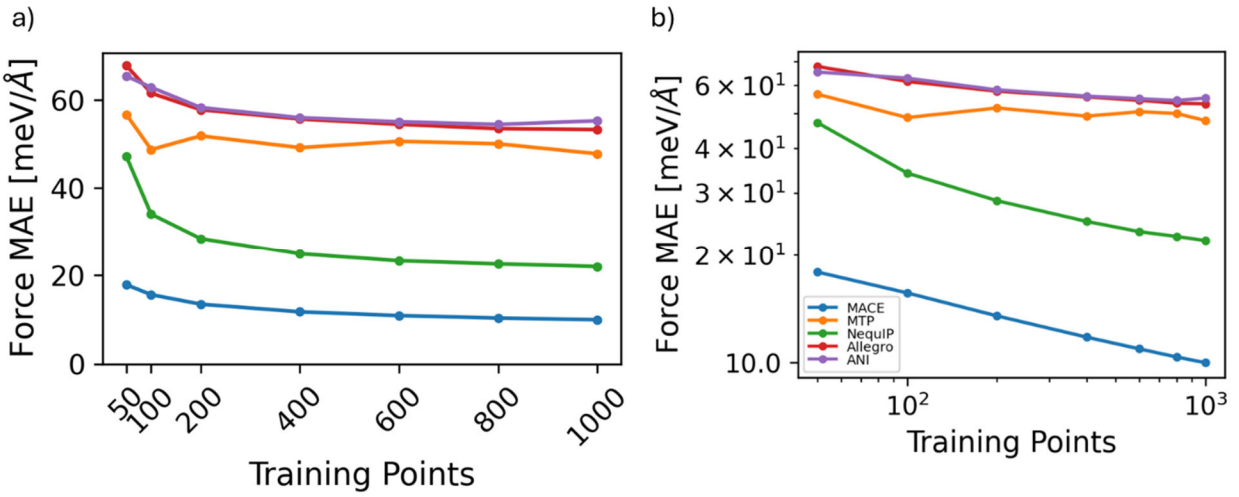


Figure 3. Training errors in force for CHA in a linear (a) and log-log (b) scale. As the number of training points increases, the error decreases non-linearly and follows an approximately log-log relationship. Additional plots for all systems are shown in the SI.

While the models are observed to be similar at 1000 training points in most cases, the data training rates in Table 8 indicate that model performance varies greatly system to system in terms of training data efficiency. We observe an example of how training progresses in Figure 3, where we can see clearly that the 2 best models are equivariant and learning at relatively similar rates, while the non-equivariant models and Allegro learn at lower rates and higher initial errors. Upon conversion of Figure 3a to a log-log scale (seen in Figure 3b) and computing the effective learning

rate, we see that MACE has the highest data learning rate (0.188) and ANI and MTP have the lowest data learning rates (0.038 and 0.036, respectively).

The relative ordering of data learning rates is dataset dependent and also depends on the quantity analyzed. Zr-O gives data learning rates for energy from best to worse in the following order: Allegro, NequIP, MACE, ANI, MTP. In contrast, the data learning rates for force from best to worst are in the following order: MTP, MACE, NequIP, Allegro, ANI. The observation that learning rates are similar across MLIP types (equivariant vs. non-equivariant) is different from some literature which reports that equivariant models are more data efficient.[87, 117, 131] However, the idea of equivariant model data efficiency is related to training on a fixed number of training images typically and does not signify the rate of improvement. Instead, we note that the similarity in training learning rates between equivariant and non-equivariant indicates that non-equivariant models start at a higher error rather than improve slower, thus why they appear less data efficient by the common definition. As equivariant MLIPs tend to be slower to train and infer than non-equivariant MLIPs, discussed later in this work, model distillation[132, 133] from equivariant MLIPs to non-equivariant MLIPs is promising as the amount of data presented to the non-equivariant MLIP can be greatly increased.

System Size Cross Validation

Modern MLIPs are size-extensive in their properties (unless they include global descriptors such as GDML[134, 135]) as predictions are made on a per-atom basis.[84, 136-139] In this section, we show energy, force, and stress errors resulting from similarly transferring between system sizes for the water system (64 molecules vs. 192 molecules) and the MgO system (2x2 cell vs. 4x4 cell) as seen in Figure 4 for 1000 training points. While size-extensivity does not strictly apply to forces directly, the forces or stresses may also deviate from their expected values if the potential energy surface changes as a function of system size. In general, we refer to errors in force as a function of size as size-extensivity in this work.

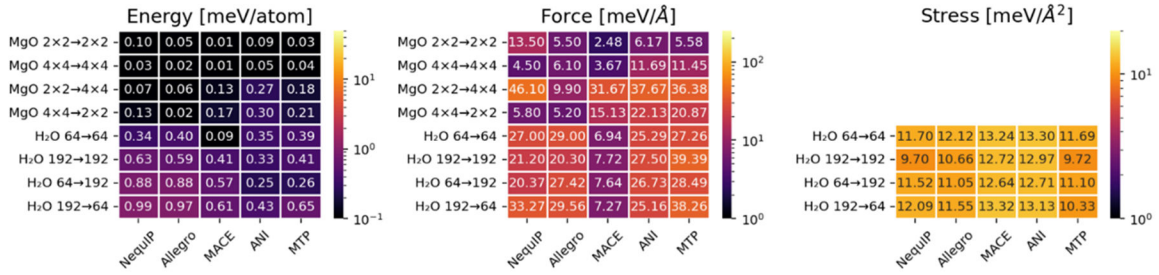


Figure 4. Training errors for the cross-validation of water (64 vs 192 molecules) and MgO (2x2 vs 4x4) with 1000 training points. Errors are shown in units of meV/atom, meV/Å, meV/Å^2 for energy, force, and stress, respectively, and colored with the inferno matplotlib mapping on a log scale to ease comparison.

Consideration of size extensivity is particularly important, as it is common in the literature to train a model on smaller systems and then extrapolate to larger systems that are never directly trained or verified via DFT simulations. Interestingly, the expected size-extensive property of MLIPs is not fully observed in this work as seen in Figure 4. Increasing the system size for MgO from 2x2 to 4x4 degrades predictions of energy and force ranging from 2X to 15X worse, with less error observed in the reverse direction. The lack of size-extensivity is concerning as localized MLIPs

that predict on a per-atom basis were developed largely to solve the size extensivity issue. The Allegro model is an outlier (9.9 meV/Å for 2x2 to 4x4) in the MgO tests and generally shows size extensivity for MgO as transferring to a different size system does not affect error.

Surprisingly, while MACE performs well for almost all other tests in this work, often giving slightly better results than other MLIPs, it produces an error 3X larger than Allegro in this case. We think the failure of size extensivity is due to the forced periodicity and the MgO system is simply too small for the radial cutoff we have chosen. The presence of periodic images of an atom in its own radial cutoff results in the model learning periodic descriptors which do not exist in the larger system or may be ill defined. When MLIPs do not see periodic images within their radial cutoff, the failure of size extensivity is mitigated naturally and no unphysical periodicity is trained into the model. We note however that reducing the radial cutoff itself is not an ideal solution as this likely harms the learning process instead of the error introduced by size extensivity. For the water system, we observe size extensivity for all models where errors are indistinguishable across all tests. The success of the Water MLIPs may indicate that size extensivity is a larger issue in systems with forced periodicity, such as MgO, and that the water systems are already large enough at 64 waters to not have any forced periodicity induced due to system size or cell size. However, the size-extensive behavior of the MLIPs appears acceptable here, as errors stay below the DFT-quality threshold (2.5 meV/atom and 35 meV/Å) except for the MgO 2×2 to 4×4 case. As such, we recommend extra ca<:Lution in training from exceptionally small systems where the radial extents of an atom are on the same scale as the cell dimensions. We recommend testing for size extensivity not only for benchmarking studies but also as part of a general workflow to verify that potentials do not contain errors showing strong size dependence that would lead to unphysical results.

Analysis of Observable Errors

Another important test of transferability is the ability to perform well in more complex tasks such as those involving multiple model inference steps, which can benefit from the phenomenon of error cancellation. Figure 5 highlights trends for chosen observables for each system: lattice constant error for HEAs, volume errors for CHA, and forwards barrier errors for Reaction.

Equation of state tests are performed on 200 HEA structures using the trained MLIPs to determine the ideal lattice at various compositions. Equation of state tests are also performed on the CHA structure using the trained MLIPs to determine the ideal volume of the CHA structure. Equation of state analysis is performed in both cases using a Murnaghan[140] fit of energies and lattices sampled at ± 10% of the EAM volume and fit using the Atomic Simulation Environment (ASE) equation of state module[141]. The Reaction dataset MLIPs are used to perform a Nudged Elastic Band[142-144] (NEB) calculation to compare with the same NEB calculated at the DFT level of theory. We compare the forward barrier of the reaction as given by the NEB spline fitting. The NEB is performed with default parameters taken from ASE starting from the DFT pathway as a starting guess to minimize the effect of the geometry optimizer and assisting in converging to the closest local minimum to the DFT pathway, if it is favorable. We note that some MLIP potentials were unstable for convergence of the NEB and in this case the point is marked with an open circle in Figure 5.

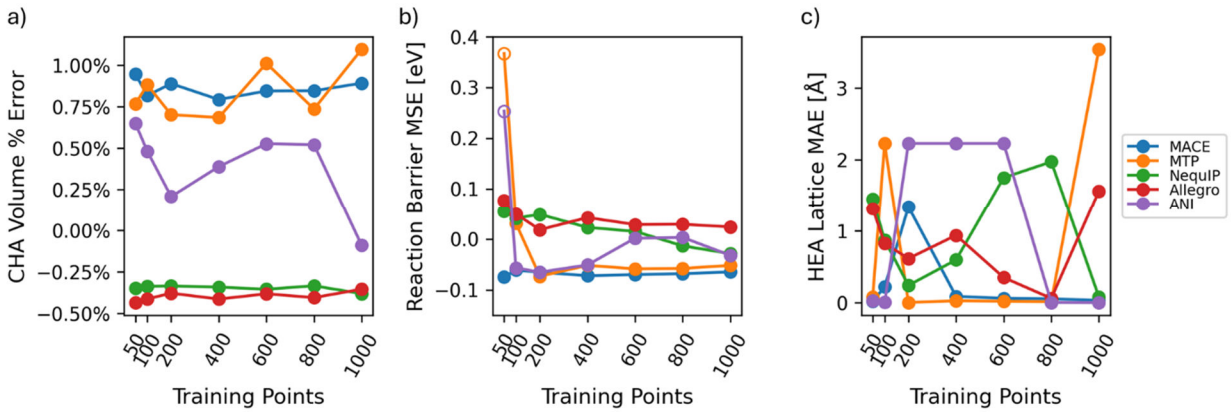


Figure 5. The CHA volume error (a), methane dissociation forward reaction barrier (b), and HEA lattice constant error (c) are shown. The CHA volume error is given as a percentage, methane dissociation barrier in eV, and HEA lattice constant in Å. Open circles represent failed NEB convergence for the reaction barrier tests.

The observable tests in Figure 5 largely indicate that the error we obtain in the energy, forces, and stresses does not fully correlate with the error of observable predictions of a given system; for example, CHA is trained to a stress error of ~6-7 meV/Å^2 but produces good predictions of volume between -0.5% and 1% error. The reverse is observed for the Reaction dataset where ~0.2 meV/atom errors are achieved, but the reaction barrier can be predicted at an error of ±0.06 eV which will be discussed in detail later in this section.

The poor correlation of error metrics to observables is well known in previous literature[88, 145], but not always tested for benchmark datasets as observable tests are not as straightforward in all cases. Additionally, we see that relatively high errors in literature (>100 meV/Å) can sometimes result in observables such as entropy predicted within 2% of the ground truth[146], demonstrating high energy or force errors can still produce useful predictions in some cases. All models appear to produce the same error consistently across all tested numbers of training points (CHA and Reaction) or produce seemingly random errors (HEA), but the error predictions are consistent in nature regardless of MLIP architecture. For example, regardless of the difficulty of the CHA dataset for the different architectures they all produce a volume error (Figure 5a) of less than 1% with ANI surprisingly producing the best prediction as a non-equivariant MLIP while being the worst predictor of energy and force as seen in Figure 2.

Similarly, all methods converge to similar barrier errors (Figure 5b) within ±0.06 eV of the actual barrier of 0.74 eV. While the barrier error may appear acceptable on a per atom basis (1.4 meV/atom), it represents an error of an order of magnitude in reaction rate at room temperature. We note, however, that parameter fitting for microkinetic models is often performed over a broader range[147-149] and this level of error may be negligible when evaluating reaction rates that are being fitted. As such, we consider the reaction barrier test to be a challenge and a potential failure of all MLIPs even though the energy/atom and corresponding barrier errors appear acceptable.

Finally, the HEA lattice constant error is shown in Figure 5c and ranges from acceptable errors of <0.05 Å to completely unreasonable and unphysical errors of >3.5 Å. As the lattice constant should be approximately 3.6 Å for most structures, the error indicates predictions are of either fully

collapsed systems or exploded in many cases similar to well-known failure modes observed in MLIPs under MD conditions or in geometry optimizations. The existence of MLIPs for all architectures which provide good predictions of lattice constant (see 800 training points for MACE, MTP, and Allegro or 1000 training points for NequIP and ANI) indicate that there is a non-systematic problem in the architectures that results in unstable potentials, but it is inconsistent and changes on a run-to-run basis.

Zeolite Transferability

As the structure of a zeolite can inform an MLIP about other frameworks in principle, we probe the transferability of the CHA dataset MLIPs to an extended test set of MFI, FAU, and LTA zeolite frameworks in Figure 6.

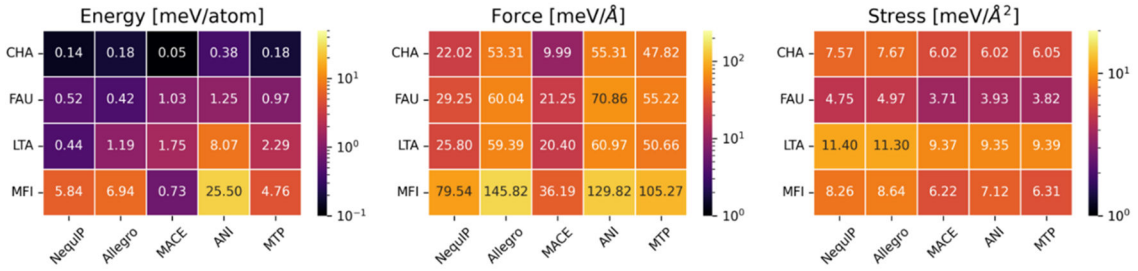


Figure 6. Training errors for the transferability of zeolite frameworks with 1000 training points, going from CHA to FAU, LTA, and MFI. Errors are shown in units of meV/atom, meV/Å, meV/Å^2 for energy, force, and stress, respectively, and colored with the inferno matplotlib mapping on a log scale to ease comparison.

Analysis of zeolite transferability suggests that energies are the hardest to predict, forces can be problematic for some frameworks, but stresses are predicted fairly reliably. We note that zeolite transferability results may be influenced by optimization processes or loss functions as energy, force, and stress are optimized concurrently as forces and stresses are produced by an automatic differentiation of energy in all tested MLIPs. MLIPs calculating forces and stresses by automatical differentiation of energy are known as conservative and non-conservative MLIPs may be a potential solution for future work if lower errors are favored over the conservative restriction. For example, CHA's best MLIP across all architectures achieves an error of just 0.05 meV/atom for MACE but increases by at least an order of magnitude for all other frameworks, although still within our acceptable range (2.5 meV/atom and 35 meV/Å).

In particular, MFI is observed to be exceedingly difficult for the MLIPs to transfer to. MFI transferability issues are likely due to the larger rings present in MFI (10 membered rings) vs CHA (8 membered rings) which contain unseen bond angles. LTA contains largely 8 membered rings, giving it some similarity to CHA. FAU on the other hand has 12 membered rings, which makes it surprising that the MLIP transfers well. One reason for the success of CHA to FAU transferability may also be the more cage-like nature of CHA and FAU, whereas MFI has one dimensional channels that are not cage-like.

In terms of force error, it is clear that MACE performs best, only increasing by a factor of 2 for the FAU and LTA and a factor of 3.5 for MFI. As the errors were already lower for MACE than any other framework for CHA, the transferability to MFI is within the DFT-like acceptable error

threshold (35 meV/Å). Our transferability results indicate that only MACE produces acceptably transferable zeolite results, but caution should be applied when considering transferability from other frameworks as our results may be specific to the frameworks chosen in this work. When looking at stress errors, confusingly the FAU is predicted more accurately (~4 meV/Å^2) than the CHA is (~6.5 meV/Å^2). All architectures also produce similarly changing stress errors error relative to the CHA error. The stress errors we observe indicates to us that stress errors are not as sensitive to transferability tests.

Inference Speed Tests

The inference speed of all MLIPs is tested using LAMMPS for a representative system of the dataset being inferred. The relative speed of each model in kAtoms/sec is shown in Figure 7. Treat the reported numbers with caution as different CPU / GPU hardware may influence obtained results considerably, as well as results can be strongly affected by improvements in the model inference calculations via PyTorch or internally to the package. As a result, results are only representative of the versions of code used in this work and we encourage readers to use the newest versions if they wish to perform similar speed analysis for their systems.

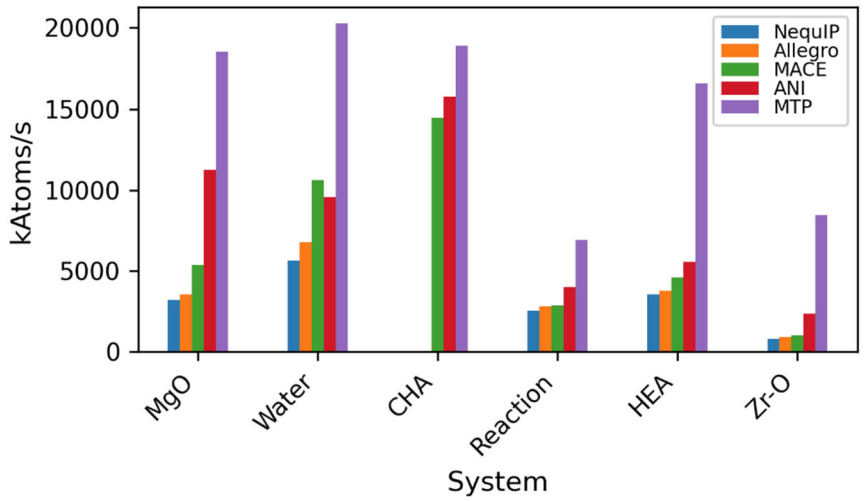


Figure 7. The inference performance of the MLIPs for each system and each architecture as trained. All codes are evaluated via LAMMPS and measured in terms of kAtoms/sec to account for differences in the size of each system. A100s are used in all cases to provide GPU acceleration. Treat the reported numbers with caution as different CPU / GPU hardware may influence obtained results considerably.

The performance tests in Figure 7 demonstrate a simple relationship between the different MLIP architectures with performances ranging from 1,000 katoms/s to 20,000 katoms/s. The equivariant MLIPs are typically within 10% of the speed of the other equivariant models, while the non-equivariant models are often double the performance. MTP is the fastest MLIP we tested with ANI being second. We then observe that MACE is often the next fastest MLIP if an equivariant MLIP is desired. While Allegro was designed for high performance with parallel GPUs[150], the systems tested here would not benefit from multi-GPU acceleration and we do not evaluate it here. As such the Allegro framework is not significantly faster than NequIP in any case, remaining within 5% of the NequIP performance in all cases except Water.

We can however produce faster or slower potentials for a given MLIP at the expense or benefit of accuracy, but we have chosen to prioritize accuracy in this work. For example, we can train non-equivariant models of NequIP and these will be faster than our equivariant models. The inference performance shown here is indicative of our chosen hyperparameters and we stress that system size, software versions, and hardware can all affect the performance observed. Additionally, our systems are still sized around typical DFT sizes (<500 atoms) and the inference speeds are likely to increase as the systems grow and GPU resources can be leveraged more effectively. With that considered, we suggest that equivariant models must improve in inference speed if they wish to become the undisputable choice for researchers utilizing MLIPs as time and resource constraints of model inference must be considered and weighed against dataset generation costs and accuracy.

CONCLUSIONS

This benchmark highlights the strengths and current limitations of machine learning interatomic potentials (MLIPs) across a diverse set of materials science-relevant systems. We find that, for well-optimized models, the modern MLIPs in this work (MACE, NequIP, Allegro, Torch-ANI, and MTP) yield comparable errors in energy, forces, and stress in the studied simpler tests (MgO, Water, CHA, and Reaction) with the HEA and Zr-O showing a reduction in error (1.5 – 2x improvement) for equivariant MLIPs.

Our results also reveal that elemental complexity and structural disorder, particularly in the HEA and Zr–O datasets, remain difficult for small training sets (<1000 structures). The HEA dataset relies on a simple underlying classical potential, which suggests there is difficulty in learning even a simple potential if given a complex enough local chemical environment. Size-extensivity tests underscore the importance of verifying this property explicitly, even in MLIPs that should be extensive by design. Observables derived from MLIPs can be unreliable even with small increases in per-atom error, reinforcing the need to test not just energies and forces, but the actual physical quantities of interest. We think observables in MLIP-based simulations must be assumed to be incorrect until tested and as many testable observables should be confirmed during production level research involving MLIPs, unlike DFT where we assume the observables are correct inherently. Importantly, we find no clear advantage in data efficiency or observable errors between equivariant and non-equivariant models. This suggests that further benchmarking should move beyond small differences in loss curves and target systems where model architectures fail to reproduce key behaviors. Semi-classical potentials (e.g., EAM[151], Tersoff[152], ReaxFF[153]) may be particularly useful for this purpose, either as baselines or augmentation tools, due to their stronger links to observables and forced physical behavior. For researchers looking to choose an MLIP architecture, we suggest selecting equivariant MLIP architectures if the complexity of the system is a challenge but for simple materials problems auxiliary features such as integration with molecular dynamics engines, trade-offs between computational dataset generation cost vs. MLIP inference speed, and framework integration may play a more important decision factor than small differences in error metrics.

Going forward, benchmarking new MLIPs should prioritize the following aspects identified in this work: (i) transferability tests to out-of-distribution datasets and to confirm expected properties such as size extensivity, (ii) hard datasets where failure modes emerge and differences between MLIPs are significant, and (iii) validation of observables rather than assuming correct predictions from energy/force agreement alone. The identification of the HEA and Zr-O datasets as difficult

datasets that appear to favor equivariant MLIPs suggests these downloadable datasets should be used for future work on benchmarking developments around equivariance (or the lack thereof) and new architectures. Development of similarly difficult datasets that encompass a wide range of compositions and at various levels of theory (classical, semi-empirical, DFT) will benefit the MLIP development community. Simple and low complexity systems appear to be a solved problem for MLIPs in general as we reach a typical energy error of ~0.2 meV/atom and ~30 meV/Å with just 1000 training images when the elemental composition does not change, implying these MLIPs would be usable for production level research projects. We strongly recommend reframing the benchmark goal in MLIP work: not to rank model architectures by small and often insignificant differences in error, but to instead identify in what modeling space (material, composition, and complexity) and why the architectures fail as this will be critical for developing robust, transferable MLIPs for real-world applications.

ACKNOWLEDGEMENTS

Solid-liquid interface related work is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, CPIMS program under grant number DE-SC0024654. Thermal catalysis related work is funded by the National Science Foundation (NSF) under grant number 2245120. The authors thank Sophia Ezendu, Gbolagade Olajide, Mustapha Iddrisu, and Ademola Soyemi for their insightful comments on the manuscript and work. Additionally, we thank Michael Waters for his assistance with the Zr-O dataset. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship, under Award DE-SC0023112. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract DE-AC02-05CH11231 using NERSC Awards BES-ERCAP0024218 and ASCR-ERCAP0033223. B. W. J. C. is grateful for the high-performance computational facilities provided by the National Supercomputing Centre (NSCC) Singapore and the A*STAR Computational Resource Centre (A*CRC). Any opinions, findings, conclusions, and/or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of funding agencies.

SUPPORTING INFORMATION

Additional validation scripts and datasets are available free of charge upon publication via the Internet at https://pubs.acs.org/jcim and at Zenodo:

https://doi.org/10.5281/zenodo.10901820

The Supporting Information includes:

- All plots of the validation of the datasets for all 3 splits and mean results (PDF)

AUTHOR CONTRIBUTIONS

Tristan Maxson: Conceptualization, Data curation, Formal analysis, Investigation, MLIP Training, Software, Writing – original draft.

Ademola Soyemi: Formal analysis, MLIP Training, Writing – review & editing.

Xinglong Zhang: MLIP Training.

Benjamin Wei Jie: Formal analysis, MLIP Training, Writing – review & editing.

Tibor Szilvasi: Conceptualization, Supervision, Formal analysis, Writing – review & editing, Funding acquisition, Project administration, Corresponding author.

CONFLICT OF INTEREST

The authors declare no competing financial interest.

REFERENCES

(1) Razali, N. A. M.; Lee, K. T.; Bhatia, S.; Mohamed, A. R. Heterogeneous catalysts for production of chemicals using carbon dioxide as raw material: A review. *Renewable and sustainable energy reviews* **2012**, *16* (7), 4951-4964.
(2) Avhad, M.; Marchetti, J. A review on recent advancement in catalytic materials for biodiesel production. *Renewable and sustainable energy reviews* **2015**, *50*, 696-718.
(3) Luo, Z.; Zhao, G.; Pan, H.; Sun, W. Strong metal–support interaction in heterogeneous catalysts. *Advanced Energy Materials* **2022**, *12* (37), 2201395.
(4) Datye, A. K.; Votsmeier, M. Opportunities and challenges in the development of advanced materials for emission control catalysts. *Nature Materials* **2021**, *20* (8), 1049-1059.
(5) Tressler, J. F.; Alkoy, S.; Newnham, R. E. Piezoelectric sensors and sensor materials. *Journal of electroceramics* **1998**, *2*, 257-272.
(6) Dhall, S.; Mehta, B.; Tyagi, A.; Sood, K. A review on environmental gas sensors: Materials and technologies. *Sensors International* **2021**, *2*, 100116.
(7) Nikolic, M. V.; Milovanovic, V.; Vasiljevic, Z. Z.; Stamenkovic, Z. Semiconductor gas sensors: Materials, technology, design, and application. *Sensors* **2020**, *20* (22), 6694.
(8) Dhilipan, J.; Vijayalakshmi, N.; Shanmugam, D.; Ganesh, R. J.; Kodeeswaran, S.; Muralidharan, S. Performance and efficiency of different types of solar cell material–A review. *Materials Today: Proceedings* **2022**, *66*, 1295-1302.
(9) Devadiga, D.; Selvakumar, M.; Shetty, P.; Santosh, M. Recent progress in dye sensitized solar cell materials and photo-supercapacitors: A review. *Journal of Power Sources* **2021**, *493*, 229698.
(10) Li, X.; Li, P.; Wu, Z.; Luo, D.; Yu, H.-Y.; Lu, Z.-H. Review and perspective of materials for flexible solar cells. *Materials Reports: Energy* **2021**, *1* (1), 100001.
(11) Tacey, S. A.; Chen, B. W.; Szilvási, T.; Mavrikakis, M. An automated cluster surface scanning method for exploring reaction paths on metal-cluster surfaces. *Computational Materials Science* **2021**, *186*, 110010.
(12) Kroes, G.-J. Computational approaches to dissociative chemisorption on metals: towards chemical accuracy. *Physical Chemistry Chemical Physics* **2021**, *23* (15), 8962-9048.
(13) Powell, A. D.; Kroes, G.-J.; Doblhoff-Dier, K. Quantum Monte Carlo calculations on dissociative chemisorption of H2+ Al (110): minimum barrier heights and their comparison to DFT values. *The Journal of Chemical Physics* **2020**, *153* (22).

(14) Rybkin, V. V. Sampling Potential Energy Surfaces in the Condensed Phase with Many-Body Electronic Structure Methods. *Chemistry–A European Journal* **2020**, *26* (2), 362-368.
(15) Wei, Q.; Chang, T.; Zeng, R.; Cao, S.; Zhao, J.; Han, X.; Wang, L.; Zou, B. Self-trapped exciton emission in a zero-dimensional (TMA) 2SbCl5· DMF single crystal and molecular dynamics simulation of structural stability. *The Journal of Physical Chemistry Letters* **2021**, *12* (30), 7091-7099.
(16) Mortazavi, B.; Novikov, I. S.; Podryabinkin, E. V.; Roche, S.; Rabczuk, T.; Shapeev, A. V.; Zhuang, X. Exploring phononic properties of two-dimensional materials using machine learning interatomic potentials. *Applied Materials Today* **2020**, *20*, 100685.
(17) Ciccotti, G.; Ferrario, M. Blue moon approach to rare events. *Molecular simulation* **2004**, *30* (11-12), 787-793.
(18) Wang, J.; Deng, Y.; Roux, B. Absolute binding free energy calculations using molecular dynamics simulations with restraining potentials. *Biophysical journal* **2006**, *91* (8), 2798-2814.
(19) Stocker, S.; Jung, H.; Csányi, G.; Goldsmith, C. F.; Reuter, K.; Margraf, J. T. Estimating free energy barriers for heterogeneous catalytic reactions with machine learning potentials and umbrella integration. *Journal of Chemical Theory and Computation* **2023**, *19* (19), 6796-6804.
(20) Foroutan-Nejad, C.; Marek, R. Potential energy surface and binding energy in the presence of an external electric field: modulation of anion–π interactions for graphene-based receptors. *Physical Chemistry Chemical Physics* **2014**, *16* (6), 2508-2514.
(21) Shan, B.; Zhao, Y.; Hyun, J.; Kapur, N.; Nicholas, J. B.; Cho, K. Coverage-dependent CO adsorption energy from first-principles calculations. *The Journal of Physical Chemistry C* **2009**, *113* (15), 6088-6092.
(22) Greeley, J.; Nørskov, J. K. A general scheme for the estimation of oxygen binding energies on binary transition metal surface alloys. *Surface science* **2005**, *592* (1-3), 104-111.
(23) Galimberti, D. R.; Sauer, J. Chemically accurate vibrational free energies of adsorption from density functional theory molecular dynamics: Alkanes in zeolites. *Journal of chemical theory and computation* **2021**, *17* (9), 5849-5862.
(24) Bernardino, K.; Ribeiro, M. C. Role of density and electrostatic interactions in the viscosity and non-newtonian behavior of ionic liquids–a molecular dynamics study. *Physical Chemistry Chemical Physics* **2022**, *24* (11), 6866-6879.
(25) Szala-Bilnik, J.; Abedini, A.; Crabtree, E.; Bara, J. E.; Turner, C. H. Molecular transport behavior of CO2 in ionic polyimides and ionic liquid composite membrane materials. *The Journal of Physical Chemistry B* **2019**, *123* (34), 7455-7463.
(26) Ensing, B.; De Vivo, M.; Liu, Z.; Moore, P.; Klein, M. L. Metadynamics as a tool for exploring free energy landscapes of chemical reactions. *Accounts of chemical research* **2006**, *39* (2), 73-81.
(27) Zubieta Rico, P. F.; Schneider, L.; Pérez-Lemus, G. R.; Alessandri, R.; Dasetty, S.; Nguyen, T. D.; Menéndez, C. A.; Wu, Y.; Jin, Y.; Xu, Y. PySAGES: Flexible, advanced sampling methods accelerated with GPUs. *npj Computational Materials* **2024**, *10* (1), 35.
(28) Dawson, W.; Gygi, F. Equilibration and analysis of first-principles molecular dynamics simulations of water. *The Journal of chemical physics* **2018**, *148* (12).
(29) Crabb, E.; France-Lanord, A.; Leverick, G.; Stephens, R.; Shao-Horn, Y.; Grossman, J. C. Importance of equilibration method and sampling for ab Initio molecular dynamics simulations of solvent–lithium-salt systems in lithium-oxygen batteries. *Journal of Chemical Theory and Computation* **2020**, *16* (12), 7255-7266.
(30) Sours, T. G.; Kulkarni, A. R. Predicting structural properties of pure silica zeolites using deep neural network potentials. *The Journal of Physical Chemistry C* **2023**, *127* (3), 1455-1463.
(31) Liu, Y.-B.; Yang, J.-Y.; Xin, G.-M.; Liu, L.-H.; Csányi, G.; Cao, B.-Y. Machine learning interatomic potential developed for molecular simulations on thermal properties of β-Ga2O3. *The Journal of Chemical Physics* **2020**, *153* (14).
(32) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Physical Review X* **2018**, *8* (4), 041048.

(33) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Physical Review B* **2017**, *95* (9), 094203.
(34) Deringer, V. L.; Bernstein, N.; Csányi, G.; Ben Mahmoud, C.; Ceriotti, M.; Wilson, M.; Drabold, D. A.; Elliott, S. R. Origins of structural and electronic transitions in disordered silicon. *Nature* **2021**, *589* (7840), 59-64. DOI: 10.1038/s41586-020-03072-z.
(35) Kapil, V.; Schran, C.; Zen, A.; Chen, J.; Pickard, C. J.; Michaelides, A. The first-principles phase diagram of monolayer nanoconfined water. *Nature* **2022**, *609* (7927), 512-516. DOI: 10.1038/s41586-022-05036-x.
(36) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **2019**, *32*, 8026-8037.
(37) Wang, H.; Zhang, L.; Han, J.; E, W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178-184.
(38) Batatia, I.; Kov'acs, D. a. P. e.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *ArXiv* **2022**, *abs/2206.07697*. (acccessed 2/19/24).
(39) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications* **2022**, *13* (1), 2453. DOI: 10.1038/s41467-022-29939-5.
(40) Gao, X.; Ramezanghorbani, F.; Isayev, O.; Smith, J. S.; Roitberg, A. E. TorchANI: A Free and Open Source PyTorch-Based Deep Learning Implementation of the ANI Neural Network Potentials. *Journal of Chemical Information and Modeling* **2020**, *60* (7), 3408-3415. DOI: 10.1021/acs.jcim.0c00451.
(41) Jinnouchi, R.; Karsai, F.; Kresse, G. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B* **2019**, *100* (1), 014105. DOI: 10.1103/PhysRevB.100.014105.
(42) Khorshidi, A.; Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Computer Physics Communications* **2016**, *207*, 310-324. DOI: https://doi.org/10.1016/j.cpc.2016.05.010.
(43) Zeng, J.; Zhang, D.; Lu, D.; Mo, P.; Li, Z.; Chen, Y.; Rynik, M.; Huang, L. a.; Li, Z.; Shi, S.; et al. DeePMD-kit v2: A software package for deep potential models. *The Journal of Chemical Physics* **2023**, *159* (5), 054801. (acccessed 3/22/2024).
(44) Unke, O. T.; Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of chemical theory and computation* **2019**, *15* (6), 3678-3693.
(45) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation* **2016**, *14* (3), 1153-1173.
(46) Bang, K.; Yeo, B. C.; Kim, D.; Han, S. S.; Lee, H. M. Accelerated mapping of electronic density of states patterns of metallic nanoparticles via machine-learning. *Scientific Reports* **2021**, *11* (1), 11604. DOI: 10.1038/s41598-021-91068-8.
(47) Sauceda, H. E.; Gálvez-González, L. E.; Chmiela, S.; Paz-Borbón, L. O.; Müller, K.-R.; Tkatchenko, A. BIGDML—Towards accurate quantum machine learning force fields for materials. *Nature Communications* **2022**, *13* (1), 3733. DOI: 10.1038/s41467-022-31093-x.
(48) Mortazavi, B.; Zhuang, X.; Rabczuk, T.; Shapeev, A. V. Atomistic modeling of the mechanical properties: the rise of machine learning interatomic potentials. *Materials Horizons* **2023**, *10* (6), 1956-1968.
(49) Mishin, Y. Machine-learning interatomic potentials for materials science. *Acta Materialia* **2021**, *214*, 116980.
(50) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials* **2019**, *31* (46), 1902765.
(51) Morrow, J. D.; Gardner, J. L.; Deringer, V. L. How to validate machine-learned interatomic potentials. *The Journal of Chemical Physics* **2023**, *158* (12).
(52) Maxson, T.; Soyemi, A.; Chen, B. W.; Szilvási, T. Enhancing the Quality and Reliability of Machine Learning Interatomic Potentials through Better Reporting Practices. *The Journal of Physical Chemistry C* **2024**.

(53) Dong, H.; Shi, Y.; Ying, P.; Xu, K.; Liang, T.; Wang, Y.; Zeng, Z.; Wu, X.; Zhou, W.; Xiong, S. Molecular dynamics simulations of heat transport using machine-learned potentials: A mini review and tutorial on GPUMD with neuroevolution potentials. *arXiv preprint arXiv:2401.16249* **2024**.
(54) Ghaffari, K.; Bavdekar, S.; Spearot, D. E.; Subhash, G. Validation Workflow for Machine Learning Interatomic Potentials for Complex Ceramics. *arXiv preprint arXiv:2402.05222* **2024**.
(55) Bandi, S.; Jiang, C.; Marianetti, C. A. Benchmarking phonon anharmonicity in machine learning interatomic potentials. *arXiv preprint arXiv:2402.18891* **2024**.
(56) Bihani, V.; Mannan, S.; Pratiush, U.; Du, T.; Chen, Z.; Miret, S.; Micoulaut, M.; Smedskjaer, M. M.; Ranu, S.; Krishnan, N. A. EGraFFBench: evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery* **2024**, *3* (4), 759-768.
(57) Kovács, D. P.; Batatia, I.; Arany, E. S.; Csányi, G. Evaluation of the MACE force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics* **2023**, *159* (4), 044118. DOI: 10.1063/5.0155322 (acccessed 4/2/2025).
(58) Morrow, J. D.; Gardner, J. L. A.; Deringer, V. L. How to validate machine-learned interatomic potentials. *The Journal of Chemical Physics* **2023**, *158* (12), 121501. DOI: 10.1063/5.0139611 (acccessed 4/2/2025).
(59) Žugec, I.; Geilhufe, R. M.; Lončarić, I. Global machine learning potentials for molecular crystals. *The Journal of Chemical Physics* **2024**, *160* (15), 154106. DOI: 10.1063/5.0196232 (acccessed 4/2/2025).
(60) Stark, W. G.; van der Oord, C.; Batatia, I.; Zhang, Y.; Jiang, B.; Csányi, G.; Maurer, R. J. Benchmarking of machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces. *Machine Learning: Science and Technology* **2024**, *5* (3), 030501.
(61) Chmiela, S.; Vassilev-Galindo, V.; Unke, O. T.; Kabylda, A.; Sauceda, H. E.; Tkatchenko, A.; Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances* **2023**, *9* (2), eadf0873. DOI: doi:10.1126/sciadv.adf0873.
(62) Tran, R.; Lan, J.; Shuaibi, M.; Wood, B. M.; Goyal, S.; Das, A.; Heras-Domingo, J.; Kolluru, A.; Rizvi, A.; Shoghi, N. The Open Catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis* **2023**, *13* (5), 3066-3084.
(63) Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W. Open catalyst 2020 (OC20) dataset and community challenges. *Acs Catalysis* **2021**, *11* (10), 6059-6072.
(64) Stark, W. G.; van der Oord, C.; Batatia, I.; Zhang, Y.; Jiang, B.; Csányi, G.; Maurer, R. J. Benchmarking of machine learning interatomic potentials for reactive hydrogen dynamics at metal surfaces. *arXiv preprint arXiv:2403.15334* **2024**.
(65) Waters, M. J.; Rondinelli, J. M. Benchmarking structural evolution methods for training of machine learned interatomic potentials. *Journal of Physics: Condensed Matter* **2022**, *34* (38), 385901.
(66) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J. r.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A. Performance and cost assessment of machine learning interatomic potentials. *The Journal of Physical Chemistry A* **2020**, *124* (4), 731-745.
(67) Fedik, N.; Zubatyuk, R.; Kulichenko, M.; Lubbers, N.; Smith, J. S.; Nebgen, B.; Messerly, R.; Li, Y. W.; Boldyrev, A. I.; Barros, K. Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nature Reviews Chemistry* **2022**, *6* (9), 653-672.
(68) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9* (2), 513-530.
(69) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **2014**, *1* (1), 1-7.
(70) Chen, G.; Chen, P.; Hsieh, C.-Y.; Lee, C.-K.; Liao, B.; Liao, R.; Liu, W.; Qiu, J.; Sun, Q.; Tang, J. Alchemy: A quantum chemistry dataset for benchmarking ai models. *arXiv preprint arXiv:1906.09427* **2019**.

(71) Riebesell, J.; Goodall, R. E.; Benner, P.; Chiang, Y.; Deng, B.; Lee, A. A.; Jain, A.; Persson, K. A. Matbench Discovery--A framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920* **2023**.
(72) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **2013**, *1* (1).
(73) Focassio, B.; M. Freitas, L. P.; Schleder, G. R. Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Applied Materials & Interfaces* **2024**.
(74) Yu, H.; Giantomassi, M.; Materzanini, G.; Wang, J.; Rignanese, G. M. Systematic assessment of various universal machine-learning interatomic potentials. *Materials Genome Engineering Advances* **2024**, *2* (3), e58.
(75) Tran, R.; Lan, J.; Shuaibi, M.; Wood, B. M.; Goyal, S.; Das, A.; Heras-Domingo, J.; Kolluru, A.; Rizvi, A.; Shoghi, N.; et al. The Open Catalyst 2022 (OC22) Dataset and Challenges for Oxide Electrocatalysts. *ACS Catalysis* **2023**, *13* (5), 3066-3084. DOI: 10.1021/acscatal.2c05426.
(76) Duriez, C.; Chapon, C.; Henry, C.; Rickard, J. Structural characterization of MgO (100) surfaces. *Surface Science* **1990**, *230* (1-3), 123-136.
(77) Wan, K.; He, J.; Shi, X. Construction of high accuracy machine learning interatomic potential for surface/interface of nanomaterials—A review. *Advanced Materials* **2024**, *36* (22), 2305758.
(78) Naber, J.; De Jong, K.; Stork, W.; Kuipers, H.; Post, M. Industrial applications of zeolite catalysis. In *Studies in surface science and catalysis*, Vol. 84; Elsevier, 1994; pp 2197-2219.
(79) Almomani, B.; Banisalman, M. J.; Elgack, O.; Syarif, J. Effects of alloying and grain boundary on primary irradiation defects in FeNiCrCoCu high entropy alloys: A molecular dynamics study. *Materials Today Communications* **2025**, *45*, 112237.
(80) Mirzoev, A.; Gelchinski, B.; Rempel, A. Neural network prediction of interatomic interaction in multielement substances and high-entropy alloys: a review. In *Doklady Physical Chemistry*, 2022; Springer: Vol. 504, pp 51-77.
(81) Cakir, Z.; Hu, L.; Wang, C.; Greeley, J. First-Principles Analysis of the Ammonia Decomposition Reaction on High Entropy Alloy Catalysts. In *The 27th North American Catalysis Society Meeting*, 2022; NAM.
(82) Deshmukh, G.; Wichrowski, N. J.; Evangelou, N.; Ghanekar, P. G.; Deshpande, S.; Kevrekidis, I. G.; Greeley, J. Active learning of ternary alloy structures and energies. *npj Computational Materials* **2024**, *10* (1), 116.
(83) Waters, M. J.; Rondinelli, J. M. Energy contour exploration with potentiostatic kinematics. *Journal of Physics: Condensed Matter* **2021**, *33* (44), 445901.
(84) Musaelian, A.; Batzner, S.; Johansson, A.; Sun, L.; Owen, C. J.; Kornbluth, M.; Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications* **2023**, *14* (1), 579.
(85) Batatia, I.; Kovacs, D. P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems* **2022**, *35*, 11423-11436.
(86) Zhai, Y.; Caruso, A.; Bore, S. L.; Luo, Z.; Paesani, F. A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions? *The Journal of Chemical Physics* **2023**, *158* (8), 084111. DOI: 10.1063/5.0142843 (acccessed 4/2/2025).
(87) Maxson, T.; Szilvási, T. Transferable water potentials using equivariant neural networks. *The Journal of Physical Chemistry Letters* **2024**, *15* (14), 3740-3747.
(88) Fu, X.; Wu, Z.; Wang, W.; Xie, T.; Keten, S.; Gomez-Bombarelli, R.; Jaakkola, T. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237* **2022**.

(89) Miret, S.; Lee, K. L. K.; Gonzales, C.; Mannan, S.; Krishnan, N. Energy & Force Regression on DFT Trajectories is Not Enough for Universal Machine Learning Interatomic Potentials. *arXiv preprint arXiv:2502.03660* **2025**.
(90) Zhou, W.; Liang, N.; Wu, X.; Xiong, S.; Fan, Z.; Song, B. Insight into the effect of force error on the thermal conductivity from machine-learned potentials. *Materials Today Physics* **2025**, *50*, 101638.
(91) Kresse, G.; Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B* **1996**, *54* (16), 11169.
(92) Kresse, G.; Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b* **1999**, *59* (3), 1758.
(93) Kresse, G.; Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational materials science* **1996**, *6* (1), 15-50.
(94) Kresse, G.; Hafner, J. Ab initio molecular dynamics for liquid metals. *Physical review B* **1993**, *47* (1), 558.
(95) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **1996**, *77* (18), 3865.
(96) Moellmann, J.; Grimme, S. DFT-D3 study of some molecular crystals. *The Journal of Physical Chemistry C* **2014**, *118* (14), 7615-7621.
(97) Evans, D. J.; Holian, B. L. The nose–hoover thermostat. *The Journal of chemical physics* **1985**, *83* (8), 4069-4074.
(98) Monkhorst, H. J.; Pack, J. D. Special points for Brillouin-zone integrations. *Physical review B* **1976**, *13* (12), 5188.
(99) Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F. CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **2020**, *152* (19).
(100) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Physical review B* **1999**, *59* (11), 7413.
(101) Goedecker, S.; Teter, M.; Hutter, J. Separable dual-space Gaussian pseudopotentials. *Physical Review B* **1996**, *54* (3), 1703.
(102) Li, W.-L.; Chen, K.; Rossomme, E.; Head-Gordon, M.; Head-Gordon, T. Optimized pseudopotentials and basis sets for semiempirical density functional theory for electrocatalysis applications. *The Journal of Physical Chemistry Letters* **2021**, *12* (42), 10304-10309.
(103) VandeVondele, J.; Hutter, J. An efficient orbital transformation method for electronic structure calculations. *The Journal of chemical physics* **2003**, *118* (10), 4365-4369.
(104) Bussi, G.; Parrinello, M. Accurate sampling using Langevin dynamics. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics* **2007**, *75* (5), 056707.
(105) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J. A foundation model for atomistic materials chemistry. *arXiv preprint arXiv:2401.00096* **2023**.
(106) Caldeweyher, E.; Mewes, J.-M.; Ehlert, S.; Grimme, S. Extension and evaluation of the D4 London-dispersion model for periodic systems. *Physical Chemistry Chemical Physics* **2020**, *22* (16), 8499-8512.
(107) Kräutler, V.; Van Gunsteren, W. F.; Hünenberger, P. H. A fast SHAKE algorithm to solve distance constraint equations for small molecules in molecular dynamics simulations. *Journal of computational chemistry* **2001**, *22* (5), 501-508.
(108) Johnson, R. Alloy models with the embedded-atom method. *Physical Review B* **1989**, *39* (17), 12554.
(109) Daw, M. S.; Foiles, S. M.; Baskes, M. I. The embedded-atom method: a review of theory and applications. *Materials Science Reports* **1993**, *9* (7-8), 251-310.
(110) Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nosé–Hoover chains: The canonical ensemble via continuous dynamics. *The Journal of chemical physics* **1992**, *97* (4), 2635-2643.

(111) Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* **2016**.
(112) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **2011**, *134* (7), 074106. DOI: 10.1063/1.3553717 (acccessed 7/20/2025).
(113) Jung, G. S.; Myung, H.; Irle, S. Artificial neural network potentials for mechanics and fracture dynamics of two-dimensional crystals. *Machine Learning: Science and Technology* **2023**, *4* (3), 035001.
(114) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**.
(115) Cortes, C.; Jackel, L. D.; Solla, S.; Vapnik, V.; Denker, J. Learning curves: Asymptotic values and rate of convergence. *Advances in neural information processing systems* **1993**, *6*.
(116) Wen, M.; Huang, W.-F.; Dai, J.; Adhikari, S. Cartesian Atomic Moment Machine Learning Interatomic Potentials. *arXiv preprint arXiv:2411.12096* **2024**.
(117) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* **2022**, *13* (1), 2453.
(118) Owen, C. J.; Xie, Y.; Johansson, A.; Sun, L.; Kozinsky, B. Low-index mesoscopic surface reconstructions of Au surfaces using Bayesian force fields. *Nature Communications* **2024**, *15* (1), 3790.
(119) Liu, Y.; He, X.; Mo, Y. Discrepancies and error evaluation metrics for machine learning interatomic potentials. *npj Computational Materials* **2023**, *9* (1), 174.
(120) Radova, M.; Stark, W. G.; Allen, C. S.; Maurer, R. J.; Bartók, A. P. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. *arXiv preprint arXiv:2502.15582* **2025**.
(121) Kovács, D. P.; Batatia, I.; Arany, E. S.; Csányi, G. Evaluation of the MACE force field architecture: From medicinal chemistry to materials science. *The Journal of Chemical Physics* **2023**, *159* (4).
(122) Musaelian, A. *NequIP Github Repository*. https://github.com/mir-group/nequip/blob/v0.6.1/configs/minimal_toy_emt.yaml (accessed 07-18-2025).
(123) F. dos Santos, L. G.; Nebgen, B. T.; Allen, A. E. A.; Hamilton, B. W.; Matin, S.; Smith, J. S.; Messerly, R. A. Improving Bond Dissociations of Reactive Machine Learning Potentials through Physics-Constrained Data Augmentation. *Journal of Chemical Information and Modeling* **2025**, *65* (3), 1198-1210. DOI: 10.1021/acs.jcim.4c01847.
(124) Ladygin, V.; Korotaev, P. Y.; Yanilkin, A.; Shapeev, A. Lattice dynamics simulation using machine learning interatomic potentials. *Computational Materials Science* **2020**, *172*, 109333.
(125) Yang, H.; Hu, C.; Zhou, Y.; Liu, X.; Shi, Y.; Li, J.; Li, G.; Chen, Z.; Chen, S.; Zeni, C. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967* **2024**.
(126) Kim, J.; Kim, J.; Kim, J.; Lee, J.; Park, Y.; Kang, Y.; Han, S. Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials. *Journal of the American Chemical Society* **2024**, *147* (1), 1042-1054.
(127) Fu, X.; Wood, B. M.; Barroso-Luque, L.; Levine, D. S.; Gao, M.; Dzamba, M.; Zitnick, C. L. Learning Smooth and Expressive Interatomic Potentials for Physical Property Prediction. *arXiv preprint arXiv:2502.12147* **2025**.
(128) Rhodes, B.; Vandenhaute, S.; Šimkus, V.; Gin, J.; Godwin, J.; Duignan, T.; Neumann, M. Orb-v3: atomistic simulation at scale. *arXiv preprint arXiv:2504.06231* **2025**.
(129) Bochkarev, A.; Lysogorskiy, Y.; Drautz, R. Graph Atomic Cluster Expansion for Semilocal Interactions beyond Equivariant Message Passing. *Physical Review X* **2024**, *14* (2), 021036. DOI: 10.1103/PhysRevX.14.021036.
(130) Fu, X.; Musaelian, A.; Johansson, A.; Jaakkola, T.; Kozinsky, B. Learning interatomic potentials at multiple scales. *arXiv preprint arXiv:2310.13756* **2023**.
(131) Yang, Z.; Wang, X.; Li, Y.; Lv, Q.; Chen, C. Y.-C.; Shen, L. Efficient equivariant model for machine learning interatomic potentials. *npj Computational Materials* **2025**, *11* (1), 49.

(132) Cheng, Y.; Wang, D.; Zhou, P.; Zhang, T. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282* **2017**.
(133) Matin, S.; Shinkle, E.; Pimonova, Y.; Craven, G. T.; Li, Y. W.; Barros, K.; Lubbers, N. Ensemble Knowledge Distillation for Machine Learning Interatomic Potentials. *arXiv preprint arXiv:2503.14293* **2025**.
(134) Achar, S. K.; Bernasconi, L.; Alvarez, J. J.; Johnson, J. K. Deep-learning potentials for proton transport in double-sided graphanol. *Journal of Materials Research* **2023**, *38* (24), 5114-5124.
(135) Sauceda, H. E.; Gastegger, M.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Molecular force fields with gradient-domain machine learning (GDML): Comparison and synergies with classical force fields. *The Journal of Chemical Physics* **2020**, *153* (12).
(136) Wang, H.; Zhang, L.; Han, J. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications* **2018**, *228*, 178-184.
(137) Niblett, S. P.; Kourtis, P.; Magdău, I.-B.; Grey, C. P.; Csányi, G. Transferability of datasets between Machine-Learning Interaction Potentials. *arXiv preprint arXiv:2409.05590* **2024**.
(138) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.
(139) Behler, J.; Csányi, G. Machine learning potentials for extended systems: a perspective. *The European Physical Journal B* **2021**, *94* (7), 142. DOI: 10.1140/epjb/s10051-021-00156-1.
(140) Jeanloz, R. Universal equation of state. *Physical Review B* **1988**, *38* (1), 805.
(141) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C. The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter* **2017**, *29* (27), 273002.
(142) Jónsson, H.; Mills, G.; Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and quantum dynamics in condensed phase simulations*, World Scientific, 1998; pp 385-404.
(143) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of chemical physics* **2000**, *113* (22), 9901-9904.
(144) Henkelman, G.; Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of chemical physics* **2000**, *113* (22), 9978-9985.
(145) Póta, B.; Ahlawat, P.; Csányi, G.; Simoncelli, M. Thermal conductivity predictions with foundation atomistic models. *arXiv preprint arXiv:2408.00755* **2024**.
(146) Gupta, G.; Bukowski, B. C. Kinetic Consequences of Quasi-Harmonic Entropies Calculated with Machine Learning Interatomic Potentials for Microkinetic Modeling. *The Journal of Physical Chemistry C* **2024**, *128* (47), 20104-20117. DOI: 10.1021/acs.jpcc.4c05841.
(147) Park, J.; Cho, J.; Lee, Y.; Park, M.-J.; Lee, W. B. Practical microkinetic modeling approach for methanol synthesis from syngas over a Cu-based catalyst. *Industrial & Engineering Chemistry Research* **2019**, *58* (20), 8663-8673.
(148) Xie, W.; Xu, J.; Chen, J.; Wang, H.; Hu, P. Achieving theory–experiment parity for activity and selectivity in heterogeneous catalysis using microkinetic modeling. *Accounts of chemical research* **2022**, *55* (9), 1237-1248.
(149) Motagamwala, A. H.; Dumesic, J. A. Microkinetic modeling: a tool for rational catalyst design. *Chemical Reviews* **2020**, *121* (2), 1049-1076.
(150) Kozinsky, B.; Musaelian, A.; Johansson, A.; Batzner, S. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2023; pp 1-12.
(151) Jacobsen, K. W.; Stoltze, P.; Nørskov, J. K. A semi-empirical effective medium theory for metals and alloys. *Surface Science* **1996**, *366* (2), 394-402. DOI: https://doi.org/10.1016/0039-6028(96)00816-3.

(152) Tersoff, J. Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. *Physical review B* **1989**, *39* (8), 5566.
(153) Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M. The ReaxFF reactive force-field: development, applications and future directions. *npj Computational Materials* **2016**, *2* (1), 1-14.