

Fine-Tuning Unifies Foundational Machine-learned Interatomic Potential Architectures at *ab initio* Accuracy

Jonas Hänseroth,^{1,*} Aaron Flötotto,¹ Muhammad Nawaz Qaisrani,¹ and Christian Dreßler¹

¹*Theoretical Solid State Physics, Institute of Physics,
Technische Universität Ilmenau, 98693 Ilmenau, Germany*

(Dated: November 10, 2025)

This work demonstrates that fine-tuning transforms foundational machine-learned interatomic potentials (MLIPs) to achieve consistent, near-*ab initio* accuracy across diverse architectures. Benchmarking five leading MLIP frameworks (MACE, GRACE, SevenNet, MatterSim, and ORB) across seven chemically diverse compounds reveals that fine-tuning universally enhances force predictions by factors of 5-15 and improves energy accuracy by 2-4 orders of magnitude. The investigated models span both equivariant and invariant, as well as conservative and non-conservative, architectures. While general-purpose foundation models are robust, they exhibit architecture-dependent deviations from *ab initio* reference data; fine-tuning eliminates these discrepancies, enabling quantitatively accurate predictions of atomistic and structural properties. Using datasets constructed from equidistantly sampled frames of short *ab initio* molecular dynamics trajectories, fine-tuning reduces force errors by an order of magnitude and harmonizes performance across all architectures. These findings establish fine-tuning as a universal route to achieving system-specific predictive accuracy while preserving the computational efficiency of MLIPs. To promote widespread adoption, we introduce the *aMACEing Toolkit*, which provides a unified and reproducible interface for fine-tuning workflows across multiple MLIP frameworks.

Introduction

The computational exploration of materials and molecular systems has long been constrained by the fundamental trade-off between accuracy and efficiency. *Ab initio* molecular dynamics (AIMD), based on density functional theory (DFT), provides chemical accuracy but limits accessible system sizes to a few hundreds of atoms and timescales to picoseconds due to prohibitive computational costs. [1–3] Empirical force-field-based molecular dynamics, while enabling simulations of millions of atoms over multiple nanoseconds, significantly lacks accuracy, transferability, and chemical fidelity beyond its parameterization domain.[4–6] This accuracy-efficiency dilemma has fundamentally restricted the scope of problems addressable through atomistic simulation.

Machine learning interatomic potentials (MLIPs) have emerged as a powerful approach, bridging near-*ab initio* accuracy with the computational efficiency approaching that of classical methods.[7–10] Early neural network potentials and Gaussian approximation potentials demonstrated the feasibility of learning potential energy surfaces directly from quantum chemical data.[11–13] The subsequent adoption of graph neural networks, equivariant architectures, and symmetry-preserving representations has dramatically improved the accuracy and transferability of MLIPs across diverse chemical systems.[14–20]

The recent development of foundation models for atomistic simulations represents a paradigm shift toward universal, pre-trained potentials capable of mod-

eling nearly the entire periodic table.[21–24] These models, trained on massive datasets spanning millions of DFT calculations from repositories such as the Materials Project, Open Materials, and Alexandria databases, offer remarkable zero-shot capabilities across diverse chemical systems.[25–29] Notable examples include MACE-MPA-0, GRACE foundation models trained on several datasets, MatterSim’s universal potentials, ORB’s v3 foundation model, and SevenNet’s multi-fidelity models.[22–24, 30, 31] However, despite their broad applicability, foundation models often fail to capture system-specific properties without further optimization.[32–39]

Fine-tuning, the process of adapting pre-trained foundation models using system-specific training data - has emerged as a critical technique for achieving quantitative accuracy in specialized applications. Recent studies have demonstrated the effectiveness of fine-tuning approaches across various domains.[32–35] Transfer learning strategies enable efficient adaptation of foundation models with relatively small datasets, typically requiring orders of magnitude less training data than training from scratch while achieving comparable accuracy.

Despite growing recognition of fine-tuning’s importance, several challenges limit its widespread adoption. First, each MLIP framework implements fine-tuning differently, with distinct procedures, hyperparameters, and data formats creating technical barriers for researchers. Second, systematic comparisons of fine-tuning effectiveness across different frameworks and chemical systems remain limited, making it difficult to establish best practices. Third, the relationship between foundation model performance and fine-tuned model accuracy, as well as the impact of different training strategies, requires comprehensive investigation.

In this work, we address these challenges through a

* jonas.haenseroth@tu-ilmenau.de

systematic evaluation of foundation model fine-tuning across five leading MLIP frameworks: MACE, GRACE, SevenNet, MatterSim, and ORB.[16, 19, 30, 40, 41] We investigate fine-tuning performance on seven diverse chemical systems: excellent solid state proton conductors such as cesium dihydrogen phosphate (CsH_2PO_4), and its derivative ($\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$) containing the unusual tetrahydroxyphosphonium cation H_4PO_4^+ , L-pyroglutamate-ammonium an organic crystal, that contains low barrier hydrogen bonds and exhibit non-aromatic intrinsic fluorescence when excited by near UV light, solvated phenol, aqueous potassium hydroxide solution, crystalline lithium silicide $\text{Li}_{13}\text{Si}_4$, and a molybdenum disulfide (MoS_2) structure containing sulfur vacancies.[42–44] These systems were selected to span different chemical environments, bonding types, and dynamical phenomena relevant to contemporary materials research.

Our comprehensive analysis reveals that fine-tuning consistently and dramatically improves model accuracy across all frameworks and systems, with force errors typically decreasing by 5-15x and energy errors by 2-4 orders of magnitude. More importantly, we demonstrate that fine-tuning enables accurate reproduction of system-specific physical properties including diffusion coefficients, hydrogen bond dynamics, and structural correlations, that foundation models fail to capture. Through systematic comparison of training times, hyperparameter requirements, and final accuracies, we provide practical guidance for selecting appropriate frameworks and strategies for different applications.

To facilitate broader adoption of these methods, we introduce the aMACEing Toolkit, which provides a unified command-line interface for fine-tuning workflows across all supported MLIP frameworks. The toolkit streamlines the process by taking care of framework-specific complexities (such as training data formatting, training setup, interference with simulation environments, model conversion, performance evaluation and documentation of the computed investigation) while still providing access to advanced features, enabling researchers to focus on scientific questions rather than implementation details. Combined with comprehensive analysis capabilities for trajectory post-processing, the toolkit significantly lowers the barrier to utilizing state-of-the-art machine learning potentials in molecular dynamics research.

Methods

Foundation Models and MLIP Frameworks

We evaluate five prominent MLIP frameworks, all based on graph neural networks, each offering foundation models trained on comprehensive quantum chemical datasets. MACE employs higher-body-order equivariant message passing.[16, 22, 45] GRACE utilizes graph extensions to the atomic cluster expansion.[19, 24] MatterSim

is an invariant graph neural network based on the M3GNet architecture.[30, 46] SevenNet offers scalable equivariant architectures with GPU-parallelism support and is based on the NequIP architecture.[15, 31, 40] ORB is non-conservative and invariant, like MD-ET framework, directly predicting forces instead of computing the gradient of an energy function.[23, 41, 47]

With the exception of MatterSim, all frameworks feature foundation models trained on combinations or subsets of the following databases: Materials Project, Alexandria Database, Open Materials 2024, and Open Molecules 2025.[25–29] The Microsoft Research AI for Science Team has trained foundation models with DFT-calculated data including a temperature range of 0-5000 K and pressure range of 0-1000 GPa.[30] This database is not publicly available. The Materials Project includes DFT calculations of over 200,000 materials.[26, 27] For training, the database is usually subsampled using pymatgen’s StructureMatcher, resulting in a dataset containing 146,000 materials and 1.5 million DFT calculations (PBE+U), referred to as MPtrj.[48, 49] The Alexandria database is composed of DFT structure relaxation trajectories of 3 million materials with 30 million DFT calculations (PBE+U), and for training, a sub-sampled dataset called sAlex is often used, including 10 million DFT calculations.[25, 28] The Open Materials 2024 and Open Molecules 2025 datasets from Meta’s FAIRchem each contain over 100 million DFT calculations (OMat24: PBE+U and OMol25: ω B97M-V).[28, 29]

All these frameworks with their respective foundation models are ranked by Matbench Discovery and MLIP Arena as among the best-performing MLIPs currently available.[50, 51]

Chemical Systems and Fine-Tuning Data Generation

Our evaluation encompasses seven chemically diverse systems selected to represent different classes of materials and dynamical phenomena. CsH_2PO_4 (CDP, 512 atoms, cubic unit cell, $a=19.82$ Å) serves as a model solid acid electrolyte exhibiting proton conductivity enabled by a strong as well as fluctuating hydrogen bond network.[52–54] $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ (CPP, 576 atoms, cubic, $a=20.20$ Å) represents a complex ionic solid with coexisting cationic and anionic phosphate groups.[55, 56] L-pyroglutamate-ammonium (144 atoms, orthorhombic, $a=5.15$ Å, $b=14.56$ Å, $c=17.05$ Å) exemplifies organic molecular crystals with short hydrogen bonds. The phenol-water system (388 atoms, cubic, $a=15.64$ Å) models a simple organic molecule with a solvent.[42–44] Aqueous KOH solution (288 atoms, cubic, $a=14.21$ Å) represents electrolyte solutions with hydroxide ion transport.[33, 57] $\text{Li}_{13}\text{Si}_4$ (204 atoms, orthorhombic, $a=15.90$ Å, $b=15.13$ Å, $c=13.40$ Å) represents a lithium silicide with lithium ion diffusion, being a material of interest for battery research.[58–60] Finally, the

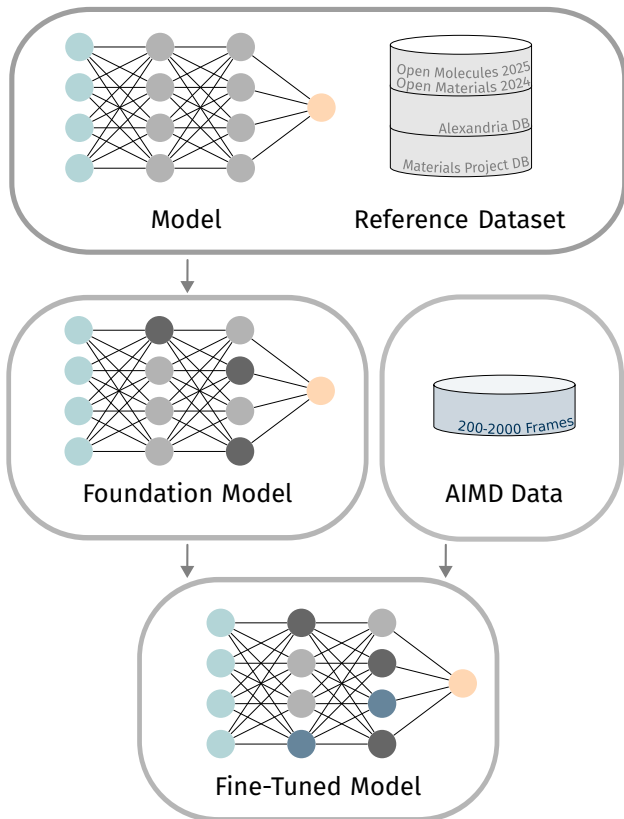


FIG. 1: Fine-tuning of a pre-trained foundation machine learning interatomic potential.

memristive and two-dimensional 1H-MoS₂ (106 atoms, hexagonal, $A=[19.15 \text{ \AA}, 0.0 \text{ \AA}, 0.0 \text{ \AA}]$, $B=[9.58 \text{ \AA}, 16.59 \text{ \AA}, 0.0 \text{ \AA}]$, $C=[0.0 \text{ \AA}, 0.0 \text{ \AA}, 40.0 \text{ \AA}]$) with sulfur vacancies exhibiting cooperative dynamics with high activation energies, completes our benchmark set.[34, 61, 62]

For each system, training data consisting of 2000 configurations was extracted from Born-Oppenheimer AIMD trajectories computed using CP2K with BLYP or PBE exchange-correlation functional, Goedecker-Teter-Hutter pseudopotentials, and DZVP-MOLOPT basis sets.[63–79] Configurations were selected every 100th frame of the AIMD to span the main part of the relevant phase space at target temperatures with structural diversity representative of dynamical processes. Using this protocol, fine-tuning datasets consisting of positions, forces, and energies were computed for all systems.

Fine-Tuning Methodology

Fine-tuning protocols were implemented individually for each MLIP framework, with hyperparameters evaluated for each framework-system combination while maintaining consistency in training data and evaluation procedures. Training utilized 70-90% of configurations for optimization, with remaining data reserved for valida-

tion and testing. To obtain fine-tuned foundation models capable of running stable MD simulations, key hyperparameter including learning rates (10^{-4} - 10^{-2}), force-to-energy loss ratios (0.5-150), batch sizes (4 or 5), and epoch counts (200-2500) were adjusted to achieve MD-ready MLIPs for each system. Training was performed on GPU clusters with careful monitoring of convergence behavior. The fine-tuning protocol was applied to first-generation foundation models: MACE-MP-0, GRACE-1L-OAM, SevenNet-0, MatterSim Large, and ORB-v2 (see Figure 1).[22, 24, 30, 40, 41] While the frameworks often offer more sophisticated foundation models that perform better on Matbench Discovery for a wide range of materials, fine-tuning these foundation models for specific systems with fewer parameters can achieve good performance while benefiting from the smaller model size, which can be used on hardware with less memory and run faster than more sophisticated models.[50] For simplicity, the fine-tuning protocol was applied without incorporating active learning.

Evaluation Metrics and Analysis

Model performance was assessed by recalculating first-principles structures excluded from the training set. In addition to force and energy mean absolute errors, models were evaluated on their ability to reproduce key physical properties derived from extended molecular dynamics simulations. Therefore, molecular dynamics simulations of 2-10 nanoseconds were performed using fine-tuned and foundation models: radial distribution functions characterizing structural correlations, mean square displacements and diffusion coefficients quantifying transport phenomena, and vector autocorrelation functions describing orientational dynamics (see Supporting Information Figures S1 - S19).

aMACEing Toolkit Implementation

To facilitate reproducible fine-tuning workflows, we developed the aMACEing Toolkit, which provides unified interfaces for all supported MLIP frameworks. The toolkit handles data format conversions, generates framework-specific input files, manages job submission for high-performance computing environments, and provides comprehensive logging for reproducibility.

Key toolkit features include interactive question-and-answer interfaces for beginners, one-line command execution for automation, systematic benchmarking capabilities across multiple frameworks, built-in analysis tools for trajectory post-processing, and comprehensive documentation with practical examples. The toolkit can create input files for the Atomic Simulation Environment (ASE) and LAMMPS.[80, 81] The modular architecture enables easy extension to additional frameworks while maintaining consistent user experiences.

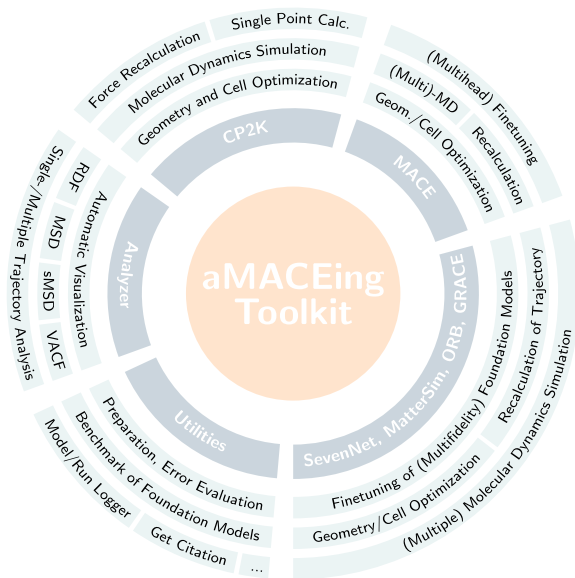


FIG. 2: Modules and functions of the aMACEing Toolkit.

Several other Python packages exist with somewhat similar functionalities, such as Janus-Core and AiIDA-TrainsPot.[82, 83] Janus-Core offers many modules for using multiple MLIPs to perform geometry optimizations, molecular dynamics, nudge elastic band calculations, and more. AiIDA-TrainsPot is a workflow that trains MLIPs automatically, currently only able to train MACE. The same limitation applies to fine-tuning with Janus-Core.

Results and Discussion

Systematic Comparison of Foundation versus Fine-Tuned Models

Our comprehensive evaluation reveals dramatic and consistent improvements achieved through foundation model fine-tuning across all tested systems and frameworks. Figure 3 presents force prediction errors for foundation models versus their fine-tuned counterparts, demonstrating the universal effectiveness of this approach. Foundation models exhibit substantial errors ranging from 0.15-0.45 eV/Å for forces reflecting their general-purpose training on diverse chemical systems rather than optimization for specific applications. The numerical values including the energy error are listed in the Table S1 in the Supporting Information.

Fine-tuning consistently reduces these errors by remarkable margins. Force accuracy improves by factors of 5-15x, with mean absolute force errors decreasing to 0.02-0.07 eV/Å across all frameworks and systems. Energy errors also decrease substantially, often by several

orders of magnitude, but their absolute values depend strongly on the underlying reference level (e.g., functional, basis set). Consequently, while the reduction in energy error highlights the overall consistency gained through fine-tuning, the improvements in force accuracy are the more physically meaningful indicator of enhanced model performance in molecular dynamics simulations. These improvements demonstrate that fine-tuning effectively adapts the broad knowledge encoded in foundation models to capture system-specific interactions with near-quantum chemical accuracy.

Notably, the magnitude of improvement shows limited dependence on the specific MLIP framework, suggesting that fine-tuning effectiveness is primarily determined by the quality and relevance of training data rather than architectural details. All frameworks: MACE, GRACE, SevenNet, MatterSim, and even the non-conservative framework ORB, achieve comparable final accuracies after fine-tuning, despite exhibiting different foundation model performance levels. These models were obtained without extensive hyperparameter optimization for every fine-tuning process; only small adjustments to the example values were needed for some systems. These findings have important practical implications, suggesting that framework selection might prioritize computational efficiency, training speed, or ease of use rather than foundation model accuracy alone. The most important step to achieve better accuracy is the fine-tuning step, as foundation models have not yet reached this level of precision. By using fine-tuned foundation models, a faster workflow requiring fewer computational resources is applied to obtain near *ab initio* accurate trajectories of large systems on nanosecond length scales.

Training Efficiency and Computational Requirements

Analysis of training times reveals significant variations across frameworks and systems, depending on system size, framework architecture, and hyperparameter choices. Table I presents a systematic comparison of the compute time for 100 epochs of fine-tuning for each framework and system, revealing framework-specific characteristics that influence practical deployment decisions.

GRACE generally exhibits the fastest training times, typically requiring less than one hour for 100 epochs of the systems studied, making it attractive for rapid prototyping and iterative refinement. MACE shows intermediate training times. SevenNet and MatterSim demonstrate variable performance depending on system characteristics, often requiring extended training periods. ORB demonstrates competitive training efficiency, particularly for system sizes where only computationally efficient non-conservative models like ORB are feasible.

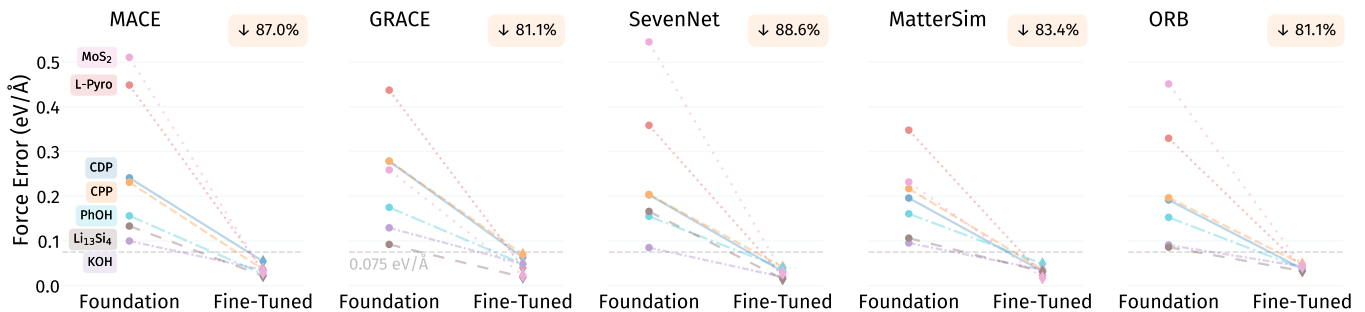


FIG. 3: Root mean squared force errors for foundation and fine-tuned models across all evaluated systems: CsH_2PO_4 (CDP), $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ (CPP), $\text{Li}_{13}\text{Si}_4$, solvated PhOH, aqueous KOH solution, L-pyroglyutamate-ammonium (L-Pyro), MoS_2 ; and frameworks with the respective foundation models: MACE-MP-0, GRACE-1L-OAM, SevenNet-0, MatterSim-Large, ORB-v2. Force errors in $\text{meV } \text{\AA}^{-1}$ and average error reduction in percent.

TABLE I: Computing time for 10,000 molecular dynamics steps and fine-tuning 100 epochs (2,000 data points) of a system containing 512 atoms on one NVIDIA A100.

Task	MACE	MACE+cueq	GRACE
Molecular Dynamics (s)	390.2	383.5	312.6
Model Fine-Tuning (min)	134.0	51.8	40.9

Task	SevenNet	MatterSim	ORB
Molecular Dynamics (s)	555.0	904.8	131.6
Model Fine-Tuning (min)	373.0	342.1	77.7

Physical Property Reproduction

Beyond conventional energy and force accuracy metrics, we evaluate the ability of fine-tuned models to recover key physical properties, such as diffusion coefficients, radial distribution functions, and energy pathways, obtained from extended molecular dynamics simulations. This analysis reveals that fine-tuning not only improves agreement with reference forces and energies, but also enables accurate prediction of structural and dynamical observables that are often inaccessible to short-timescale *ab initio* simulations and poorly captured by foundation models.

The solid acids CsH_2PO_4 and $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ (Figure 4a,b) are inorganic crystalline compounds that exhibit a superprotonic phase transition at elevated temperatures, accompanied by a drastic increase in proton conductivity. In the high-temperature phases of these compounds, the hydrogen-bond network becomes highly disordered, and the rotational dynamics of the anions approach those of a liquid state. This strong and dynamically fluctuating hydrogen-bond network enables efficient proton transfer through the Grotthuss mechanism. The overall proton diffusivity in these materials arises from a combination of the anion rotational rate and the proton transfer rate between neighboring anions. While

proton transfer events within individual hydrogen bonds occur on the picosecond timescale, the rotational motion of the anions typically occurs on the order of several hundred picoseconds. Consequently, diffusion coefficients are challenging to converge in *ab initio* molecular dynamics simulations. Experimental studies indicate that proton diffusion is faster in CsH_2PO_4 than in $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$.^[55, 84] However, due to the limited timescales accessible to AIMD, even *ab initio* simulations often fail to reproduce this qualitative difference in diffusion coefficients (Figure 4a,b).^[32, 56] Similarly, many foundation models incorrectly predict the ratio of diffusion coefficients between the two compounds. In contrast, all fine-tuned foundation force fields correctly reproduced the experimental trend (see Supporting Information, Figures S1-S7).

The diffusion coefficient for the lithium ions in the lithium silicide $\text{Li}_{13}\text{Si}_4$ obtained with first principle methods is reproduced in the trajectories computed by the fine-tuned foundation model, while the foundation models consistently underestimates this value (see Figure 4c and Supporting Information Figures S8 and S9).

Given the critical role of the O–H stretch in determining phenol’s vibrational response and hydrogen-bonding behavior, the accuracy of various machine learning interatomic potentials in reproducing this structural feature was assessed. Figure 4d presents the O–H distance distributions in phenol, showing that the fine-tuned models yield distributions closely aligned with the *ab initio* molecular dynamics reference, effectively capturing interactions with the surrounding solvent environment. In contrast, the foundation models produce broader and excessively delocalized distributions, reflecting an unrealistically soft potential along the O–H stretching coordinate. This artificial softening results in an overrepresentation of elongated O–H configurations, potentially biasing both infrared (IR) peak positions and intensities. Furthermore, the water structure surrounding the hydroxyl hydrogen of phenol is accurately reproduced by the fine-tuned model (see Supporting Information Figure S12).

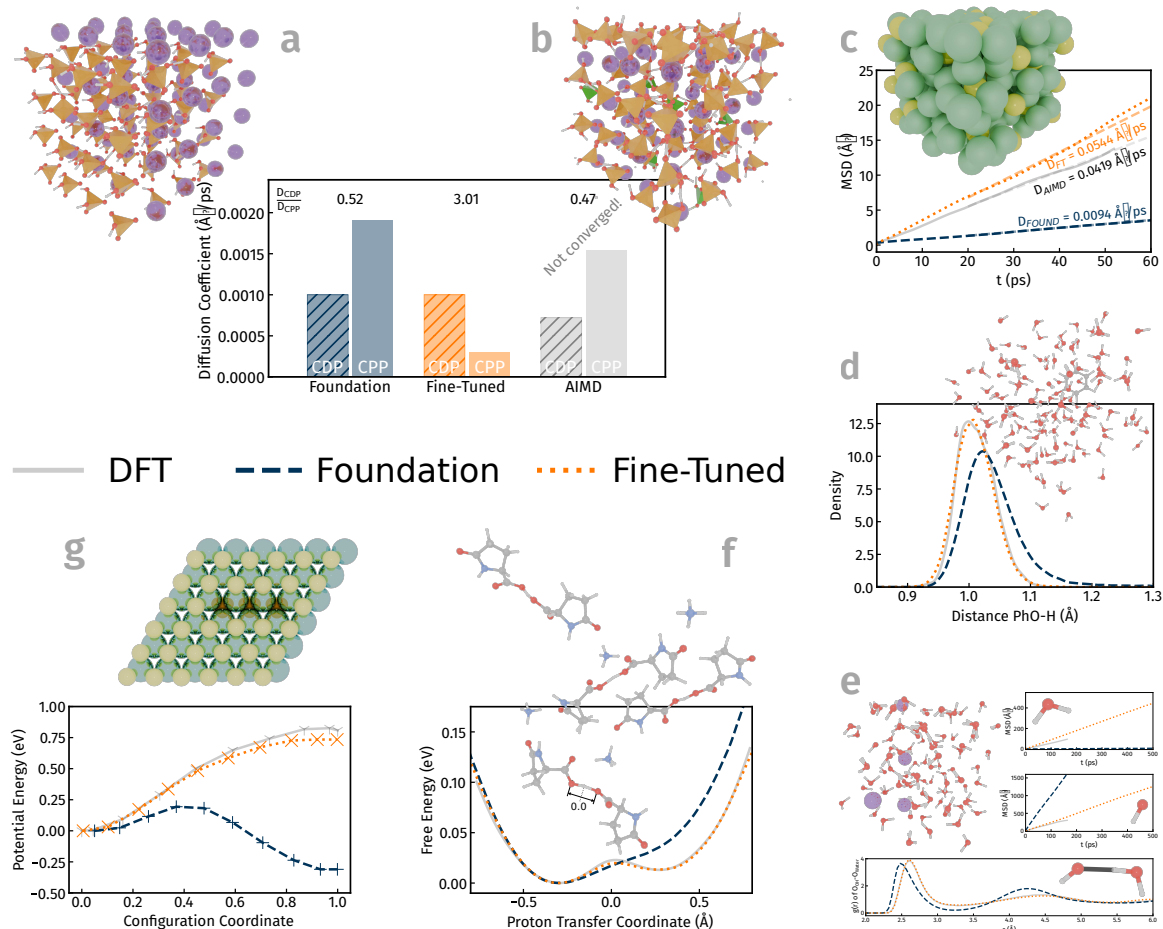


FIG. 4: Comparison of different physical properties obtained with first principles methods, foundation models and fine-tuned foundation models: (a & b) CDP and CPP, proton diffusion coefficients ratios of $D(\text{CDP})/D(\text{CPP})$ (MatterSim), (c) $\text{Li}_{13}\text{Si}_4$, lithium ion mean-squared displacements and diffusion coefficients (ORB), (d) Phenol in water, $(\text{O-H})_{\text{Hydroxyl-Group}}$ bond length distribution (SevenNet), (e) KOH in water, water molecule and hydroxide ion mean-squared displacements and $\text{O}_{\text{Hydroxide-Ion}}-\text{O}_{\text{Water}}$ radial distribution function (GRACE), (f) L-pyroglutamate-ammonium, free energy profiles along the proton transfer coordinate (ORB), (g) MoS_2 , potential energy curves for a sulfur jump into a neighboring line of sulfur vacancies (MACE).

Figure 4e compares the mobilities of hydroxide ions and water molecules in aqueous potassium hydroxide solution. The foundation models fail to accurately reproduce the diffusion coefficients, underestimating or overestimating that of H_2O , K^+ and overestimating that of OH^- (see Supporting Information Figure S13 - S17). In contrast, the fine-tuned models show excellent agreement with the AIMD data. Moreover, the fine-tuned models more accurately capture the solvation environment of the hydroxide ion compared to the foundation models.

L-pyroglutamate-ammonium is another interesting system, an organic crystal that features a short hydrogen bond (SHB) with a donor-acceptor distance below $2.5 \text{ \AA}$. In prior work, we showed that this SHB exhibit a low-barrier, asymmetric proton transfer (PT) poten-

tial in classical Born-Oppenheimer molecular dynamics (BOMD) simulations. The PT free energy barrier in these simulation is approximately 30 meV . Although this barrier is shallow, it plays an important role in mediating the optical properties of this system.[43, 44] When nuclear quantum effects are included via path-integral molecular dynamics, this barrier disappears and the SHB becomes symmetric and delocalized.[42] However, the reference data for training machine-learned potentials in this work are derived from classical BOMD trajectories, which retain the asymmetric low-barrier profile. This distinction is crucial for evaluating ML model performance. As shown in Figure 4f, the correct classical reference profile is asymmetric with a shallow minimum. Most foundation models, however, fail to reproduce this structure. With the exception of MACE-MP-0, they instead pre-

dict flat or symmetric free energy surfaces, incorrectly mimicking quantum behavior that is absent from the training data. This results in inaccurate SHB dynamics and misleading structural interpretations. In contrast, all fine-tuned models across all frameworks recover the correct asymmetric profile and reproduce the low barrier observed in classical BOMD simulations (Figure 5). These findings demonstrate that subtle but chemically important features such as shallow PT barriers in SHBs are not captured by general-purpose models and require system-specific fine-tuning.

In MoS₂, the potential energy curves predicted by the foundation models substantially underestimate the vacancy jump barrier and exhibit an overall qualitatively incorrect trend. In contrast, the fine-tuned models accurately reproduce the DFT energy profile (see Figure 4g).

A broader comparison is provided in the Supporting Information, where the performance of all investigated foundation and fine-tuned models is shown for every analysis presented here. Additional radial distribution function comparisons and other analyses are also listed there. These results demonstrate that fine-tuning enables not merely improved numerical accuracy but faithful reproduction of physical phenomena, making fine-tuned models suitable for quantitative prediction of experimentally observable properties. To illustrate this effect more concretely, a representative example of the exceptional performance achieved by fine-tuning is provided for the material L-pyroglutamate-ammonium in Figure 5. The free energy profiles predicted by the foundation models (despite MACE) deviate from the AIMD reference in a non-systematic manner. In contrast, fine-tuning substantially mitigates these discrepancies: all profiles obtained from molecular dynamics simulations with fine-tuned foundation models show excellent agreement with the AIMD reference data.

A comprehensive assessment across all investigated properties in the 70 multi-nanosecond MLIP simulations allows us to generalize the observations from Figure 4 and the figures in the Supporting Information:

1. The performance of the foundation models is noteworthy. In particular, these models are well suited for predicting non-dynamic properties in inorganic solids, such as radial distribution functions, where fine-tuning is sometimes unnecessary.
2. For organic solids and general liquids, foundation models perform reasonably well but still show significant deviations from AIMD and experimental reference data.
3. The differences in between the different foundation models are often substantial; none of the investigated models performs best in all cases, and the most accurate model is system-dependent.
4. Fine-tuning systematically enhances force and energy predictions, yielding property predictions that

are virtually indistinguishable from AIMD reference data across all MLIP frameworks. (Only one exception was identified: the potential energy curve for a sulfur jump in MoS₂ predicted with SevenNet.)

5. In all cases, fine-tuning significantly reduces the spread in accuracy observed among foundation models for both property and force predictions.

Framework Comparison and Recommendations

All evaluated MLIP frameworks exhibit substantial performance improvements upon fine-tuning, while their corresponding foundation models already demonstrate remarkable versatility. With minimal fine-tuning effort, performed without active learning, all frameworks accurately reproduce first-principles trajectories and frequently achieve near-*ab initio* precision. The fine-tuned models derived from conservative frameworks produce stable molecular dynamics simulations extending over multiple-nanosecond timescales for all investigated systems. Overall, the differences between the frameworks are minor and do not lead to significant variations in their practical applicability. Out of 35 fine-tuning attempts, only one, MoS₂ simulated with SevenNet, failed to reproduce physical properties with near-*ab initio* accuracy. This finding underscores both the robustness of the evaluated approaches and the practical advantage of the aMACEing_toolkit, which enables efficient testing and comparison of multiple MLIP frameworks, in contrast to other packages with limited model support. Nevertheless, subtle distinctions among the frameworks may still inform their selection for specific research objectives. MACE offers an excellent balance between training time, accuracy, and the availability of robust foundation models, making it particularly suitable for exploratory studies. GRACE combines outstanding accuracy with the fastest training and inference performance, enabling simulations over extended temporal and spatial scales. Through the integration of the new cuEquivariance package, which replaces the computational routines of the widely used equivariant neural network library e3nn, MACE achieves computation times comparable to GRACE, emerging as the most robust framework in our study.[85] ORB, owing to its non-conservative architecture, also delivers high computational speed; however, during extended molecular dynamics simulations, this same characteristic can sometimes cause instabilities that lead to the simulation box exploding. SevenNet and MatterSim achieve reliable accuracy, though their fine-tuning and inference stages are somewhat slower during molecular dynamics simulations. In summary, all investigated frameworks provide satisfactory accuracy and computational performance across the studied systems, indicating that the choice of MLIP framework for fine-tuning does not constitute a critical

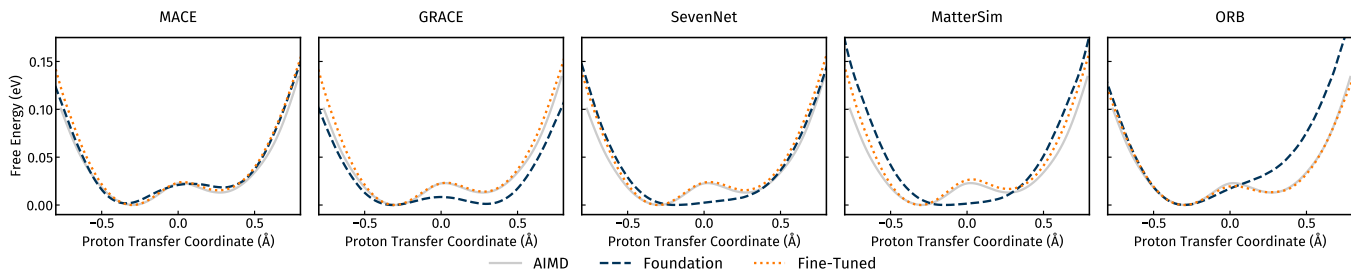


FIG. 5: Free energy profiles along the proton transfer coordinate of the short-hydrogen-bond in L-pyroglutamate-NH₄ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

limiting factor in practice.

Conclusions

This comprehensive evaluation demonstrates that fine-tuning foundation models represents a transformative approach for achieving near-*ab initio* accuracy in specialized molecular dynamics applications. Our systematic study across five MLIP frameworks and diverse chemical systems establishes several key findings with broad implications for computational chemistry and materials science.

Fine-tuning consistently and dramatically improves model accuracy regardless of framework choice, with typical improvements of 5-15x for forces and 2-4 orders of magnitude for energies. This universality suggests that fine-tuning effectiveness depends primarily on training data quality and relevance rather than architectural details, providing flexibility in framework selection based on computational requirements and user preferences.

More importantly, fine-tuning enables accurate reproduction of system-specific physical properties that foundation models often fail to capture at this level of detail, including transport coefficients, structural correlations, and other dynamical phenomena. This capability transforms MLIPs from approximate simulation tools to predictive methods at near-*ab initio* accuracy suitable for direct comparison with experimental measurements.

Given the observed independence of fine-tuning accuracy with respect to the underlying MLIP architecture and considering that no active learning protocol was employed for training data selection, we suggest that future community development efforts should prioritize inference speed, even at the cost of a minor loss in accuracy.

The development of the aMACEing Toolkit addresses critical barriers limiting widespread adoption by providing unified workflows across multiple frameworks. By abstracting technical complexities while maintaining flexibility, the toolkit enables researchers to leverage state-of-the-art methods without extensive specialized knowledge, potentially accelerating scientific discovery across diverse applications.

The universality of fine-tuning improvements across frameworks suggests that standardized benchmarking protocols and high-quality datasets could facilitate systematic comparison of different approaches. Such initiatives would benefit from the unified interfaces provided by tools like the aMACEing Toolkit, enabling large-scale collaborative evaluation studies.

Ultimately, this work establishes fine-tuning as an essential component of modern molecular simulation workflows, providing a practical pathway to near-quantum chemical accuracy for extended simulations. As foundation models continue to evolve and training datasets expand, fine-tuning approaches will likely become increasingly sophisticated, offering exciting opportunities for advancing our understanding of complex chemical systems across diverse applications in energy storage, catalysis, biological systems, and materials design.

Computational Details

All fine-tuning calculations were performed using the respective MLIP framework implementations: MACE-torch 0.3.10, GRACE tensorpotential, SevenNet 0.11.2, MatterSim 1.1.2, and ORB 0.3.2 through their official APIs.[16, 19, 23, 24, 30, 31, 40, 41, 45] *Ab initio* reference calculations were performed using CP2K 2025.1 with PBE and BLYP exchange-correlation functionals and GTH pseudopotentials.[63–79] Training data consisted of 2000 configurations per system extracted from AIMD trajectories at relevant temperatures (300-600 K depending on system). Molecular dynamics simulations for property evaluation were performed using LAMMPS and ASE with system-specific temperatures using Nosé-Hoover chain thermostats.[77–81] Calculations were performed on the compute cluster of Technische Universität Ilmenau using NVIDIA A100 GPUs for training and MD simulations.

The aMACEing Toolkit is available at https://github.com/jhaens/amaceing_toolkit with comprehensive documentation at <https://amaceing-toolkit.readthedocs.io>.

Acknowledgments

We gratefully acknowledge funding from the Thüringer Aufbaubank (TAB) and the European Social Fund Plus (ESF+): KapMemLyse, grant no. 2024 FGR 0081 / 0082, the Carl-Zeiss-Stiftung through project SustEnt-Mat and computational resources provided by the Compute Center of Technische Universität Ilmenau. We thank Henning Schwanbeck for technical support and system administration.

Data Availability

The fine-tuning, evaluation, and production run workflows are available through the aMACEing Toolkit repository: https://github.com/jhaens/amaceing_toolkit. The complete production input, evaluation data, training datasets and the fine-tuned models are available at <https://doi.org/10.5281/zenodo.17438087>. The large trajectory data is upon request from the authors.

References

- [1] D. Marx and J. Hutter, Ab initio molecular dynamics: Theory and implementation, *Modern methods and algorithms of quantum chemistry* **1**, 141 (2000).
- [2] M. E. Tuckerman, Ab initio molecular dynamics: basic concepts, current trends and novel applications, *Journal of Physics: Condensed Matter* **14**, R1297 (2002).
- [3] R. Iftimie, P. Minari, and M. E. Tuckerman, Ab initio molecular dynamics: Concepts, recent developments, and future trends, *Proceedings of the National Academy of Sciences* **102**, 6654 (2005).
- [4] S. Plimpton, Computational limits of classical molecular dynamics simulations, *Computational Materials Science* **4**, 361 (1995).
- [5] G. Sutmann, *Classical molecular dynamics* (2002).
- [6] C. L. Brooks, D. A. Case, S. Plimpton, B. Roux, D. Van der Spoel, and E. Tajkhorshid, Classical molecular dynamics, *The Journal of chemical physics* **154**, 10.1063/5.0045455 (2021).
- [7] A. Kabylda, J. T. Frank, S. Suárez-Dou, A. Khabibrakhmanov, L. Medrano Sandonas, O. T. Unke, S. Chmiela, K.-R. Müller, and A. Tkatchenko, Molecular simulations with a pretrained neural network and universal pairwise force fields, *Journal of the American Chemical Society* **147**, 10.1021/jacs.5c09558 (2025).
- [8] I. Poltavsky, M. Puleva, A. Charkin-Gorbulin, G. Fonseca, I. Batatia, N. J. Browning, S. Chmiela, M. Cui, J. T. Frank, S. Heinen, *et al.*, Crash testing machine learning force fields for molecules, materials, and interfaces: molecular dynamics in the tea challenge 2023, *Chemical Science* **16**, 3738 (2025).
- [9] E. Prašnikar, M. Ljubič, A. Perdih, and J. Borišek, Machine learning heralding a new development phase in molecular dynamics simulations, *Artificial intelligence review* **57**, 102 (2024).
- [10] Y. Wang, J. M. L. Ribeiro, and P. Tiwary, Machine learning approaches for analyzing and enhancing molecular dynamics simulations, *Current opinion in structural biology* **61**, 139 (2020).
- [11] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Physical Review Letters* **98**, 146401 (2007).
- [12] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Physical Review Letters* **104**, 136403 (2010).
- [13] P. Friederich, F. Häse, J. Proppe, and A. Aspuru-Guzik, Machine-learned potentials for next-generation matter simulations, *Nature Materials* **20**, 750 (2021).
- [14] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds, *arXiv preprint arXiv:1802.08219* 10.48550/arXiv.1802.08219 (2018).
- [15] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nature communications* **13**, 2453 (2022).
- [16] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, *Advances in neural information processing systems* **35**, 11423 (2022).
- [17] O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, and K. R. Müller, Machine Learning Force Fields, *Chemical Reviews* **121**, 10142 (2021).
- [18] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, and P. Friederich, Graph neural networks for materials science and chemistry, *Communications Materials* **3**, 10.1038/s43246-022-00315-6 (2022).
- [19] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing, *Phys. Rev. X* **14**, 021036 (2024).
- [20] R. Drautz, Atomic cluster expansion for accurate and transferable interatomic potentials, *Physical Review B* **99**, 014104 (2019).
- [21] R. Jacobs, D. Morgan, S. Attarian, J. Meng, C. Shen, Z. Wu, C. Y. Xie, J. H. Yang, N. Artrith, B. Blaiszik, *et al.*, A practical guide to machine learning interatomic potentials—status and future, *Current Opinion in Solid State and Materials Science* **35**, 101214 (2025).
- [22] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, W. J. Baldwin, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. D. Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O’Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen,

- L. L. Schaaf, C. Schran, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, A foundation model for atomistic materials chemistry, arXiv preprint arXiv:2401.00096 10.48550/arXiv.2401.00096 (2023).
- [23] B. Rhodes, S. Vandenhaute, V. Šimkus, J. Gin, J. Godwin, T. Duignan, and M. Neumann, Orb-v3: atomistic simulation at scale, arXiv preprint arXiv:2504.06231 10.48550/arXiv.2504.06231 (2025).
- [24] Y. Lysogorskiy, A. Bochkarev, and R. Drautz, Graph atomic cluster expansion for foundational machine learning interatomic potentials, arXiv preprint arXiv:2508.17936 10.48550/arXiv.2508.17936 (2025).
- [25] J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, Machine-learning-assisted determination of the global zero-temperature phase diagram of materials, *Advanced Materials* **35**, 2210788 (2023).
- [26] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, *APL Mater.* **1**, 011002 (2013).
- [27] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, *Comput. Mater. Sci.* **97**, 209–215 (2015).
- [28] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi, Open materials 2024 (omat24) inorganic materials dataset and models, arXiv preprint arXiv:2410.12771 10.48550/arXiv.2410.12771 (2024).
- [29] D. S. Levine, M. Shuaibi, E. W. C. Spotte-Smith, M. G. Taylor, M. R. Hasyim, K. Michel, I. Batatia, G. Csányi, M. Dzamba, P. Eastman, *et al.*, The open molecules 2025 (omol25) dataset, evaluations, and models, arXiv preprint arXiv:2505.08762 10.48550/arXiv.2505.08762 (2025).
- [30] H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, *et al.*, Mattersim: A deep learning atomistic model across elements, temperatures and pressures, arXiv preprint arXiv:2405.04967 10.48550/arXiv.2405.04967 (2024).
- [31] J. Kim, J. Kim, J. Kim, J. Lee, Y. Park, Y. Kang, and S. Han, Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials, *J. Am. Chem. Soc.* **147**, 1042 (2024).
- [32] M. Grunert, M. Großmann, J. Hänseroth, A. Flötotto, J. Oumard, J. L. Wolf, E. Runge, and C. Dreßler, Modeling complex proton transport phenomena - exploring the limits of fine-tuning and transferability of foundational machine-learned force fields, *The Journal of Physical Chemistry C* **129**, 9662 (2025).
- [33] J. Hänseroth and C. Dreßler, Optimizing machine learning interatomic potentials for hydroxide transport: Surprising efficiency of single-concentration training, *The Journal of Chemical Physics* **163**, 10.1063/5.0284063 (2025).
- [34] A. Flötotto, B. Spetzler, R. von Stackelberg, M. Ziegler, E. Runge, and C. Dreßler, Large-scale cooperative sulfur vacancy dynamics in two-dimensional mos2 from machine learning interatomic potentials, arXiv preprint arXiv:2508.13790 10.48550/arXiv.2508.13790 (2025).
- [35] H. Weiske, R. Barrett, R. Tonner-Zech, P. Melix, and J. Westermayr, Statistics makes a difference: Machine learning adsorption dynamics of functionalized cyclooctene on si (001) at dft accuracy, arXiv preprint arXiv:2509.14828 10.48550/arXiv.2509.14828 (2025).
- [36] P.-Y. Chen, K. Shibata, and T. Mizoguchi, High precision machine learning force field development for batio3 phase transitions, amorphous, and liquid structures, *APL Machine Learning* **3**, 10.1063/5.0268149 (2025).
- [37] M. Radova, W. G. Stark, C. S. Allen, R. J. Maurer, and A. P. Bartók, Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning, *npj Computational Materials* **11**, 237 (2025).
- [38] X. Liu, K. Zeng, Z. Luo, Y. Wang, T. Zhao, and Z. Xu, Fine-tuning universal machine-learned interatomic potentials: A tutorial on methods and applications, arXiv preprint arXiv:2506.21935 10.48550/arXiv.2506.21935 (2025).
- [39] H. Kaur, F. Della Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides, and V. Kapil, Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies, *Faraday Discussions* **256**, 120 (2025).
- [40] Y. Park, J. Kim, S. Hwang, and S. Han, Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations, *J. Chem. Theory Comput.* **20**, 4857 (2024).
- [41] M. Neumann, J. Gin, B. Rhodes, S. Bennett, Z. Li, H. Choubisa, A. Hussey, and J. Godwin, Orb: A fast, scalable neural network potential, arXiv preprint arXiv:2410.22570 10.48550/arXiv.2410.22570 (2024).
- [42] M. N. Qaisrani, N. Kumar, C. Dreßler, R. Gebauer, and A. Hassanali, Acid–base chemistry of short hydrogen bonds: A tale of schrödinger’s cat in glutamine-derived crystals, *The Journal of Physical Chemistry Letters* **16**, 8588 (2025).
- [43] G. D. Mirón, J. A. Semelak, L. Grisanti, A. Rodriguez, I. Conti, M. Stella, J. Velusamy, N. Seriani, N. Došlić, I. Rivalta, *et al.*, The carbonyl-lock mechanism underlying non-aromatic fluorescence in biological matter, *Nature Communications* **14**, 7325 (2023).
- [44] A. D. Stephens, M. N. Qaisrani, M. T. Ruggiero, G. Díaz Mirón, U. N. Morzan, M. C. González Lebrero, S. T. Jones, E. Poli, A. D. Bond, P. J. Woodhams, *et al.*, Short hydrogen bonds enhance nonaromatic protein-related fluorescence, *Proceedings of the National Academy of Sciences* **118**, e2020389118 (2021).
- [45] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, The design space of e (3)-equivariant atom-centred interatomic potentials, *Nature Machine Intelligence* **7**, 56 (2025).
- [46] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, *Nature Computational Science* **2**, 718 (2022).
- [47] M. Eissler, T. Korjakow, S. Gansch, O. T. Unke, K.-R. Mázler, and S. Gugler, How simple can you go? an off-the-shelf transformer approach to molecular dynamics, arXiv preprint arXiv:2503.01431 10.48550/arXiv.2503.01431 (2025).

- [48] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, *Computational Materials Science* **68**, 314 (2013).
- [49] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling, *Nature Machine Intelligence* **5**, 1031 (2023).
- [50] J. Riebesell, R. E. Goodall, P. Benner, Y. Chiang, B. Deng, A. A. Lee, A. Jain, and K. A. Persson, Matbench discovery—a framework to evaluate machine learning crystal stability predictions, arXiv preprint arXiv:2308.14920 10.1038/s42256-025-01055-1 (2023).
- [51] Y. Chiang, T. Kreiman, E. Weaver, M. Kuner, C. Zhang, A. Kaplan, D. Chrzan, S. M. Blau, A. S. Krishnapriyan, and M. Asta, MLIP arena: Advancing fairness and transparency in machine learning interatomic potentials through an open and accessible benchmark platform, in *AI for Accelerated Materials Design - ICLR 2025* (2025).
- [52] K. Yamada, T. Sagara, Y. Yamane, H. Ohki, and T. Okuda, Superprotonic conductor csh2po4 studied by ¹h, ³¹p nmr and x-ray diffraction, *Solid State Ionics* **175**, 557 (2004).
- [53] D. A. Boysenand, S. M. Haile, H. Liu, and R. A. Secco, High-temperature behavior of csh2po4 under both ambient and high pressure conditions, *Chemistry of materials* **15**, 727 (2003).
- [54] C. Dreßler and D. Sebastiani, Effect of anion reorientation on proton mobility in the solid acids family cshyxo4 (x= s, p, se, y= 1, 2) from ab initio molecular dynamics simulations, *Physical Chemistry Chemical Physics* **22**, 10738 (2020).
- [55] L. S. Wang, S. V. Patel, S. S. Sanghvi, Y.-Y. Hu, and S. M. Haile, Structure and properties of cs7 (h4po4)(h2po4) 8: A new superprotonic solid acid featuring the unusual polycation (h4po4)⁺, *Journal of the American Chemical Society* **142**, 19992 (2020).
- [56] C. Dreßler, J. Hänseroth, and D. Sebastiani, Coexistence of cationic and anionic phosphate moieties in solids: Unusual but not impossible, *The Journal of Physical Chemistry Letters* **14**, 7249 (2023).
- [57] J. Hänseroth, D. Sebastiani, J. A. Jimenez Siegert, J. Scholl, K. Skadell, and C. Dreßler, Hydroxide mobility in aqueous systems: Combining ab initio accuracy with millisecond timescales, *Small*, 2500931 (2025).
- [58] C. Kirsch, C. Dreßler, and D. Sebastiani, Atomistic diffusion pathways of lithium ions in crystalline lithium silicides from ab initio molecular dynamics simulations, *The Journal of Physical Chemistry C* **126**, 12136 (2022).
- [59] C. Kirsch, C. Dreßler, and D. Sebastiani, Li⁺ diffusion in crystalline lithium silicides: influence of intrinsic point defects, *Journal of Physics: Energy* **7**, 025003 (2025).
- [60] M. Zeilinger and T. F. Fässler, Revision of the li13si4 structure, *Structure Reports* **69**, i81 (2013).
- [61] D. Li, B. Wu, X. Zhu, J. Wang, B. Ryu, W. D. Lu, W. Lu, and X. Liang, MoS₂ memristors exhibiting variable switching characteristics toward biorealistic synaptic emulation, *ACS Nano* **12**, 9240 (2018).
- [62] B. Spetzler, D. Abdel, F. Schwierz, M. Ziegler, and P. Farrell, The role of vacancy dynamics in two-dimensional memristive devices, *Advanced Electronic Materials* **10**, 10.1002/aelm.202300635 (2024).
- [63] J. Lippert, G. Hutter and M. Parrinello, A hybrid gaussian and plane wave density functional scheme, *Molecular Physics* **92**, 477 (1997).
- [64] J. Hutter, M. Iannuzzi, F. Schiffmann, and J. VandeVondele, cp2k: atomistic simulations of condensed matter systems, *WIREs Computational Molecular Science* **4**, 15 (2014).
- [65] U. Borštnik, J. VandeVondele, V. Weber, and J. Hutter, Sparse matrix multiplication: The distributed block-compressed sparse row library, *Parallel Computing* **40**, 47 (2014).
- [66] T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, *et al.*, CP2K: An electronic structure and molecular dynamics software package-Quickstep: Efficient and accurate electronic structure calculations, *The Journal of Chemical Physics* **152**, 10.1063/5.0007045 (2020).
- [67] M. Iannuzzi, J. Wilhelm, F. Stein, A. Bussy, H. Elgabarty, D. Golze, A. Hehn, M. Graml, S. Marek, B. S. Gökmen, *et al.*, The cp2k program package made simple, arXiv preprint arXiv:2508.15559 10.48550/arXiv.2508.15559 (2025).
- [68] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach, *Computer Physics Communications* **167**, 103 (2005).
- [69] J. VandeVondele and J. Hutter, Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases, *The Journal of Chemical Physics* **127**, 114105 (2007).
- [70] J. VandeVondele and J. Hutter, An efficient orbital transformation method for electronic structure calculations, *The Journal of Chemical Physics* **118**, 4365 (2003).
- [71] C. Hartwigsen, S. Goedecker, and J. Hutter, Relativistic separable dual-space gaussian pseudopotentials from h to rn, *Physical Review B* **58**, 3641 (1998).
- [72] M. Krack, Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals, *Theoretical Chemistry Accounts* **114**, 145 (2005).
- [73] S. Goedecker, M. Teter, and J. Hutter, Separable dual-space gaussian pseudopotentials, *Physical Review B* **54**, 1703 (1996).
- [74] A. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Physical Review A* **38**, 3098 (1988).
- [75] C. Lee, W. Yang, and R. Parr, Development of the colle-salvetti correlation-energy formula into a functional of the electron density, *Physical Review A* **37**, 785 (1988).
- [76] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [77] S. Nosé, A Unified Formulation of the Constant Temperature Molecular Dynamics Methods, *The Journal of Chemical Physics* **81**, 511 (1984).
- [78] S. Nosé, A Molecular Dynamics Method for Simulations in the Canonical Ensemble, *Molecular Physics* **52**, 255 (1970).
- [79] G. J. Martyna, M. L. Klein, and M. Tuckerman, Nosé-Hoover chains: The Canonical Ensemble via Continuous Dynamics, *The Journal of Chemical Physics* **97**, 2635 (1992).

- [80] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dulak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, *Journal of Physics: Condensed Matter* **29**, 273002 (2017).
- [81] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, *Computer Physics Communications* **271**, 108171 (2022).
- [82] E. Kesoar, P. Austin, H. Devereux, K. Harris, D. Mason, J. Wilkins, F. Zanca, and A. Elena, *janus-core* (2025).
- [83] D. Bidoggia, N. Manko, M. Peressi, and A. Marrazzo, Automated training of neural-network interatomic potentials, *arXiv preprint arXiv:2509.11703* 10.48550/arXiv.2509.11703 (2025).
- [84] L. S. Wang, S. V. Patel, E. Truong, Y.-Y. Hu, and S. M. Haile, Phase behavior and superprotonic conductivity in the system (1-x) csh_2po_4 -x h_3po_4 : Discovery of off-stoichiometric α -[$\text{cs}_{1-x}\text{h}_x$] h_2po_4 , *Chemistry of Materials* **34**, 1809 (2022).
- [85] M. Geiger and T. Smidt, e3nn: Euclidean neural networks, *arXiv preprint arXiv:2207.09453* 10.48550/arXiv.2207.09453 (2022).

Supporting Information:

Fine-Tuning Unifies Foundational

Machine-learned Interatomic Potential

Architectures at ab initio Accuracy

Jonas Hänseroth,^{*} Aaron Flötotto, Muhammad Nawaz Qaisrani, and Christian
Dreßler

*Theoretical Solid State Physics, Institute of Physics, Technische Universität Ilmenau,
98693 Ilmenau, Germany*

E-mail: jonas.haenseroth@tu-ilmenau.de

Table S1: Absolute force and energy errors for foundation and fine-tuned models across all evaluated systems. Errors: force (eV Å⁻¹), energy per atom (eV).

System	Model	MACE-MP-0		GRACE-1L-OAM		SevenNet-0		MatterSim-Large		ORB-v2	
		Force	Energy	Force	Energy	Force	Energy	Force	Energy	Force	Energy
CsH ₂ PO ₄	Foundation	0.2411	307.68	0.2782	307.69	0.2031	307.69	0.1960	307.69	0.1916	307.69
	Fine-tuned	0.0543	0.00041	0.0631	0.00017	0.0316	0.00096	0.032	0.00050	0.0378	0.00192
Cs ₇ (H ₄ PO ₄)(H ₂ PO ₄) ₈	Foundation	0.2310	292.84	0.2787	292.85	0.2039	292.84	0.2169	292.85	0.1966	292.85
	Fine-tuned	0.0364	0.00015	0.0699	0.00017	0.0414	0.00077	0.0363	0.00237	0.0480	0.00249
L-pyrogutamate-NH ₄	Foundation	0.4484	139.42	0.4370	139.43	0.3584	139.44	0.3476	139.44	0.3295	139.43
	Fine-tuned	0.0333	0.00153	0.0403	0.00013	0.0211	0.00183	0.0232	0.00320	0.0404	0.00075
PhOH in H ₂ O	Foundation	0.1562	149.64	0.1750	149.64	0.1551	149.64	0.1607	149.64	0.1528	149.63
	Fine-tuned	0.0261	0.00052	0.0485	0.00022	0.0388	0.00312	0.0490	0.00940	0.0383	0.00150
KOH in H ₂ O	Foundation	0.0999	161.61	0.1294	161.62	0.0851	161.62	0.0954	161.62	0.0910	161.61
	Fine-tuned	0.0351	0.00341	0.0485	0.00213	0.0193	0.00216	0.0354	0.00594	0.0426	0.00272
Li ₁₃ Si ₄	Foundation	0.1333	177.21	0.0923	177.23	0.1660	177.22	0.1063	177.23	0.0861	177.22
	Fine-tuned	0.0220	0.00310	0.0190	0.00143	0.0151	0.00186	0.0313	0.00234	0.0327	0.00405
MoS ₂	Foundation	0.5103	807.51	0.2587	807.52	0.5448	807.51	0.2317	807.53	0.4510	807.55
	Fine-tuned	0.0299	0.00109	0.02299	0.00350	0.0284	0.00013	0.0175	0.00023	0.0431	0.00136

Table S2: Computing time in minutes for fine-tuning across different MLIP frameworks and chemical systems per 100 epochs on one NVIDIA A100.

System	MACE	GRACE	SevenNet	MatterSim	ORB
CDP	134.0	40.9	373.0	342.1	77.7
CPP	167.0	36.7	364.7	456.0	108.0
L-PyroNH ₄	37.5	12.2	188.5	128.0	44.0
PhOH	42.3	16.8	45.5	45.4	8.8
KOH	85.0	26.1	359.2	338.0	150.5
Li ₁₃ Si ₄	67.0	21.0	164.0	178.6	58.7

Table S3: Computing time for 10,000 molecular dynamics steps and fine-tuning 100 epochs (2,000 data points) of a system containing 512 atoms on one NVIDIA A100.

Task	MACE	MACE+cueq	GRACE
MD foundation (s)	412.6	385.1	292.2
MD fine-tuned (s)	390.2	383.5	312.6
Fine-tuning model (min)	134.0	51.8	40.9

Task	SevenNet	MatterSim	ORB
MD foundation (s)	549.0	915.6	131.6
MD fine-tuned (s)	555.0	904.8	131.6
Fine-tuning model (min)	373.0	342.1	77.7

Table S4: Hyperparameters used for fine-tuning across different MLIP frameworks and chemical systems. Learning rates, force weights, and epoch counts show both framework-specific preferences and system-dependent requirements.

System	Framework	Learning Rate	Force Weight	Batch Size	Epochs
CsH ₂ PO ₄	MACE	0.01	100	5	200
	GRACE	0.002	150	4	2000
	SevenNet	0.01	1	5	250
	MatterSim	0.0005	0.5	5	500
	ORB	0.0003	0.5	4	1650
Cs ₇ (H ₄ PO ₄)(H ₂ PO ₄) ₈	MACE	0.01	100	5	200
	GRACE	0.002	50	4	2500
	SevenNet	0.004	1	4	300
	MatterSim	0.0005	0.5	5	500
	ORB	0.0003	1	4	400
L-pyroglutamate-NH ₄	MACE	0.01	10	5	200
	GRACE	0.002	100	4	1000
	SevenNet	0.01	100	4	200
	MatterSim	0.0005	0.5	5	500
	ORB	0.0003	0.5	4	400
PhOH in H ₂ O	MACE	0.01	10	5	200
	GRACE	0.002	100	4	500
	SevenNet	0.004	100	4	400
	MatterSim	0.0005	0.25	5	500
	ORB	0.0002	0.25	8	800
KOH in H ₂ O	MACE	0.01	100	5	200
	GRACE	0.001	5	4	500
	SevenNet	0.01	100	4	200
	MatterSim	0.0005	0.5	5	500
	ORB	0.0003	1	4	200
Li ₁₃ Si ₄	MACE	0.01	10	5	200
	GRACE	0.002	100	4	500
	SevenNet	0.004	50	4	200
	MatterSim	0.0005	0.5	5	350
	ORB	0.0003	0.75	4	1250
MoS ₂	MACE	0.01	100	5	200
	GRACE	0.001	100	4	1000
	SevenNet	0.01	1	5	400
	MatterSim	0.001	10	5	500
	ORB	0.0003	0.5	4	750

System A: CsH_2PO_4

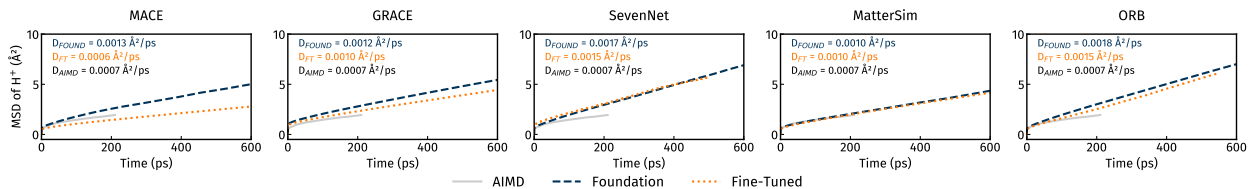


Figure S1: Mean-squared displacements of H^+ in CsH_2PO_4 computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

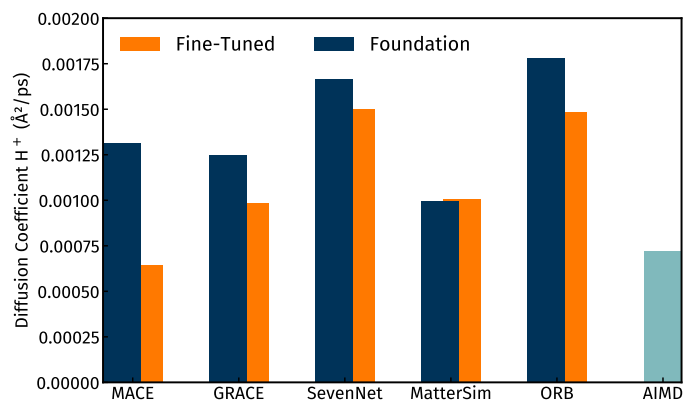


Figure S2: Diffusion coefficients of H^+ in CsH_2PO_4 computed using different MLIP frameworks from the mean-square displacements (see Figure S1). Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

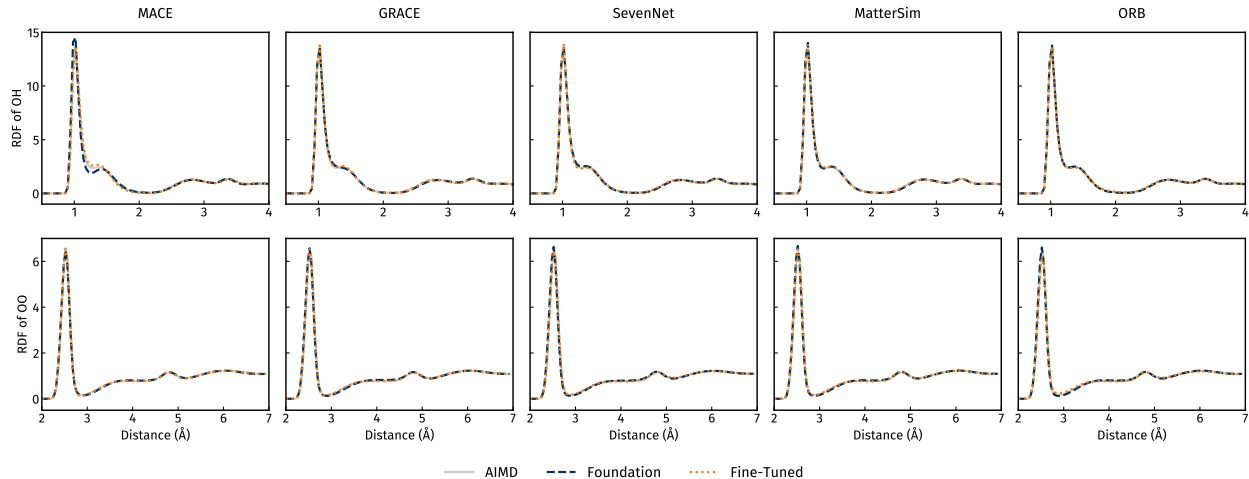


Figure S3: Radial-distribution functions of O-H and O-O in CsH_2PO_4 computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

System B: $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$

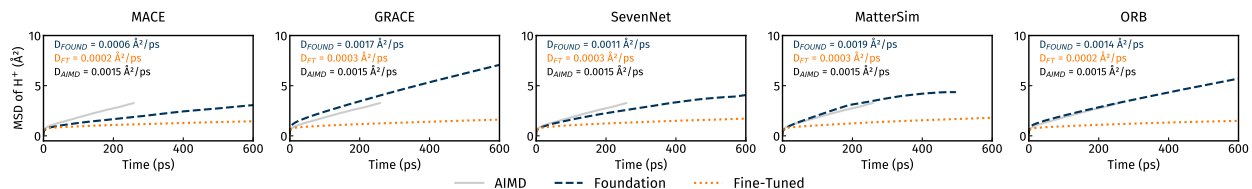


Figure S4: Mean-squared displacements of H^+ in $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

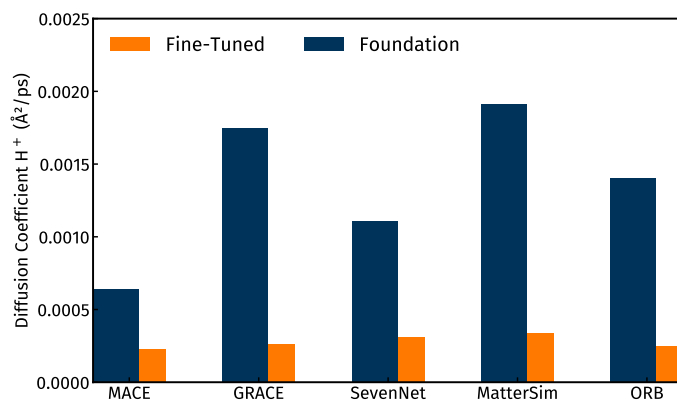


Figure S5: Diffusion coefficients of H^+ in $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ computed using different MLIP frameworks from the mean-square displacements (see Figure S4). Results are shown for the foundation model and the fine-tuned foundation model. Reference AIMD data are not available, as AIMD simulations cannot provide converged diffusion coefficients for this system. For comparison, a recent MLIP study reported a diffusion coefficient of $0.0004 \text{ Å}^2/\text{ps}$ at 510 K .^{S1}

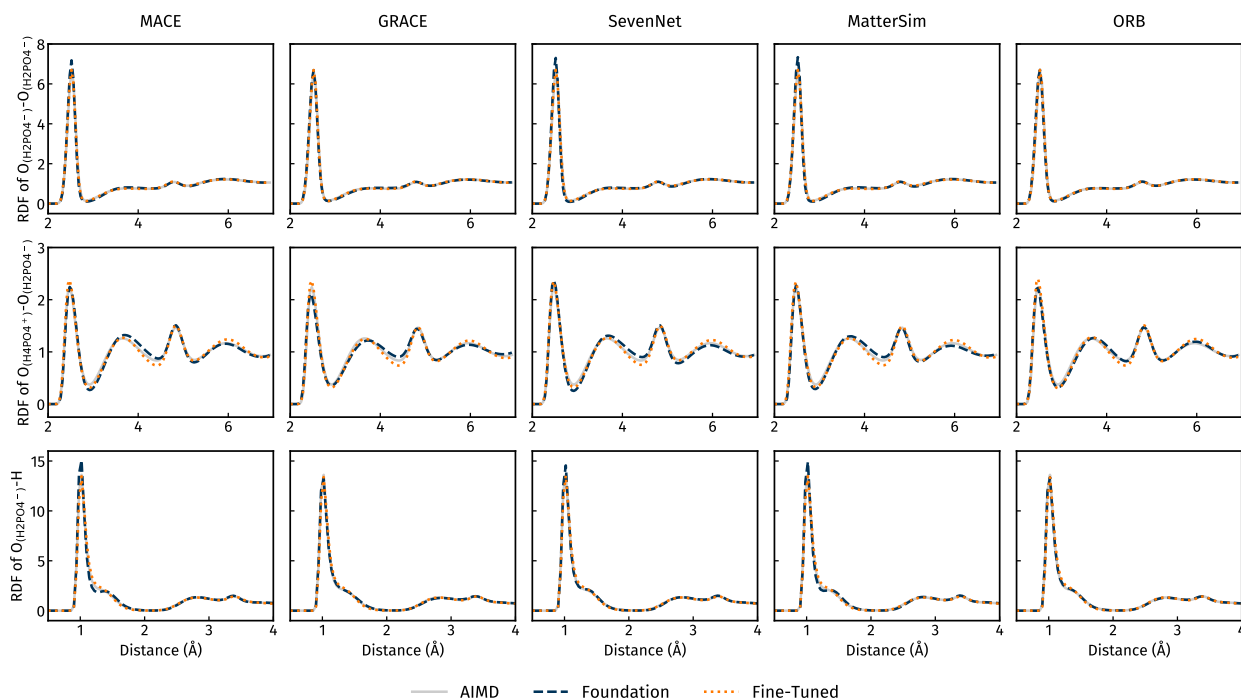


Figure S6: Radial-distribution functions of $\text{O}_{\text{H}_2\text{PO}_4}-\text{O}_{\text{H}_2\text{PO}_4^-}$, $\text{O}_{\text{H}_4\text{PO}_4^+}-\text{O}_{\text{H}_2\text{PO}_4^-}$ and $\text{O}_{\text{H}_2\text{PO}_4}-\text{H}$ in $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

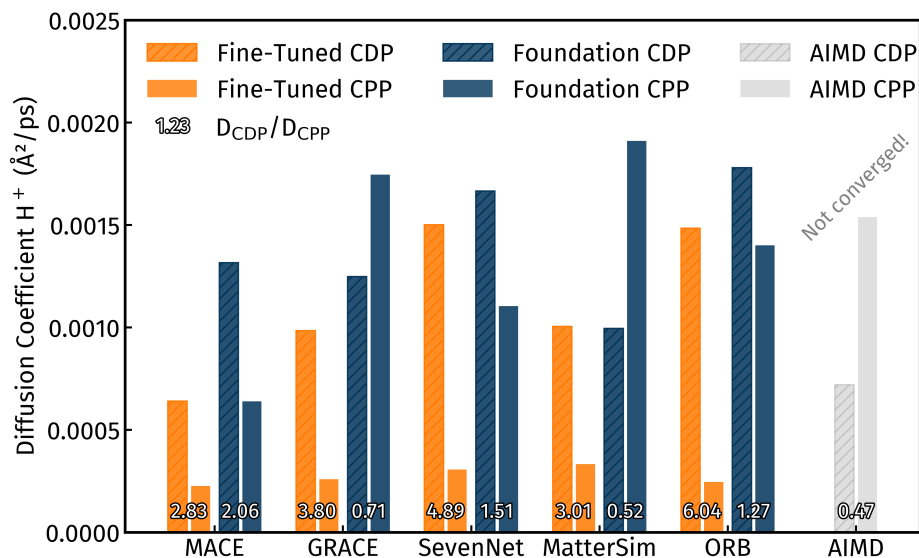


Figure S7: Diffusion coefficients comparison of H^+ in CsH_2PO_4 and $\text{Cs}_7(\text{H}_4\text{PO}_4)(\text{H}_2\text{PO}_4)_8$ computed using different MLIP frameworks from the mean-square displacements (see Figure S2 and Figure S5). Results are shown for the foundation model and the fine-tuned foundation model. The AIMD data result in non-converged diffusion coefficients. For comparison, a recent MLIP study reported a diffusion coefficient ratio of 4.^{S1}

System C: $\text{Li}_{13}\text{Si}_4$

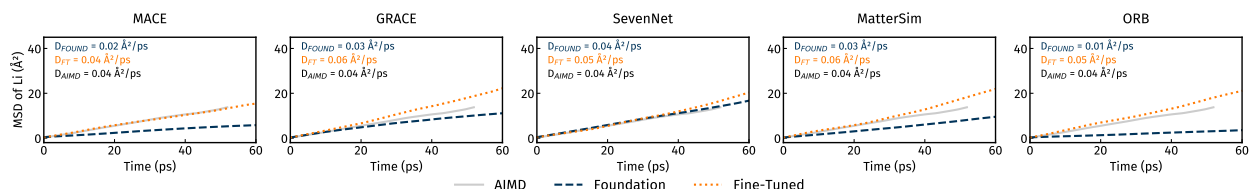


Figure S8: Mean-squared displacements of Li^+ in $\text{Li}_{13}\text{Si}_4$ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

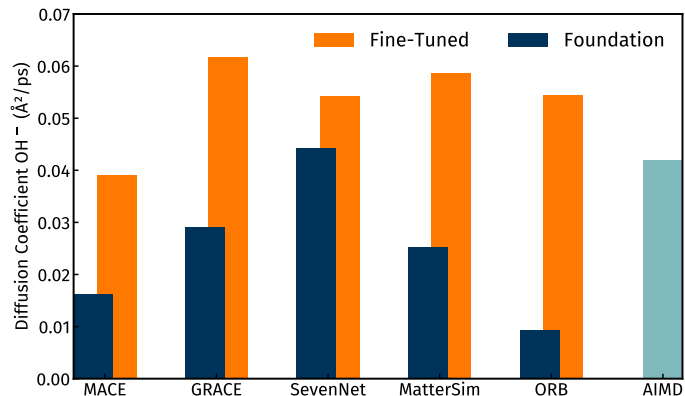


Figure S9: Diffusion coefficients of Li⁺ in Li₁₃Si₄ computed using different MLIP frameworks from the mean-square displacements (see Figure S8). Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

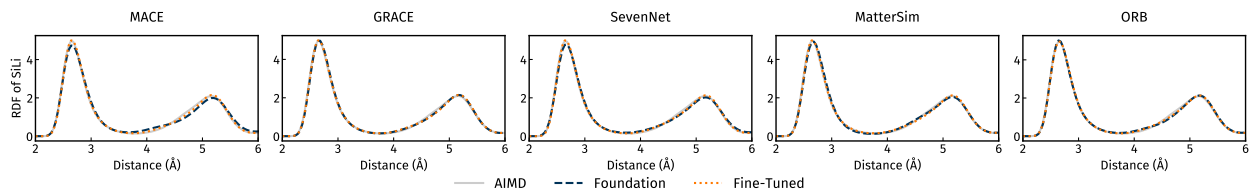


Figure S10: Radial-distribution functions of Si-Li in Li₁₃Si₄ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

System D: PhOH in H₂O

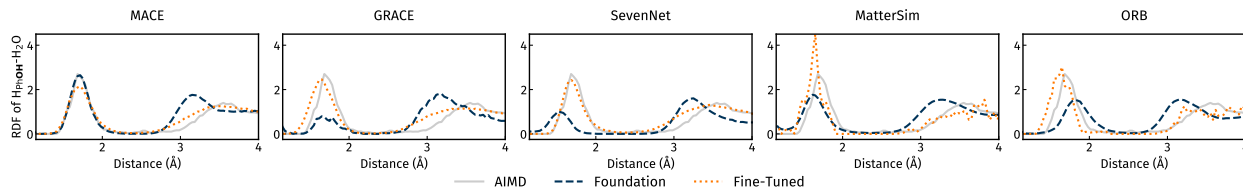


Figure S11: Radial-distribution functions of $H_{\text{Hydroxyl-Group-O}_{\text{Water}}}$ in PhOH in water computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

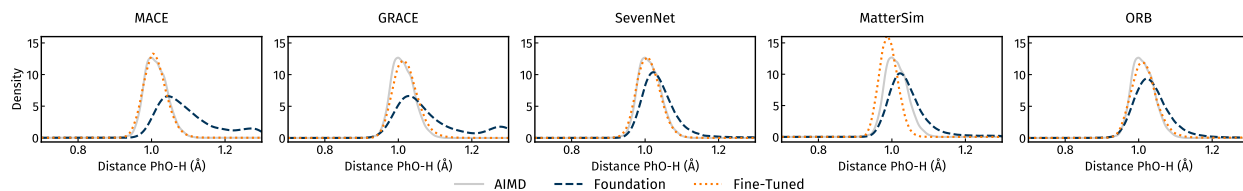


Figure S12: Distribution of the hydroxyl group O-H bond length in the phenol in water computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

System E: KOH in H₂O

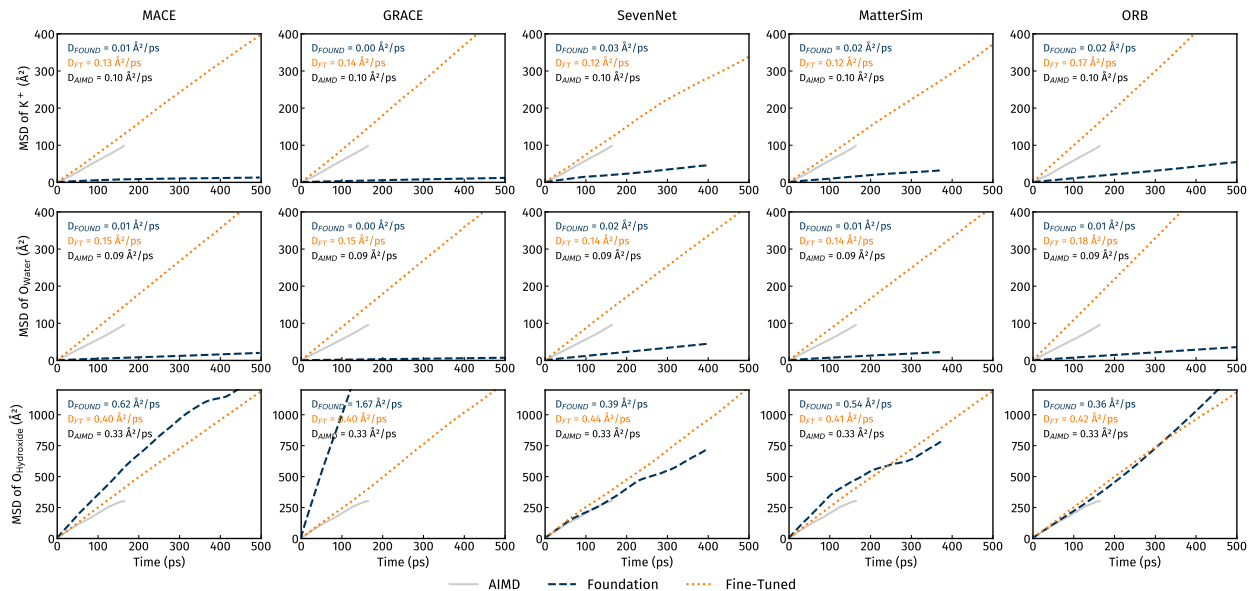


Figure S13: Mean-squared displacements of K⁺, H₂O and OH⁻ in aqueous potassium hydroxide solution computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

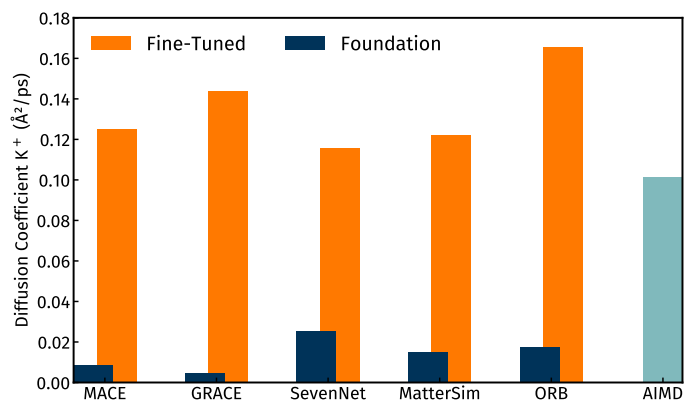


Figure S14: Diffusion coefficients of K^+ in aqueous potassium hydroxide solution computed using different MLIP frameworks from the mean-square displacements (see Figure S13). Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

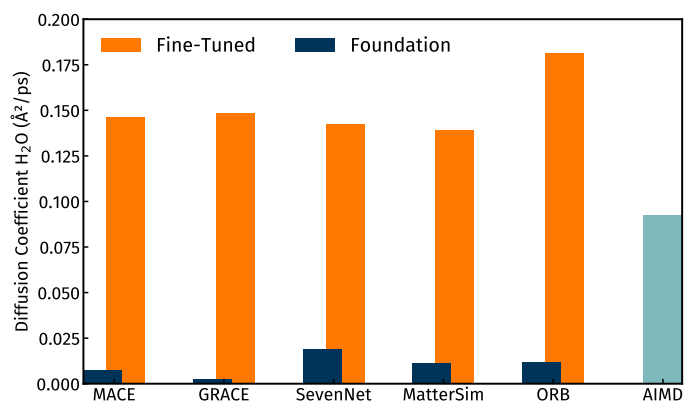


Figure S15: Diffusion coefficients of H_2O in aqueous potassium hydroxide solution computed using different MLIP frameworks from the mean-square displacements (see Figure S13). Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

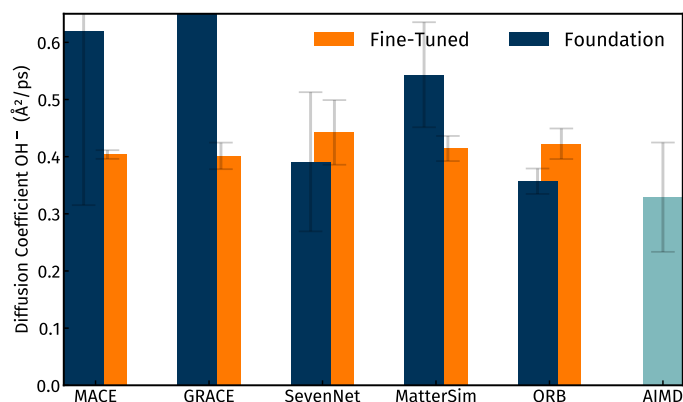


Figure S16: Diffusion coefficients of OH⁻ in aqueous potassium hydroxide solution computed using different MLIP frameworks from the mean-square displacements (see Figure S13). Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

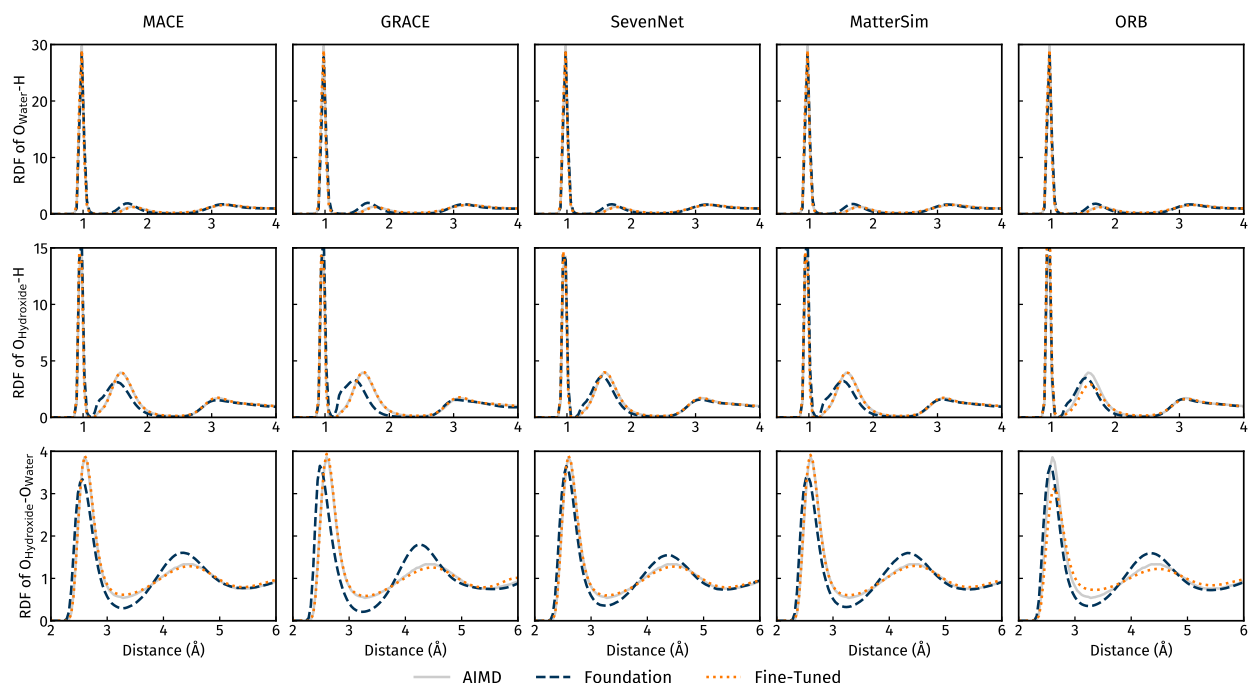


Figure S17: Radial-distribution functions of O_{Water}-H, O_{Hydroxide}-H and O_{Hydroxide}-O_{Water} in aqueous potassium hydroxide solution computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

System E: L-pyroglutamate-NH₄

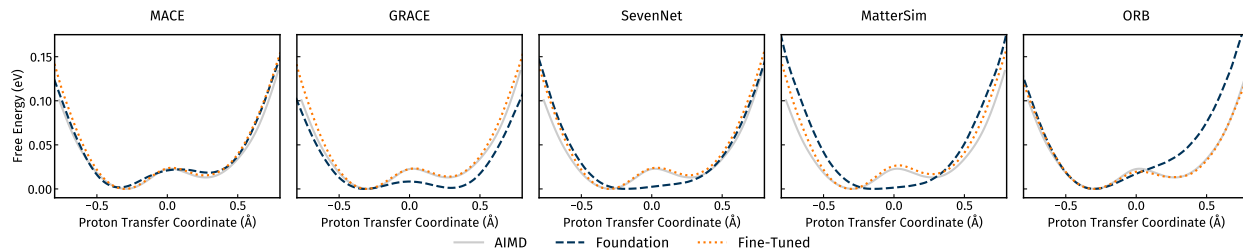


Figure S18: Free energy profiles along the proton transfer coordinate of the short-hydrogen-bond in L-pyroglutamate-NH₄ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

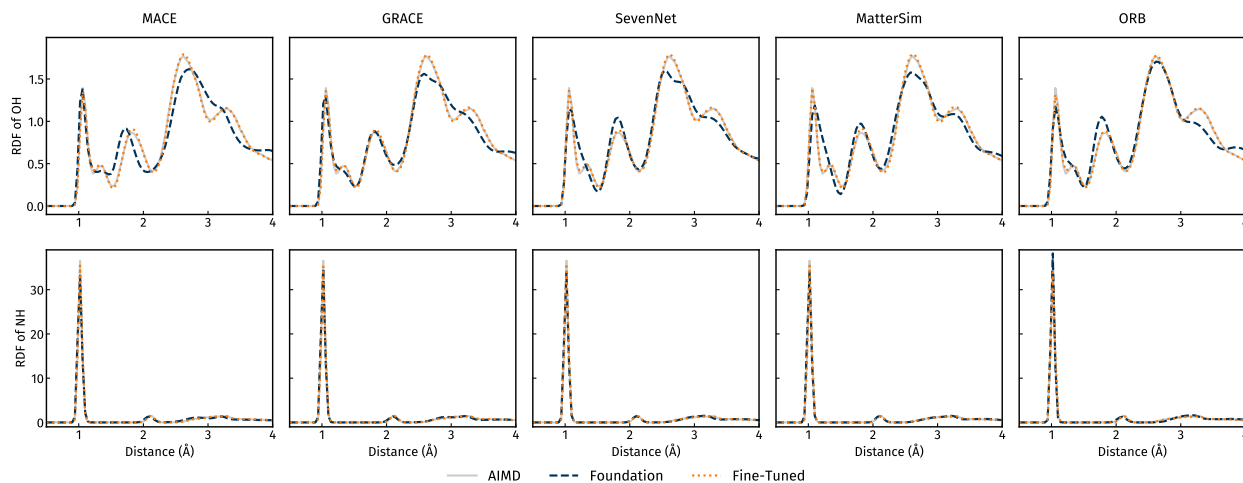


Figure S19: Radial-distribution functions of O-H and N-H in L-pyroglutamate-NH₄ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against AIMD reference data.

System E: MoS₂

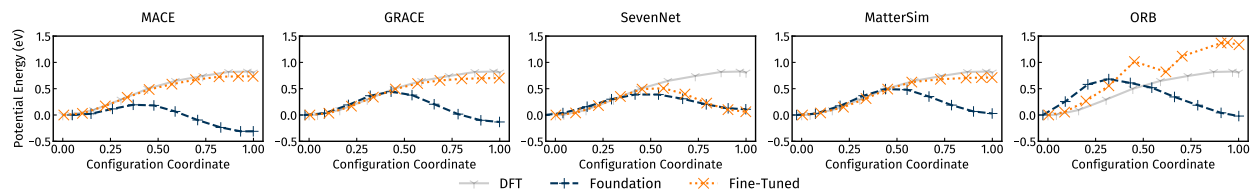


Figure S20: Potential energy curves for a sulfur jump into a sulfur vacancy cluster in MoS₂ computed using different MLIP frameworks. Results from the foundation model and the fine-tuned foundation model are compared against DFT reference data. Note: Fine-tuning attempts for the SevenNet foundation model did not yield models capable of reproducing the reference potential energy curve, even after hyperparameter optimization.

References

- (S1) Grunert, M.; Großmann, M.; Hänseroth, J.; Flötotto, A.; Oumard, J.; Wolf, J. L.; Runge, E.; Dreßler, C. Modeling Complex Proton Transport Phenomena - Exploring the Limits of Fine-Tuning and Transferability of Foundational Machine-Learned Force Fields. *The Journal of Physical Chemistry C* **2025**, *129*, 9662–9669.