Migration as a Probe: A Generalizable Benchmark Framework for Specialist vs. Generalist Machine-Learned Force Fields

Yi Cao
Johns Hopkins University
Baltimore, MD 21218
ycao73@jh.edu

Paulette Clancy*
Johns Hopkins University
Baltimore, MD 21218
pclancy3@jhu.edu

Abstract

Machine-learned force fields (MLFFs), particularly pre-trained foundation models, are revolutionizing computational materials science by enabling *ab initio*-level accuracy at the length- and time-scales of classical molecular dynamics (MD). However, their rapid proliferation presents a critical strategic question: Should researchers train bespoke "specialist" models from scratch, fine-tune large "generalist" foundation models, or employ hybrid approaches? The trade-offs in data efficiency, predictive accuracy, computational cost, and susceptibility to out-of-distribution failure remain poorly understood, as does the fundamental question of how different training paradigms affect learned physical representations.

Here, we introduce a systematic benchmarking framework that addresses this question using defect migration pathways, evaluated via Nudged Elastic Band trajectories, as diagnostic probes that simultaneously test interpolation and extrapolation capabilities. Using Cr-doped Sb_2Te_3 as a technologically relevant 2D material case study, we benchmark multiple training strategies within the MACE architecture across equilibrium, kinetic (atomic migration), and mechanical (interlayer sliding) properties.

Our key findings reveal that while all models adequately capture equilibrium structures, their predictions for non-equilibrium processes diverge dramatically. Targeted fine-tuning substantially outperforms both from-scratch and zero-shot approaches for kinetic properties, but induces catastrophic forgetting of long-range physics. Critically, analysis of learned representations shows that different training paradigms produce fundamentally distinct, non-overlapping latent space encodings, suggesting they capture different aspects of the underlying physics.

This work provides practical guidelines for MLFF development and establishes migration-based probes as an efficient, broadly applicable strategy for distinguishing model quality. We hope that this approach will offer a diagnostic framework that links performance to learned representations, paving the way for more intelligent, uncertainty-aware active learning strategies.

1 Introduction

The discovery and design of novel two-dimensional (2D) van der Waals (vdW) materials [Novoselov et al., 2016, Manzeli et al., 2017, Guo et al., 2021, Liu et al., 2024] continues to drive innovation in spintronics [He et al., 2022], quantum devices [Qian, 2024, Song and Gabor, 2018, Nayak, 2025], and energy technologies [Jin et al., 2019, Gautam et al., 2024]. A particularly powerful strategy for tuning the properties of these layered materials involves doping—the insertion of guest atoms between

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Discovery (AI4Mat) Workshop.

vdW layers [Novoselov et al., 2016, Whittingham, 2004]. This approach has enabled remarkable advances, from lithium-ion batteries [Winter et al., 1998, Hu et al., 2024] to the engineering of magnetic semiconductors [Gibertini et al., 2019] and quantum anomalous Hall systems [Tokura et al., 2019].

Among 2D materials, topological insulators such as antimony telluride (Sb_2Te_3) [Zhang et al., 2009, Guo et al., 2015] represent an especially promising class for doping studies. The doping of transition metals has proven effective across multiple topological systems: iron in Bi_2Se_3 creates ferromagnetic order [Checkelsky et al., 2014, Kulbachinskii et al., 2012, Haazen et al., 2012], manganese in Bi_2Te_3 enables tunable magnetic properties [Klimovskikh et al., 2020, Fu et al., 2021], and vanadium in Sb_2Te_3 modifies electronic structure [Sun et al., 2023, Zhang et al., 2021]. The doping of chromium (Cr) into Sb_2Te_3, the focus of this work, offers a compelling route to engineer its magnetic and topological properties for spintronic applications [Cortie et al., 2020, Han et al., 2013, de A. Deus et al., 2021]. Predicting the stability and dynamics of these complex systems is essential for guiding experimental synthesis, a task for which large-scale Molecular Dynamics (MD) simulations are indispensable [Zhang et al., 2023, Thompson et al., 2022], provided a suitably representative force field is available. However, the atomic-scale processes governing functionality—including dopant migration and thermal stability—occur on time and length scales far beyond the reach of first-principles methods like Density Functional Theory (DFT) [Kohn and Sham, 1965, Perdew et al., 1996a, Sohier et al., 2017, Choudhary et al., 2017].

Machine-Learned Force Fields (MLFFs) [Unke et al., 2021, Smith et al., 2017, Chen et al., 2017, Batatia et al., 2022, Vandermause et al., 2022, Batzner et al., 2022, Bartók et al., 2010, Zeng et al., 2023, Wines and Choudhary, 2025, Musaelian et al., 2023, Chen and Ong, 2022] have emerged as a powerful solution to this accuracy-versus-cost dilemma. The recent development of large-scale, pre-trained "foundation models" such as CHGNet [Deng et al., 2023], M3GNet/MEGNet [Chen et al., 2019], MACE-MP [Batatia et al., 2022], MatterSim [Yang et al., 2024], and UMA [Wood et al., 2025] marks a paradigm shift [Merchant et al., 2023]. This shift presents researchers with a critical and unresolved dilemma: for a specific system, is it more effective to invest significant resources to build a robust *specialist* model from scratch [Zhang, 2018, Smith et al., 2019, Chmiela et al., 2019], or to adapt a *generalist* foundation model through fine-tuning? While generic benchmarks like Matbench [Dunn et al., 2020] validate their broad utility, their reliability for specific, high-fidelity applications is an open question [Friederich et al., 2021, Deringer et al., 2019]. A valid concern is that these models may fail to capture nuanced interactions governing critical processes [Schran et al., 2021, Kapil et al., 2022, Grisafi et al., 2019]. Fine-tuning is a promising solution, but naïve strategies can lead to *catastrophic forgetting*, where general knowledge is erased. Key questions of data efficiency [Janet et al., 2019, Lookman et al., 2019], stability [Zuo et al., 2020, Jinnouchi et al., 2019], and transferability [Zeni et al., 2021, Vandermause et al., 2020] for non-equilibrium processes remain [Bernstein et al., 2019, Zhang et al., 2019].

Furthermore, when an MLFF-driven simulation fails, the underlying cause is often treated as a "black box." The default response—indiscriminately adding more data—is a brute-force approach that wastes resources. We propose that the key to overcoming this is to diagnose failure using physical probes. Specifically, we argue that migration pathways, evaluated via Nudged Elastic Band (NEB) trajectories [Miskin et al., 2025, Ruttinger et al., 2022], provide a powerful and generalizable probe that simultaneously tests interpolation and extrapolation performance. Such migration-based probes impose stricter requirements than equilibrium-only tests, offering a sharper distinction between models and a path toward more intelligent, uncertainty-aware active learning.

This work addresses these gaps by presenting a systematic benchmark of bespoke training and fine-tuning strategies for the case study of Cr-doped Sb_2Te_3 using the MACE architecture. By combining latent-space analysis with migration-based probes, we move beyond traditional accuracy metrics to evaluate the dynamic stability and kinetic predictions of each model. Our study aims to serve as both a practical guide for selecting MLFF training strategies and a diagnostic framework for designing more data-efficient learning loops. To this end, we address the following key questions:

1. How do bespoke MLFFs, trained from scratch, compare against a large pre-trained foundation model in terms of accuracy, stability, and data efficiency for simulating Cr-doped Sb_2Te_3?

2. What are the performance and stability trade-offs for fine-tuning strategies?

3. Can migration pathways serve as a generalizable and efficient probe to benchmark specialist versus generalist models, and can latent-space analysis reveal signatures of failure to guide data acquisition?

2 Methodology

We employed a systematic approach to benchmark Machine-Learned Force Fields (MLFFs) for Cr-doped Sb_2Te_3 [1], comparing specialist models trained from scratch against fine-tuned foundation models.

2.1 Data Generation and Model Training

Reference data were generated using Density Functional Theory (DFT) calculations employing the PBE exchange-correlation functional [Perdew et al., 1996b]. The dataset comprises approximately 20,000 atomic configurations of a $4 \times 2 \times 1$ supercell of Sb_2Te_3 obtained from *ab initio* molecular dynamics (AIMD) simulations conducted at three temperatures (300 K, 600 K, and 1200 K) using a 1 fs integration timestep.

All machine learning force field models were trained using the MACE architecture, [Batatia et al., 2022] which is an equivariant message-passing neural network designed for accurate representation of atomic interactions and many-body correlations.

Four distinct training strategies were evaluated:

1. **Scratch**: MACE model trained exclusively on our AIMD dataset
2. **Foundation**: Pre-trained MACE-MP model used without fine-tuning
3. **FT-600K**: Foundation model fine-tuned on 5% of data from 600K trajectories
4. **FT-Multi_T**: Foundation model fine-tuned on multi-temperature data

2.2 Evaluation Framework

Models were assessed through: (i) MD simulations at 300K and 600K to evaluate structural stability and transport properties, (ii) nudged elastic band calculations for migration barriers, and (iii) representation learning analysis using t-SNE visualization of physically interpretable descriptors. Performance metrics included force/energy accuracy, dynamic stability, barrier prediction errors, and latent space clustering quality. Detailed computational parameters are provided in Appendix B.

3 Results and Discussions

3.1 Equilibrium Properties: Foundation for Comparison

We considered four candidate models: a "zero-shot" foundation model, a bespoke model trained from scratch, and two fine-tuned variants of the foundation model. To make consistent evaluations to benchmark the performance of these models, we first assessed the performance of each training strategy via their performance in MD simulations (Fig. 2a). This provides a strong test of each model's ability to generate not only thermodynamic and structural properties but also accurately reproduce key dynamic, and transport properties. Each model was used to drive a 200 ps simulation, with the resulting trajectories then subjected to the same set of post-processing analyses to evaluate their resulting key physical properties; shown in Figure 2.

All MACE models—foundation, scratch, and fine-tuned variants—successfully reproduce the equilibrium structure of Cr-doped Sb_2Te_3, with RDF analysis showing excellent agreement with AIMD references for all atomic pair correlations (Fig. 7, see Appendix A for detailed analysis). While all

[1] We chose this material system because it serves as an ideal testbed: its anisotropic 2D structure offers distinct migration environments—in-gap diffusion within quintuple layers (in-distribution) and interlayer migration across vdW gaps (out-of-distribution). This duality enables efficient sampling from equilibrium to high-energy transition states, addressing technologically relevant processes where both thermodynamic stability and kinetics are critical for doping engineering.

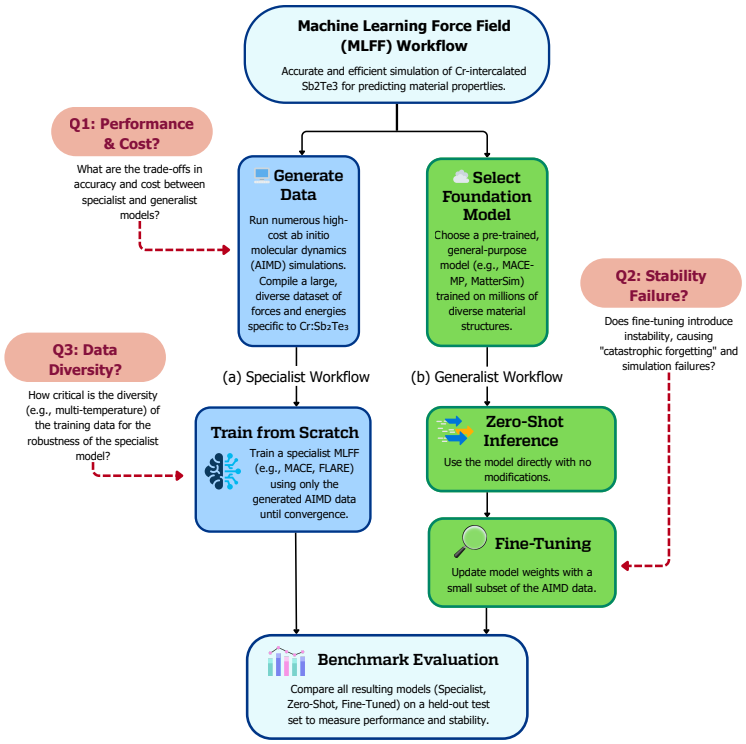

Figure 1: **Conceptual Diagram.** A flowchart illustrating two competing workflows: (a) training a specialist MLFF from scratch, and (b) fine-tuning a generalist foundation model. The diagram highlights the benchmarking questions this paper addresses.

models maintain stable thermodynamic ensembles at 600K (Fig. 2c), the zero-shot foundation model exhibits a persistent pressure offset, likely due to its training on 0K equilibrium structures rather than finite-temperature configurations.

Despite similar performance on local structural (RDF) and short-time dynamic (VACF) properties (Fig. 2e), the models diverge significantly in their predictions of long-timescale transport phenomena. Fine-tuned models predict higher diffusion coefficients than both foundation and scratch models (Fig. 2d), with the multi-temperature fine-tuned variant showing the largest enhancement. This suggests that exposure to high-temperature training data creates a flatter potential energy (PES) that persists even at lower temperatures.

Notably, thermal conductivity calculations reveal differences in how models capture collective vibrational modes. The foundation model's thermal conductivity decays rapidly toward zero, failing to sustain the long-range phonon correlations characteristic of the crystalline structure. In contrast, models trained or fine-tuned on system-specific data maintain these correlations, though the 600K-only model exhibits an anomalous peak suggesting potential structural instability. These results demonstrate that validating MLFFs solely on static structures and short-time dynamics is insufficient—a comprehensive assessment must include transport properties that probe long-range, collective phenomena.

3.2 Non-Equilibrium Diffusion Pathways: The Critical Test of Generalization

A rigorous test of the generalizability of a machine learning force field extends beyond its ability to reproduce thermodynamic properties. It must also accurately describe the potential energy surface (PES) in regions far from the training data's energetic minima, such as the high-energy transition states that govern kinetic processes. To this end, we evaluated the performance of our MACE model on the challenging task of predicting the diffusion barrier for a fundamental migration event in the Cr-doped Sb_2Te_3 system. This task is substantially more demanding than constant-temperature MD simulations (e.g., at 600 K), as it requires the model to capture not only accurate energies, but also

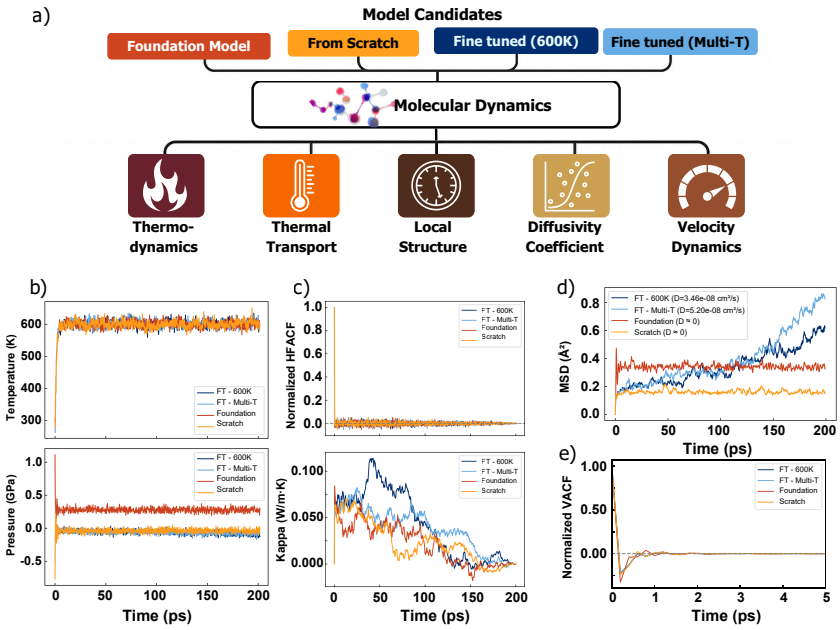

Figure 2: **Comprehensive Benchmarking of MLFF Performance in Molecular Dynamics Simulations.** (a) A schematic of the unified evaluation protocol. Four candidate models—a zero-shot foundation model, a bespoke model trained from scratch, and two fine-tuned variants—are each used to drive a 200 ps MD simulation. The resulting trajectories are then subjected to a uniform set of post-processing analyses to evaluate key physical properties. (b) Thermodynamic stability, demonstrated by the evolution of temperature and pressure over the simulation, which remain stable around their target values for all models. (c) Thermal transport properties, showing the Heat Flux Autocorrelation Function (HFACF) and its running integral to compute thermal conductivity (κ). (d) The Mean Squared Displacement (MSD), used to assess atomic mobility and calculate the diffusion coefficient. (e) The Velocity Autocorrelation Function (VACF), which describes the system's underlying dynamics.

the subtle curvature of the PES at a first-order saddle point—a stringent test of the model's learned representation of interatomic interactions.

3.2.1 Local Migration Events

To perform a direct and controlled evaluation of the learned PES for each model, we used the minimum energy pathway (MEP) obtained from our reference DFT nudged elastic band (NEB) calculation as a fixed geometric trajectory. By calculating the single-point energy for each of these pre-defined images, we can decouple the model's energetic accuracy from its structural relaxation behavior, providing a direct probe of its ability to describe the reaction pathway.

Our DFT calculations, serving as the ground-truth reference, establish a migration energy barrier (E_{a}) of 0.34 eV for this process. The results, presented in Fig. 3d, reveal a clear hierarchy in performance and highlight the considerable impact of training strategy on the prediction of kinetic barriers. The two baseline strategies—training from scratch and using the zero-shot foundation model—demonstrate the fundamental challenges of specialization and generalization. The MACE Scratch model, despite being trained on a comprehensive in-house dataset, exhibits a catastrophic failure in predicting the barrier, overestimating the activation energy by over 4 eV. This is a classic signature of poor extrapolation. Even with tens of thousands of AIMD frames, the high-energy, low-probability configurations corresponding to the transition state are insufficiently sampled. Consequently, the model overfits to the more prevalent, near-equilibrium states and fails dramatically when asked to evaluate this critical, out-of-distribution rare event.

Conversely, the MACE Foundation model provides a qualitatively plausible prediction, capturing the smooth, convex shape of the energy barrier. This is a testament to the power of pre-training on

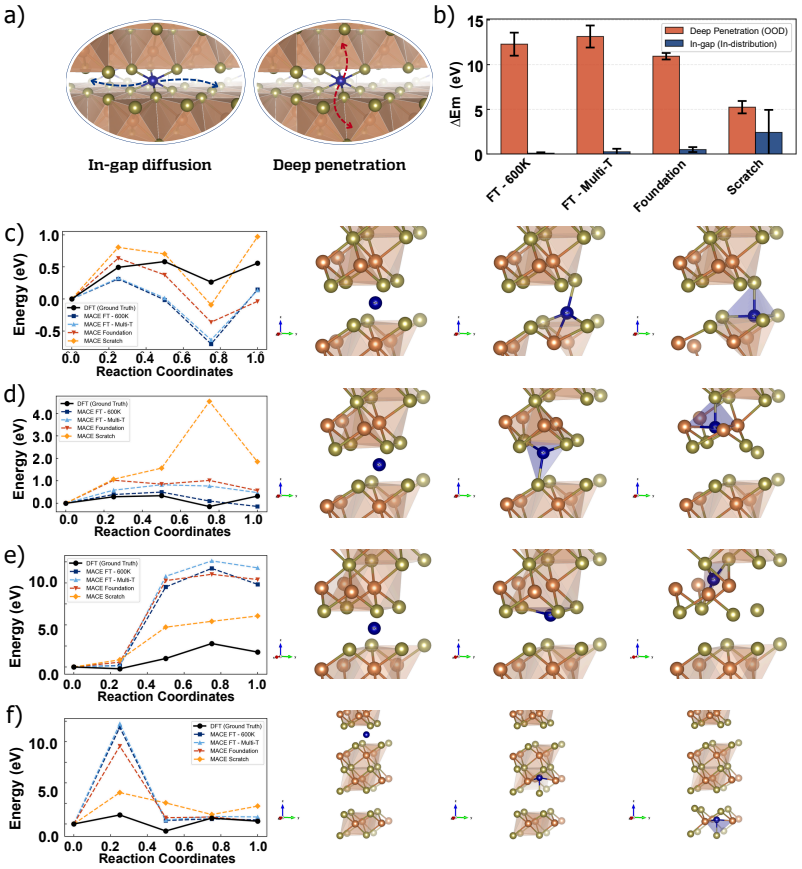

Figure 3: **Comparative analysis of MLFF training strategies for predicting atomic migration barriers.** (a) Schematic illustration of the simulated Cr atom migration pathway between two stable sites within the Sb_2Te_3 bilayer. (b) Bar plot showing the migration energy prediction errors for various models, spanning bespoke training to advanced active learning approaches. (c–f) Comparison of the minimum energy pathway (MEP) profiles for the Cr migration process. The solid black line denotes the ground-truth DFT reference, while dashed lines represent predictions from MACE models trained with different strategies.

millions of diverse structures; the model has learned a general understanding of physical interactions. However, it is quantitatively inaccurate, overestimating the barrier by approximately 0.7 eV—an unacceptably high amount—and misplacing the transition state along the reaction coordinate. This behavior is characteristic of a "softened" or "averaged" potential energy surface, a known trait of foundation models trained to generalize across vast chemical spaces. While broadly correct, the foundation model lacks the sharp, system-specific features of the true PES for Cr:Sb_2Te_3, effectively smoothing over the precise details required for high-fidelity kinetic predictions. The pronounced failures of these two baseline strategies underscore the necessity of a hybrid approach, motivating our investigation into fine-tuning.

The Critical Role of Task-Specific Fine-Tuning The dramatic improvement seen in the fine-tuned models underscores the necessity of specializing the foundation model's knowledge. The MACE FT–600K model, fine-tuned specifically on data generated at the target temperature of the migration event, achieves remarkable accuracy, with a barrier error of only 0.16 eV. This demonstrates that targeted fine-tuning effectively "sharpens" the softened PES of the foundation model. By providing a small but highly relevant set of in-domain data, we enable the model to learn the specific local atomic interactions required to accurately resolve the transition state structure and energy.

Intriguingly, the MACE FT–Multi-T model, which was fine-tuned on a larger and more diverse dataset spanning multiple temperatures, yields a less accurate barrier than the model trained at a specific

temperature, here 600 K. While its predictions for the initial and final states (the thermodynamic endpoints) are accurate, its description of the kinetic barrier is a compromise, retaining some of the "softened" character of the original foundation model. This leads to an important insight: for MLFFs, more data is not axiomatically better. The relevance of the fine-tuning data to the specific task is paramount. For predicting thermodynamic properties, a multi-temperature dataset is superior. For predicting a specific kinetic process, a dataset rich in configurations relevant to that process's temperature regime is more effective. These findings collectively highlight a critical challenge in the practical application of MLFFs. The failure of the from-scratch model illustrates the difficulty of adequately sampling rare events, while the inaccuracies of the foundation model reveal the limitations of a purely generalist approach.

The success of targeted fine-tuning points the way forward, but the divergent results of the 600K and Multi-T models prove that the data selection process is non-trivial. This motivates the need for more advanced methods, such as the uncertainty-guided active learning explored in our work, which can intelligently identify and acquire the most informative data to improve model robustness and accuracy in a data-efficient manner.

To probe the generalization capabilities of the different MLFF training strategies, we evaluated their performance on two distinct Cr migration pathways (Fig. 3a) with fundamentally different characteristics. The first pathway, *in-gap diffusion*, involves Cr migration within the vdW gap between Sb_2Te_3 quintuple layers, with low energy barriers and configurations similar to the training data, thereby testing interpolative accuracy.

In contrast, the second pathway, *deep penetration*, requires the Cr atom to move vertically from the vdW gap and penetrate directly into the interior of a QL, ultimately reaching a deeply buried site within the layer. This migration path is associated with significantly higher energy barriers and accesses high-distortion configurations that are not well-represented in the training data—making it an out-of-distribution (OOD) scenario with potentially more metastable configurations along the pathway. Testing both pathways allows us to evaluate model accuracy on familiar configurations and extrapolative power in challenging regions of the configuration space.

In-Gap Diffusion: A Test of Interpolative Accuracy For the in-gap diffusion pathway (Fig. 3c–d), where the Cr atom moves within the vdW gap or just shallowly penetrates the interface, the atomic environments along the minimum energy path (MEP) are reasonably similar to the near-equilibrium states sampled during the AIMD simulations. In this regime, the performance hierarchy is clear. The fine-tuned models, particularly MACE FT–600K, demonstrate excellent agreement with the DFT reference, accurately predicting both the thermodynamic endpoints and the kinetic barrier (Fig. 3d). This success highlights that, when the task lies within the domain of the training data, fine-tuning effectively "sharpens" the generalist foundation model's potential energy surface (PES) to capture system-specific details. In contrast, the MACE Scratch model, despite being trained on the same data, fails significantly, underscoring its inability to learn the subtle energy differences required to resolve the transition state from a limited dataset.

Deep Penetration: A Test of Extrapolative Robustness The "deep penetration" pathway provides a much more stringent test of model robustness. This path involves significant lattice distortion as the Cr atom pushes its way through the covalently bonded quintuple layer (QL), creating atomic environments that are far from those seen in the training data (Fig. 3a).

While the fine-tuned and foundation models remain accurate for the stable initial and final states (an interpolative task), they fail catastrophically in predicting the energy of the transition state, overestimating the barrier by a large margin. This is a classic example of out-of-distribution failure, where the inductive biases learned from near-equilibrium data do not generalize to highly distorted, high-energy configurations. The models have learned to be a "materials expert" for stable structures but remain a "naïve physicist" for unseen, strained states.

The MACE Scratch model, which performed least well of the four models on the in-gap pathway, yields the lowest barrier error for this out-of-distribution task (Fig. 3e, f). This is not because the scratch model is "better"; rather, its globally inaccurate and likely unphysical PES happens, by chance, to be less pathologically incorrect in this specific high-energy region than the foundation model's PES. The pre-trained model's failure suggests that its supposedly general knowledge contains

strong implicit assumptions about near-equilibrium physics that break down dramatically when extrapolating.

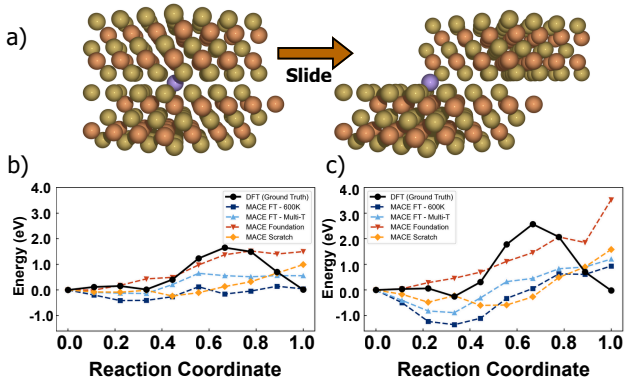

Figure 4: **Evaluation of MACE models on collective lattice displacement: interlayer sliding in Sb_2Te_3.** (a) Schematic illustration of the bilayer sliding process in Sb_2Te_3, where the top layer (yellow/orange atoms) slides relative to the bottom layer along the crystallographic direction. The purple sphere indicates the position of a Cr dopant when present. (b) Energy barriers for interlayer sliding in pristine Sb_2Te_3 as calculated by DFT (black, ground truth) and various MACE models. The reaction coordinate represents the normalized sliding distance from the initial to final configuration. (c) Energy barriers for the same sliding process in Cr-doped Sb_2Te_3.

3.3 Interlayer Sliding: Probing Robustness to Non-Local, Collective Displacements

Having tested the models' response to local, high-distortion events, we evaluated their robustness to a different class of out-of-distribution challenge: large-scale, collective lattice displacements. To this end, we simulated the shearing of one Sb_2Te_3 layer relative to the other, a process governed by the subtle corrugations of the interlayer vdW potential energy surface (Fig. 4a). This task is particularly challenging for MLFFs as the local atomic environment of any given atom changes only minimally, while the global configuration undergoes a significant transformation that crosses periodic boundaries. The configurations along this sliding path were not present in the training data.

We first examined the case of pure, undoped Sb_2Te_3 (Fig. 4b). The results reveal a clear trade-off between the models. The MACE Foundation model provides the most reasonable estimate of the energy barrier's shape and magnitude, suggesting its vast pre-training on bulk crystals has endowed it with a better implicit understanding of such periodic, mechanical deformations. However, it fails to maintain translational symmetry, incorrectly predicting the final state to be higher in energy than the identical initial state. This energy drift is a clear artifact, indicating a failure to perfectly respect the periodic nature of the simulation cell under large displacements. The MACE Scratch model exhibits a similar version of this artifact.

Conversely, the fine-tuned models (MACE FT–600K and Multi-T) significantly underestimate the energy barrier. This suggests a compelling hypothesis: the process of fine-tuning, while "sharpening" the PES for local chemistry around the Cr dopant, has degraded the model's learned representation of the weaker, long-range interlayer physics inherited from the foundation model. The optimization has prioritized local accuracy at the expense of non-local, collective interactions.

This trend is largely mirrored in the Cr-doped system (Fig. 4c), confirming that this is a fundamental behavior of the models rather than an effect specific to the dopant. The failure of all models to perfectly capture both the barrier height and the endpoint periodicity highlights a key limitation of local-descriptor-based MLFFs. Phenomena like shearing, stacking faults, and dislocation glide are inherently non-local. While the models excel at describing local bonding and coordination, they can struggle to enforce long-range physical constraints that extend beyond their cutoff radius. This underscores the need for careful validation when using standard MLFFs to study mechanical properties and points towards future work in developing training sets that explicitly include these collective deformation modes or exploring architectures designed to capture long-range physics more effectively.

3.4 Representation Learning Analysis

To elucidate the origins of the observed performance differences, we conducted a multi-faceted analysis of the learned representations. We projected the high-dimensional atomic environment descriptors from each model's 600 K MD trajectory into a two-dimensional space using two complementary techniques: t-distributed stochastic neighbor embedding (t-SNE) [Maaten and Hinton, 2008] to assess representational dissimilarity, and Potential of Heat-diffusion for Affinity-based Trajectory Embedding (PHATE) [Moon et al., 2019] to reveal the underlying geometric structure of the system's dynamics (Fig. 5).

The t-SNE projection (Fig. 5a) confirms that models trained with different strategies learn qualitatively distinct encodings. The MACE Foundation and MACE Scratch models occupy well-separated regions of the latent space, a finding quantitatively supported by their high average silhouette scores (Fig. 5d). This highlights the fundamental dissimilarity between the generalist prior and the specialist model trained from scratch. Crucially, the two fine-tuned models occupy an intermediate region, demonstrating that fine-tuning acts as a bridge between these two representational paradigms.

While t-SNE shows *that* the representations are different, PHATE reveals *why* it matters for physical prediction. The PHATE projection (Fig. 5b) maps the continuous time-evolution of the system to a low-dimensional "dynamical manifold"—an arc-like structure whose geometry is intrinsically linked to the potential energy surface (Fig. 5c). Along the primary axis of this manifold (PHATE 1), all specialist models (Scratch and FT) trace a similar region, indicating they have all learned the dominant, low-energy dynamics of the system.

The key insight comes from the separation along the second axis (PHATE 2). The scratch-trained model's representation is isolated from fine-tuned models in a distinct region (higher PHATE 2 values). This suggests it has learned a brittle, overfit representation, effectively memorizing a narrow path through the energy landscape without understanding the broader physical context. In stark contrast, the fine-tuned models are constrained to a different region of the manifold. This is the signature of regularization by pre-training; the foundation model's robust physical prior prevents the fine-tuned models from collapsing into the brittle state of the scratch model. Instead, they learn the system-specific dynamics *within the context of a general and smooth physical manifold*.

This geometric difference in the latent space provides a direct mechanistic explanation for the models' performance on the diffusion task. Diffusion is a rare-event process that requires the model to generalize to out-of-distribution transition states. The scratch model's isolated manifold corresponds to a noisy and untrustworthy PES outside its training domain, leading to unphysical dynamics and a flattened MSD plot. Conversely, the fine-tuned models' robust, regularized manifold corresponds to a globally smoother and more reliable PES. This enables them to accurately predict the energy barriers and forces along the diffusion pathway, resulting in physically correct, linear MSD behavior. Fine-tuning succeeds not by creating a simple hybrid, but by aligning a general physical prior with the specific dynamical manifold of a target system, thereby enabling robust generalization to complex, long-timescale phenomena.

4 Conclusion

We benchmarked specialist (from-scratch) and generalist (foundation-based) machine-learned force fields (MLFFs) for Cr-doped Sb_2Te_3, emphasizing migration pathways as a diagnostic probe. NEB trajectories revealed that, while all models reproduce equilibrium structures, their performance diverges sharply for kinetic tasks. Migration thus provides a generalizable benchmark that tests both interpolation and extrapolation, exposing weaknesses invisible to equilibrium-only validation.

Our results address the core questions posed in the introduction. First, both from-scratch and zero-shot foundation models fail for migration barriers: the former struggles with extrapolation, while the latter produces overly averaged potentials. Second, task-specific fine-tuning recovers kinetic accuracy, but at the cost of degraded performance for long-range physics such as interlayer sliding. Third, latent-space analysis shows these paradigms encode fundamentally distinct, non-overlapping physical representations, explaining hidden extrapolation failures.

These findings reframe the specialist vs. generalist debate: the choice is not just about efficiency but about qualitatively different physics learned by each model. Foundation models excel at capturing broad chemical trends but require careful adaptation for system-specific kinetics, while specialist

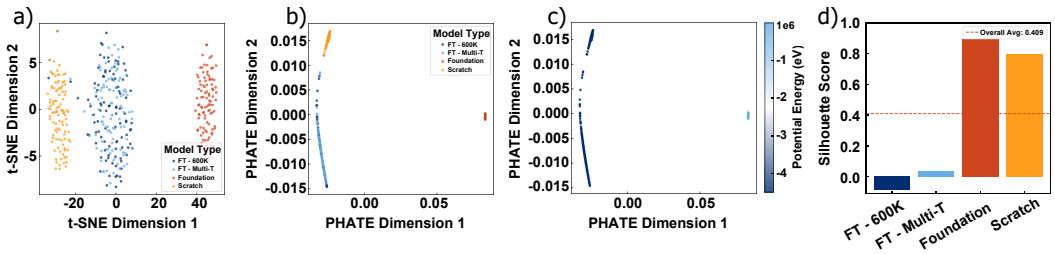

Figure 5: **Representation Analysis Reveals How Fine-Tuning Aligns a Physical Manifold to Enable Generalization.** Projections of atomic environment descriptors from 600 K MD trajectories. **(a)** t-SNE projection shows that Foundation (red) and Scratch (orange) models produce highly separated representations, while fine-tuned models (blue) are intermediate. **(b)** PHATE projection reveals the system's continuous dynamical manifold. Note the clear separation between the brittle scratch model representation and the regularized fine-tuned models along the vertical axis (PHATE 2). **(c)** The same PHATE embedding colored by potential energy confirms that the manifold's geometry corresponds to the physical energy landscape. The foundation model's large energy offset indicates its nature as an uncalibrated prior. **(d)** Average silhouette scores (from t-SNE) quantify the representational dissimilarity. The results demonstrate that fine-tuning succeeds by constraining a generalist representation to the specific physical manifold of the target system, a process that regularizes the model and enables accurate prediction of complex dynamics like diffusion.

models offer more reliable local accuracy at the cost of transferability. Migration-based probes, coupled with latent-space diagnostics, offer a practical route to uncover hidden failures and guide uncertainty-aware active learning.

Limitations and Future Work Our current study focuses on a single material system and migration mechanism as a proof-of-concept case study. Future work should extend this framework to multiple material classes, investigate uncertainty quantification methods for identifying unreliable predictions, and develop hybrid approaches that combine the transferability of foundation models with the precision of specialist training. Additionally, exploring active learning strategies guided by migration-based diagnostics could enable more efficient data collection for kinetic properties. This framework has the potential to generalize beyond the present case, enabling more robust and uncertainty-aware MLFF development for accelerated materials discovery.

Acknowledgments and Disclosure of Funding

This work was supported by the United States Department of Defense-funded Center of Excellence for Advanced Electro-photonics with 2D materials—Morgan State University, under Grant No. W911NF2120213. The authors thank Frank Gardea and Owen Vail, the cooperative agreement managers of the Center, for their interest and support. They also thank Dr. Ramesh Budhani for his guidance on the project. Yi thanks Johns Hopkins University for their support in AY 24–25. Computational resource support was provided by the petascale Hopkins facility, Advanced Research Computing at Hopkins (ARCH) (rockfish.jhu.edu), supported by National Science Foundation award OAC 1920103. Partial funding for ARCH's infrastructure was originally provided by the State of Maryland.

References

K. S. Novoselov, A. Mishchenko, A. Carvalho, and A. H. Castro Neto. 2D materials and van der Waals heterostructures. *Science*, 353(6298):aac9439, July 2016. doi: 10.1126/science.aac9439. URL `https://www.science.org/doi/full/10.1126/science.aac9439`. Publisher: American Association for the Advancement of Science.

Sajedeh Manzeli, Dmitry Ovchinnikov, Diego Pasquier, Oleg V. Yazyev, and Andras Kis. 2D transition metal dichalcogenides. *Nature Reviews Materials*, 2(8):1–15, June 2017. ISSN 2058-8437. doi: 10.

1038/natrevmats.2017.33. URL https://www.nature.com/articles/natrevmats201733. Publisher: Nature Publishing Group.

Hao-Wei Guo, Zhen Hu, Zhi-Bo Liu, and Jian-Guo Tian. Stacking of 2D Materials. *Advanced Functional Materials*, 31(4):2007810, 2021. ISSN 1616-3028. doi: 10.1002/adfm.202007810. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adfm.202007810. _eprint: https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/adfm.202007810.

Anhan Liu, Xiaowei Zhang, Ziyu Liu, Yuning Li, Xueyang Peng, Xin Li, Yue Qin, Chen Hu, Yanqing Qiu, Han Jiang, Yang Wang, Yifan Li, Jun Tang, Jun Liu, Hao Guo, Tao Deng, Songang Peng, He Tian, and Tian-Ling Ren. The Roadmap of 2D Materials and Devices Toward Chips. *Nano-Micro Letters*, 16(1):1–96, February 2024. ISSN 2150-5551. doi: 10.1007/s40820-023-01273-5. URL https://link.springer.com/article/10.1007/s40820-023-01273-5. Publisher: Springer.

Qing Lin He, Taylor L. Hughes, N. Peter Armitage, Yoshinori Tokura, and Kang L. Wang. Topological spintronics and magnetoelectronics. *Nature Materials*, 21(1):15–23, January 2022. ISSN 1476-4660. doi: 10.1038/s41563-021-01138-5. URL https://www.nature.com/articles/s41563-021-01138-5. Publisher: Nature Publishing Group.

Qi Qian. Van der Waals integration: Enables quantum explorations and innovative devices. *MRS Bulletin*, 49(4):385–390, February 2024. ISSN 1938-1425. doi: 10.1557/s43577-024-00668-y. URL https://link.springer.com/article/10.1557/s43577-024-00668-y. Number: 4 Publisher: Springer.

Justin C. W. Song and Nathaniel M. Gabor. Electron quantum metamaterials in van der Waals heterostructures. *Nature Nanotechnology*, 13(11):986–993, November 2018. ISSN 1748-3395. doi: 10.1038/s41565-018-0294-9. URL https://www.nature.com/articles/s41565-018-0294-9. Publisher: Nature Publishing Group.

Chetan Nayak. Microsoft unveils Majorana 1, the world's first quantum processor powered by topological qubits, February 2025. URL https://azure.microsoft.com/en-us/blog/quantum/2025/02/19/microsoft-unveils-majorana-1-the-worlds-first-quantum-processor-powered-by-topological-qubits/

Qun Jin, Song Jiang, Yang Zhao, Dong Wang, Jianhang Qiu, Dai-Ming Tang, Jun Tan, Dong-Ming Sun, Peng-Xiang Hou, Xing-Qiu Chen, Kaiping Tai, Ning Gao, Chang Liu, Hui-Ming Cheng, and Xin Jiang. Flexible layer-structured Bi2Te3 thermoelectric on a carbon nanotube scaffold. *Nature Materials*, 18(1):62–68, January 2019. ISSN 1476-4660. doi: 10.1038/s41563-018-0217-z. URL https://www.nature.com/articles/s41563-018-0217-z. Publisher: Nature Publishing Group.

Ravi Gautam, Takamasa Hirai, Abdulkareem Alasli, Hosei Nagano, Tadakatsu Ohkubo, Ken-ichi Uchida, and Hossein Sepehri-Amin. Creation of flexible spin-caloritronic material with giant transverse thermoelectric conversion by nanostructure engineering. *Nature Communications*, 15(1):2184, March 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46475-6. URL https://www.nature.com/articles/s41467-024-46475-6. Publisher: Nature Publishing Group.

M. Stanley Whittingham. Lithium Batteries and Cathode Materials. *Chemical Reviews*, 104(10):4271–4302, October 2004. ISSN 0009-2665. doi: 10.1021/cr020731c. URL https://doi.org/10.1021/cr020731c. Publisher: American Chemical Society.

Martin Winter, Jürgen O. Besenhard, Michael E. Spahr, and Petr Novák. Insertion Electrode Materials for Rechargeable Lithium Batteries. *Advanced Materials*, 10(10):725–763, 1998. ISSN 1521-4095. doi: 10.1002/(SICI)1521-4095(199807)10:10<725::AID-ADMA725>3.0.CO;2-Z. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291521-4095%28199807%2910%3A10%3C725%3A%3AAID-ADMA725%3E3.0.CO%3B2-Z. _eprint: https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/%28SICI%291521-4095%28199807%2910%3A10%3C725%3A%3AAID-ADMA725%3E3.0.CO%3B2-Z.

Wanghua Hu, Jinbo Shen, Tao Wang, Zishun Li, Zhuokai Xu, Zhefeng Lou, Haoyu Qi, Junjie Yan, Jialu Wang, Tian Le, Xiaorui Zheng, Yunhao Lu, and Xiao Lin. Lithium Ion Intercalation-Induced

Metal-Insulator Transition in Inclined-Standing Grown 2D Non-Layered Cr2S3 Nanosheets. *Small Methods*, 8(12):2400312, 2024. ISSN 2366-9608. doi: 10.1002/smtd.202400312. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/smtd.202400312. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smtd.202400312.

M. Gibertini, M. Koperski, A. F. Morpurgo, and K. S. Novoselov. Magnetic 2D materials and heterostructures. *Nature Nanotechnology*, 14(5):408–419, May 2019. ISSN 1748-3395. doi: 10.1038/s41565-019-0438-6. URL https://www.nature.com/articles/s41565-019-0438-6. Publisher: Nature Publishing Group.

Yoshinori Tokura, Kenji Yasuda, and Atsushi Tsukazaki. Magnetic topological insulators. *Nature Reviews Physics*, 1(2):126–143, February 2019. ISSN 2522-5820. doi: 10.1038/s42254-018-0011-5. URL https://www.nature.com/articles/s42254-018-0011-5. Publisher: Nature Publishing Group.

Haijun Zhang, Chao-Xing Liu, Xiao-Liang Qi, Xi Dai, Zhong Fang, and Shou-Cheng Zhang. Topological insulators in Bi2Se3, Bi2Te3 and Sb2Te3 with a single Dirac cone on the surface. *Nature Physics*, 5(6):438–442, June 2009. ISSN 1745-2481. doi: 10.1038/nphys1270. URL https://www.nature.com/articles/nphys1270. Publisher: Nature Publishing Group.

Yunfan Guo, Zhongfan Liu, and Hailin Peng. A Roadmap for Controlled Production of Topological Insulator Nanostructures and Thin Films. *Small*, 11(27):3290–3305, 2015. ISSN 1613-6829. doi: 10.1002/smll.201403426. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/smll.201403426. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/smll.201403426.

J. G. Checkelsky, R. Yoshimi, A. Tsukazaki, K. S. Takahashi, Y. Kozuka, J. Falson, M. Kawasaki, and Y. Tokura. Trajectory of the anomalous Hall effect towards the quantized state in a ferromagnetic topological insulator. *Nature Physics*, 10(10):731–736, October 2014. ISSN 1745-2481. doi: 10.1038/nphys3053. URL https://www.nature.com/articles/nphys3053. Publisher: Nature Publishing Group.

V. A. Kulbachinskii, V. G. Kytin, A. A. Kudryashov, and P. M. Tarasov. Thermoelectric properties of Bi2Te3, Sb2Te3 and Bi2Se3 single crystals with magnetic impurities. *Journal of Solid State Chemistry*, 193:47–52, September 2012. ISSN 0022-4596. doi: 10.1016/j.jssc.2012.03.042. URL https://www.sciencedirect.com/science/article/pii/S0022459612002071.

P. P. J. Haazen, J.-B. Laloë, T. J. Nummy, H. J. M. Swagten, P. Jarillo-Herrero, D. Heiman, and J. S. Moodera. Ferromagnetism in thin-film Cr-doped topological insulator Bi2Se3. *Applied Physics Letters*, 100(8):082404, February 2012. ISSN 0003-6951. doi: 10.1063/1.3688043. URL https://doi.org/10.1063/1.3688043.

Ilya I. Klimovskikh, Mikhail M. Otrokov, Dmitry Estyunin, Sergey V. Eremeev, Sergey O. Filnov, Alexandra Koroleva, Eugene Shevchenko, Vladimir Voroshnin, Artem G. Rybkin, Igor P. Rusinov, Maria Blanco-Rey, Martin Hoffmann, Ziya S. Aliev, Mahammad B. Babanly, Imamaddin R. Amiraslanov, Nadir A. Abdullayev, Vladimir N. Zverev, Akio Kimura, Oleg E. Tereshchenko, Konstantin A. Kokh, Luca Petaccia, Giovanni Di Santo, Arthur Ernst, Pedro M. Echenique, Nazim T. Mamedov, Alexander M. Shikin, and Eugene V. Chulkov. Tunable 3D/2D magnetism in the (MnBi2Te4)(Bi2Te3)m topological insulators family. *npj Quantum Materials*, 5(1):54, August 2020. ISSN 2397-4648. doi: 10.1038/s41535-020-00255-9. URL https://www.nature.com/articles/s41535-020-00255-9. Publisher: Nature Publishing Group.

Jia Fu, Jiaxuan Huang, and Fabrice Bernard. Electronic structure, elastic and optical properties of Bi2Te3/Sb2Te3 thermoelectric composites in the periodic-superlattice thin films. *Composites Communications*, 28:100917, December 2021. ISSN 2452-2139. doi: 10.1016/j.coco.2021.100917. URL https://www.sciencedirect.com/science/article/pii/S245221392100293X.

Yuxin Sun, Haixu Qin, Chenglong Zhang, Hao Wu, Li Yin, Zihang Liu, Shengwu Guo, Qian Zhang, Wei Cai, Haijun Wu, Fengkai Guo, and Jiehe Sui. Sb2Te3 based alloy with high thermoelectric and mechanical performance for low-temperature energy harvesting. *Nano Energy*, 107:108176, March 2023. ISSN 2211-2855. doi: 10.1016/j.nanoen.2023.108176. URL https://www.sciencedirect.com/science/article/pii/S2211285523000125.

Liguo Zhang, Toni Helm, Haicheng Lin, Fengren Fan, Congcong Le, Yan Sun, Anastasios Markou, and Claudia Felser. Quantum Oscillations in Ferromagnetic (Sb, V)2Te3 Topological Insulator Thin Films. *Advanced Materials*, 33(41):2102107, 2021. ISSN 1521-4095. doi: 10.1002/adma.202102107. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202102107. _eprint: https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/adma.202102107.

David Cortie, Weiyao Zhao, Zengji Yue, Zhi Li, Abuduliken Bake, Olexandra Marenych, Zeljko Pastuovic, Mitchell Nancarrow, Zhaoming Zhang, Dong-Chen Qi, Peter Evans, David R. G. Mitchell, and Xiaolin Wang. Creating thin magnetic layers at the surface of Sb2Te3 topological insulators using a low-energy chromium ion beam. *Applied Physics Letters*, 116(19):192410, May 2020. ISSN 0003-6951. doi: 10.1063/5.0006447. URL https://doi.org/10.1063/5.0006447.

Mi-Kyung Han, Hyein Ryu, and Sung-Jin Kim. Effect of Chromium Doping on the Thermoelectric Properties of Bi2Te3: CrxBi2Te3 and CrxBi2-xTe3. *Journal of Electronic Materials*, 42(9):2758–2763, September 2013. ISSN 1543-186X. doi: 10.1007/s11664-013-2670-4. URL https://doi.org/10.1007/s11664-013-2670-4.

Dominike P. de A. Deus, Igor S. S. de Oliveira, João B. Oliveira, Wanderlã L. Scopel, and R. H. Miwa. Magnetic switch and electronic properties in chromium-intercalated two-dimensional ${\mathrm{GeP}}_{3}$. *Physical Review Materials*, 5(5):054002, May 2021. doi: 10.1103/PhysRevMaterials.5.054002. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.5.054002. Publisher: American Physical Society.

Pan Zhang, Wenkai Liao, Ziyang Zhu, Mi Qin, Zhenhua Zhang, Dan Jin, Yong Liu, Ziyu Wang, Zhihong Lu, and Rui Xiong. Tuning the lattice thermal conductivity of Sb2Te3 by Cr doping: a deep potential molecular dynamics study. *Physical Chemistry Chemical Physics*, 2023. ISSN 14639076. doi: 10.1039/d3cp00999h.

Aidan P. Thompson, H. Metin Aktulga, Richard Berger, Dan S. Bolintineanu, W. Michael Brown, Paul S. Crozier, Pieter J. in 't Veld, Axel Kohlmeyer, Stan G. Moore, Trung Dac Nguyen, Ray Shan, Mark J. Stevens, Julien Tranchida, Christian Trott, and Steven J. Plimpton. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Physics Communications*, 271:108171, February 2022. ISSN 0010-4655. doi: 10.1016/j.cpc.2021.108171. URL https://www.sciencedirect.com/science/article/pii/S0010465521002836.

W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, November 1965. doi: 10.1103/PhysRev.140.A1133. URL https://link.aps.org/doi/10.1103/PhysRev.140.A1133. Publisher: American Physical Society.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996a. doi: 10.1103/PhysRevLett.77.3865. URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865. Publisher: American Physical Society.

Thibault Sohier, Matteo Calandra, and Francesco Mauri. Density functional perturbation theory for gated two-dimensional heterostructures: Theoretical developments and application to flexural phonons in graphene. *Physical Review B*, 96(7):075448, August 2017. doi: 10.1103/PhysRevB.96.075448. URL https://link.aps.org/doi/10.1103/PhysRevB.96.075448. Publisher: American Physical Society.

Kamal Choudhary, Irina Kalish, Ryan Beams, and Francesca Tavazza. High-throughput Identification and Characterization of Two-dimensional Materials using Density functional theory. *Scientific Reports*, 7(1):5179, July 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-05402-0. URL https://www.nature.com/articles/s41598-017-05402-0. Publisher: Nature Publishing Group.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine Learning Force Fields. *Chemical Reviews*, 121(16):10142–10186, August 2021. ISSN 0009-2665. doi: 10.1021/acs.chemrev.

0c01111. URL https://doi.org/10.1021/acs.chemrev.0c01111. Publisher: American Chemical Society.

Justin S Smith, Olexandr Isayev, and Adrian E Roitberg. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017.

Chi Chen, Zhi Deng, Richard Tran, Hanmei Tang, Iek-Heng Chu, and Shyue Ping Ong. Accurate force field for molybdenum by machine learning large materials data. *Physical Review Materials*, 1(4): 043603, September 2017. doi: 10.1103/PhysRevMaterials.1.043603. URL https://link.aps.org/doi/10.1103/PhysRevMaterials.1.043603. Publisher: American Physical Society.

Ilyes Batatia, David P Kovacs, Gregor Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in neural information processing systems*, 35:11423–11436, 2022.

Jonathan Vandermause, Yu Xie, Jin Soo Lim, Cameron J. Owen, and Boris Kozinsky. Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt. *Nature Communications*, 13(1):5183, September 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32294-0. URL https://www.nature.com/articles/s41467-022-32294-0.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, May 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL https://www.nature.com/articles/s41467-022-29939-5. Publisher: Nature Publishing Group.

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters*, 104(13): 136403, April 2010. doi: 10.1103/PhysRevLett.104.136403. URL https://link.aps.org/doi/10.1103/PhysRevLett.104.136403. Publisher: American Physical Society.

Jinzhe Zeng, Duo Zhang, Denghui Lu, Pinghui Mo, Zeyu Li, Yixiao Chen, Marián Rynik, Li'ang Huang, Ziyao Li, Shaochen Shi, Yingze Wang, Haotian Ye, Ping Tuo, Jiabin Yang, Ye Ding, Yifan Li, Davide Tisi, Qiyu Zeng, Han Bao, Yu Xia, Jiameng Huang, Koki Muraoka, Yibo Wang, Junhan Chang, Fengbo Yuan, Sigbjørn Løland Bore, Chun Cai, Yinnian Lin, Bo Wang, Jiayan Xu, Jia-Xin Zhu, Chenxing Luo, Yuzhi Zhang, Rhys E. A. Goodall, Wenshuo Liang, Anurag Kumar Singh, Sikai Yao, Jingchao Zhang, Renata Wentzcovitch, Jiequn Han, Jie Liu, Weile Jia, Darrin M. York, Weinan E, Roberto Car, Linfeng Zhang, and Han Wang. DeePMD-kit v2: A software package for deep potential models. *The Journal of Chemical Physics*, 159(5):054801, August 2023. ISSN 0021-9606. doi: 10.1063/5.0155600. URL https://doi.org/10.1063/5.0155600.

Daniel Wines and Kamal Choudhary. CHIPS-FF: Evaluating Universal Machine Learning Force Fields for Material Properties. *ACS Materials Letters*, pages 2105–2114, May 2025. doi: 10.1021/acsmaterialslett.5c00093. URL https://doi.org/10.1021/acsmaterialslett.5c00093. Publisher: American Chemical Society.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14(1):579, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36329-y. URL https://www.nature.com/articles/s41467-023-36329-y. Publisher: Nature Publishing Group.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00349-3. URL https://www.nature.com/articles/s43588-022-00349-3. Publisher: Springer Science and Business Media LLC.

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

Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31 (9):3564–3572, May 2019. ISSN 0897-4756. doi: 10.1021/acs.chemmater.9b01294. URL https://doi.org/10.1021/acs.chemmater.9b01294. Publisher: American Chemical Society.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, Matthew Horton, Robert Pinsler, Andrew Fowler, Daniel Zügner, Tian Xie, Jake Smith, Lixin Sun, Qian Wang, Lingyu Kong, Chang Liu, Hongxia Hao, and Ziheng Lu. MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures, May 2024. URL http://arxiv.org/abs/2405.04967. arXiv:2405.04967 [cond-mat].

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco Cohen, Abhishek Das, Ammar Rizvi, Sushree Jagriti Sahoo, Zachary W. Ulissi, and C. Lawrence Zitnick. UMA: A Family of Universal Models for Atoms, June 2025. URL http://arxiv.org/abs/2506.23971. arXiv:2506.23971 [cs].

Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06735-9. URL https://www.nature.com/articles/s41586-023-06735-9. Publisher: Nature Publishing Group.

Linfeng Zhang. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Physical Review Letters*, 120(14), 2018. doi: 10.1103/PhysRevLett.120.143001.

Justin S Smith, Benjamin T Nebgen, Roman Zubatyuk, Nicholas Lubbers, Christian Devereux, Kipton Barros, Sergei Tretiak, Olexandr Isayev, and Adrian E Roitberg. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nature communications*, 10(1):2903, 2019.

Stefan Chmiela, Huziel E. Sauceda, Igor Poltavsky, Klaus-Robert Müller, and Alexandre Tkatchenko. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications*, 240:38–45, July 2019. ISSN 0010-4655. doi: 10.1016/j.cpc.2019.02.007.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1):138, September 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00406-3. URL https://www.nature.com/articles/s41524-020-00406-3. Publisher: Nature Publishing Group.

Pascal Friederich, Florian Häse, Jonny Proppe, and Alán Aspuru-Guzik. Machine-learned potentials for next-generation matter simulations. *Nature Materials*, 20(6):750–761, June 2021. ISSN 1476-4660. doi: 10.1038/s41563-020-0777-6.

Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Advanced Materials*, 31(46):1902765, 2019. ISSN 1521-4095. doi: 10.1002/adma.201902765.

Christoph Schran, Fabian L. Thiemann, Patrick Rowe, Erich A. Müller, Ondrej Marsalek, and Angelos Michaelides. Machine learning potentials for complex aqueous systems made simple. *Proceedings of the National Academy of Sciences*, 118(38):e2110077118, September 2021. doi: 10.1073/pnas.2110077118.

Venkat Kapil, Christoph Schran, Andrea Zen, Ji Chen, Chris J. Pickard, and Angelos Michaelides. The first-principles phase diagram of monolayer nanoconfined water. *Nature*, 609(7927):512–516, September 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05036-x.

Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M. Wilkins, Clemence Corminboeuf, and Michele Ceriotti. Transferable Machine-Learning Model of the Electron Density. *ACS Central Science*, 5(1):57–64, January 2019. ISSN 2374-7943. doi: 10.1021/acscentsci.8b00551.

Jon Paul Janet, Fang Liu, Aditya Nandy, Chenru Duan, Tzuhsiung Yang, Sean Lin, and Heather J. Kulik. Designing in the Face of Uncertainty: Exploiting Electronic Structure and Machine Learning Models for Discovery in Inorganic Chemistry. *Inorganic Chemistry*, 58(16):10592–10606, August 2019. ISSN 0020-1669. doi: 10.1021/acs.inorgchem.9b00109.

Turab Lookman, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1):21, February 2019. ISSN 2057-3960. doi: 10.1038/s41524-019-0153-8.

Yunxing Zuo, Chi Chen, Xiangguo Li, Zhi Deng, Yiming Chen, Jörg Behler, Gábor Csányi, Alexander V. Shapeev, Aidan P. Thompson, Mitchell A. Wood, and Shyue Ping Ong. Performance and Cost Assessment of Machine Learning Interatomic Potentials. *The Journal of Physical Chemistry A*, 124(4):731–745, January 2020. ISSN 1089-5639. doi: 10.1021/acs.jpca.9b08723.

Ryosuke Jinnouchi, Jonathan Lahnsteiner, Ferenc Karsai, Georg Kresse, and Menno Bokdam. Phase Transitions of Hybrid Perovskites Simulated by Machine-Learning Force Fields Trained on the Fly with Bayesian Inference. *Physical Review Letters*, 122(22):225701, June 2019. doi: 10.1103/PhysRevLett.122.225701.

Claudio Zeni, Kevin Rossi, Theodore Pavloudis, Joseph Kioseoglou, Stefano de Gironcoli, Richard E. Palmer, and Francesca Baletto. Data-driven simulation and characterisation of gold nanoparticle melting. *Nature Communications*, 12(1):6056, October 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-26199-7.

Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials*, 6(1):20, March 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-0283-z.

Noam Bernstein, Bishal Bhattarai, Gábor Csányi, David A. Drabold, Stephen R. Elliott, and Volker L. Deringer. Quantifying Chemical Structure and Machine-Learned Atomic Energies in Amorphous and Liquid Silicon. *Angewandte Chemie*, 131(21):7131–7135, 2019. ISSN 1521-3757. doi: 10.1002/ange.201902625.

Linfeng Zhang, De-Ye Lin, Han Wang, Roberto Car, and Weinan E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials*, 3(2):023804, February 2019. doi: 10.1103/PhysRevMaterials.3.023804.

Kumar Miskin, Yi Cao, Madaline Marland, Farhan Shaikh, David T Moore, John A Marohn, and Paulette Clancy. Low-energy pathways lead to self-healing defects in cspbbr 3. *Physical Chemistry Chemical Physics*, 27(29):15446–15459, 2025.

Andrew W. Ruttinger, Divya Sharma, and Paulette Clancy. Protocol for Directing Nudged Elastic Band Calculations to the Minimum Energy Pathway: Nurturing Errant Calculations Back to Convergence. *Journal of Chemical Theory and Computation*, 18(5):2993–3005, May 2022. ISSN 1549-9618. doi: 10.1021/acs.jctc.1c00926. URL `https://doi.org/10.1021/acs.jctc.1c00926`. Publisher: American Chemical Society.

John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, October 1996b. doi: 10.1103/PhysRevLett.77.3865.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Kevin R Moon, David Van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

A Detailed Computational Methods

A.1 First-Principles Reference Calculations

All reference data were generated using Density Functional Theory (DFT) as implemented in the Quantum Espresso simulation package. We employed ultrasoft pseudopotentials for all elements (Cr, Sb, Te) with a plane-wave kinetic energy cutoff of 400 eV. The Brillouin zone was sampled using a Monkhorst-Pack k-point grid of $4 \times 4 \times 1$ for structural relaxations and Nudged Elastic Band (NEB) calculations. The Perdew-Burke-Ernzerhof (PBE) exchange-correlation functional was used, which we have found in prior investigations to provide a reliable balance of computational efficiency and accuracy for this class of materials.

The training dataset was generated from AIMD simulations performed in the NVT ensemble using a Langevin thermostat. These simulations covered a range of temperatures (300 K, 600 K, 1200 K) and Cr doping concentrations to ensure the model was exposed to a diverse set of thermodynamic and structural configurations. Each AIMD trajectory was run for 10 ps on a system of 120 atoms. To investigate the nature of atomic migration pathways, minimum energy paths (MEPs) and energy barriers were calculated using the Nudged Elastic Band (NEB) method. All initial, final, and intermediate configurations were considered to be converged when the forces on all unconstrained atoms fell below 0.01 eV/Å.

A.2 MACE Model Training Protocols

A.2.1 Training Hyperparameters

All training and fine-tuning procedures were executed with a consistent set of hyperparameters to ensure fair comparison:

- Optimizer: Adam
- Initial learning rate: 1×10^{-3}
- Batch size: 4
- Early stopping: Implemented based on validation set Force MAE
- Maximum epochs: 1000
- Validation split: 10% of training data

A.2.2 Fine-Tuning Details

For the FT-600K model, we selected a representative 5% subset (approximately 1,000 configurations) from the 600K AIMD trajectories. The subset was chosen to capture the full range of structural variations observed at this temperature, including both equilibrium fluctuations and transitional configurations.

For the FT-Multi_T model, the training subset was composed equally from 300K, 600K, and 1200K trajectories, ensuring exposure to diverse thermal conditions. The multi-temperature dataset was designed to test whether broader thermodynamic sampling improves generalization.

The choice of 600K for single-temperature fine-tuning represents typical thermoelectric operating temperatures, reflecting realistic usage conditions. At roughly two-thirds of Sb_2Te_3's melting point, this temperature captures significant thermal dynamics without structural deterioration.

Fine-tuning foundation models for specific chemical systems presents a fundamental challenge: how to adapt to new domains while preserving learned representations. This document discusses our fine-tuning strategy.

A.2.3 Fine-tuning Underlying Mechanism

In fine-tuning, the pre-trained model parameters θ_0 are directly optimized on the target dataset $\mathcal{D}_{\text{target}}$:

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\theta; \mathcal{D}_{\text{target}}) \tag{1}$$

Characteristics

- **Simple implementation**: Direct optimization on new data
- **Fast convergence**: High learning rate ($\alpha \sim 10^{-2}$)
- **Catastrophic forgetting**: Loss of original capabilities
- **Single objective**: Optimizes only for target domain performance

A.2.4 Implementation Details

A.2.5 Fine-tuning Algorithm

Algorithm 1 Fine-tuning

Initialize: $\theta \leftarrow \theta_{\text{foundation}}$
for epoch = 1 to N_{epochs} **do**
  for batch $\in \mathcal{D}_{\text{target}}$ **do**
    $\mathcal{L} \leftarrow \text{MSE}(f_\theta(\text{batch}), y_{\text{batch}})$
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}$
  end for
end for

A.2.6 Practical Considerations

When to Use Fine-tuning

- Limited computational resources
- Domain-specific applications only
- Rapid prototyping requirements
- No need for generalization beyond the target dataset

Take-away Fine-tuning offers rapid adaptation to new datasets with relatively low computational cost. However, this comes at the expense of catastrophic forgetting, where the model loses its original generalization capabilities.

B Supplementary Materials

B.1 MACE Model Performance Evaluation by RMSE

All developed Machine-Learned Force Fields (MLFFs) demonstrate high fidelity in predicting energies and forces, as illustrated by the parity plots in Figure 6, which compare model predictions to the reference DFT calculations. The performance of each model across the training, validation, and test sets is quantitatively summarized in Table 1. As detailed in the table, the models fine-tuned from the foundation model show a marked improvement over the model trained from scratch, with energy and force Root Mean Square Errors (RMSEs) being significantly reduced on the independent test set.

Table 1: Performance metrics for MACE models across training, validation, and test sets. The FT-600K metrics are the mean $\pm$ standard deviation across three independent training runs with different random seeds.

Model	**Train RMSE F** (meV/Å)	**Valid RMSE F** (meV/Å)	**Test RMSE F** (meV/Å)	**Test RMSE E** (meV/atom)
From Scratch	67.1	76.1	75.2	1.0
FT-600K	20.7 ± 11.8	44.6 ± 2.9	37.2 ± 0.6	0.5 ± 0.0
FT-MultiT	20.3	49.1	45.5	0.5

A deeper analysis of the training dynamics reveals important characteristics of the models. The model trained from scratch exhibits a relatively small gap between training (67.1 meV/Å) and validation (76.1 meV/Å) force RMSEs. While this suggests only moderate overfitting, its high error across all datasets indicates that it underfits the true physical interactions.

In contrast, the fine-tuned models display a more pronounced overfitting behavior, characterized by a large gap between their low training RMSEs and higher validation RMSEs. For instance, the FT-MultiT model shows a 142% increase in force RMSE from the training (20.3 meV/Å) to the validation set (49.1 meV/Å). This aggressive fitting is expected when fine-tuning on smaller, more specialized datasets. The high standard deviation in the FT-600K model's training RMSE (11.8 meV/Å) also suggests a sensitivity to initialization during the training process.

Despite the strong overfitting signals, the fine-tuned models generalize well to the test set, outperforming the scratch model by a significant margin. This analysis reinforces the central argument of our work: standard accuracy metrics like RMSE are insufficient for a comprehensive evaluation of MLFFs. While these metrics and the parity plots in Figure 6 confirm the models' general accuracy, they do not capture critical performance on physical properties such as energy barriers and transport phenomena, necessitating the more extensive, property-driven benchmarks presented in the main text.

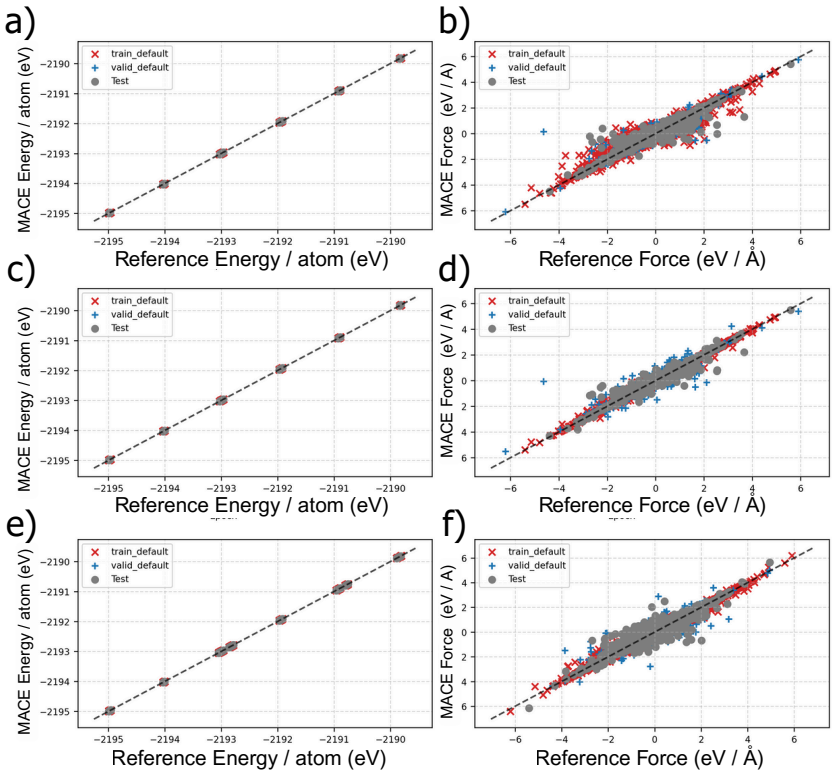

Figure 6: **Parity plots comparing MACE-predicted energies and forces against DFT reference calculations for the test set.** (a, b) Model trained from scratch, (c, d) fine-tuned 600 K model ("FT-600K"), and (e, f) fine-tuned multi-temperature model ("FT-MultiT"). Parity plots for energies (a, c, e) demonstrate excellent agreement for all models, while force parity plots (b, d, f) show strong overall correlation but subtle differences in error distributions that are not distinguishable visually. These results highlight the necessity of quantitative and property-based benchmarks for robust model evaluation.

B.2 Detailed Structural and Thermodynamic Analysis

The RDF analysis reveals that all MACE models, regardless of training strategy, successfully reproduce the key structural features of the Cr-doped Sb_2Te_3 system when compared to the AIMD ground truth (Fig. 7b).

All models correctly capture the primary coordination shells, with the first peak positions for Cr–Cr, Cr–Sb, and Cr–Te pairs occurring at approximately 3.0 Å, 3.2 Å, and 2.8 Å, respectively, in excellent agreement with the AIMD reference. The peak heights and positions remain consistent across all training strategies (Figures 7c–f), indicating that the local atomic structure is well preserved regardless of whether the model was trained from scratch, used as a foundation model, or fine-tuned with temperature-specific data.

The primary observable difference between models lies in the smoothness of the RDF curves rather than their fundamental features. The scratch-trained model (Fig. 7d) exhibits slightly smoother RDF profiles compared to the fine-tuned variants (Figures 7e–f). This difference arises from practical computational considerations rather than fundamental accuracy: the scratch model, being more compact with fewer parameters, allowed for longer MD trajectories within the same computational budget, resulting in better statistical sampling. In contrast, the fine-tuned models, while maintaining larger parameter counts from their foundation architectures, required more computational resources per MD step, limiting the total simulation time and resulting in slightly noisier RDF profiles.

Notably, both fine-tuning strategies—whether using single-temperature (600 K) or multi-temperature AIMD data—produce nearly identical RDF profiles, suggesting that the structural representation of the material is robust to the temperature range of the training data. This indicates that, for structural properties, the choice of training strategy has minimal impact, with all approaches converging to similar descriptions of the local atomic environment. The preservation of accurate structural features across all models provides confidence that the learned interatomic potentials correctly capture the fundamental bonding characteristics of the Cr-doped Sb_2Te_3 system.

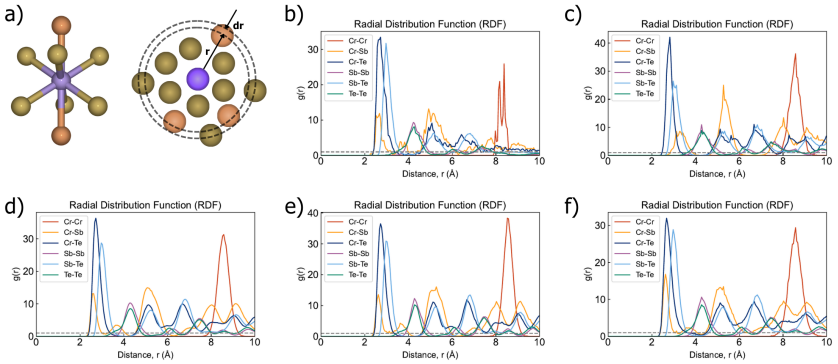

Figure 7: **Radial distribution function (RDF) analysis of Cr-doped Sb_2Te_3 at 600 K.** (a) Schematic illustration of the local atomic environment around a Cr dopant (purple) in the Sb_2Te_3 lattice, showing the first coordination shell of Te atoms (gold) and second-shell Sb atoms (orange). The dashed circles indicate the radial distances used for RDF calculation. (b-f) Computed RDFs for different atom pairs from MD simulations: (b) Ab initio molecular dynamics (AIMD) ground truth reference, (c) MACE foundation model without fine-tuning, (d) MACE model trained from scratch on Cr-Sb_2Te_3 data, (e) MACE model fine-tuned with 600 K AIMD data ("FT - 600K"), and (f) MACE model fine-tuned with multi-temperature AIMD data ("FT - Multi-T"). All simulations were performed at 600 K with $5\times5\times1$ supercells for 200 ps. The RDF peaks correspond to characteristic interatomic distances in the doped structure, with the first peak representing nearest-neighbor correlations.

B.3 Thermodynamic Ensemble Stability

A detailed analysis of the pressure profiles reveals subtle but important differences between models. The zero-shot MACE Foundation model equilibrates to a slightly different average pressure than the models trained or fine-tuned on our in-house AIMD data. This is likely a consequence of the foundation model being trained on a vast dataset of materials at their 0K equilibrium volumes. A

minor mismatch between the foundation model's predicted equilibrium lattice parameters and the true DFT values for our specific Cr:Sb_2Te_3 system at 600K manifests as a persistent non-zero average pressure in an NVT simulation. The fine-tuned and scratch models, having been trained explicitly on data from this system's true ensemble, do not exhibit this deviation.

B.4 Transport Property Analysis

The divergence in transport properties provides deeper insights into model behavior. The enhanced diffusivity in multi-temperature fine-tuned models likely stems from their exposure to high-temperature configurations approaching disordered or liquid-like states during training. By learning from these states, the model may have developed a potential energy surface that is slightly "flatter" or has lower barriers to diffusion, an effect that persists even at the 600K simulation temperature.

The thermal conductivity analysis reveals even more fundamental differences. The rapid decay of thermal conductivity in the foundation model indicates a failure to sustain long-range heat-carrying vibrational modes (phonons) specific to this crystalline structure. The anomalous peak observed in the 600K-only fine-tuned model's HFACF during the first 50 ps suggests potential structural instability or abrupt structural changes that alter phonon behavior. This highlights how different training strategies can lead to qualitatively different representations of collective phenomena, even when local properties appear identical.

B.5 Molecular Dynamics Simulations

B.5.1 Simulation Protocol

MD simulations were performed using the Atomic Simulation Environment (ASE) with the following protocol:

- **Integrator**: Langevin dynamics for NVT ensemble or Nosé-Hoover for NPT ensemble
- **Timestep**: 1.0 fs
- **Friction coefficient**: $\gamma = 0.01$ fs^{-1} (Langevin)
- **Barostat parameters**: $\tau_p = 1000$ fs, $P = 1$ bar (NPT only)
- **Equilibration**: 50,000 steps (50 ps)
- **Production**: 200,000 steps (200 ps) for property calculations
- **Sampling interval**: Every 100 steps for analysis
- **System size**: $5\times5\times1$ supercell (2050 atoms)

Simulations were conducted at two temperatures: 300 K and 600 K, to assess model performance under different thermal conditions. Initial velocities were assigned according to the Maxwell-Boltzmann distribution with removal of center-of-mass motion and angular momentum.

B.6 Property Calculations

B.6.1 Structural Properties

The radial distribution function (RDF) $g(r)$ was computed for all unique atom pairs:

$$g_{\alpha\beta}(r) = \frac{V}{4\pi r^2 \Delta r N_\alpha N_\beta} \left\langle \sum_{i\in\alpha} \sum_{j\in\beta} \delta(r - r_{ij}) \right\rangle \tag{2}$$

where α and β denote atomic species, V is the system volume, N_α is the number of atoms of species α, and the angle brackets denote time averaging.

B.6.2 Dynamic Properties

The mean square displacement (MSD) was calculated for each atomic species:

$$\text{MSD}_\alpha(t) = \left\langle \frac{1}{N_\alpha} \sum_{i \in \alpha} |\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2 \right\rangle \tag{3}$$

Diffusion coefficients were extracted from the linear regime of MSD using the Einstein relation:

$$D_\alpha = \lim_{t \to \infty} \frac{\text{MSD}_\alpha(t)}{6t} \tag{4}$$

The velocity autocorrelation function (VACF) was computed as:

$$C_v(t) = \frac{\langle \mathbf{v}(t) \cdot \mathbf{v}(0) \rangle}{\langle \mathbf{v}(0) \cdot \mathbf{v}(0) \rangle} \tag{5}$$

B.6.3 Thermodynamic Properties

Average potential energy per atom, temperature, and volume (for NPT simulations) were calculated over the production phase:

$$\langle E \rangle = \frac{1}{N_{\text{frames}}} \sum_{i=1}^{N_{\text{frames}}} \frac{E_{\text{pot}}^{(i)}}{N_{\text{atoms}}} \tag{6}$$

with corresponding standard deviations to assess thermal fluctuations.

B.6.4 Transport Properties

For thermal conductivity calculations, the heat flux vector was computed:

$$\mathbf{J} = \frac{1}{V} \left[\sum_i e_i \mathbf{v}_i + \frac{1}{2} \sum_{i \neq j} (\mathbf{F}_{ij} \cdot \mathbf{v}_j) \mathbf{r}_{ij} \right] \tag{7}$$

where e_i is the per-atom energy, $\mathbf{v}_i$ is the velocity, $\mathbf{F}_{ij}$ is the force between atoms i and j, and $\mathbf{r}_{ij}$ is their separation vector. The heat flux autocorrelation function (HFACF) was then computed for subsequent Green-Kubo analysis.

B.7 Representation Learning Feature Extraction

The following physically interpretable descriptors were extracted from MD trajectories:

1. **Energy landscape**: Total potential energy per atom
2. **Force fields**: 3N-dimensional force vectors for all atoms
3. **Structural order parameters**: Radial distribution histograms computed with 100 bins up to 5.0 Å cutoff
4. **SOAP descriptors**: Smooth Overlap of Atomic Positions with:
   - Cutoff radius: 5.0 Å
   - Number of radial basis functions: 8
   - Maximum angular momentum: 4
   - Gaussian width: 0.3 Å
5. **Mechanical response**: Numerical force derivatives with respect to 0.01 Å atomic displacements

These features were concatenated into a single vector per configuration, normalized, and projected using t-SNE (perplexity=30, learning rate=200) for visualization.

B.8 Evaluation Metrics Definitions

- **Force MAE**: $\frac{1}{3N}\sum_{i=1}^{N}\sum_{\alpha=x,y,z}|F_{i,\alpha}^{\text{MLFF}} - F_{i,\alpha}^{\text{DFT}}|$
- **Energy MAE**: $\frac{1}{N}|E^{\text{MLFF}} - E^{\text{DFT}}|$
- **RMSD Growth Rate**: Linear fit slope of $\text{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|\mathbf{r}_i(t) - \mathbf{r}_i(0)|^2}$
- **Silhouette Score**: Average of $\frac{b_i - a_i}{\max(a_i, b_i)}$ where a_i is mean intra-cluster distance and b_i is mean nearest-cluster distance

B.9 Detailed Migration Pathway Analysis

B.9.1 In-Gap Diffusion

This pathway tests interpolative accuracy, as the atomic environments remain similar to equilibrium configurations in the training data. The clear performance hierarchy—with fine-tuned models excelling and scratch failing—validates that fine-tuning successfully "sharpens" the foundation model's PES for in-domain predictions.

B.9.2 Deep Penetration

This pathway creates severe lattice distortions as Cr pushes through covalently bonded layers, accessing configurations far from the training distribution. The universal failure at transition states, despite accurate endpoint predictions, reveals how models trained on near-equilibrium data develop strong inductive biases that break down for strained configurations.

The scratch model's accidentally lower error on this OOD task is particularly instructive. Its globally inaccurate PES happens to be less catastrophically wrong in this specific high-energy region—not through physical insight but random chance. This underscores that the foundation model's "general" knowledge contains implicit assumptions about equilibrium physics that fail dramatically under extrapolation.

B.9.3 Implications for MLFF Development

These findings highlight critical considerations for practical MLFF deployment:

- **Validation Insufficiency**: Testing only on stable configurations masks critical failures at transition states
- **Hidden Extrapolation**: Models can appear accurate at trajectory endpoints while failing at crucial transition states
- **Data Quality over Quantity**: Task-specific training data outweighs larger but less relevant datasets
- **Foundation Model Limitations**: Pre-trained models require careful adaptation for processes involving significant atomic rearrangement

The results motivate advanced strategies like uncertainty-guided active learning that can identify high-uncertainty regions and intelligently augment training sets for improved extrapolative performance.

B.10 Nudged Elastic Band (NEB) Benchmark for Cr Migration

To directly assess the practical performance of each model on a physically critical task, we performed a Nudged Elastic Band (NEB) calculation for a Cr atom migration event, as shown in Figure 8. This calculation serves as an effective and challenging probe to distinguish the models' stability and predictive accuracy when exploring unseen transition-state geometries. The results reveal a dramatic difference in performance that is not apparent from the RMSE metrics alone.

Both the model trained from scratch and the fine-tuned models (FT-600K and FT-MultiT) exhibit explosive behavior during the NEB optimization, as evidenced by the sudden spike in the maximum force (f_{max}) shown in Figure 8c. This instability forced the early termination of the calculations. This failure suggests that the interpolated images along the NEB path introduced metastable configurations,

such as an unphysical separation of the material layers, which were outside the manifold of the training data. For the fine-tuned models, this indicates a form of catastrophic forgetting, where the models lost their generalized stability after being trained on a narrow dataset.

In contrast, the MACE foundation model, without any system-specific fine-tuning, successfully converged the NEB calculation. It predicts a migration barrier of 0.41 eV (Figure 8b), a value in good agreement with DFT calculations for similar in-gap diffusion pathways (0.3 eV). This level of accuracy, close to the bounds of chemical accuracy, demonstrates the foundation model's exceptional capability to generalize to complex transition-state configurations. This benchmark underscores that evaluating performance on dynamic, physically relevant processes is a critical and necessary step for validating the true capabilities of MLFFs.

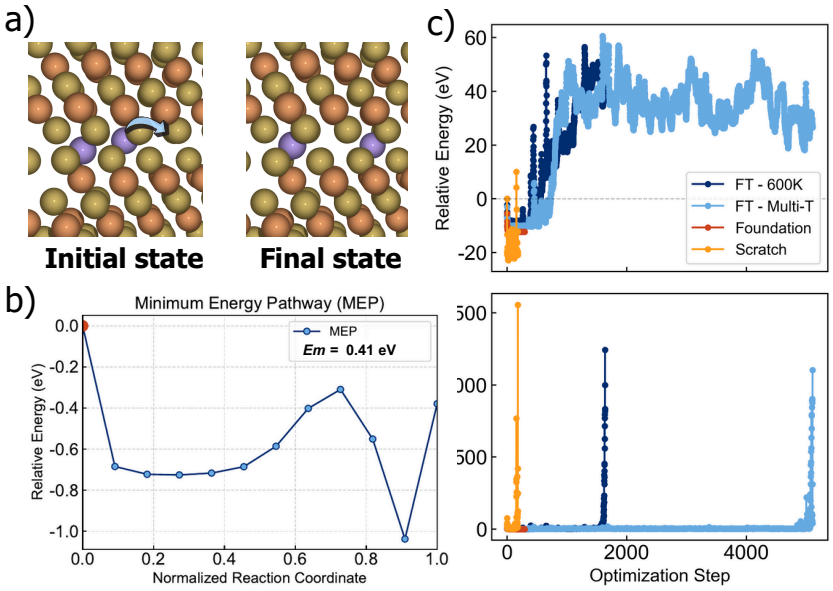

Figure 8: Nudged Elastic Band (NEB) benchmark of MACE models for a Cr atom migration. (a) Visualization of the initial and final states of the diffusion pathway. (b) The converged Minimum Energy Pathway (MEP) calculated with the MACE foundation model, yielding a migration barrier of 0.41 eV. (c) The evolution of the relative system energy (top panel) and the maximum force (f_{max}, bottom panel) during the NEB optimization for each model. The foundation model (red) converges smoothly. In contrast, the scratch (orange), FT-600K (dark blue), and FT-MultiT (light blue) models all exhibit explosive behavior, where a rapid increase in f_{max} indicates instability and leads to the termination of the calculation.

NeurIPS Paper Checklist

Acknowledgments and Disclosure of Funding

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

- **Claims**

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

  Answer: [Yes]

  Justification: The abstract and introduction clearly state our claims about benchmarking specialist vs. generalist MLFF training strategies for Cr-doped Sb_2Te_3, and these are fully supported by our systematic evaluation of five distinct training approaches across equilibrium and non-equilibrium properties.

  Guidelines:

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

- **Limitations**

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

  Answer: [Yes]

  Justification: We explicitly discuss limitations throughout the paper, including the single test case nature of our study (Cr-Sb_2Te_3), the computational constraints that limited MD trajectory lengths, and the fundamental limitations of local-descriptor-based MLFFs for capturing long-range physics (Section 4.5).

  Guidelines:

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

- **Theory assumptions and proofs**

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

  Answer: [NA]

  Justification: This paper is an empirical benchmarking study of machine learning force fields and does not contain theoretical results or proofs.

  Guidelines:

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

- **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We provide comprehensive details including DFT parameters (Section 3.1), MACE architecture specifications, training hyperparameters (Section 3.2), MD simulation protocols (Section 3.3), and detailed descriptions of all evaluation metrics and analysis methods.

Guidelines:

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

- **Open access to data and code**

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

Answer: [Yes]

Justification: We provide detailed methodology for reproducibility. The code and datasets are released on our Github repository (mlff-benchmark-cr-sb2te3), which is actively maintained and will be updated with the newest code and results. Additionally, we use publicly available tools (MACE, Quantum Espresso, ASE) and provide sufficient detail in both the main text and appendix for independent reproduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

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

- **Experimental setting/details**

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

  Answer: [Yes]

  Justification: Section 3.2 and the Appendix provide complete training details including optimizer (Adam), learning rate (1×10^{-3}), batch size (4), early stopping criteria, temperature settings for AIMD (300K, 600K, 1200K), and dataset sizes (20,000 configurations).

  Guidelines:

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

- **Experiment statistical significance**

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

  Answer: [Yes]

  Justification: For model training, I used three different random seeds to verify the consistency of model behavior. The performance was not affected by the seeds, with all models showing nearly identical results.

  For the NEB diffusion pathways, we conducted multiple independent calculations with different initial perturbations of the migration pathway to establish statistical significance. These results are presented with proper error bars in Figure 3b, quantifying the uncertainty in migration energy predictions across all MLFF training strategies.

  For thermodynamic and structural properties, we analyzed extended trajectories (200 ps) to ensure proper equilibration and sampling. While computational constraints limited us to single training runs for each MLFF strategy, we verified consistency by computing time-averaged properties.

  Guidelines:

  - The answer NA means that the paper does not include experiments.
  - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
  - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

- **Experiments compute resources**

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

  Answer: [Yes]

  Justification: The paper provides detailed information on the compute resources used for all experiments, including hardware type (GPU), execution time, system sizes, and number of trajectories/configurations. The compute requirements are summarized in Table 2. This includes AIMD, nudged elastic band (NEB), and machine-learning force field (MLFF) MD runs. Both per-trajectory costs and total project costs are explicitly stated, allowing independent researchers to reproduce the experiments and estimate compute requirements. Furthermore, the sbatch script for the MLFF-MD run explicitly discloses the resource request (1 A100 GPU, 4 MPI tasks, 8 CPUs per task), ensuring clarity in execution environment specification. The disclosure includes total wall-time estimates across configurations and temperatures, thereby providing a transparent account of the actual compute footprint of the research project.

Table 2: Summary of compute resources required for different experiments.

Experiment	System / Tasks	Hardware	Compute Cost
AIMD	120 atoms; 10 configs $\times$ 3 temps	1 GPU (A100)	9 days / traj (270 runs; $\sim$2430 GPU-days)
NEB	60 atoms (8 traj); 120 atoms (2 traj)	1 GPU (A100)	6 days / traj (10 runs; $\sim$60 GPU-days)
MLFF-MD	2050 atoms; 4 models $\times$ 2 tasks	1 GPU (A100)	$\sim$2 hrs / run (8 runs; $\sim$16 GPU-hours)

  Guidelines:

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

- **Code of ethics**

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

  Answer: [Yes]

  Justification: Our research on machine learning force fields for materials science applications adheres to all ethical guidelines, involves no human subjects, and presents honest assessments of both capabilities and limitations.

  Guidelines:

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

- **Broader impacts**

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

  Answer: [Yes]

  Justification: The paper explicitly addresses both positive and potential negative societal impacts. On the positive side, our work contributes to accelerating materials discovery by providing a benchmark framework that combines MD simulations under operational temperatures with nudged elastic band migration studies. This integrated approach enables simultaneous access to thermodynamic and kinetic insights, while offering a rigorous platform to test interpolation and extrapolation capabilities of machine-learned force fields (MLFFs). Importantly, the benchmark pipeline is not limited to the 2D materials investigated here: it can be extended to a wide range of systems with different dopants or migrating species. Such generalizability allows for faster and more reliable evaluation of MLFFs, thereby shortening the model improvement cycle and supporting the development of more robust and flexibly tunable fine-tuning strategies. Ultimately, this work has the potential to significantly boost computational efficiency and accelerate the pace of scientific discovery in materials science.

  On the negative side, while direct societal risks are minimal given the fundamental nature of the research, there are indirect considerations. More accurate and efficient MLFFs could accelerate the discovery of materials with dual-use potential, including those relevant for defense or energy storage in sensitive contexts. Acknowledging this possibility underscores the importance of ensuring that such computational advancements are guided by responsible dissemination and ethical application.

  Guidelines:

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

- **Safeguards**

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

Answer: [NA]

Justification: Our work on force fields for materials simulation poses no direct risks for misuse. The models are specific to Cr-Sb_2Te_3 systems and have no applications outside scientific research.

Guidelines:

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

- **Licenses for existing assets**

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

Answer: [Yes]

Justification: We properly cite MACE (ref 12), MACE-MP foundation model, Quantum Espresso, and ASE. We use these tools according to their open-source licenses, though specific license details could be made more explicit.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

- **New assets**

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

Answer: [NA]

Justification: Although we train new models as part of the study, the primary contribution of the paper lies in the benchmarking methodology and the systematic evaluation of training strategies. The intent is not to introduce a new publicly released asset, but rather to provide a reproducible and extensible evaluation framework. Thus, no new standalone assets are released, and the answer is not applicable in the context of this work.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

- **Crowdsourcing and research with human subjects**

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

  Answer: [NA]

  Justification: This paper involves only computational materials science research with no human subjects or crowdsourcing components.

  Guidelines:

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

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

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

  Answer: [NA]

  Justification: This paper does not involve research with human subjects.

  Guidelines:

  - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
  - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
  - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
  - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

- **Declaration of LLM usage**

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

  Answer: [NA]

  Justification: No large language models (LLMs) were used as part of the research methodology. The core contributions rely exclusively on physics-based machine learning force fields (MACE architecture) trained on density functional theory (DFT) data. Since LLMs did not play any role in method development, experimentation, benchmarking, or scientific analysis, their declaration is not applicable in the context of this work.

  Guidelines:

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