The Good, the Bad, and the Ugly of Atomistic Learning for "Clusters-to-Bulk" Generalization

Mikołaj J. Gawkowski,[1,2] Mingjia Li,[1,2] Benjamin X. Shi,[3] and Venkat Kapil[1,2,a)]

[1)] *Department of Physics and Astronomy, University College London, 7-19 Gordon St, London WC1H 0AH, UK*

[2)] *Thomas Young Centre and London Centre for Nanotechnology, 9 Gordon St, London WC1H 0AH, UK*

[3)] *Initiative for Computational Catalysis, Flatiron Institute, New York, New York 10010, USA*

Training machine learning interatomic potentials (MLIPs) on total energies of molecular clusters using differential or transfer learning is becoming a popular route to extend the accuracy of correlated wave-function theory to condensed phases. A key challenge, however, lies in validation, as reference observables in finite-temperature ensembles are not available at the reference level. Here, we construct synthetic reference data from pretrained MLIPs and evaluate the generalizability of cluster-trained models on ice-Ih, considering scenarios where both energies and forces and where only energies are available for training. We study the accuracy and data-efficiency of differential, single-fidelity transfer, and multi-fidelity transfer learning against ground-truth thermodynamic observables. We find that transferring accuracy from clusters to bulk requires regularization, which is best achieved through multi-fidelity transfer learning when training on both energies and forces. By contrast, training only on energies introduces artefacts: stable trajectories and low energy errors conceal large force errors, leading to inaccurate microscopic observables. More broadly, we show that accurate reproduction of microscopic structure correlates strongly with low force errors but only weakly with energy errors, whereas global properties such as energies and densities correlate with low energy errors. This highlights the need to incorporate forces during training or to apply careful validation before production. Our results highlight the promise and pitfalls of cluster-trained MLIPs for condensed phases and provide guidelines for developing – and critically, validating – robust and data-efficient MLIPs.

I. INTRODUCTION

Machine learning interatomic potentials (MLIPs) have emerged as powerful tools to reduce the computational cost of first-principles-level simulations across chemistry[1], materials science[2–4], and biology[5]. Traditionally, MLIPs are trained on total energies and forces derived from density functional theory (DFT), inheriting the accuracy of DFT at a significantly reduced computational cost. While this strategy has been highly successful, DFT lacks a systematic path to quantitative improvement, limiting its accuracy for many classes of systems[6–10]. Correlated wave-function theory (cWFT), by contrast, provides a hierarchy of systematically improvable methods that can approach experimental accuracy[11–16]. Training MLIPs directly on cWFT data is therefore highly desirable, but poses many challenges: the steep computational cost of cWFT calculations, the high cost and complexity of computing energy gradients, and the limited availability of periodic implementations, which often requires training MLIPs for bulk systems on molecular clusters[17–19]. We refer to this process as "cluster-to-bulk" generalization.

Several strategies have been proposed to reach cWFT-level accuracy in MLIPs, with liquid water and ice serving as prototypical systems. Early "machine-learning" approaches were rooted in the many-body expansion, with the water molecule as the basic unit, leading to accurate water models such as MBpol[20,21] and, more recently, q-AQUA(-pol)[22]. These models are based on permutationally invariant polynomials[23], which learn up to three- or four-body contributions, and a semi-empirical polarizable force field which approximates higher-order terms[20]. In contrast, molecule-agnostic approaches have directly applied MLIPs. In this context, Lan et al.[24] carried out a *tour de force* study, based on the Behler–Parrinello framework[25], directly training on resolution-of-identity MP2[26] total energies and forces, which required more than 20,000 periodic configurations.

Considering the high data requirement of training on cWFT data with standard MLIPs, many strategies have been proposed. The first class of strategies eases the learning task by using DFT as an intermediate level of theory. For example, Chen *et al.*[27] adopted a transfer learning strategy[28]: the MLIP was first pretrained on DFT-level energies and forces, then fine-tuned on a much smaller dataset of cWFT-level energies using the pretrained weights for initialization. This approach proved data-efficient for CCSD[29], CCSD(T)[30], and auxiliary-field quantum Monte Carlo[31], requiring fewer than 200 total-energy evaluations for periodic configurations with up to 100 atoms. However, validation was restricted to a single thermodynamic state point in the NVT ensemble. In a complementary effort, Daru *et al.*[17] introduced a differential learning framework[32], which was further supplemented by a fixed point-charge Coulomb baseline and short-range repulsion terms. Here, one model was trained on total energies and forces at a local MP2 level, while

[a)] Electronic mail: v.kapil@ucl.ac.uk

a second captured the energy differences between local MP2[33] and local CCSD(T)[34]. Crucially, training relied on nearly 20,000 small water clusters carved from periodic MLIP simulations, substantially reducing the computational and memory demands of generating cWFT data. Validation was again limited to liquid water in the NVT ensemble at a single state point.

A second class of approaches focuses on scalable MLIP architectures that intrinsically require less training data. Iterative schemes that construct high-body-order features from tensor products of low-body-order terms[35], most prominently within the atomic cluster expansion (ACE) framework[36], enable systematically improvable and data-efficient models. Graph neural network architectures such as NequIP[37], MACE[38], and GRACE[39] extend the scalability of ACE features across chemical space through tensor reduction[40]. Building on this, Kaur *et al.*[41] demonstrated that transfer learning, via fine-tuning of the so-called MACE-MP-0 model[2] pretrained on the diverse "MPTrj" dataset[42], further reduces data requirements compared to earlier architectures and direct training with the MACE architecture. Using fewer than 50 periodic configurations with up to 100 molecules, they achieved sub-1% density errors across four ice phases at finite thermodynamic conditions, enabling RPA-quality simulations in the NPT ensemble. More recently, Mészáros, Szabó, and Daru [18] combined ACE models with differential learning to train MLIPs for liquid water on water clusters carved from the bulk liquid. The ACE descriptors reduced the required training data from the 20,000 clusters reported in their earlier work[17] to just 1,000 clusters, while working in the NVT ensemble. Building upon this, O'Neill *et al.*[19] incorporated clusters of increasing radii from liquid water structures from the NPT ensemble and fitting a differential learning model to the difference between local CCSD(T) and DFT energies, achieving good agreement with experiments. Independently, more advanced transfer learning strategies, notably multi-fidelity fine-tuning amongst others[43–46], have been studied and demonstrated to outperform differential learning, separately for both molecular[47] and solid-state[48] systems; however, this approach remains to be tested on the cluster-to-bulk generalization of MLIPs.

These studies suggest that an affordable route to cWFT-level MLIPs combines scalable architectures with training on clusters, while leveraging differential or transfer learning in conjunction with an intermediate level of theory. However, a challenge associated with these approaches is their validation. Unlike DFT-based work, the ground-truth observables in finite thermodynamic ensembles – against which the models should ultimately be benchmarked – are prohibitive. As a result, validation relies either on reproducing the energies of clusters extracted from the bulk trajectory[17], local convergence of observables with respect to the dataset, or on comparison with experiments[19]. All these routes have certain limitations. The first assumes that a model that yields stable trajectories and low energy errors on clusters also yields low errors on forces and other microscopic observables. Similarly, the criterion based on the local convergence of observables with respect to the dataset assumes that the MLIPs trained on clusters can be systematically improved to the ground truth. And finally, a comparison with experiments doesn't allow for disentangling error cancellation between the MLIP and the cWFT reference. A second important question is understanding how sensitive the cluster-to-bulk generalization of MLIPs is to the choice of training strategies, whether energies alone or both energies and forces are used, the volume of training data, and the size of clusters employed and their interplay.

In this work, we assess the ability of MLIPs trained on molecular clusters to generalize to bulk phases, using ice-Ih as a prototypical molecular solid. We evaluate four training strategies: (i) transfer learning via fine-tuning a pretrained model, (ii) differential learning with respect to a pretrained model, (iii) transfer learning with multi-fidelity fine-tuning, and (iv) direct learning on cluster data, which serves as a baseline. We benchmark these strategies across two scenarios: training on energies and forces versus training only on energies, assessing how well both global and microscopic observables are recovered in the NVT and NPT ensembles. Overall, we find that incorporating forces into training is highly beneficial, as it improves stability and enables simultaneous accuracy in both global properties (energies and density) and microscopic structure. Notably, multi-fidelity transfer learning performs best in this setting, reaching sub-meV/atom energy errors, densities within 1% of the reference, and excellent agreement with microscopic structure using as few as 100 clusters. When trained only on energies, however, models reproduce global properties such as energy and density reasonably well, but consistently fail to capture microscopic structure accurately. In this case, differential learning performs best among the strategies considered. Our findings critically evaluate the cluster-to-bulk generalization of MLIPs and provide practical guidelines for developing and validating data-efficient MLIPs that can reliably generalize to condensed-phase systems.

The paper is organized as follows: Section II introduces the benchmark system, training strategies, datasets, and test metrics. Section III examines model performance by analyzing how well global and microscopic observables are recovered as a function of training data volume in the NVT and NPT ensembles. Section IV summarizes the strengths and limitations of the different strategies in the context of the cluster-to-bulk generalization, while also considering extrapolation behavior and systematic improvement with increasing cluster size, and outlines directions for future work.

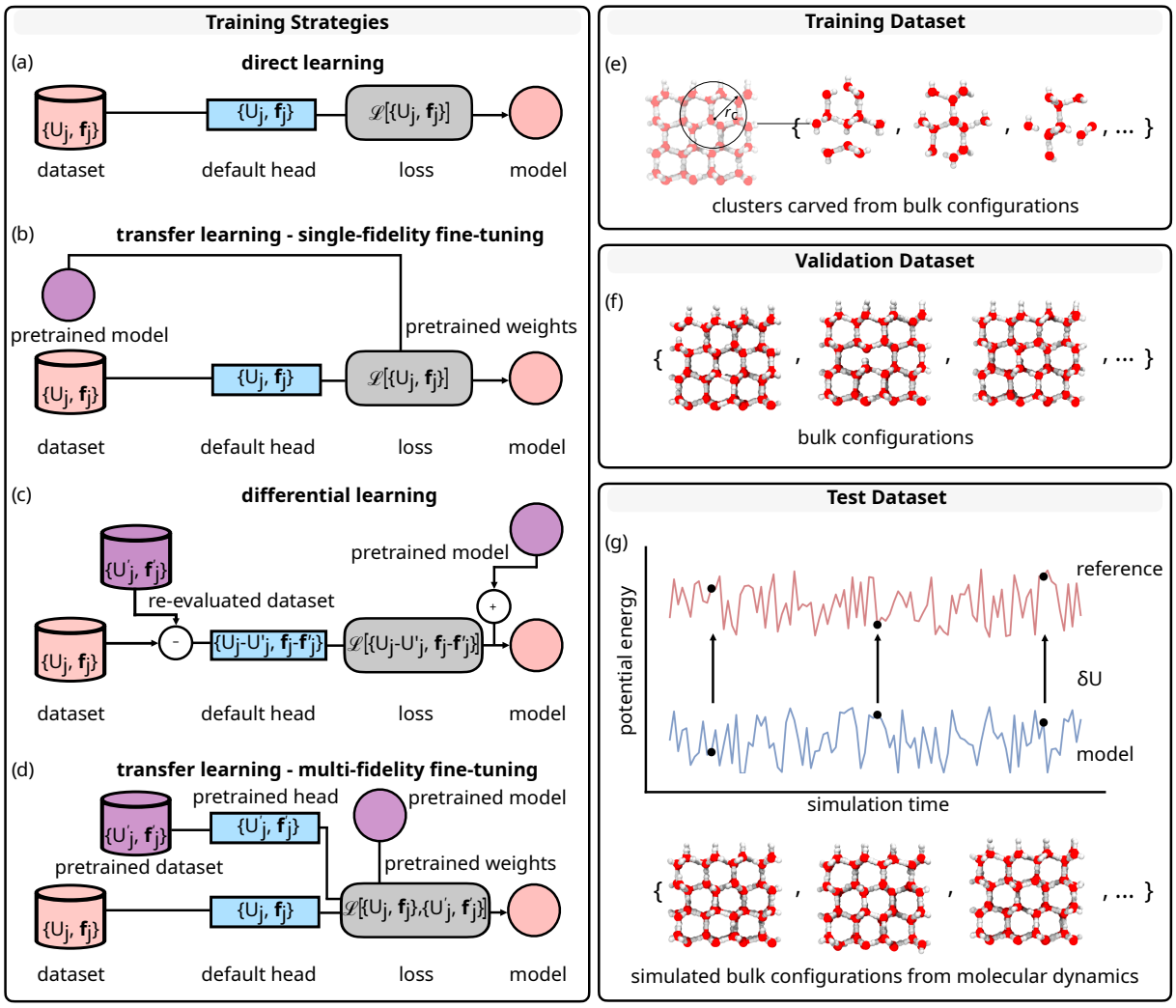

FIG. 1. Schematic overview of the training strategies and datasets used in this work. Panels (a–d) illustrate the training strategies: direct learning, single-fidelity fine-tuning, differential learning, and multi-fidelity transfer learning, respectively. Panels (e-g) respectively show the datasets: clusters carved from bulk ice-Ih for training, bulk configurations for validation, and bulk configurations generated from molecular dynamics by the models for testing. In panel (c), the pretrained model serves as the baseline.

II. METHODS

A. System

To assess the generalizability of MLIPs for bulk phases when trained on molecular clusters, we use ice-Ih as a benchmark system. This system has proven to be a suitable testbed for studying the data efficiency of transfer learning[41] via fine-tuning of pretrained MLIPs. The guidelines derived from benchmarks for ice-Ih were shown to apply to ice phases and the X23 dataset of rigid molecular crystals[49]. Thus, we expect the insights gained from this study to apply to other molecular systems.

B. Training strategies for MLIPs

We discuss four training strategies that combine high- and low-fidelity data or levels of theory differently. A qualitative overview of these strategies is shown in Fig. 1(a)-(d). Here, high-fidelity level refers to the target level of theory, and the corresponding high-fidelity data consists of labelled energies and forces at this level. Analogously, the low-fidelity level refers to an intermediate theory – ideally much less expensive yet semi-quantitatively accurate– that can be leveraged to improve data efficiency relative to training solely on high-fidelity data.

Direct learning corresponds to standard regression, where the MLIP is trained from randomly initial-

ized weights solely on high-fidelity data as depicted in Fig. 1(a). This is achieved by minimizing a loss function that measures the discrepancy between the reference values and the model predictions. For the set of M training configurations $\{\mathbf{q}_j\}_{j=1}^{M}$ and the set of the corresponding high-fidelity total energies and forces, $\{U_j, \mathbf{f}_j\}_{j=1}^{M}$, the loss is defined as

$$\mathcal{L}^{\text{direct}} = \mathcal{L}\left[\{U_j, \mathbf{f}_j\}_{j=1}^{M}\right] = \frac{1}{M}\sum_{j=1}^{M}\left[\lambda_U\left(U_j - \hat{U}\left(\mathbf{q}_j\right)\right)^2 + \lambda_{\mathbf{f}}\left|\left|\mathbf{f}_j - \hat{\mathbf{f}}(\mathbf{q}_j)\right|\right|_2^2\right] \quad (1)$$

with λ_U and $\lambda_{\mathbf{f}}$ are referred to as the energy and the force weights respectively, $\hat{U}$ is the potential energy predicted by the MLIP (the explicit dependence on the unit cell, the atomic identities, and the learnable weights are not shown for brevity), $\hat{\mathbf{f}}$ is the force predicted by the model (the MLIPs considered in this work are conservative with $\hat{\mathbf{f}} = -\nabla_{\mathbf{q}}\hat{U}$), and $||\cdot||_2$ corresponds to the l^2 norm. In this work, the high-fidelity training data comprises total energies and forces estimated on water cluster configurations carved from bulk structures.

Transfer learning with single-fidelity fine-tuning involves two steps: pretraining of the MLIP on low-fidelity training data, followed by fine-tuning of the same model on high-fidelity training data. For a low-fidelity dataset consisting of a set of M' configurations $\{\mathbf{q}'_j\}_{j=1}^{M'}$ and their total energies and forces $\{U'_j, \mathbf{f}'_j\}_{j=1}^{M'}$, the pretraining step involves direct training of an MLIP, by minimizing the low-fidelity loss

$$\mathcal{L}^{\text{direct}}_{\text{low-fidelity}} = \mathcal{L}\left[\{U'_j, \mathbf{f}'_j\}_{j=1}^{M'}\right].$$

The final weights of this pretrained model will be referred to as pretrained weights. Low-fidelity data can comprise larger and more complex configurations, including the target bulk phase.

In the single-fidelity fine-tuning step, as depicted in Fig. 1(b), the MLIP is initialized to the pretrained weights and optimized on a new loss defined for the high-fidelity training data comprising molecular clusters

$$\mathcal{L}^{\text{direct}}_{\text{high-fidelity}} = \mathcal{L}\left[\{U_j, \mathbf{f}_j\}_{j=1}^{M}\right].$$

Differential learning learns the difference between high- and low-fidelity levels for a given dataset[32], as depicted in Fig. 1(c). It requires a low-fidelity model, which need not share the same architecture as the MLIP being trained, or even be an MLIP. In the first step, the M high-fidelity configurations $\{\mathbf{q}_j\}_{j=1}^{M}$ are evaluated using the low-fidelity model to obtain the corresponding low-fidelity energies and forces, $\{U'_j, \mathbf{f}'_j\}_{j=1}^{M}$. A new set of target values is constructed

$$\{\delta U_j, \delta\mathbf{f}_j\}_{j=1}^{M} \equiv \{U_j - U'_j, \mathbf{f}_j - \mathbf{f}'_j\}_{j=1}^{M}$$

representing the difference between the high- and low-fidelity predictions. The MLIP is trained, starting from randomly initialized model weights, to minimize the loss

$$\mathcal{L}^{\text{differential}} = \mathcal{L}\left[\{\delta U_j, \delta\mathbf{f}_j\}_{j=1}^{M}\right]$$

The final model prediction is obtained by summing the low-fidelity model and the MLIP: $\hat{U} = U' + \delta\hat{U}$ with $\delta\hat{U}$ representing the differential model.

Transfer learning with multi-fidelity fine-tuning utilizes low-fidelity data during both the pretraining and fine-tuning stage, as depicted schematically in Fig. 1(d). In the first step, the MLIP is first pretrained on low-fidelity training data. To facilitate the explanation of the rest of the approach, we introduce terminology for different components of the MLIPs. The output block of the pretrained model, a small multilayer perceptron that predicts low-fidelity energies and forces $\hat{U}', \hat{\mathbf{f}}'$, will be referred to as the pretraining (pt) head. The remainder of the network will be referred to as the shared layers. During the fine-tuning stage, the model is augmented with an additional output block tasked with predicting high-fidelity energies and forces $\hat{U}, \hat{\mathbf{f}}$, referred to as the default head. With two output heads, the model can be trained simultaneously on low- and high-fidelity data by minimizing a weighted sum of their corresponding direct losses

$$\mathcal{L}^{\text{multi-fidelity}} = \lambda_{\text{pt}}\,\mathcal{L}^{\text{direct}}_{\text{low-fidelity}} + \lambda_{\text{default}}\,\mathcal{L}^{\text{direct}}_{\text{high-fidelity}} \quad (2)$$

where λ_{pt} and λ_{default} control the relative importance of the two loss contributions. The model weights are initialized using the pretrained values for the shared layers and both the pretraining and default heads. The multi-fidelity approach offers flexibility in dataset selection – for example, enabling pretraining on a diverse low-fidelity dataset and fine-tuning on a more specialized low-fidelity dataset that closely resembles the target system. In the context of the cluster-to-bulk generalization of MLIPs, a desirable option that leverages the computational advantages of both fidelity levels is to use low-fidelity bulk configurations with total energies and forces alongside high-fidelity total energies for clusters.

C. Datasets and test metrics

Levels of theory: Although the ultimate goal of this work is to evaluate training strategies for cWFT methods, we employ pretrained MLIPs as inexpensive proxies. This step is essential, since extensive benchmarking in finite thermodynamic ensembles against known ground truth values would be computationally prohibitive with cWFT methods, making it otherwise impossible to assess the accuracy of the cluster-to-bulk training paradigms.

We select the MACE-MP-0 model, trained at the generalized gradient approximation density functional

PBE, as the low-fidelity level, and the MACE-OFF-23 model, trained at the van der Waals (vdW)-inclusive hybrid-functional level ωB97M-D3(BJ), as the high-fidelity level. We recognize that pretrained MLIP proxies may introduce artifacts, owing to their locality and the use of training data restricted to clusters or small unit cells. To address this, we validate our conclusions with explicit DFT calculations, which confirm that the trends observed with proxy MLIPs also hold for true electronic structure calculations, as shown in Fig. S3 of the Supporting Information (SI).

Training and validation datasets: The high-fidelity training sets consist of total energies and forces for molecular clusters. As shown in Fig. 1(e), these clusters are carved from bulk configurations. To systematically explore data efficiency, we construct subsets of increasing size, containing 50, 100, 200, 400, 800, and up to 1000 clusters for each cluster size defined by the cluster radius $r_c \in \{4.0, 4.5, 6.0\}$ Å. Most of the results presented in this work are based on 4.5 Å sized clusters containing 15 water molecules on average.

The low-fidelity training data comprises two types. The first type corresponds to a subset of configurations sampled from the MPTrj – the dataset used to train the MACE-MP-0 model. The second type corresponds to a set of ice-Ih configurations extracted from molecular dynamics[41]. For differential learning – which does not require low-fidelity training data per se, but rather a low-fidelity model – we use the MACE-MP-0 model as the reference baseline. The validation dataset consists of 25 periodic ice-Ih configurations evaluated at the high-fidelity level of theory (see Fig. 1(f)). For the isolated atomic energies, we use the MACE-OFF-23 isolated atomic energy values of U_H = -13.571964772646918 eV, U_O = -2043.933693071156 eV where U_H is the energy of hydrogen and U_O of oxygen and in the case of differential learning, where one trains on differences in energies and forces, we use δU_H = -12.412976535265937 eV, δU_O = -2041.8522316415601 eV, where each energy is the difference between MACE-OFF-23 and MACE-MP-0 references.

Test ensembles and metrics: Building on our earlier observation that validation errors do not reflect out-of-domain performances in molecular dynamics ensembles[41], we evaluate each model's ability to generalize by performing molecular dynamics simulations in the NVT and NPT thermodynamic ensembles. In the NVT ensemble, we report the energy errors on configurations sampled from trajectories generated by the MLIPs, as depicted in Fig. 1(g). While these configurations are not drawn from the high-fidelity distribution, the resulting energy error can be interpreted – via thermodynamic perturbation theory – as a measure of the MLIP's ability to represent the true ensemble[50], provided it is interpreted with care. In the NPT ensemble, in addition to energy errors, we also report the predicted density. Density errors are computed relative to the reference density obtained from molecular dynamics simulations performed using the high-fidelity MACE-OFF-23 model under identical thermodynamic conditions. In Fig. S1 of the SI, we report the density, the O–O pair correlation function, and the O–H bond distribution at both high- and low-fidelity levels.

D. Computational details

Machine learning: The MACE architecture[51] is used to represent the MLIPs, and all numerical experiments are performed using version 0.3.9. Unless stated otherwise, the MACE models used in this work correspond to the "medium" architecture from Ref. 2 and are trained in 64-bit floating-point precision. Training is performed using the AMSGrad variant of the Adam optimizer with default parameters: $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, for 150 epochs. We use a learning rate of 0.001, an exponential moving average scheduler with a decay factor of 0.995, and apply gradient clipping with a maximum norm of 100. After 150 epochs, and 500 for differential learning, the energy loss weight was increased by 20%, and training continued for an additional 150 and 500 epochs, respectively. The larger number of epochs for differential learning was used to account for its slower learning dynamics. The final model is selected from the checkpoints (across both training stages) that minimize the validation loss.

Molecular dynamics: All molecular dynamics simulations are performed using the i-PI code [52], using an ASE client[53], for calculating total energy and forces for the MLIPs. All simulations are thermalized using a Langevin thermostat with a time constant of 100 fs; for NPT simulations, we utilize the fully flexible Martyna-Tuckerman-Tobias-Klein barostat with a relaxation time of 1 ps, and both the system and the barostat "piston" are coupled with a Langevin thermostat with a time constant of 100 fs. We use a timestep of 0.5 fs, and sample positions, potential energies, and densities every 200 timesteps for 50 ps. The first 10 ps of the simulations are discarded as equilibration.

III. RESULTS

Before testing the performance of the models in the thermodynamic ensembles, we report total energy and force validation RMSEs across all strategies as a function of the training set size (number of 4.5 Å clusters) in Fig. S2 of the SI. The left panels correspond to training with high-fidelity forces included, and the right panels to training without forces. To arrive at these results, we perform for each strategy, a hyperparameter sweep over energy and force weights, with $\lambda_U, \lambda_\mathbf{f} \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, and report the RMSEs

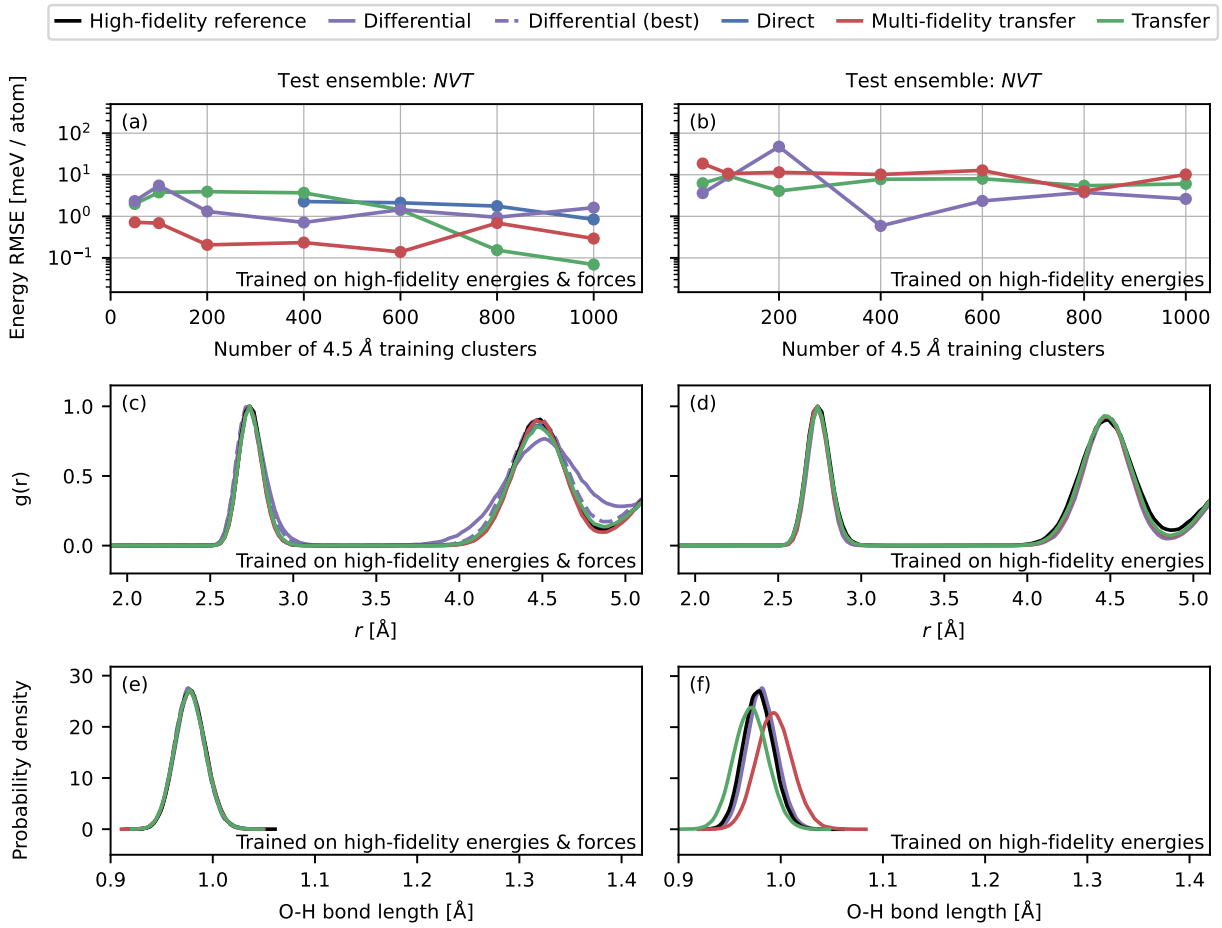

FIG. 2. Panels (a,c,e) show NVT energy RMSEs, O-O pair correlation functions, and O-H bond distributions for models trained on high-fidelity energies and forces. Panels (b,d,f) show the same quantities for models trained on high-fidelity energies only. 'High-fidelity reference' corresponds to the MACE-OFF-23 model. 'Differential (best)' refers to the delta learning model that achieved the lowest energy RMSE and was trained on 400 clusters. 'Transfer' is the single-fidelity fine-tuning method.

corresponding to the models that minimize the validation error. Hyperparameter optimization yields more than an order-of-magnitude improvement in energy and force RMSEs, with the optimal weights differing across training strategies. Table S1 of the SI reports the optimal weights used across the different strategies.

A. Test errors in the NVT ensemble

The accuracy and data efficiency of the training strategies in the NVT ensemble are summarized in Fig. 2. Results are shown separately for two scenarios: training on high-fidelity energies and forces, and training on high-fidelity energies (i.e., with forces omitted). As a first step, we report an error metric on a global observable – the average RMSE associated with potential energy. This metric is fundamental: in practice, the total energies of bulk systems or clusters carved from them at cWFT levels are the only reference data for validating a model[17,19,27]. We further examine microscopic observables with respect to the ground truth thermodynamic averages, namely the first two peaks of the O–O pair correlation function and the distribution of O–H bond lengths. This assessment is important, as average structural observables are often inaccessible at high-fidelity levels; it is typically assumed that low energy RMSEs are a sufficient condition for the identification of an accurate model.

Global observable: In panels (a) and (b) of Fig. 2, we report the energy RMSE across NVT trajectories generated by the MLIPs, as a function of the number of clusters in the training set. We first consider the scenario when high-fidelity forces are included in training. As shown in Fig. 2(a), all methods achieve sub- or near-meV/atom energy RMSE for the largest training set volume. Direct learning yields unstable trajectories and high RMSEs for small training sets, but improves systematically and achieves a sub-meV/atom RMSE with 1000 configurations. Differential learning yields unstable trajectories for 100 clusters, but with more training data,

shows improvement in simulation stability, although the energy RMSEs remain saturated around 1 meV/atom. Transfer learning approaches lead to stable simulations across the full data range, with single-fidelity fine-tuning leading to larger errors for small dataset sizes and eventually reaching the lowest energy error for 1000 clusters. Multi-fidelity learning yields near-constant sub-meV/atom errors, even with just 50 clusters.

We next consider the scenario when high-fidelity forces are not included in the training. Results from the direct learning are omitted in Fig. 2(b), as the trajectories were unstable. This occurs despite these models reporting systematically improving learning curves on the validation set, and achieving a sub-meV/atom energy RMSE for 1000 clusters (see Fig. S2 of the SI). Differential learning is unstable when training with 100 clusters, but improves with more data, saturating at ∼2 meV/atom with 1000 configurations. Standard transfer learning maintains stable performance across all training set sizes, reaching ∼5 meV/atom error. Multi-fidelity transfer learning, by contrast, yields stable simulations but consistently higher errors of ∼10 meV/atom.

Microscopic observables: We now examine if low-energy RMSEs correlate with a good description of the microscopic structure. We begin with the scenario in which high-fidelity forces are included in training. Figure 2(c) shows the oxygen–oxygen pair correlation functions up to the second peak for models trained on 1000 clusters. Direct learning and both transfer learning approaches (single-fidelity and multi-fidelity fine-tuning) reproduce the reference distribution within statistical error. By contrast, differential learning yields broadened first and second peaks. It is important to note that the differential learning model trained with 1000 configurations is not the one that achieves the lowest energy RMSE. For this reason, we also examine the best-performing differential learning model trained on 400 configurations. This model shows an improved agreement with the reference but still exhibits a mild systematic broadening. Turning to the O–H bond length distributions in Fig. 2(e), all models—including both differential learning models—accurately reproduce the reference distribution.

We next consider the scenario in which forces are not included in training. As shown in Fig. 2(d), all strategies (with the exception of direct learning, which yields unstable trajectories) reproduce the first peak of the O–O pair correlation function with quantitative accuracy. The differential learning model trained only on energies outperforms its counterpart, which is trained on both energies and forces. This suggests that, for differential learning, including forces in the training set leads to overfitting on clusters, hindering generalization to the bulk phase. For the second peak, however, all methods exhibit a mild over-structuring. The O–H bond length distributions in Fig. 2(f) reveal further issues. All approaches perform poorly for this O–H mean values and fluctuations. Here, both transfer learning approaches predict unphysical bond lengths, with single-fidelity fine-tuning yielding shorter and multi-fidelity fine-tuning yielding longer bond lengths. In contrast, differential learning shows the best performance but systematically overestimates the bond length by 0.003 Å. As an absolute, this deviation is small, but it is in the same ballpark as the changes in the O–H bond length when nuclear quantum effects are incorporated[54]. These differences can modulate O–H autoionization[55,56] and lead to changes in the vibrational spectra by over 100 cm^{-1}[57]. A possible explanation for the poor description of O–H modes is that when training on forces – which fluctuate strongly for high-frequency modes as opposed to the energy – the loss effectively assigns greater weight to these modes[58]. When training only on energies, this weighting is absent, which explains the systematically worse performance for high-frequency O–H compared to low-frequency O–O distributions.

These results may seem puzzling given the low energy RMSEs and stable trajectories; however, they are readily explained by the force RMSEs (calculated across the trajectories) shown in Fig. S4(a) and S4(b). Training only on high-fidelity energies (a global quantity) can lead to low-energy RMSEs but high-force RMSEs, which correlates with the poor description of microscopic properties. These findings highlight several caveats for the cluster-to-bulk generalization of MLIPs, which are further tested in the NPT ensemble. (1) Stable trajectories and low energy errors do not imply accurate forces, (2) Agreement on global properties, such as total energy, does not guarantee correct microscopic structure.

B. Test errors in the NPT ensemble

We next evaluate model performance in the NPT ensemble. This represents a more stringent test ensemble, as energy and force errors directly affect the instantaneous pressure, which in turn can impact the stability of the simulation as well as global and microscopic observables. Moreover, it allows us to assess another important thermodynamic quantity, the density.

Global observables: In Fig. 3(a) and Fig. 3(b), we report the energy RMSEs across NPT trajectories generated by the MLIPs, as a function of the number of clusters used for training. We first consider the scenario in which training includes both energies and forces. As shown in Fig. 3(a), direct models are unstable up to 400 clusters but improve systematically with larger training sets, reaching sub-meV/atom energy RMSEs. Differential learning models do not exhibit instability in the NPT ensemble but clearly show signs of overfitting: the model trained on 200 clusters achieves the lowest RMSE, while adding more data leads to unstable simulations. Both transfer learning approaches achieve the lowest-energy RMSEs, with multi-fidelity fine-tuning performing better in the low-data regime, and single-fidelity fine-tuning

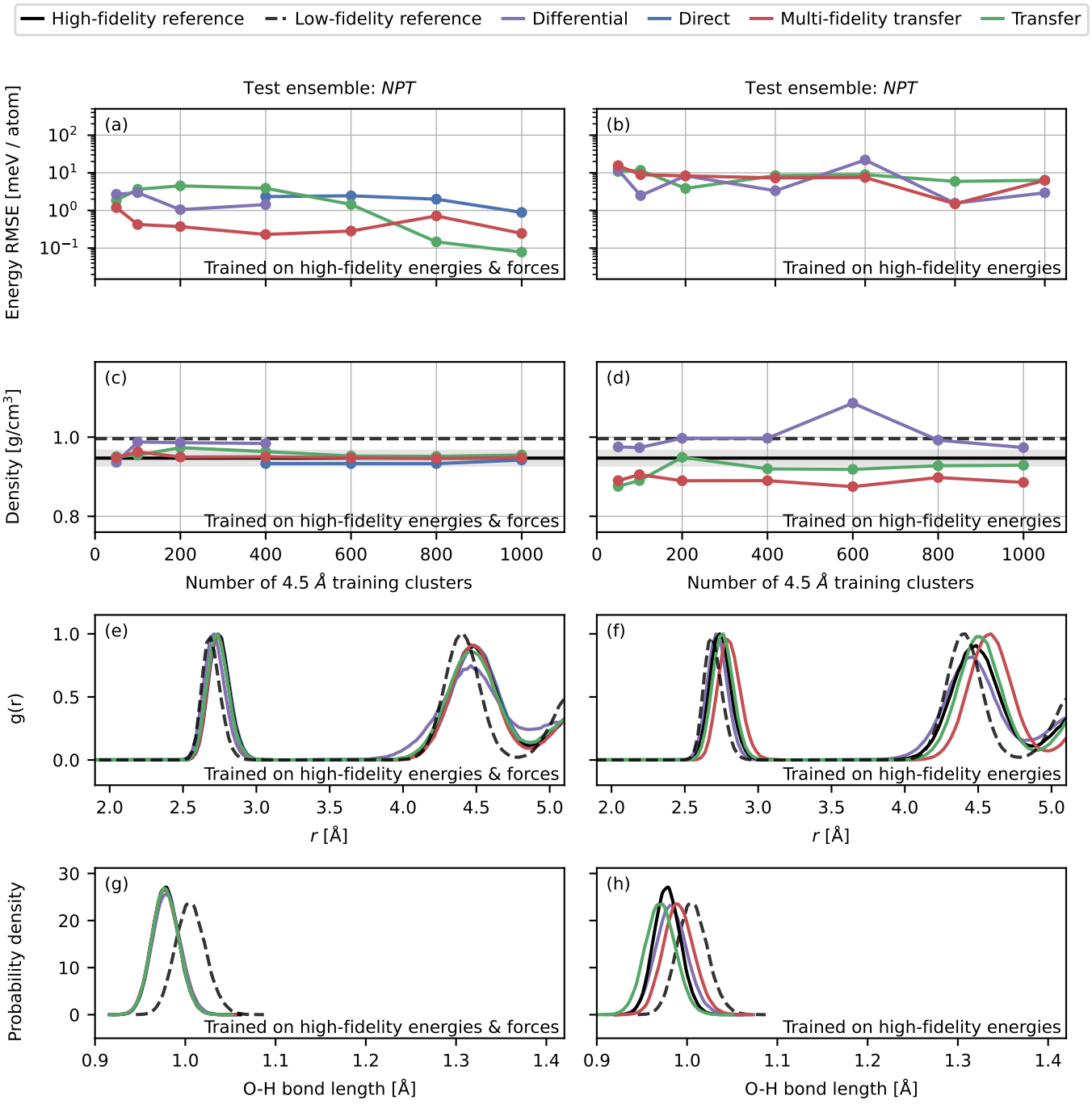

FIG. 3. Panels (a,c,e,g) show *NPT* energy RMSEs, average densities, O-O pair correlation functions, and O–H bond length distributions for models trained on high-fidelity energies and forces. Panels (b,d,f,h) show the same quantities for models trained on high-fidelity energies only. 'Low-fidelity reference' is the MACE-MP-0 reference.

outperforming at larger training set sizes. We also consider the density as another global observable. In this case, we also contextualize the performance of the models with respect to the low-fidelity reference density. As shown in Fig. 3(c), direct learning and both transfer learning strategies predict densities to within 2% of the reference. Differential learning, on the other hand, leads to a systematic overestimation of the density up to 400 clusters (this value is close to the low-fidelity reference), and unstable simulations occur beyond this point.

We next consider the scenario when high-fidelity forces are not included in the training. As shown in Fig. 3(b), the overall energy RMSEs are higher compared to the scenario where forces are incorporated in the training. This secondary effect is induced by errors in sampling, as understood by the high force RMSEs

in Figs. S4(c) and S4(d). Direct learning leads to unstable simulations across the full range of training data. Both types of transfer learning saturate in the range of 5-10 meV/atom errors. Differential learning leads to energy RMSEs in the 2-11 meV/atom range. We next analyze the density in Fig. 3(d), while comparing with the low-fidelity ground truth density estimate to understand if the errors arise from an undercorrection or an overcorrection. The description of the density is overall worse compared to the scenario when forces are included in the training. Differential learning leads to an overestimation of the density, to nearly the same extent as when training on both energies and forces, and close to the low-fidelity estimate. This suggests that differential learning undercorrects. Single-fidelity fine-tuning leads to sub-2% agreement with the reference, while multi-fidelity fine-tuning leads to an underestimation of the density, which suggests both these approaches (over)correct for the density difference between the two fidelity levels.

Microscopic observables: We first consider the scenario in which forces are included in the training. Fig. 3(e) shows the oxygen–oxygen radial distribution functions along with the low- and high-fidelity references. For the first peak of the distribution, both direct and transfer learning approaches yield a quantitative agreement with the reference. On the other hand, the best differential learning model (trained on 200 clusters) leads to a small underestimation of the nearest O–O distance. This deviation is in the same direction as the low-fidelity distribution, again implying that differential learning undercorrects. For the second peak, direct learning and multi-fidelity fine-tuning yield a quantitative agreement, while single-fidelity fine-tuning and differential learning yield a distribution skewed towards the low-fidelity estimate. Fig. 3(g) shows the O–H bond distribution with all methods yielding quantitative agreement with the expectation of differential learning, which exhibits mild overstructuring. These results show that the accuracies achieved by the different approaches when training with forces in the NVT ensemble carry over to the NPT ensemble, and the artifacts are more pronounced.

We next consider the scenario in which high-fidelity forces are excluded from the training, starting the O–O distribution. From the NVT analysis, we already anticipate a poor performance, but we check further if the low-fidelity level biases the microscopic structure. As shown in 3(f), the distribution of both the first and second peaks deviates from the reference for all the approaches. Single-fidelity fine-tuning skews the distributions to longer O–O distances, implying an overcorrection. The multi-fidelity fine-tuning method skews the distributions towards longer distances. Differential learning shows the smallest errors, but systematically skews the distributions towards the low-fidelity reference. For the O–H distances, shown in Fig. 3(h), we note that differential learning and the multi-fidelity fine-tuning models are skewed towards the low-fidelity reference. Single-fidelity fine-tuning model overcorrects towards the high-fidelity reference. In summary, the NPT ensemble highlights the same caveats identified in NVT, but in a more pronounced form. It further shows that, when models are trained only on energies, differential learning and multi-fidelity fine-tuning tend to inherit the bias of the low-fidelity reference, whereas single-fidelity fine-tuning tends to overcorrect.

IV. DISCUSSION

In summary, we investigate how well MLIPs generalize to bulk properties when trained with direct, differential, and transfer learning on high-fidelity total energies and/or forces estimated for molecular clusters. We utilize pretrained MLIPs as proxies to the low- and high-fidelity levels, enabling us to estimate ground-truth observables in NVT and NPT ensembles. We consider both global properties, such as the total energy and density, as well as microscopic properties, including the intermolecular (O-O) and intramolecular (O–H) distributions. We first summarize the identified strengths and limitations of the four strategies and then discuss the best recipes for the cluster-to-bulk generalization of MLIPs. It should be noted, however, that the ability of an MLIP to generalise depends on how the implicit atomic body order is represented, which in turn is sensitive to the underlying architecture[59]. Accordingly, our conclusions are limited to ACE-based MLIPs.

Direct learning: Direct learning is feasible with only a few hundred clusters, provided both total energies and forces are included in the training set, yielding an excellent description of both global and microscopic properties. The main drawback of direct learning is that MLIPs are not robust to extrapolation. We arrive at this conclusion by performing temperature extrapolations, summarized in Fig. S9, in which NPT simulations were performed up to 300 K in 50 K increments in the metastable solid phase. These show that, despite excellent performance at 100 K, direct learning on total energies and forces fails to produce stable simulations at higher temperatures.

Transfer learning – single-fidelity fine-tuning: Single-fidelity fine-tuning is found to be data-efficient and robust with respect to the stability of the simulations. When trained on both energies and forces, this strategy yields quantitative accuracy for both global and microscopic properties in the NVT and NPT ensembles. When trained only on energies, its accuracy is sensitive to the observable and the thermodynamic ensemble. For the total energy and O–O distribution functions, quantitative agreement is achieved in the NVT ensemble, while higher energy RMSEs and skewed (overcorrected) O–O distributions are observed in the

NPT ensemble. For O–H bond distributions, both ensembles exhibit shorter (overcorrected) bond lengths. We also investigated whether using larger cluster sizes in the training set fixes these limitations when training only on energies. As summarized in Figs. S6 and S7, we do not observe systematic improvements with larger clusters: energy and force errors are saturated nearly to the same extent as for 4.5 Å clusters. In fact, the good agreement with the density obtained for 4.5 Å clusters is degraded when larger clusters are used.

Differential learning: We observe artifacts associated with differential learning in the context of the cluster-to-bulk generalization. When trained on both energies and forces, differential learning is noted to overfit to cluster data, which results in an overall poor description of global and microscopic properties and unphysical simulations with 400 clusters. Overfitting is reduced when training only on energies, but the accuracy of the MLIPs is sensitive to the observables. Considering global properties, the total energy RMSEs decrease by ∼2 meV/atom in both ensembles, and the density error is reduced to ∼2% with 1000 clusters. In contrast, the O–O and O–H distributions remain skewed toward the low-fidelity reference, with the deviations significantly larger in the NPT ensemble than in NVT. It is important to note, however, that all strategies performed poorly when trained only on energies, with differential learning yielding the best results among them. We also tested for the temperature extrapolation of MLIPs trained with differential learning on energies only, and note that while all simulations are stable, the errors in the predicted density increase with temperature towards the low-fidelity reference. Thus, the results of differential learning simulations must be interpreted with caution, as stable trajectories do not necessarily imply accuracy with respect to the reference. A positive aspect of this approach is that its performance can be improved by increasing the cluster size in the training set. As shown in Fig. S6, the description of global properties in the NPT ensemble shows a modest improvement with 6 Å clusters compared to 4.5 Å clusters. For microscopic properties (Fig. S7), the accuracy of O–H distributions improves using larger clusters, whereas the error in the O–O distributions is essentially unchanged. Given the systematically improvable nature of differential learning, we anticipate that data augmentation strategies could help mitigate its limitations.

Transfer learning – multi-fidelity fine-tuning: Multi-fidelity fine-tuning emerges as the most data-efficient and accurate approach for both global and microscopic properties across both ensembles when trained on energies and forces. As shown in Fig. S8, multi-fidelity models also display excellent extrapolation of the density up to 300 K, despite training clusters being representative of 100 K only. There, it should be the method of choice, then the training data comprises both total energies and forces. Alongside these strengths, we identify two areas where multi-fidelity fine-tuning can be improved. First, we do not observe systematic improvements with increasing cluster size. When trained on 4 Å clusters comprising only five water molecules, multi-fidelity fine-tuning is the only approach that yields stable simulations with semi-quantitative agreement for the density and O–O and O–H bond lengths, as shown in Fig. S5. In contrast, training on 6 Å clusters results in slightly worse descriptions of these observables as shown in Figs. S6 and S7. This suggests that the fine-tuning dynamics considered here do not exhibit "size consistency" and may require a dedicated hyperparameter optimization per dataset. Second, when trained only on energies, the method performs poorly for microscopic observables, showing significant deviations in O–H bond distributions, and thus, the recipe considered here is not recommended. Overall, multi-fidelity fine-tuning remains the most flexible approach in terms of combining low- and high-fidelity data. Given its demonstrated success on both clusters[48] and bulk systems[47], we expect that future work focused on data augmentation could yield a robust recipe for the cluster-to-bulk generalization when training on energies.

As far as the cluster-to-bulk generalization of MLIPs is concerned, our results show that this approach can be successfully applied to a range of molecular systems, provided validation is performed thoroughly. For an MLIP to be robust, it must accurately describe atomic forces in addition to energies. We reach this conclusion by noting that quantitative agreement with microscopic properties correlates strongly with low force errors, but only weakly with low energy errors or stable simulations. This highlights that in scenarios where forces are unavailable, it is crucial to obtain indirect probes of force accuracy, for example, by examining multiple global and microscopic observables associated with high-frequency intramolecular modes – which, by definition, are governed by large forces. We further find that generalization to the NPT ensemble is considerably more demanding than to NVT, making careful validation essential. Finally, the strong correlation between accurate microscopic properties and low-force errors across trajectory ensembles motivates two key directions for future work: (i) developing MLIP strategies capable of accurate force prediction when training only on energies, and (ii) advancing efficient, mainstream implementations of gradients[60–62] for cWFT methods, both for clusters (training) and bulk phases (validation), to enable robust and cost-effective cluster-to-bulk generalization.

ACKNOWLEDGMENTS

We thank Michele Ceriotti, Ilyes Batatia, and Sander Vandenhaute for their valuable comments on the manuscript. MJG acknowledges the UCL Department of Physics and Astronomy's PhD studentship, and VK acknowledges support from UCL's startup funds. ML

further acknowledges the CCP summer bursary. The Flatiron Institute is a division of the Simons Foundation. We are grateful for computational support from the Swiss National Supercomputing Centre (CSCS) under project s1288 on Alps and UCL Myriad High Performance Computing Facility (Myriad@UCL). This work used computing equipment funded by the Research Capital Investment Fund (RCIF) provided by UKRI, and partially funded by the UCL Cosmoparticle Initiative.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

DATA AVAILABILITY

All the training data, code, and scripts required to reproduce the results will be made available upon publication.

REFERENCES

[1]D. P. Kovács, J. H. Moore, N. J. Browning, I. Batatia, J. T. Horton, Y. Pu, V. Kapil, W. C. Witt, I.-B. Magdău, D. J. Cole, and G. Csányi, Journal of the American Chemical Society **147**, 17598 (2025).
[2]I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S. M. Blau, V. Cărare, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A. C. Ferrari, A. Genreith-Schriever, J. George, R. E. A. Goodall, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T. D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, F. Zills, and G. Csányi, "A foundation model for atomistic materials chemistry," (2024), arXiv:2401.00096 [cond-mat, physics:physics].
[3]A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Nature **624**, 80 (2023).
[4]H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, C. Zeni, M. Horton, R. Pinsler, A. Fowler, D. Zügner, T. Xie, J. Smith, L. Sun, Q. Wang, L. Kong, C. Liu, H. Hao, and Z. Lu, "MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures," (2024), arXiv:2405.04967 [cond-mat].
[5]O. T. Unke, M. Stöhr, S. Ganscha, T. Unterthiner, H. Maennel, S. Kashubin, D. Ahlin, M. Gastegger, L. Medrano Sandonas, J. T. Berryman, A. Tkatchenko, and K.-R. Müller, Science Advances **10**, eadn4397 (2024).
[6]P. J. Feibelman, B. Hammer, J. K. Nørskov, F. Wagner, M. Scheffler, R. Stumpf, R. Watwe, and J. Dumesic, The Journal of Physical Chemistry B **105**, 4018 (2001).
[7]J. L. Bao, L. Gagliardi, and D. G. Truhlar, The Journal of Physical Chemistry Letters **9**, 2353 (2018).
[8]B. X. Shi, V. Kapil, A. Zen, J. Chen, A. Alavi, and A. Michaelides, The Journal of Chemical Physics **156**, 124704 (2022).
[9]K. R. Bryenton, A. A. Adeleke, S. G. Dale, and E. R. Johnson, WIREs Computational Molecular Science **13**, e1631 (2023).
[10]B. X. Shi, D. J. Wales, A. Michaelides, and C. W. Myung, Journal of Chemical Theory and Computation **20**, 5306 (2024).
[11]L. Schimka, J. Harl, A. Stroppa, A. Grüneis, M. Marsman, F. Mittendorfer, and G. Kresse, Nature Materials **9**, 741 (2010).
[12]J. Yang, W. Hu, D. Usvyat, D. Matthews, M. Schütz, and G. K.-L. Chan, Science **345**, 640 (2014).
[13]J. Sauer, Accounts of Chemical Research **52**, 3502 (2019).
[14]H.-Z. Ye and T. C. Berkelbach, Faraday Discussions **254**, 628 (2024).
[15]F. Della Pia, A. Zen, D. Alfè, and A. Michaelides, Physical Review Letters **133**, 046401 (2024).
[16]B. X. Shi, A. S. Rosen, T. Schäfer, A. Grüneis, V. Kapil, A. Zen, and A. Michaelides, Nature Chemistry , 1 (2025).
[17]J. Daru, H. Forbert, J. Behler, and D. Marx, Physical Review Letters **129**, 226001 (2022).
[18]B. B. Mészáros, A. Szabó, and J. Daru, Journal of Chemical Theory and Computation **21**, 5372 (2025).
[19]N. O'Neill, B. X. Shi, W. Baldwin, W. C. Witt, G. Csányi, J. D. Gale, A. Michaelides, and C. Schran, "Towards Routine Condensed Phase Simulations with Delta-Learned Coupled Cluster Accuracy: Application to Liquid Water," (2025), arXiv:2508.13391 [physics].
[20]G. R. Medders, A. W. Götz, M. A. Morales, P. Bajaj, and F. Paesani, The Journal of Chemical Physics **143**, 104102 (2015).
[21]S. K. Reddy, S. C. Straight, P. Bajaj, C. Huy Pham, M. Riera, D. R. Moberg, M. A. Morales, C. Knight, A. W. Götz, and F. Paesani, The Journal of Chemical Physics **145**, 194504 (2016).
[22]C. Qu, Q. Yu, P. L. Houston, R. Conte, A. Nandi, and J. M. Bowman, Journal of Chemical Theory and Computation **19**, 3446 (2023).
[23]Z. Xie and J. M. Bowman, Journal of Chemical Theory and Computation **6**, 26 (2010).
[24]J. Lan, V. Kapil, P. Gasparotto, M. Ceriotti, M. Iannuzzi, and V. V. Rybkin, Nature Communications **12**, 766 (2021).
[25]J. Behler, International Journal of Quantum Chemistry **115**, 1032 (2015).
[26]M. Del Ben, J. Hutter, and J. VandeVondele, Journal of Chemical Theory and Computation **9**, 2654 (2013).
[27]M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman, and T. E. Markland, Journal of Chemical Theory and Computation **19**, 4510 (2023).
[28]S. Thrun, in *Advances in Neural Information Processing Systems*, Vol. 8, edited by D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (MIT Press, 1995).
[29]G. D. Purvis, III and R. J. Bartlett, The Journal of Chemical Physics **76**, 1910 (1982).
[30]K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon, Chemical Physics Letters **157**, 479 (1989).
[31]S. Zhang and H. Krakauer, Physical Review Letters **90**, 136401 (2003).
[32]R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, Journal of Chemical Theory and Computation **11**, 2087 (2015).
[33]P. Pinski, C. Riplinger, E. F. Valeev, and F. Neese, The Journal of Chemical Physics **143**, 034108 (2015).
[34]C. Riplinger, P. Pinski, U. Becker, E. F. Valeev, and F. Neese, The Journal of Chemical Physics **144**, 024109 (2016).
[35]J. Nigam, S. Pozdnyakov, and M. Ceriotti, The Journal of Chemical Physics **153**, 121101 (2020).
[36]R. Drautz, Physical Review B **99**, 014104 (2019).
[37]S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nature Communications **13**, 2453 (2022).

[38] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, (2022), 10.48550/ARXIV.2205.06643.
[39] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, Physical Review X **14**, 021036 (2024).
[40] J. P. Darby, D. P. Kovács, I. Batatia, M. A. Caro, G. L. Hart, C. Ortner, and G. Csányi, Physical Review Letters **131**, 028001 (2023).
[41] H. Kaur, F. D. Pia, I. Batatia, X. R. Advincula, B. X. Shi, J. Lan, G. Csányi, A. Michaelides, and V. Kapil, Faraday Discussions (2024), 10.1039/D4FD00107A.
[42] B. Deng, "Materials project trajectory (mptrj) dataset," (2023).
[43] M. Bocus, S. Vandenhaute, and V. Van Speybroeck, Angewandte Chemie International Edition **64**, e202413637 (2025).
[44] P. Novelli, G. Meanti, P. J. Buigues, L. Rosasco, M. Parrinello, M. Pontil, and L. Bonati, "Fast and Fourier Features for Transfer Learning of Interatomic Potentials," (2025), arXiv:2505.05652 [physics].
[45] M. Radova, W. G. Stark, C. S. Allen, R. J. Maurer, and A. P. Bartók, "Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning," (2025), arXiv:2502.15582 [cond-mat].
[46] A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche, and M. Ceriotti, "PET-MAD, a lightweight universal interatomic potential for advanced materials modeling," (2025), arXiv:2503.14118 [cond-mat].
[47] M. Messerly, S. Matin, A. E. A. Allen, B. Nebgen, K. Barros, J. S. Smith, N. Lubbers, and R. Messerly, "Multi-fidelity learning for interatomic potentials: Low-level forces and high-level energies are all you need," (2025), arXiv:2505.01590 [physics].
[48] M. Cui, K. Reuter, and J. T. Margraf, Machine Learning: Science and Technology **6**, 015071 (2025).
[49] F. D. Pia, B. X. Shi, V. Kapil, A. Zen, D. Alfè, and A. Michaelides, Chemical Science **16**, 11419 (2025).
[50] M. Tuckerman, *Statistical Mechanics: Theory and Molecular Simulation* (OUP Oxford, 2010) google-Books-ID: Lo3Jqc0pgrcC.
[51] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, Advances in Neural Information Processing Systems **35**, 11423 (2022).
[52] Y. Litman, V. Kapil, Y. M. Y. Feldman, D. Tisi, T. Begušić, K. Fidanyan, G. Fraux, J. Higer, M. Kellner, T. E. Li, E. S. Pós, E. Stocco, G. Trenins, B. Hirshberg, M. Rossi, and M. Ceriotti, The Journal of Chemical Physics **161**, 062504 (2024), https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0215869/20111898/062504_1_5.0215869.pdf.
[53] A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, J. Phys. Condens. Matter **29**, 273002 (2017).
[54] N. Stolte, J. Daru, H. Forbert, J. Behler, and D. Marx, The Journal of Physical Chemistry Letters **15**, 12144 (2024).
[55] M. Ceriotti, J. Cuny, M. Parrinello, and D. E. Manolopoulos, Proceedings of the National Academy of Sciences **110**, 15591 (2013).
[56] S. Dasgupta, G. Cassone, and F. Paesani, The Journal of Physical Chemistry Letters **16**, 2996 (2025).
[57] V. Kapil, D. P. Kovacs, G. Csányi, and A. Michaelides, Faraday Discussions (2023), 10.1039/D3FD00113J.
[58] F. Bigi, M. Langer, and M. Ceriotti, "The dark side of the forces: assessing non-conservative force models for atomistic machine learning," (2025), arXiv:2412.11569 [physics].
[59] S. Chong, T. Jiang, M. Domina, F. Bigi, F. Grasselli, J. Lee, and M. Ceriotti, "Resolving the Body-Order Paradox of Machine Learning Interatomic Potentials," (2025), arXiv:2509.14146 [physics].
[60] P. Pinski and F. Neese, The Journal of Chemical Physics **150**, 164102 (2019).
[61] X. Zhang, C. Li, H.-Z. Ye, T. C. Berkelbach, and G. K.-L. Chan, (2024), 10.48550/arXiv.2404.03129, arXiv:2404.03129 [physics].
[62] E. Slootman, I. Poltavsky, R. Shinde, J. Cocomello, S. Moroni, A. Tkatchenko, and C. Filippi, Journal of Chemical Theory and Computation **20**, 6020 (2024).

The Good, the Bad, and the Ugly of Atomistic Learning for "Clusters-to-Bulk" Generalization

Mikołaj J. Gawkowski,[1,2] Mingjia Li,[1,2] Benjamin X. Shi,[3] and Venkat Kapil[1,2,a)]

[1)] *Department of Physics and Astronomy, University College London, 7-19 Gordon St, London WC1H 0AH, UK*

[2)] *Thomas Young Centre and London Centre for Nanotechnology, 9 Gordon St, London WC1H 0AH, UK*

[3)] *Initiative for Computational Catalysis, Flatiron Institute, New York, New York 10010, USA*

[a)]Electronic mail: v.kapil@ucl.ac.uk

CONTENTS

S1. REFERENCE OBSERVABLES AT LOW- AND HIGH-FIDELITY

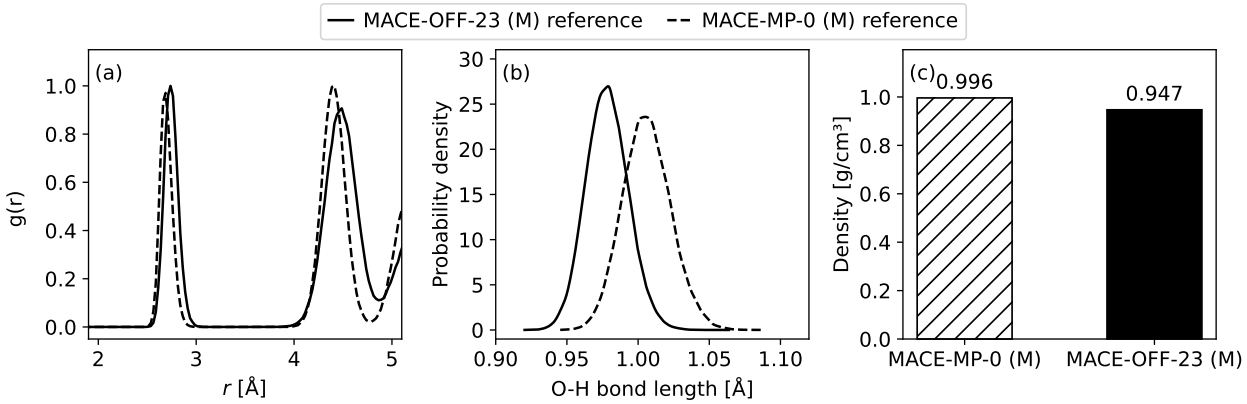

FIG. S1. Comparison of microscopic properties from NPT simulations of hexagonal ice using the medium (M) MACE-MP-0 and MACE-OFF-23 models. (a) O-O pair correlation functions, (b) distributions of O-H bond lengths, and (c) average densities from the NPT trajectories.

S2. VALIDATION ERRORS

A. Energy and force RMSEs

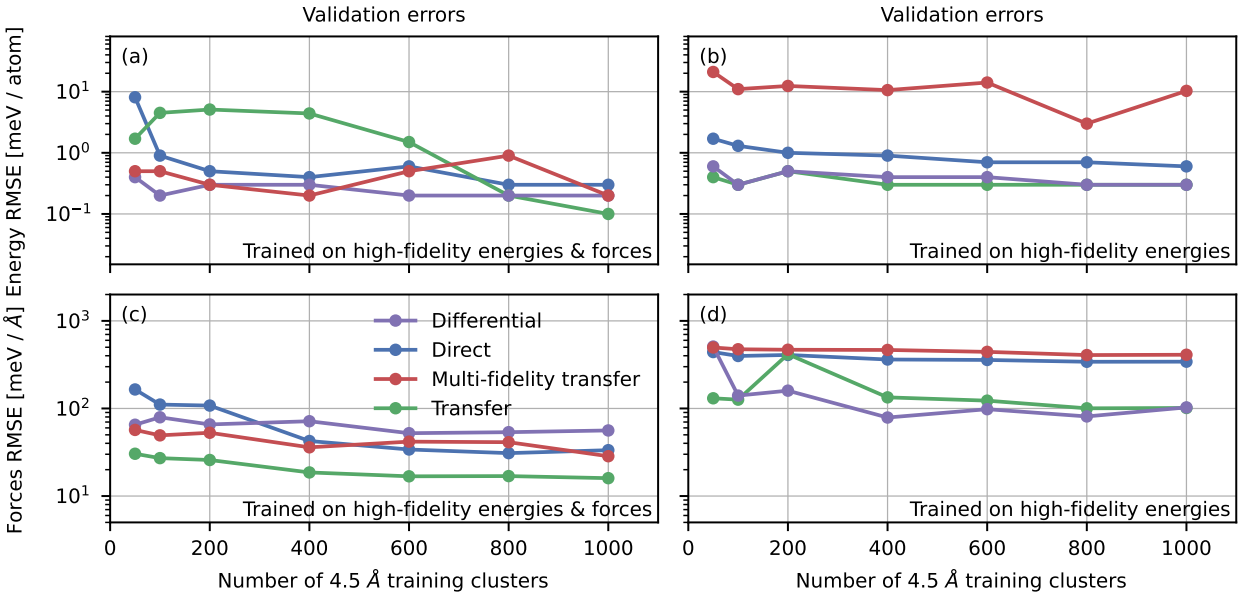

FIG. S2. RMSEs for ice-Ih on the validation set as a function of the number of 4.5 Å clusters in the training set. Panels (a) and (c) show the energy-per-atom and mean force RMSEs, respectively, for models trained on both high-fidelity energies and forces. Panels (b) and (d) show the corresponding results for models trained on high-fidelity energies only.

B. Hyperparameter optimization

Training protocol	$(E^{4.5}_{\mathrm{RMSE}}, F^{4.5}_{\mathrm{RMSE}})$	$(\lambda^{4.5}_{U}, \lambda^{4.5}_{\mathbf{f}})$	$(E^{4.5}_{\mathrm{RMSE}}, F^{4.5}_{\mathrm{RMSE}})$	$(\lambda^{4.5}_{U}, \lambda^{4.5}_{\mathbf{f}})$
Direct	(0.3, 33.4)	(1000, 10)	(0.6, 342.2)	(1000, 0)
Differential	(0.2, 56.1)	(1000, 1)	(0.3,102.8)	(1000,0)
Transfer	(0.1, 16)	(10, 10)	(0.3,101.3)	(100,0)
Multi-fidelity transfer	(0.2, 28.5)	(100, 10)	(0.3, 250.9)	(10, 0)

TABLE S1. Hyperparameter optimization results for energy and force errors, as well as the optimal training weights, across four training strategies. $(E^{4.5}_{\mathrm{RMSE}}, F^{4.5}_{\mathrm{RMSE}})$ denote the energy and force RMSEs obtained when training on clusters of size 4.5 Å. $(\lambda^{i}_{U}, \lambda^{i}_{\mathbf{f}})$ denote the corresponding optimal energy and force weights when training on 4.5 Å clusters.

S3. VALIDATION WITH DFT DATA

The high-fidelity DFT calculations were revPBE-D3, with zero-damping in the D3 dispersion, and the low-fidelity was a medium MACE-MP-0 model with the pretrained head set to the MPTrj dataset. The reference bulk density was calculated from periodic calculations in the Vienna *ab initio* Simulations Package (VASP)[1,2]. We used hard projector augmented wave (PAW) potentials (version 54) and an energy cutoff of 1000 eV on the Γ-point. The cluster calculations were performed in ORCA v5.0.3[3] for the def2-QZVPPD basis set. The isolated atomic energies were U_H = -13.72710738642251 eV and U_O = -2042.922011881597 eV for the revPBE-D3 data.

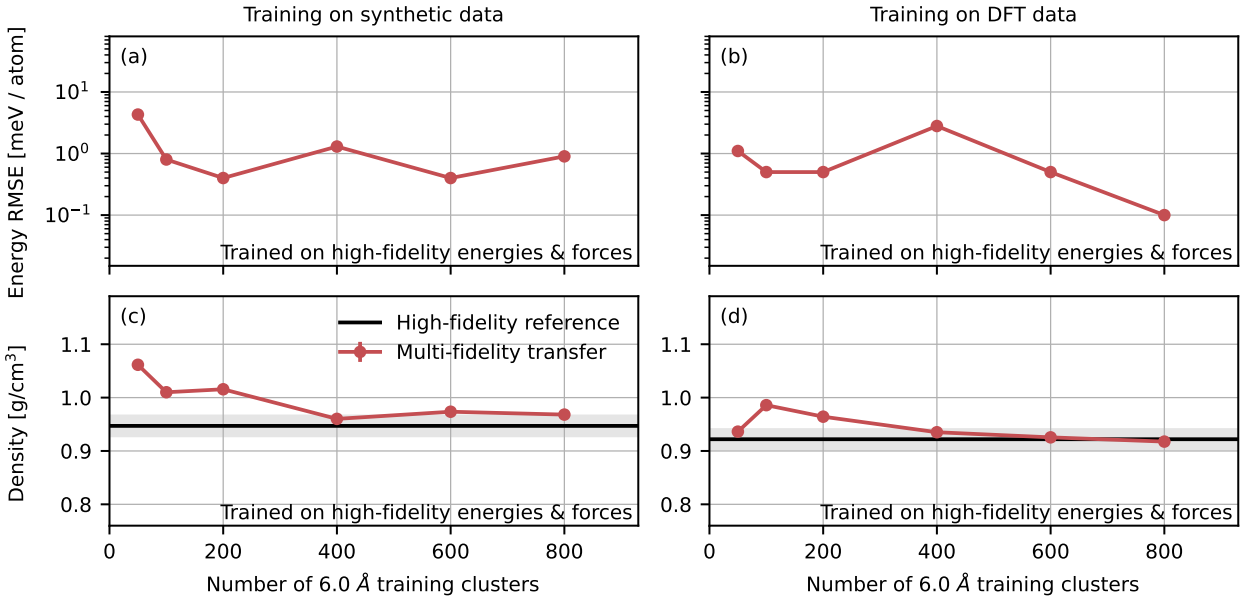

FIG. S3. Validation energy RMSEs and average densities of NPT trajectories for models trained on 6.0 Å clusters: panels (a,c) show results using synthetic data, and panels (b,d) show results using DFT data. 'High-fidelity reference' in panel (c) corresponds to the MACE-OFF-23 reference density, whereas the high-fidelity reference in panel (d) is the revPBE-D3 reference density.

S4. FORCE RMSE IN THE NVT AND NPT ENSEMBLES

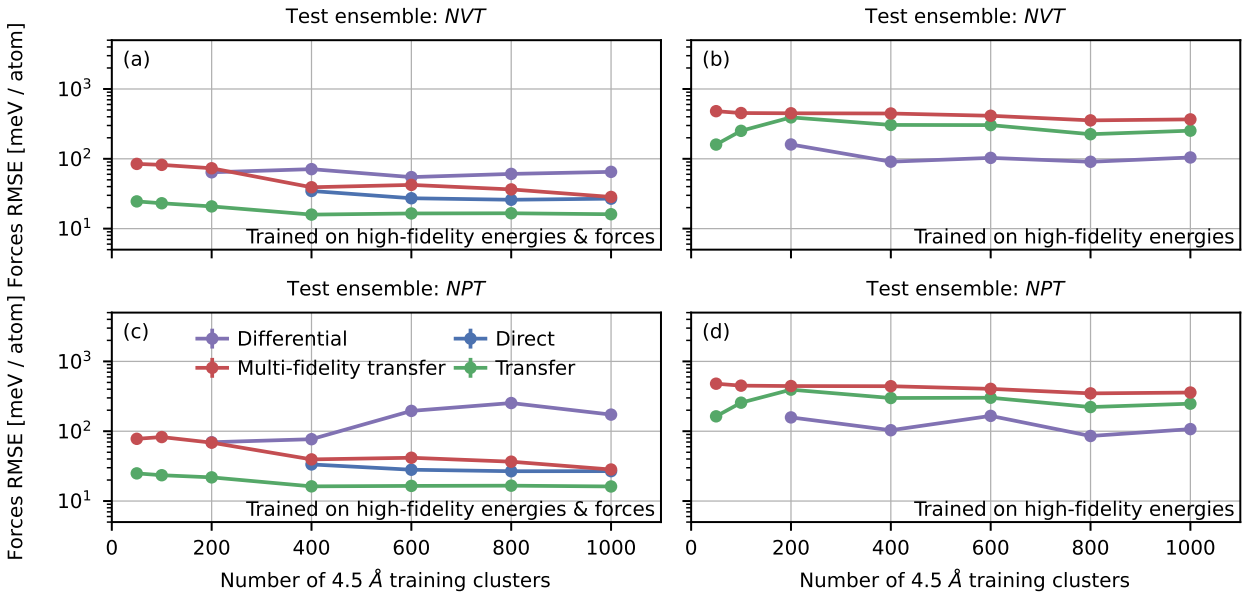

FIG. S4. Force RMSEs under different training conditions: (a,c) trained on high-fidelity energies and forces; (b,d) trained on high-fidelity energies only. Results are shown for testing in the NVT (a,b) and NPT (c,d) ensembles.

S5. TRAINING ON 4 Å AND 6 Å CLUSTERS

A. 4 Å clusters

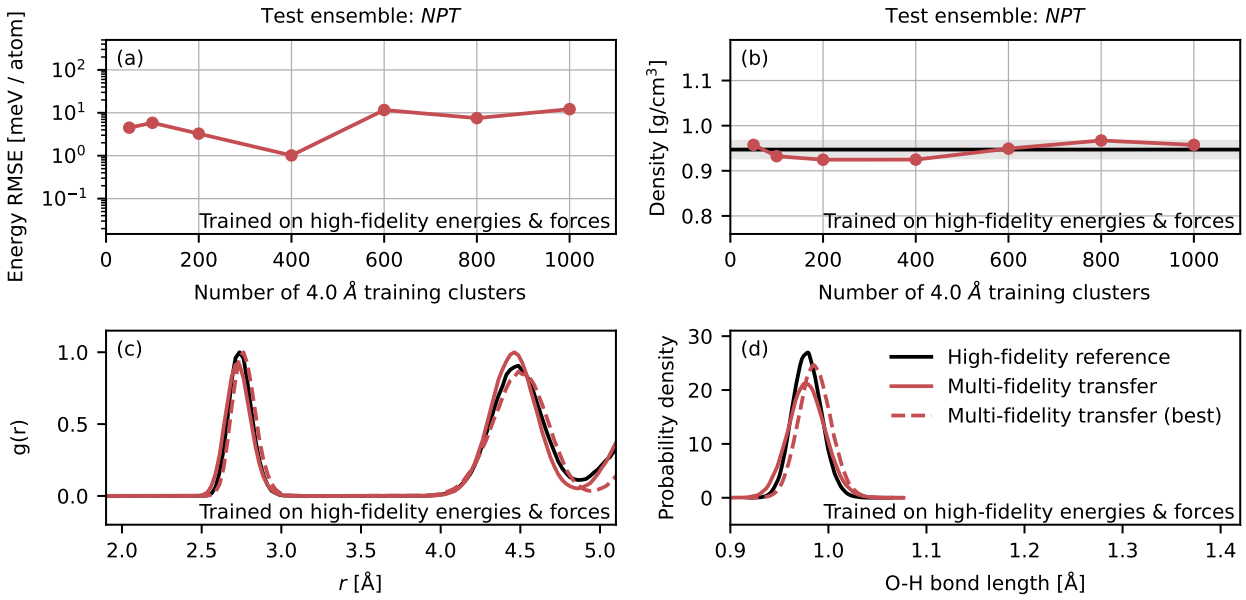

FIG. S5. (a) Energy RMSEs, (b) densities of NPT trajectories, (c) O-O pair correlation functions, and (d) O-H bond distributions obtained with the multi-fidelity transfer learning approach trained on 4.0 Å clusters. 'Multi-fidelity transfer (best)' refers to the multi-fidelity transfer model that achieved the lowest energy RMSE and was trained on 400 clusters.

B. 6 Å clusters

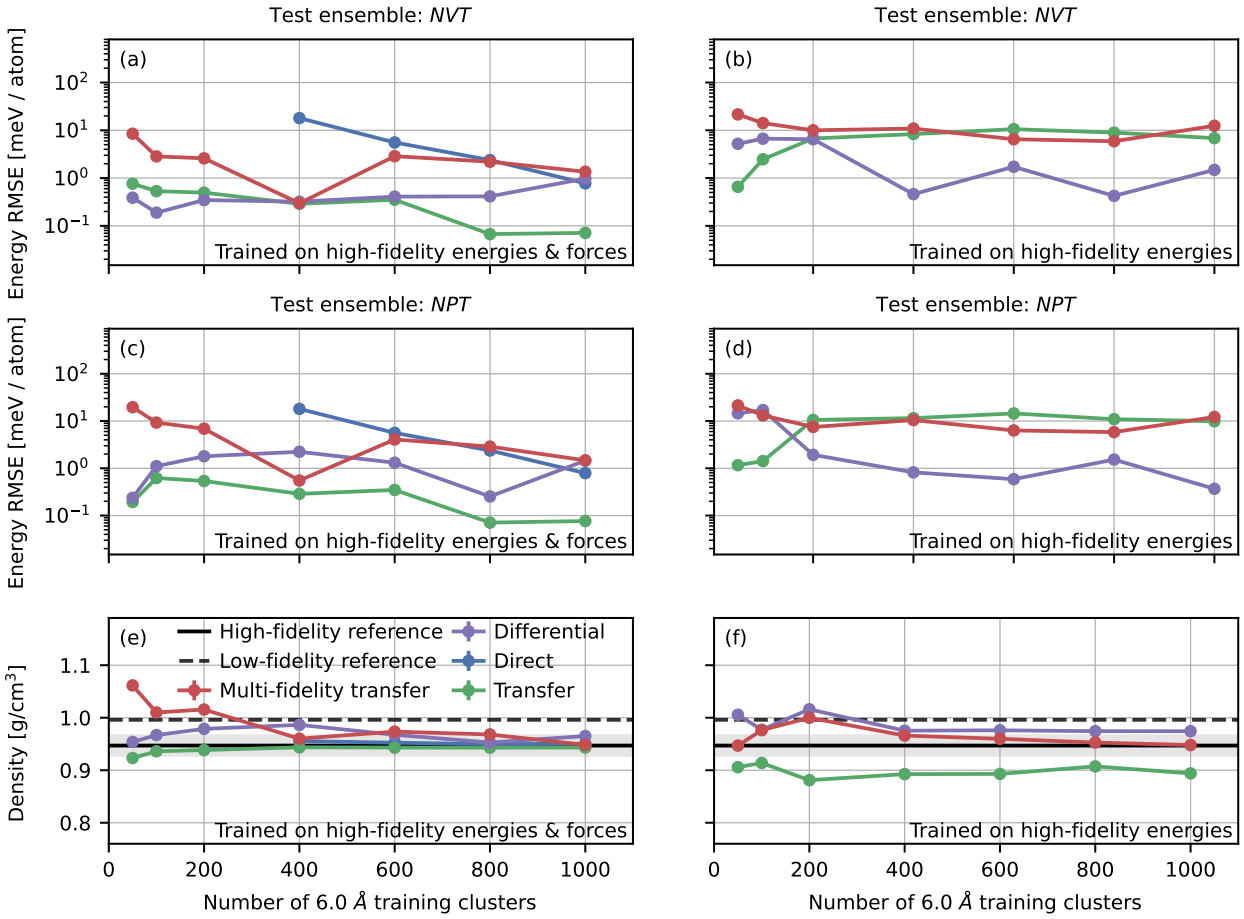

FIG. S6. Energy RMSEs and average densities from NPT simulations for models trained on 6.0 Å clusters. Left-hand panels show results for models trained on both high-fidelity energies and forces, tested in the NVT (a) and NPT (c,e) ensembles. Right-hand panels show results for models trained on high-fidelity energies only, tested in the NVT (b) and NPT (d,f) ensembles.

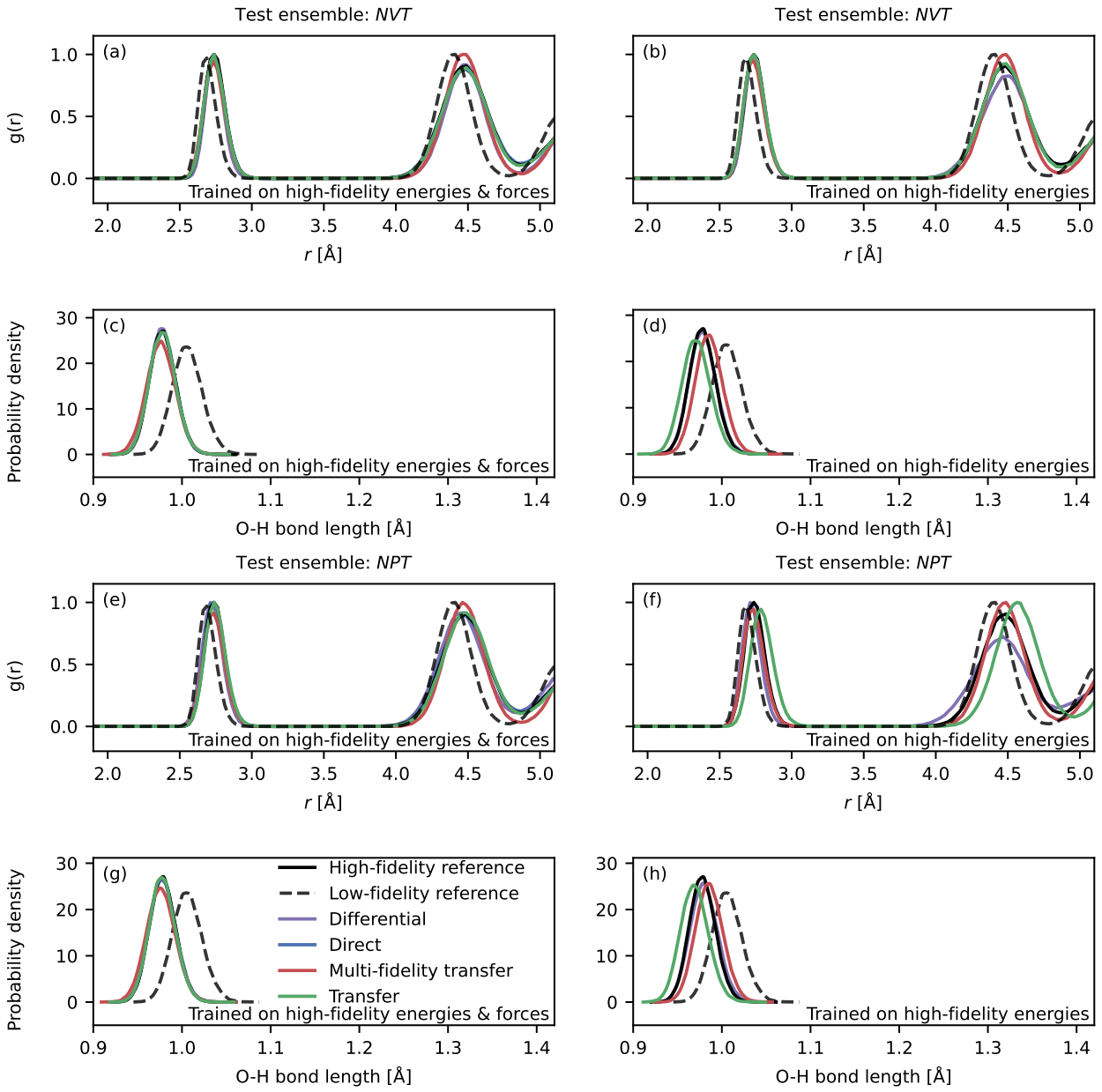

FIG. S7. Structural properties from NPT simulations using models trained on 6.0 Å clusters. Left-hand panels show results for models trained on both high-fidelity energies and forces, tested in the NVT (a,c) and NPT (e,g) ensembles. Right-hand panels show results for models trained on high-fidelity energies only, tested in the NVT (b,d) and NPT (f,h) ensembles.

S6. TEMPERATURE EXTRAPOLATIONS

A. Density

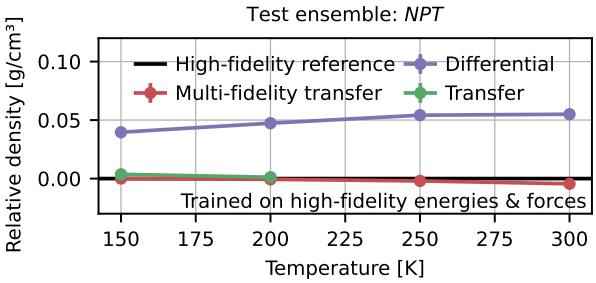

FIG. S8. Relative densities of structures sampled from NPT simulations at four temperatures shown for different training methods relative to the high-fidelity reference values.

B. Microscopic observables

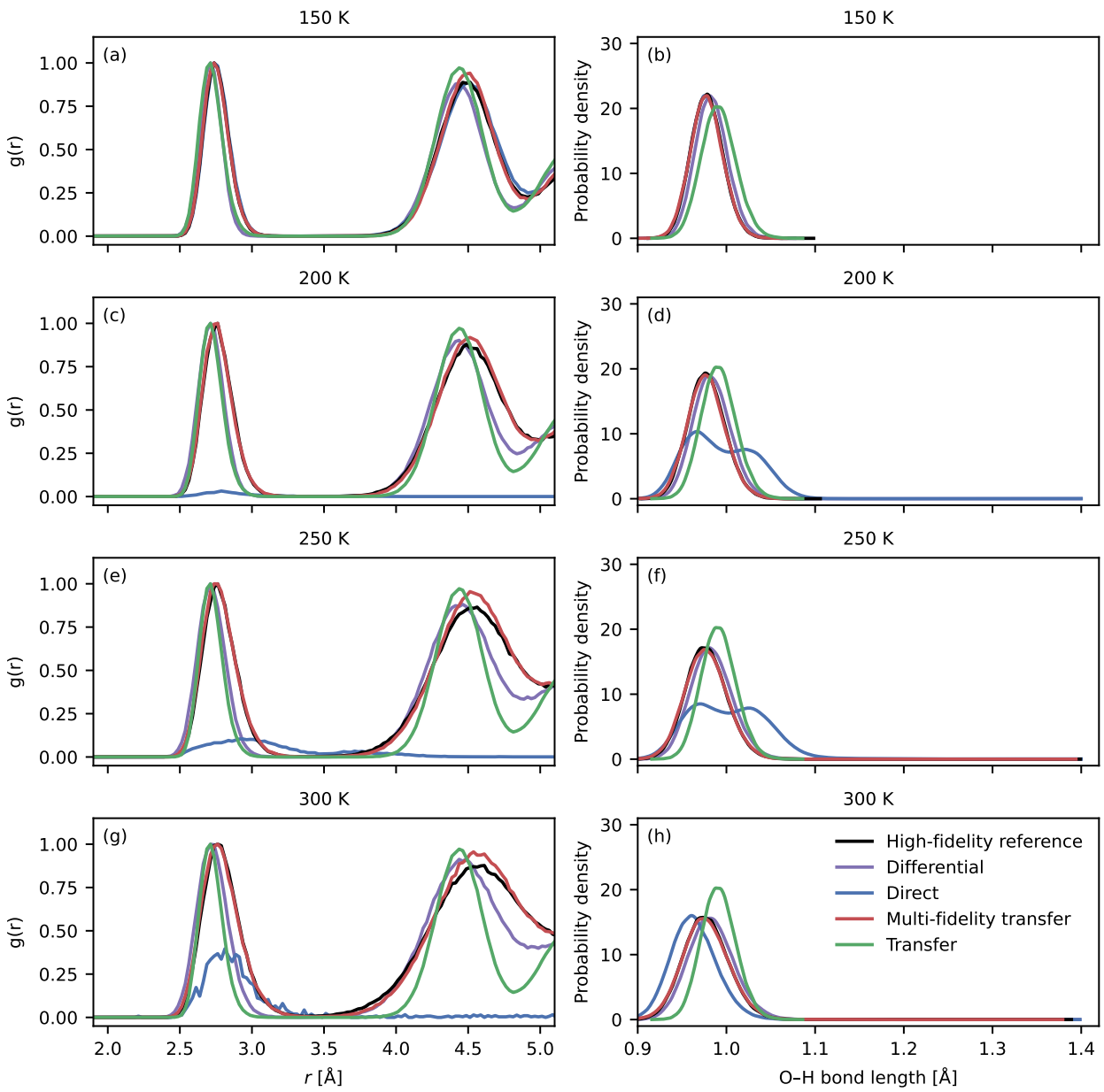

FIG. S9. O-O pair correlation functions (a,c,e,g) and O-H bond-length distributions (b,d,f,h) of NPT trajectories for the temperature extrapolation simulations.

REFERENCES

[1]G. Kresse and J. Furthmüller, Computational Materials Science **6**, 15 (1996).

[2]G. Kresse and J. Furthmüller, Physical Review B **54**, 11169 (1996).

[3]F. Neese, F. Wennmohs, U. Becker, and C. Riplinger, The Journal of Chemical Physics **152**, 224108 (2020).