Accelerating the Training and Improving the Reliability of Machine-Learned Interatomic Potentials for Strongly Anharmonic Materials through Active Learning

Kisung Kang,[1] Thomas A. R. Purcell,[1] Christian Carbogno,[1,*] and Matthias Scheffler[1]

[1] *The NOMAD Laboratory at the FHI of the Max-Planck-Gesellschaft, Faradayweg 4-6, 14195 Berlin, Germany*

(Dated: September 19, 2024)

Molecular dynamics (MD) employing machine-learned interatomic potentials (MLIPs) serve as an efficient, urgently needed complement to *ab initio* molecular dynamics (aiMD). By training these potentials on data generated from *ab initio* methods, their averaged predictions can exhibit comparable performance to *ab initio* methods at a fraction of the cost. However, insufficient training sets might lead to an improper description of the dynamics in strongly anharmonic materials, because critical effects might be overlooked in relevant cases, or only incorrectly captured, or hallucinated by the MLIP when they are not actually present. In this work, we show that an active learning scheme that combines MD with MLIPs (MLIP-MD) and uncertainty estimates can avoid such problematic predictions. In short, efficient MLIP-MD is used to explore configuration space quickly, whereby an acquisition function based on uncertainty estimates and on energetic viability is employed to maximize the value of the newly generated data and to focus on the most unfamiliar but reasonably accessible regions of phase space. To verify our methodology, we screen over 112 materials and identify 10 examples experiencing the aforementioned problems. Using CuI and $AgGaSe_2$ as archetypes for these problematic materials, we discuss the physical implications for strongly anharmonic effects and demonstrate how the developed active learning scheme can address these issues.

I. INTRODUCTION

Machine-learned interatomic potentials (MLIPs) have an immense promise to accelerate molecular dynamics (MD) simulations since, in principle, they provide an accuracy nearing that of *ab initio* calculations at a fraction of the cost [1–4]. In material science, important applications for instance include thermal transport [5–14], the dynamics of amorphous structures [15–18], and ionic diffusion [19–22]. Recent works in this field have focused on improving these potentials by either decreasing their cost (faster running) or increasing their reliability for materials science applications (more reliable training). Further accelerating the potentials for large-scale applications [23, 24] is necessary given that they are –albeit faster than first-principles approaches– still considerably more costly than traditional force fields. For the latter case, this includes improvements in the MLIPs such as using the Euclidean group equivariant neural networks (e3nn) [25–28], long-range interacting physics [29–33], or an enormously augmented amount of training data [34].

The key aspects of MLIP applications are the training process and the creation and selection of the data used for the training. Given that data production –often performed via *ab initio* molecular dynamics (aiMD)– can easily become computationally limiting, it is desirable to avoid redundancy, i.e., the creation of additional data for areas that are already well covered in the training set. Concurrently, it is pivotal, but impossible to guarantee, that the training data appropriately covers the relevant configurational space that will later be explored in the MD simulations. This reflects the finding that MLIPs are most reliable in those areas for which enough training data is provided [25, 35–38]. In machine learning, this is phrased that the training data need to be independent and identically distributed (iid). This is hard, if not impossible, to know. Therefore, the active-learning scheme explained in the following paragraphs is critical.

To streamline training data production, several active-learning ($\mathcal{AL}$) approaches have been proposed in recent years. These have been devised to address the aforementioned challenges associated with domain applicability, leading to a more effective and faster process of MLIP training [39, 40]. The key idea of $\mathcal{AL}$ is an iterative training of MLIP models by explicitly augmenting the training set with "unfamiliar" data, to achieve uniform reliability across the configurational space and avoid redundant learning for the well-trained areas. Thus, the *exploration* of the configurational space and *sampling* of "unfamiliar" data are significant steps in the $\mathcal{AL}$ scheme. To cope with the issue concerning exploration coverage, the MD simulation employing an efficient MLIP model (MLIP-MD) has been recently introduced to complement expensive aiMD [41–49]. It enables rapid exploration of vast spaces, leading to ample coverage of configurational space. Furthermore, various methods for sampling these configurations as unfamiliar have been developed using metrics such as the similarity of the atomic environment [50], a density-based hierarchical clustering [18], and uncertainty estimates of MLIP models [13, 41–46, 51]. The combination of exploration and data-sampling methods allows for the retraining of MLIP models with unfamiliar data through on-the-fly [18, 50] or iterative procedures [41–49]. In addition, novel exploration methods with uncertainty-biased dynamics have appeared in recent years to expedite the exploration of regions with high uncertainty [47–49].

The power of the described $\mathcal{AL}$ approaches in accel-

[*] carbogno@fhi-berlin.mpg.de

erating the training of more accurate MLIPs has been demonstrated for several applications. To this end, it has been shown that such $\mathcal{AL}$ improved MLIPs and typically achieved a better description of microscopic quantities, e.g., mean absolute errors (MAE) of total energies, forces, stresses, etc. [34, 52–55] compared to *ab initio* reference data. In turn, predictions of thermodynamic equilibrium properties, e.g., temperature and pressure-dependent elastic constants, bulk moduli, phonon dispersions, radial-distribution functions, etc. also improve. [18, 24, 41–44, 46–48, 50].

In principle, $\mathcal{AL}$ is expected to improve the prediction of thermodynamic equilibrium and non-equilibrium properties. For instance, the key aspect of transport coefficient calculations is how well MLIPs capture anharmonic lattice dynamics. As described previously, MLIPs well describe the equilibrium lattice vibrations around the ground-state position, offering the long-term dynamics of the system and its memory. From this, transport properties in most cases can be well determined by the time-autocorrelation of the respective fluxes in thermodynamic equilibrium as formulated in the fluctuation-dissipation theorem viz. the Green-Kubo formalism [56–58]. However, in practice, materials undergo rare events that may disruptively impact phase transitions, local (phase) changes, and transport phenomena. As examples of such non-equilibrium cases, it has been recently shown that the spontaneous formation of defects in CuI and of phase transition precursors in $KCaF_3$ are important dynamical phenomena that induce strong anharmonic behaviors in materials [59] resulting, e.g., in a reduction of the thermal conductivity of CuI by a factor of 3.5. An MLIP must be able to reproduce these anharmonic effects, but the ability of the $\mathcal{AL}$ addressing rare events is still elusive.

Naturally, one would assume that $\mathcal{AL}$ schemes would also improve predictions with respect to such rare events, as demonstrated, e.g., for bond-breaking events in simple molecules. [60] However, a systematic quantification of the benefits of $\mathcal{AL}$ for complex materials has, so far, remained elusive. In principle, a systematic comparison of *ab initio* and MLIP-MD calculations for transport calculations would be able to shed light on these questions. However, this would require prohibitive numerical efforts for the first-principles MD to be able to reach the necessary statistics and time- and length scales relevant to strongly anharmonic events. Similarly, one cannot explicitly target and monitor the phase space region associated with strongly anharmonic events since their occurrence and the associated path on the potential-energy surface are usually unknown *a priori*. This is further aggravated by the fact that such events are typically short-lived, so that their influence on average properties like the MAE of microscopic quantities or predictions on thermodynamic equilibrium properties is small and hence hardly stands out against statistical noise.

As demonstrated in this work, the problematic prediction of MLIPs regarding strongly anharmonic effects can happen in various contexts:

- Rare events may be missing in the training data.
- Rare events may be present but with insufficient information about probabilities and lifetimes.
- Rare events may be present but smoothened away by regularization.
- MLIPs may well exhibit fake rare vents.

Alarmingly, such problems can be easily overlooked, since neither checking average predictions for testing data, such as mean absolute errors or R^2, nor inspecting close-to-equilibrium properties such as quasi-harmonic phonons allows to reliably detect these problems.

In this study, we build on existing $\mathcal{AL}$ ideas and adapt them for the description of strongly anharmonic materials by combining an ensemble uncertainty metric with thermodynamically meaningful acquisition functions. Our benchmark on 112 materials for which extensive aiMD data is available from literature reveals that 10 out of these 112 materials require an $\mathcal{AL}$ approach to achieve a physically correct description of the PES. We carefully analyze the failure of standard MLIP training for two representative examples. In Sec. IV A, we discuss CuI, the anharmonicity of which is severely underestimated procedures and in Sec. IV B we discuss $AgGaSe_2$, the anharmonicity of which is strongly overestimated by standard MLIP training procedures. Furthermore, we analyze how the proposed $\mathcal{AL}$ scheme rectifies these failures and propose best practices to achieve stable anharmonic MLIPs with $\mathcal{AL}$. Eventually, Sec. IV C discusses the physical implication of the proposed approach for actual material-science predictions and shows that only the usage of $\mathcal{AL}$ guarantees to correctly identify the correct transport regime in actual MLIP-MD simulations.

II. METHODOLOGY

In this section, we will first give a concise overview of the different established $\mathcal{AL}$ strategies employed in literature for the training of MLIPs in Sec. II A. In the following, Sec. II B explains how we build on these concepts and adapt them to specifically target and benchmark strongly anharmonic materials.

A. State-of-the-Art $\mathcal{AL}$ strategies

1. *The standard $\mathcal{AL}$ workflow*

Here we summarize a step-by-step description of the five steps in a typical $\mathcal{AL}$ workflow, as illustrated in Fig. 1.

1. **Initialization**: An initial training set of configurations for a material is generated by either a

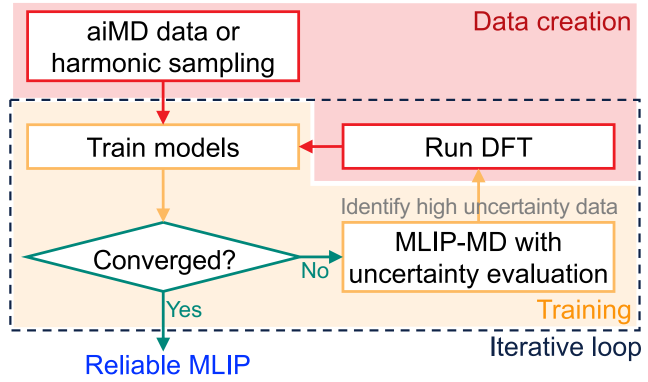
aiMD data or harmonic sampling
Data creation
Train models
Run DFT
Identify high uncertainty data
Converged?
No
MLIP-MD with uncertainty evaluation
Yes
Training
Iterative loop
Reliable MLIP

FIG. 1. A workflow plot depicting the active learning ($\mathcal{AL}$) scheme.

short aiMD trajectory [14], stochastic sampling of phonon eigenvectors (harmonic sampling) [61, 62], or random displacements of randomly chosen atoms [43, 63]. Because the $\mathcal{AL}$ process will improve their reliability anyway, even using pretrained MLIPs, such as recent universal MLIPs [27, 52, 64] (See Fig. S1 in the supplementary material (SM) [65] for more details), will work, unless their descriptions are incompatible with MD simulations.

2. **(Re)Training**: Initial(augmented) training data are utilized to (re)train MLIP models.

3. **Convergence test**: Once trained, the prediction quality of the MLIP is evaluated using a test set obtained from unseen parts of aiMD trajectories and error metrics such as the mean absolute errors (MAE). If the desired reliability is achieved, the iterative $\mathcal{AL}$ terminates, otherwise it proceeds to the next step. However, it is not able to ensure that a rare event pops up later. In addition, this reliability during the $\mathcal{AL}$ does not hold for temperatures higher than the trained temperature.

4. **Exploration and data-sampling**: The most critical step of the $\mathcal{AL}$ is the exploration of configurational space to find unfamiliar regions. Three popular methods to perform the exploration are aiMD [18, 50], explorative MLIP-MD [41–46], or uncertainty-biased MD [47, 48]. For those approaches using MLIP, models trained with the data from previous iterations data are used in this step. Popular choices to sample each new configuration as familiar or not include the evaluation of extrapolation grades [66], the analysis of the structural similarity of samples [50], the examination of correlations among samples [18], and the uncertainty estimates of MLIP predictions [41–48].

5. **Data acquisition**: Full *ab initio* calculations are performed on the snapshots sampled as unfamiliar to obtain their genuine force, energy, and stresses. This data is then added to the training set for the subsequent retraining. The workflow then goes back to the step 2.

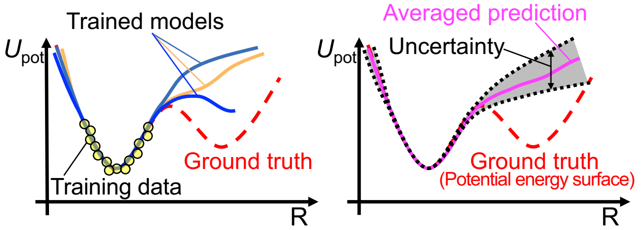

FIG. 2. A schematic plot illustrating the definition of an ensemble uncertainty estimate in terms of the potential energy.

Steps 2-5 form a closed loop. Whenever MLIP-MD, even during the practical applications, shows a high uncertainty estimate, the iterative $\mathcal{AL}$ scheme should resume. The major difference among various $\mathcal{AL}$ schemes originates from the choice of exploration and data-sampling methods depending on their target systems. For the exploration approaches, aiMD snapshots can be directly used as training data, whereas explorative MLIP-MD and uncertainty-based MD can effectively travel to unseen areas with their efficient implementations. The choice of the exploration method is also influenced by the trade-off between cost and accuracy: aiMD trajectories are computationally expensive, but MLIP-MD may explore physically unfavorable parts of configurational space. The choice of data-sampling methods also depends on training environments and target properties, and more details are discussed in the following section.

2. *Ensemble uncertainty estimates*

Selecting an efficient and reliable *data-sampling* method is critical to the performance of the $\mathcal{AL}$ scheme as that determines how effectively it identifies unfamiliar data among the substantial volume of new data generated. In this study, the uncertainty estimates of the MLIP predictions serve as a qualitative signal for when the MD trajectory leaves the well-trained regions. Although concerns persist regarding its quantitative usage in extrapolation regions [67], uncertainty estimates still hold remarkable value as a qualitative indicator. Various uncertainty techniques have been developed such as uncertainty estimates from the probabilistic framework of Bayesian linear regression [13, 51] or an internal principled uncertainty quantification mechanism to evaluate the prediction uncertainty [44, 45]. These approaches can quickly assess the uncertainty estimates and do not require training multiple models. However, this approach can only be applicable to the MLIP models, which can internally provide a probabilistic model, e.g., the Gaussian mixture model [46].

Instead, this section introduces an ensemble uncertainty estimate in detail due to its simple implementation and wide applicability across various MLIP architectures. The ensemble uncertainty estimate is assessed as a standard deviation of predictions from multiple MLIP models that are trained using different training sets (subsampling) and different initial structures and parameters of the neural networks (deep ensemble), as depicted in Fig. 2. The resulting MLIP models consistently predict reliable atomic motions within well-trained areas but begin to diverge beyond these areas, as illustrated in Fig. 2. Despite the fact that such uncertainty estimates do not allow for a quantitatively exact prediction of the error and, hence, of the ground-truth potential-energy surface [67–69], higher than average uncertainty estimates can be used to qualitatively identify unfamiliar regions so to steer further sampling for retraining MLIP models.

Here we describe how to obtain ensemble uncertainties of the potential energy and forces employed as the data-sampling method in the $\mathcal{AL}$ scheme. The uncertainty estimate for a target property (X) of a MLIP-MD snapshot is determined by calculating the standard deviation of predicted target properties using the following equation:

$$\mathrm{UCE}_X = \sqrt{\frac{1}{N}\sum_{I}^{N}(X_I - \mu_X)^2}. \tag{1}$$

Here, X_I represents the target energy predicted by the I^{th} MLIP model, and μ_X corresponds to the mean value of target properties from the ensemble of all N different MLIP models. Thus, the potential-energy uncertainty estimate (UCE_U) is obtained from the standard deviation with respect to a predicted mean value of potential energies (U). The maximum of the predicted forces uncertainty estimates ($\mathrm{UCE}_F^{\max}$) is evaluated as the largest one among the uncertainties estimate $\left(\mathrm{UCE}_F^i\right)$ of each force for all different i^{th} atoms in the material ($\mathbf{F}_I^i$).

B. Adaption of $\mathcal{AL}$ to strongly anharmonic systems

1. Exploration and data-sampling methods

Our $\mathcal{AL}$ scheme is designed to iteratively retrain MLIP models with unfamiliar strongly anharmonic events selectively sampled during configurational space exploration. Since such events can occur infrequently during MD explorations, aiMD might not be able to reach the relevant time and length scales to sample them properly or capture them at all. Instead, explorative MLIP-MD is used as an efficient complement to aiMD, enabling simulations to reach the needed time and length scales to observe also rare events and estimate the model's uncertainty. In this study, we implement MLIP-MD based on the mean value of the forces from MLIP models, resulting in one MLIP-MD trajectory with target uncertainty estimates.

Ensemble uncertainty estimates are selected as the data-sampling method in the present study due to its generalized application regardless of MLIP types. Ensemble uncertainties of total energy and maximum atomic force in Sec. II A 2 are examined to verify its ability to sense the unfamiliar strongly anharmonic events. We also extend this approach to the degree of anharmonicity because this quantity can effectively identify the anharmonic rare event using a single value. The degree of anharmonicity (A) at the MD simulation time (t) is defined based on a concept devised by Knoop *et al.* as follows [70].

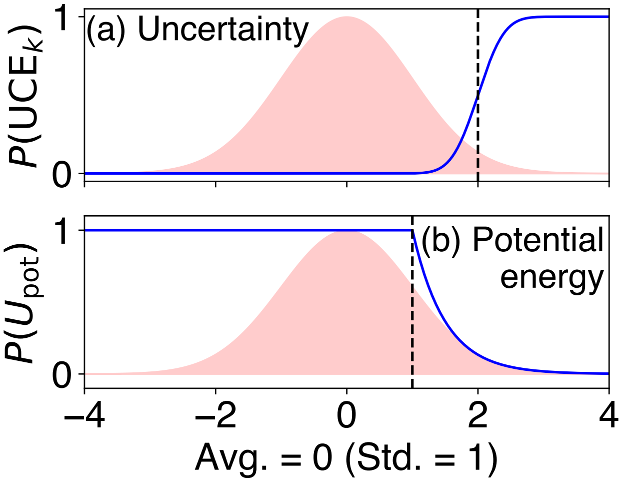

FIG. 3. (Color online.) Sampling probability distribution (blue solid lines) regarding (a) uncertainty estimates and (b) potential energy in relation to a reference average set at 0, with each integer representing a standard deviation of 1. The pink bell-shaped plots represent schematic normal distributions.

$$A(t) = \sqrt{\frac{\sum_{i,\alpha}\left(F_{\mathrm{Aha}}^{i,\alpha}(t)\right)^2}{\sum_{i,\alpha}\left(F^{i,\alpha}(t)\right)^2}}, \tag{2}$$

where $F^{i,\alpha}(t)$ represents the α ($= x, y, z$) component of the force vector on the i-th atom at the MD time (t). $F_{\mathrm{Aha}}^{i,\alpha}(t)$ means the α component of the anharmonic atomic force vector on the i-th atom at the time (t), evaluated as $F_{\mathrm{Aha}}^{i,\alpha}(t) = F^{i,\alpha}(t) - F_{\mathrm{Ha}}^{i,\alpha}(t)$, where $F_{\mathrm{Ha}}^{i,\alpha}(t)$ represents the α component of the force that would be obtained at the same geometry but using a harmonic (parabolic) potential. Accordingly, the uncertainty of the degree of anharmonicity of a configuration (UCE_A) is defined as the standard deviation with respect to the degree of anharmonicity (A) via Eq. 1.

2. Sampling probability

For retraining, it is necessary to balance between maximizing the uncertainty estimate of the targeted regions of phase-space and ensuring those regions are at least accessible. To this end, we propose a sampling probability that serves as an acquisition function.

First, data-sampling new training data for subsequent rounds of MLIP training necessitates the definition of high uncertainty estimates. The uncertainty estimate fluctuates across different $\mathcal{AL}$ steps due to the changes in training data and training areas. Hence, high uncertainty estimates in our $\mathcal{AL}$ scheme mean a *relatively higher* uncertainty estimate compared to other data points. At the beginning of each $\mathcal{AL}$ step, we evaluate uncertainties in 300 testing data randomly selected from aiMD trajectories with 12001 snapshots (not included in the training set), recording their mean value and standard deviation. These quantities serve as reference points to identify unfamiliar MD snapshots from preliminary MLIP-MD. Then, we set a soft selection criterion by building a sampling probability function in terms of uncertainty estimates, shown as $P(\mathrm{UCE}_k)$, creating a smooth curve, a cumulative normal distribution function, at the criterion limit at two standard deviations away from the average (dashed line in Fig. 3 (a)). The mathematical form of this criterion, $P(\mathrm{UCE}_k)$, is described below:

$$P(\mathrm{UCE}_k) = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{\mathrm{UCE}_k - \mu_{\mathrm{UCE}_k^{\mathrm{test}}} - 2\sigma_{\mathrm{UCE}_k^{\mathrm{test}}}}{\sigma_{\mathrm{UCE}_k^{\mathrm{test}}}\sqrt{0.2}}\right)\right]. \tag{3}$$

Here, UCE_k ($= \mathrm{UCE}_U, \mathrm{UCE}_F^{\max}$, or UCE_A) signifies the uncertainty estimate at the current step. $\mu_{\mathrm{UCE}_k^{\mathrm{test}}}$ and $\sigma_{\mathrm{UCE}_k^{\mathrm{test}}}$ mean the average and standard deviation of uncertainties in testing data, respectively. Fig. 3 (a) illustrates that $\mathcal{AL}$ mostly collects data with uncertainty estimates larger than two times the standard deviation away from the reference average of testing data. The probability function shape is adopted from a cumulative normal distribution function, yielding a soft limit rather than a rigid threshold.

Second, we considered the potential energy of sampled data to prevent sampling unphysical configurations that have enormously high energy. This criterion is implemented by a different probability function, similar to the one for uncertainty estimates. The sampling probability function regarding the potential energy, $P(U)$, is derived from a modified probability distribution of the canonical ensemble by decaying the probability from a criterion limit at one standard deviation away from the average (dashed line in Fig. 3 (b)). In the mathematical expression,

$$P(U) = \begin{cases} 1 & \text{if } U \leq \mu_{U^{\mathrm{test}}} + \sigma_{U^{\mathrm{test}}}, \\ \exp\left(\frac{\ln 0.2}{0.8}\cdot\frac{U-\mu_{U^{\mathrm{test}}}-\sigma_{U^{\mathrm{test}}}}{\sigma_{U^{\mathrm{test}}}}\right) & \\ & \text{if } U > \mu_{U^{\mathrm{test}}} + \sigma_{U^{\mathrm{test}}}, \end{cases} \tag{4}$$

where U represents the potential energy at the current step of the MD simulation. $\mu_{U^{\mathrm{test}}}$ and $\sigma_{U^{\mathrm{test}}}$ mean the average and standard deviation of potential energies predicted by MLIP models in testing data. The coefficients are determined to ensure that $P(U)$ starts decaying from the criterion limit, i.e., at one standard deviation away from the average. Fig. 3 (b) displays that this criterion starts to exclude data with potential energy beyond one standard deviation away from the reference average. Finally, the actual sampling process is executed based on the combined criteria of both probability functions, concluding $P = P(\mathrm{UCE}_k)\cdot P(U)$.

Accordingly, the proposed $\mathcal{AL}$ scheme augments state-of-the-art $\mathcal{AL}$ workflows with a soft sampling criterion that ensures that strongly anharmonic effects are correctly captured, even if they occur infrequently.

III. COMPUTATIONAL DETAILS

For comprehensive analysis and comparison, we applied our $\mathcal{AL}$ workflow to 112 different bulk materials from the NOMAD aiMD repository, which contains extended MD simulations up to 60 ps for thermal transport studies previously conducted by Knoop *et al.* [59, 71]. Since the consistent training of MLIP models should be ensured to attain their reliable predictions, the sampled data during the $\mathcal{AL}$ iterations are calculated employing exactly the same computational frameworks. In detail, all DFT calculations were executed by the `FHI-aims` code packages employing an all-electron formalism [72, 73]. The utilized supercells with 160-256 atoms are consistently chosen to describe the dynamics of 112 materials during the $\mathcal{AL}$ workflow. The identical **k**-point sampling densities were applied to integrate the Brillouin zone. The basis sets are set to *light* default in the `FHI-aims`, and PBEsol [74] is selected as exchange-correlation functional. All quasi-harmonic and anharmonic vibrational properties of of CuI and $\mathrm{AgGaSe_2}$ are computed using `FHI-vibes` [75]; whereby the perturbative formalism implemented in `phonopy` [76, 77] and `phono3py` [76, 78] is used for the harmonic phonon dispersions and phonon lifetimes. For more details, we refer to the original aiMD paper [59] and its data repository [71].

For the MLIP architecture, we adopt a recent graph neural network potential with a message-passing scheme implemented by NequIP version 0.5.6 [25]. Throughout convergence tests for 300 training data, the values of hyperparameters are carefully determined through the convergence test for potential energy and forces predictions and they are consistently applied for all $\mathcal{AL}$ benchmarks. Our hyperparameters are a local cutoff radius for the atomic environment (`rmax`) of 5 Å, a maximum rotation order for the neural network features (`lmax`) of 3, and a feature multiplicity of 32, respectively. Four layers of the neural network are employed, and eight basis functions are used in the radial basis. To balance the prediction accuracy, the loss function ratio between potential energy per atom (`PerAtomMSELoss`) and atomic forces is chosen to be 1:1. The `float64` precision is adopted to maintain the high accuracy. All MLIP training with NequIP is implemented via a single GPU core from the NVIDIA Tesla A100. A total of six MLIP models are utilized to evaluate the ensemble uncertainty by training three different MLIP models using *subsampled* training datasets with two distinct random initializations with a *deep ensemble* approach, which provides the optimal performance by balancing more converged uncertainty estimates and the computational costs for running multiple MLIP models. In the initialization, each MLIP model is trained with 25 training data and 5 validation data from the aiMD trajectory. Sequentially, the DFT results of 30 new MLIP-MD data from the exploration and data-sampling step are added to the previous training data with a consistent ratio of training and validation.

Our iterative $\mathcal{AL}$ workflow is automatically implemented by a Python code, `ALmoMD` [79], that interfaces a DFT code (`FHI-aims` [72, 73]) and a MLIP code (`NequIP` [25]) via the Atomic Simulation Environment (`ASE`) [80] and the `FHI-vibes` [75]. In this study, the $\mathcal{AL}$ workflow is conducted based on the initial settings for NequIP and MLIP-MD as described above. In practical applications, human intervention may be possible to tune the loss function ratio of NequIP or adjust the sampling criteria. However, this study keeps the initial setting without any intervention during the $\mathcal{AL}$. The MLIP-MD in the $\mathcal{AL}$ is implemented for the NVT ensemble employing the Langevin thermostat to explore the configurational space with a target temperature of 300 K using the `ASE` library [80]. The friction parameter and timestep of MLIP-MD are set to 0.03 and 5 fs, respectively, to follow the original MD setting in the referenced repository [71]. The phonon lifetimes from the MD trajectory are extracted by the workflow designed in the `FHI-vibes` [75]. Its MD simulation is implemented based on the NVE ensemble, and the detailed MD parameters are set the same as the one implemented in the aiMD data repository [59, 71]. Nudged elastic band (NEB) calculations [81–83] are performed with the implementation available within `ASE` [80] to extract the potential energy surface between the ground state structure and the structure with a defect.

IV. RESULTS AND DISCUSSION

To substantiate our argument regarding erroneous MLIP predictions associated with rare events and test our concept and the proposed $\mathcal{AL}$ approach, we first investigated 112 bulk materials with MLIP training via random sampling for the aiMD trajectory, i.e., without $\mathcal{AL}$. Fig. 4 shows the mean absolute error results of energy and atomic force predictions from MLIP models using the ordinary training approach. All results for 112 bulk materials show high-accuracy predictions for the energy ($< 1\,\mathrm{meV/atom}$) and atomic forces ($< 10^{-2}\,\mathrm{eV/Å}$) of testing data, which was not seen during training. At first, this erroneously suggests that the trained MLIP models will provide reliable predictions for all these materials.

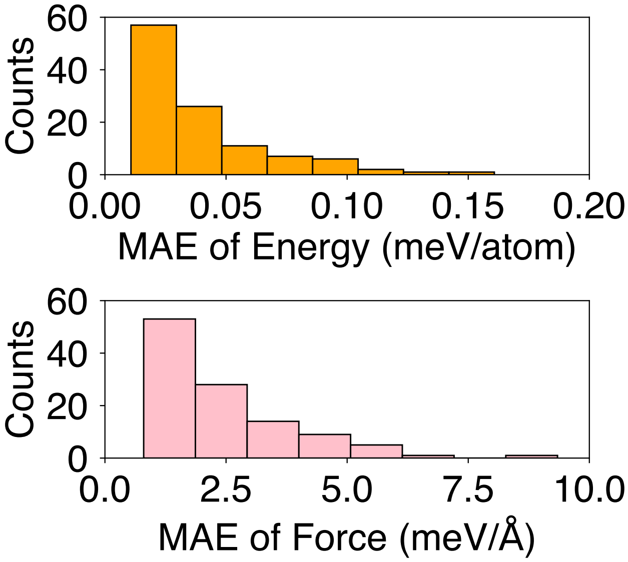

FIG. 4. Mean absolute errors (MAE) statistics of energy (top) and atomic force (bottom) predicted by MLIP models from standard training metrics for 112 bulk materials. The bin width is determined based on Sturges's rule [84].

However, at the considered temperature, 10 of these materials (≈ 9%) required multiple $\mathcal{AL}$ iterations before reaching convergence, despite the fact that the MAE of these material were comparable to all other materials in the initial training. A deeper analysis revealed that the MLIPs for these 10 materials featured erroneous predictions before AL in extrapolated regions not covered by the training set, as illustrated in Fig. 5. Here, we can distinguish between two distinct cases: The MLIP model may either overlook the presence of a metastable state (Fig. 5 (a)) or predict a false metastable state (Fig. 5 (b)) that is not present in actual first-principles calculations. Both scenarios are associated with *exploration* and *sampling* during the standard training method. A low visiting frequency for high-energy training boundary areas yields a low sampling of configurations in these regions. Obviously, this could in principle be mitigated running extensively long aiMD simulations, which come with impractically increased computational effort. But even in this case, human inspection would be needed to ensure that the respective phase-space regions associated with the rare event are appropriately covered in test and training sets. In the following, we discuss the two scenarios using representative examples and show that the proposed $\mathcal{AL}$ is able to correct these erroneous predictions in an automatic fashion with modest computational overhead. Furthermore, materials that do not suffer under the described problems are not affected by the $\mathcal{AL}$ approach, as exemplarily shown for $KCaF_3$ in the SM [65] (See Fig. S2 and S3).

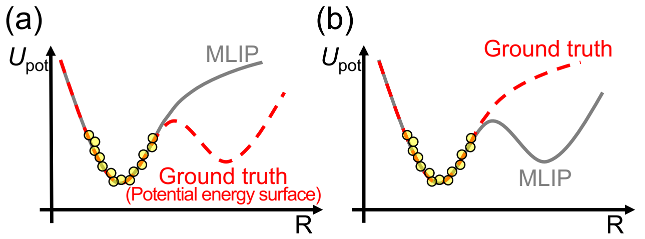

FIG. 5. Schematic plots representing problematic scenarios: (a) the absence of metastable states in MLIP and (b) the prediction of erroneous metastable states in MLIP. The yellow circles represent the training data of MLIP models.

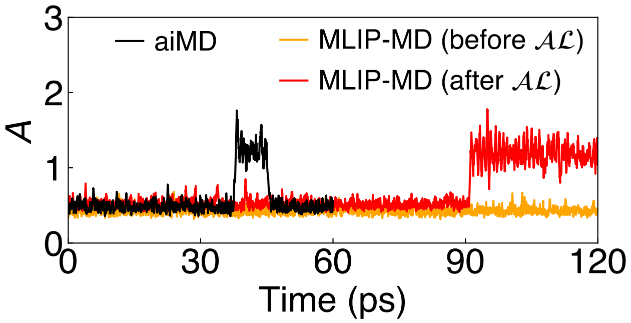

FIG. 6. The degree of anharmonicity (A) of MD trajectories of CuI from aiMD (Ref. [59]) and MD using MLIPs before/after applying the $\mathcal{AL}$ scheme.

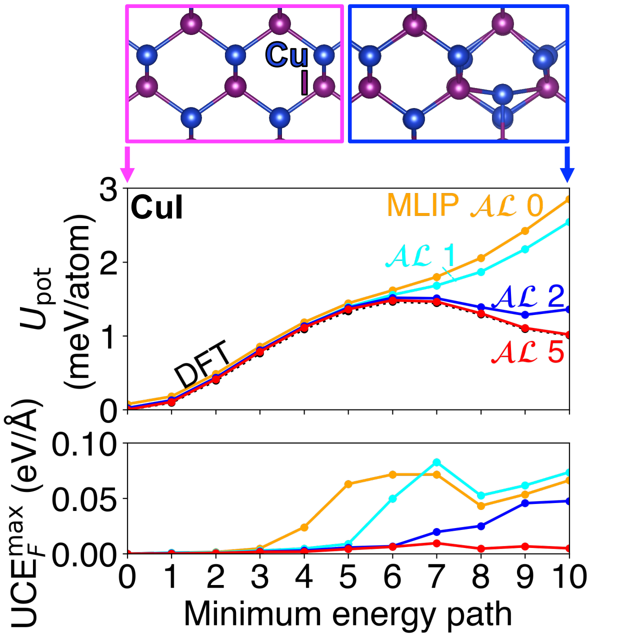

FIG. 7. The PES and corresponding maximum uncertainty estimates of atomic forces (UCE$_F^{\max}$) predicted by MLIP from $\mathcal{AL}$ using UCE$_F^{\max}$ as a function of the gradual structural evolution of CuI, transitioning from its ground state structure (top left outset structure) to a defect-bearing configuration (top right outset structure) via NEB calculations. The black line shows the DFT ground truth and the colored points with solid lines represent PES results predicted by MLIP models. The $\mathcal{AL}$ N means the PES snapshots from MLIP models obtained after the N-th step of the $\mathcal{AL}$ iterations.

A. CuI: The Case of Missing Minima

For the first scenario, in which a metastable state associated with a strongly anharmonic effect is overlooked by the MLIP, we discuss the case of copper iodide (CuI). Already in a previous aiMD study, it was reported that CuI features spontaneous defect creation at 300 K, which has a significant impact on the resulting thermal conductivity [59]. However, defect creation occurs infrequently on aiMD time-scales, e.g., it is observed –when at all– only after more than 35 ps in Ref. [59] as illustrated as a black solid line in Fig. 6 and can happen at different any point for other trajectories. When the degree of anharmonicity jumps from 0.5 to 1.2, it implies the occurrence of the defect creation. This defect creation causes a strongly anharmonic effect, which impacts the transport properties; e.g., the phonon lifetime is drastically reduced by this effect. Accordingly, an MLIP training procedure that only focuses, e.g., on then first 10 ps of the trajectory or fitting by regularization, is prone to miss this mechanism, as discussed below. Qualitatively, this strongly anharmonic effect leads to a breakdown of the phonon picture [85, 86] as quantified by the Ioffe-Regel limit [87]. When a phonon lifetime becomes shorter than the oscillation period, the vibrational quasi-particle becomes invalid, as for instance observed for in the case of spontaneous defect creation [59]. Conversely, the quasi-particle picture holds when the lifetimes are longer and transport theories based on the phonon picture can be used to evaluate heat transport. Therefore, it is important to verify whether MLIP can reproduce such physics associated with strong anharmonicity.

Although, in the present case, we *a priori* know about the existence of the metastable defect state, we do not exploit this information for the initial training of the MLIP, so to mimic the typical application case in which little is known about the different processes that might be active in a material at the beginning. Accordingly, we train the initial MLIP on one of the aiMD trajectories, which does not contain any explicit defect formation. Accordingly, the resulting MLIP fails to predict a metastable state, as checked by running 30 independent MLIP-MD before $\mathcal{AL}$ for 1 ns. Energetically, this is rationalized in Fig. 7, in which the minimum-energy path between a pristine structure and a metastable defect is plotted both for the ground-truth DFT data and the different iterations N of the MLIP training labeled with $\mathcal{AL}$ N. While the initial MLIP model ($\mathcal{AL}$ 0) erroneously misses the presence of a metastable state, subsequent $\mathcal{AL}$ iterations incorporate information about the respective phase-space region in the training data, so to correctly reproduce the DFT PES at the fifth iteration, at which convergence is achieved.

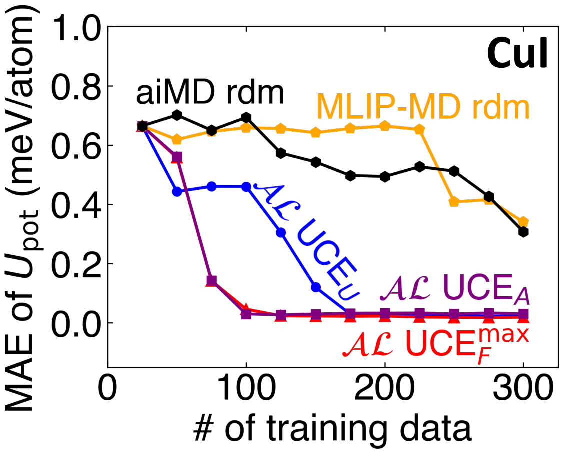

FIG. 8. MAE of potential energies in testing data from aiMD trajectory of CuI as a function of the number of training data. The testing data also includes configurations with defects that only occur rarely in the dynamics. `aiMD rdm` (black hexagon) involves MLIP training with random sampling from aiMD trajectories. `MLIP-MD rdm` (orange pentagon) represents MLIP training with random sampling from MD trajectories using MLIP at each step. $\mathcal{AL}$ UCE$_U$ (blue circle), $\mathcal{AL}$ UCE$_F^{\mathrm{max}}$ (red triangle), and $\mathcal{AL}$ UCE$_A$ (purple square) exhibit MLIP training with the active learning approach utilizing energy uncertainty estimates, maximum uncertainty estimates of atomic forces, and uncertainty estimates in the degree of anharmonicity, respectively.

This is further substantiated by the fact that also the respective uncertainty estimates in the maximum forces shows a well-balanced behavior over the complete minimum energy path, cf. the bottom panel in Fig 7. In the MD trajectory of CuI, the defect creation missed in MLIP-MD before $\mathcal{AL}$ is now observed in MLIP-MD after $\mathcal{AL}$, as illustrated in Fig. 6. Let us emphasize that no information about the presence of the defect state was fed to the $\mathcal{AL}$ procedure; the observed improvements solely result automatically from the designed acquisition function that iteratively enables sampling high uncertainty regions, even if they appear to be energetically inaccessible at first, see Fig. S3.

The acceleration of MLIP training via the $\mathcal{AL}$ scheme is also observed from the mean absolute error (MAE) check of testing results, as shown in Fig. 8. Fig. 8 depicts how the MAE of potential energies in testing data varies when we augment training data in five different methods. `aiMD rdm` stands for random sampling of training data from aiMD trajectory, while `MLIP-MD rdm` follows the $\mathcal{AL}$ workflow using random sampling instead of data-sampling via uncertainty estimates. The actual $\mathcal{AL}$ implementations utilize three different uncertainties (UCE$_U$, UCE$_F^{\mathrm{max}}$, and UCE$_\mathrm{A}$) introduced in Sec. II A 2 and Sec. II B 1. Testing data comprise MD snapshots from aiMD trajectories, including strongly anharmonic events. $\mathcal{AL}$ UCE$_F^{\mathrm{max}}$ and $\mathcal{AL}$ UCE$_A$ reach the convergence first, followed by $\mathcal{AL}$ UCE$_U$. This difference stems from the fact that UCE$_F^{\mathrm{max}}$ has an atomic resolution of uncertainty evaluation and the degree of anharmonicity (A) used in UCE$_A$ is sensitive to defect creation whereas UCE$_U$ utilizes the potential energy into that all atomic information are merged. `aiMD rdm` and `MLIP-MD rdm` trained with up to 300 training data could not get similar reliability and convergence behavior for energy predictions because they could not properly sample rare events from their trajectory. From this, we could draw the lesson that implementing $\mathcal{AL}$ leads to effective MLIP training, even further effective when the data-sampling method has atomic resolution or structural change sensitivity.

B. AgGaSe$_2$: The Case of Fictitious Minima

As a second scenario, we discuss the case of AgGaSe$_2$, for which the initial MLIP incorrectly predicts a metastable state that induces fictitious, strongly anharmonic effects that are not active on the ground-truth DFT PES. The $\mathcal{AL}$ of the other eight materials, including AgGaS$_2$, InNaO$_2$, CsBr, CsCl, LiBr, LiCl, LiI, and Na$_2$Te, are illustrated in SM [65] (See Fig. S6-S15). For this purpose, we train an initial MLIP as described in Sec. III. Let us emphasize that no notable artifacts are observed during the training procedure and that further augmenting the training set with data points from the initial aiMD trajectory does not further improve the initial MLIP, see Fig. S16 in the SM [65]. Subsequent MD runs with this initial MLIP predict the occurrence of strongly anharmonic effects associated with spontaneous defect creation, see Fig. 9, which displays a MLIP-MD example for AgGaSe$_2$ at 300 K. Here, the defect creation is observed at 480 ps, as indicated by a jump in the degree of anharmonicity (A).

To better rationalize the impact of the $\mathcal{AL}$ scheme, we again inspect the minimum-energy path between the pristine and the defective structures, whereby the latter is obtained from the initial MLIP potential. As shown in Fig. 10 the initial MLIP ($\mathcal{AL}$ 0) predicts a wrong, very likely appeared metastable state, which is not at all present in the ground-truth DFT PES, and there it has a very unfavorable energy. Again, the devised $\mathcal{AL}$ scheme iteratively improves on the prediction and corrects the erroneous topology of the PES. These rectifications stem from the sampling of MLIP-MD trajectories traveling beyond the energy barrier during the $\mathcal{AL}$ scheme, featured by the jumps of the degree of anharmonicity in MLIP-MD trajectories illustrated as Fig. S16. Whenever unfamiliar events occur during MLIP-MD, our data-sampling method effectively samples these configurational snapshots as sequential training data, resulting in the improvement of MLIP description. However, since the occurrence of rare events during MLIP-MD is based on chance, corrections of erroneous predictions for such events do not happen in each $\mathcal{AL}$ step, but only in those steps in which such a dynamics is actually observed in Fig. 10. The uncertainty estimates of maximum atomic force (UCE$_F^{\mathrm{max}}$) become smaller after the sixth step but

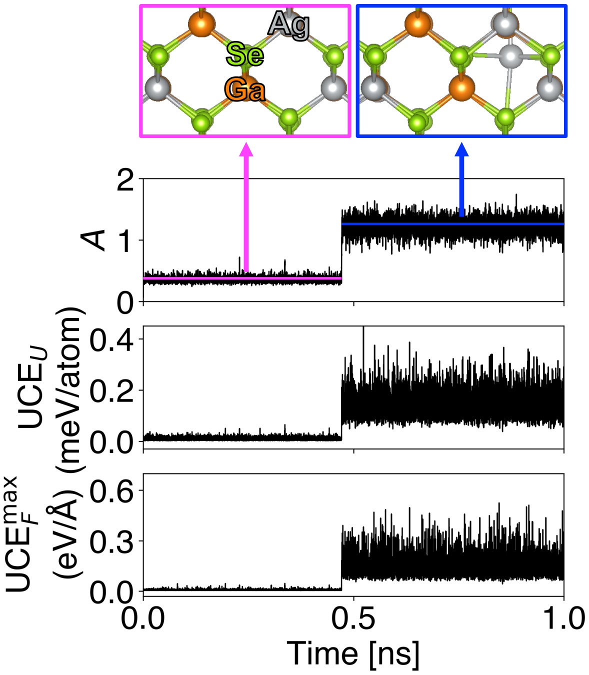

FIG. 9. The MD trajectory using MLIP trained on 300 aiMD data, showing the evolution of anharmonicity degree (A), energy uncertainty estimates (UCE$_U$), and maximum uncertainty estimates of atomic force (UCE$_F^{\max}$) over a 1 ns simulation.

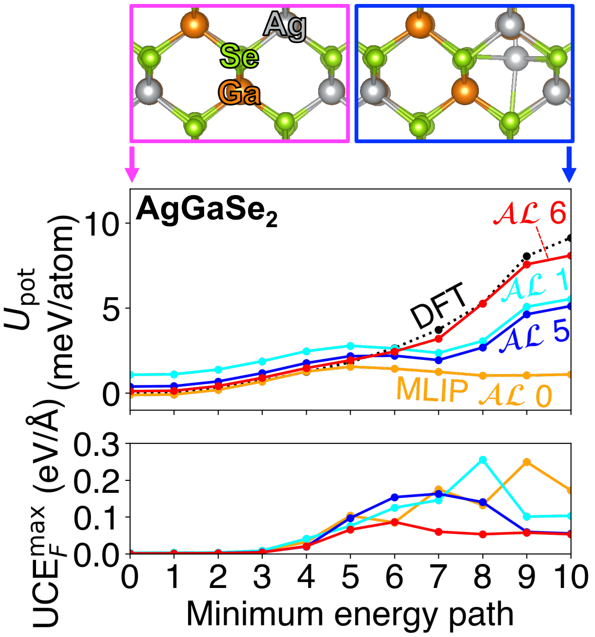

FIG. 10. The PES and corresponding maximum uncertainty estimates of atomic forces (UCE$_F^{\max}$) predicted by MLIP from $\mathcal{AL}$ using UCE$_F^{\max}$ as a function of the gradual structural evolution of AgGaSe$_2$, transitioning from its ground state structure to a defect-bearing configuration via NEB calculations.

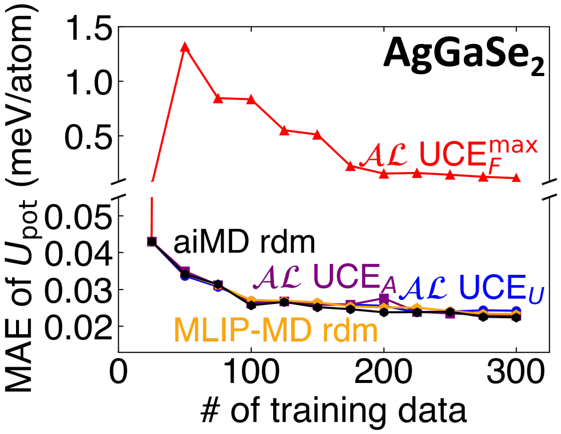

FIG. 11. MAE of potential energies in testing data from aiMD trajectory of AgGaSe$_2$ as a function of the number of training data. `aiMD rdm` (black hexagon) involves MLIP training with random sampling from aiMD trajectories. `MLIP-MD rdm` (orange pentagon) represents MLIP training with random sampling from MD trajectories using MLIP at each step. $\mathcal{AL}$ UCE$_U$ (blue circle), $\mathcal{AL}$ UCE$_F^{\max}$ (red triangle), and $\mathcal{AL}$ UCE$_A$ (purple square) exhibit MLIP training with the active learning approach utilizing energy uncertainty estimates, maximum uncertainty estimates of atomic forces, and uncertainty estimates in the degree of anharmonicity, respectively.

are still not uniform across the PES, implying that there are still some uncertainty estimates in the high-energy region.

Figure 11 illustrates the MAE and its converging behavior for potential energies in testing data with different approaches. `aiMD rdm` and `MLIP-MD rdm` exhibit the typical convergence behavior with excellent accuracy for testing data because MLIP predictions for vibrational motions in structure at equilibrium get improved with an increased number of training data. However, the MLIP models `aiMD rdm` and `MLIP-MD rdm` trained with 300 data still suffer from the occurrence of erroneous, incorrect events in the MLIP-MD simulation (Fig. S18 [65] for the `aiMD rdm` cases). In addition, the $\mathcal{AL}$ using uncertainties in the energy (UCE$_U$) and the degree of anharmonicity (UCE$_A$) exhibits similar converging behavior compared to `aiMD rdm`. This is because $\mathcal{AL}$ UCE$_U$ and $\mathcal{AL}$ UCE$_A$ did not experience any incorrect events during its exploration steps, resulting in no PES corrections in this $\mathcal{AL}$. This implies that the respective MLIP-MD may likely undergo false events, emphasizing the importance of their appearance during MLIP-MD explorations in $\mathcal{AL}$. On the other hand, $\mathcal{AL}$ UCE$_F^{\max}$ have a large jump in the MAE of potential energy. The correction of predictions for the region far from the ground state worsens the prediction for the overall PES. This deterioration stems from poor regressions due to insufficient training points in the far regions. Additional sampling is required in these regions, which is not easily conducted due to their high potential energy. However, this correction is crucial to prevent erroneous events in MLIP-MD, which can seriously affect the dynamic properties of materials, even when happening seldomly. As long as such incorrect events are prevented, the MAE of below 0.5 meV/atom is still acceptable for applications.

C. Physical Implications of the AL scheme for Practical Simulations

In the two scenarios above, we have analyzed the effectiveness of the proposed $\mathcal{AL}$ scheme to correct for the erroneously predicted absence and presence of metastable states that induce strongly anharmonic effects. To judge the overall stability of this approach, we will now consider further long-term MLIP-MD runs. For the $AgGaSe_2$ case, we ran 30 independent MLIP-MD simulations for up to 1 ns, as depicted in Fig. S19 [65], and no strongly anharmonic effects could be detected anymore. Similarly, 30 independent 1 ns MLIP-MD simulations of CuI from our $\mathcal{AL}$ scheme exhibit defect creations observed in the aiMD trajectory with the correct probabilities and lifetimes. For the latter, the aiMD trajectory was too short to determine the actual frequency of defect creations. Instead, we set a pretrained MLIP model as a reference, i.e. the ground truth, and conducted a $\mathcal{AL}$ scheme, yielding new MLIP models. 50 independent 1 ns MLIP-MD simulations with the new model and the reference model demonstrate that the MLIP models from our $\mathcal{AL}$ scheme can capture the reliable dynamics for these strongly anharmonic events.

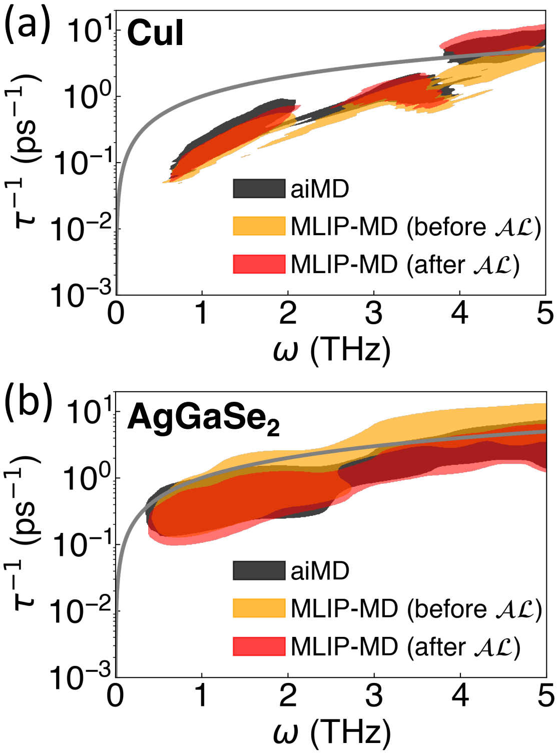

FIG. 12. The inverse phonon lifetimes population distribution(τ^{-1}) as a function of phonon frequency of (a) CuI and (b) $AgGaSe_2$, computed by aiMD (Ref. [59]) and MD using MLIPs (MLIP-MD) before/after applying the $\mathcal{AL}$ scheme. Shaded areas are illustrated for comparison purposes, and their actual, individual point distributions are plotted in Fig S20. These areas are determined by population distribution with Gaussian convolution using a bandwidth of 0.1, and normalized distributions of more than 2 % of their maximum populations are shown for each case. The gray line shows the Ioffe-Regel limit, i.e. a phonon lifetime that corresponds to just one single oscillation.

Further insights can be gained by analyzing the phonon lifetimes obtained from MLIP MD via the fully anharmonic procedure described in Ref. [88]. As shown in Fig. 12 (a), the phonon lifetimes of CuI extracted from the first-principles MD and MLIP MD after $\mathcal{AL}$ are in excellent agreement with each other. However, the MLIP before $\mathcal{AL}$ significantly overestimates the phonon lifetimes. Not surprisingly, $AgGaSe_2$ exhibits an inverse effect, in which the lifetimes are underestimated before $\mathcal{AL}$ as illustrated in Fig. 12 (b). Clearly, these trends are related to the fact that the respective MLIPs before $\mathcal{AL}$ either erroneously miss a metastable state or predict one that is actually not there. In turn, this misses or fictitiously induces strongly anharmonic effects that increase viz. lower the lifetimes, respectively, for CuI and $AgGaSe_2$. Let us emphasize that this change in lifetimes does not depend on whether a spontaneous defect creation is actually observed during the MD trajectory used for the lifetime extraction or not. The sheer presence of additional minimum results in incessant attempts to overcome the barrier and reach this metastable state. Even if unsuccessful, these continuous attempts to induce strongly anharmonic effects to massively lower the lifetime. Notably, this even induces a qualitative change: For CuI, the lifetimes before $\mathcal{AL}$ are low, but still larger than the Ioffe-Regel limit, as plotted in Fig. 12 (a). In DFT and after $\mathcal{AL}$, the strongly anharmonic effects induce a massive reduction of the lifetimes beyond the Ioffe-Regel limit. Since the phonon picture can break down already when one is close to this Ioffe-Regel limit, it implies that the quasi-particle picture is no longer valid [85]. As demonstrated in Ref. [59], this results in a strong reduction of the thermal conductivity that is only accessible with fully anharmonic MD simulations, but not with perturbative phonon-based transport equations. For $AgGaSe_2$, fictitious anharmonic effects are induced due to the erroneous minima in MLIP before $\mathcal{AL}$, inducing the significant underestimation of phonon lifetime as displayed in Fig. 12 (b). This wrong description is corrected in MLIP after $\mathcal{AL}$, closely reproducing the phonon lifetimes from the aiMD trajectory. In turn, this demonstrates that $\mathcal{AL}$ is absolutely necessary for this system not just for quantitative reasons, but even just to predict the correct qualitative transport regime.

In this context, let us emphasize the importance of using fully anharmonic phonon lifetimes extracted from MD as a metric for judging the reliability of the MLIP.

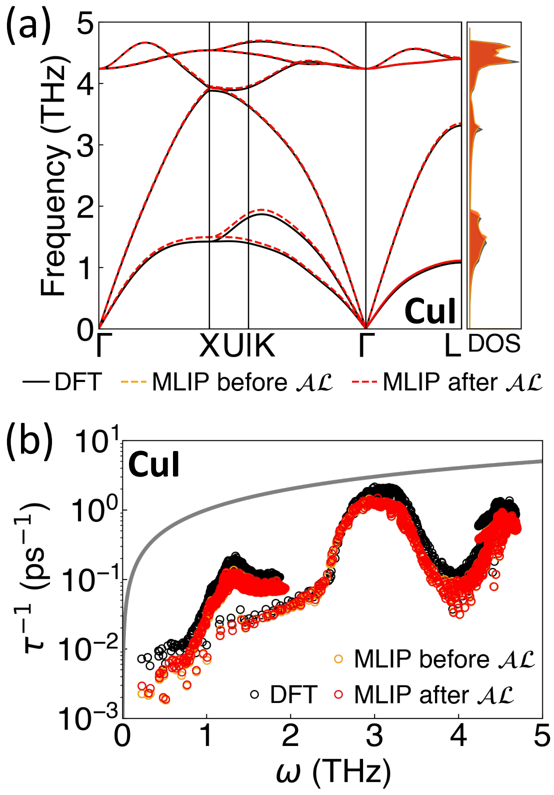

FIG. 13. Perturbative treatment of (a) Phonon dispersion and density of state (DOS) of CuI and (b) inverse phonon lifetimes (τ^{-1}) as a function of phonon frequency of CuI (compare to the non-perturbative results of Fig. 12 (a)). The gray solid line in (b) is the Ioffe-Regel limit. The phonon lifetimes of MLIP before and after $\mathcal{AL}$ (Yellow and red markers) in (b) are almost overlapping.

In the literature, this metric is seldom used because a one-to-one comparison of DFT and MLIP data would require extensive aiMD runs with significant computational costs. Instead, perturbative techniques for calculating phonon band structures and lifetimes are often used to validate the MLIP against DFT reference data. However, these techniques only probe the near-equilibrium portion of the PES and can thus lead to erroneously confident conclusions.

For instance, the harmonic phonon band structures obtained properties via perturbation theory are always in excellent agreement between DFT and MLIP, even before $\mathcal{AL}$, as shown in Fig. 13 for CuI and Fig. S21 [65] for $AgGaSe_2$. Similarly, phonon lifetimes calculated via perturbation theory from third-order force constant [76, 78] are fairly close between DFT and MLIP, regardless of $\mathcal{AL}$. This results from the fact that perturbation theory is "short-sighted", i.e., it only probes small displacements from equilibrium, but not the full PES accessible in thermodynamic equilibrium. Accordingly, such techniques are largely insensitive to the presence of additional, metastable minima and to the associated occurrence of strongly anharmonic effects. This can be further rationalized and substantiated by comparing the perturbative lifetimes in Fig. 13 and Fig. S21 [65] to the fully anharmonic ones in Fig. 12, which exhibit massively different qualitative and quantitative behavior. In particular, the perturbative lifetimes of CuI lie mostly well below the Ioffe-Regel limit as shown in Fig. 13 (b), in sharp contrast to the fully anharmonic ones as illustrated in Fig. 12 (b). With that, perturbative approaches tend to severely underestimate anharmonicity and so to serve as "self-fulfilling prophecy", since only the short-range equilibrium dynamics is probed.

V. CONCLUSIONS

In this work, we adapted existing concepts from literature to develop and test an $\mathcal{AL}$ scheme that is suited to train MLIPs that consistently capture strongly anharmonic effects, even if these occur very rarely and are not present in or regularized away from the initial training data. Our benchmark on available literature data reveals that, at room temperature, the proposed approach is decisive for 10 out of 112 materials. In these problematic cases, standard MLIP training procedures either erroneously predict the absence of strongly anharmonic effects or erroneously predict fictitious strongly anharmonic effects. Using CuI and $AgGaSe_2$ as examples, we show that this is related to the presence or absence of meta-stable configurations on the PES that are only seldom explored. Despite not providing quantitative errors, uncertainty estimates allow to qualitatively detect such problems and can serve as a warning when the MLIP-MD probes regions uncharted in the training data. By exploiting that, the proposed $\mathcal{AL}$ scheme iteratively includes more and more data associated with the phase-space regions featuring problematic predictions, even if those areas are not easily thermodynamically accessible from the start. With that, the $\mathcal{AL}$ scheme is able to train more precise and accurate MLIPs that correctly account for strongly anharmonic effects with modest computational overhead. Obviously, the resulting MLIP is not universally valid for the whole structural and thermodynamic phase space. Rather, the proposed $\mathcal{AL}$ scheme can and should be applied whenever different systems and/or thermodynamic conditions are explored, e.g., at new temperatures or pressures, when introducing impurities or defects, under stress or strain, and when introducing new interfaces.

If MLIP models are utilized for systems away from the trained regions, uncertainty estimates must be implemented during MLIP-MD for slightly different systems, e.g., at new temperatures, with impurity atoms, at new interfaces, or under strain. They serve as the warning alarm to detect when it goes beyond the trained area.

From a physical point of view, our analysis reveals that the proposed $\mathcal{AL}$ procedure is able to produce reliable MLIPs that accurately predict strongly anharmonic effects **without** prior knowledge about the actuating mechanism or about if strongly anharmonic effects are active at all. In fact, our study also reveals that usual metrics used to monitor the accuracy of MLIPs during training are actually not sensitive to strongly anharmonic effects: Average quantities, like MAE of energies and forces or thermodynamic equilibrium averages, are typically insensitive to such strong anharmonic effects that can be short-lived and occur rarely. Similarly, standard vibrational properties like phonon frequencies and lifetimes obtained from perturbational ansatzes only probe the near-to-equilibrium region and are, hence, blind to strongly anharmonic effects. As demonstrated by computing fully anharmonic phonon lifetimes from MD, such approaches are overconfident when it comes to anharmonicity. This results not just in quantitative errors, but even in the wrong qualitative transport picture. Conversely, the devised $\mathcal{AL}$ approach reliably accounts for these effects and is able to correctly reproduce all transport regimes.

ACKNOWLEDGMENTS

This project was supported by the ERC Advanced Grant TEC1p (European Research Council, Grant Agreement No. 740233). This research used the resources of the Max Planck Computing and Data Facility.

[1] S. Lorenz, A. Groß, and M. Scheffler, Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks, Chemical Physics Letters **395**, 210 (2004).

[2] V. L. Deringer, M. A. Caro, and G. Csányi, Machine learning interatomic potentials as emerging tools for materials science, Advanced Materials **31**, 1902765 (2019).

[3] Y. Mishin, Machine-learning interatomic potentials for materials science, Acta Materialia **214**, 116980 (2021).

[4] J. Behler, Four generations of high-dimensional neural network potentials, Chemical Reviews **121**, 10037 (2021).

[5] G. C. Sosso, D. Donadio, S. Caravati, J. Behler, and M. Bernasconi, Thermal transport in phase-change materials from atomistic simulations, Phys. Rev. B **86**, 104301 (2012).

[6] X. Qian, S. Peng, X. Li, Y. Wei, and R. Yang, Thermal conductivity modeling using machine learning potentials: application to crystalline and amorphous silicon, Materials Today Physics **10**, 100140 (2019).

[7] P. Korotaev, I. Novoselov, A. Yanilkin, and A. Shapeev, Accessing thermal conductivity of complex compounds by machine learning interatomic potentials, Phys. Rev. B **100**, 144308 (2019).

[8] B. Mortazavi, E. V. Podryabinkin, S. Roche, T. Rabczuk, X. Zhuang, and A. V. Shapeev, Machine-learning interatomic potentials enable first-principles multiscale modeling of lattice thermal conductivity in graphene/borophene heterostructures, Mater. Horiz. **7**, 2359 (2020).

[9] C. Mangold, S. Chen, G. Barbalinardo, J. Behler, P. Pochet, K. Termentzidis, Y. Han, L. Chaput, D. Lacroix, and D. Donadio, Transferability of neural network potentials for varying stoichiometry: Phonons and thermal conductivity of MnxGey compounds, Journal of Applied Physics **127**, 244901 (2020).

[10] R. Li, E. Lee, and T. Luo, A unified deep neural network potential capable of predicting thermal conductivity of silicon in different phases, Materials Today Physics **12**, 100181 (2020).

[11] R. Li, Z. Liu, A. Rohskopf, K. Gordiz, A. Henry, E. Lee, and T. Luo, A deep neural network interatomic potential for studying thermal conductivity of β-Ga2O3, Applied Physics Letters **117**, 152102 (2020).

[12] H. Liu, X. Qian, H. Bao, C. Y. Zhao, and X. Gu, High-temperature phonon transport properties of snse from machine-learning interatomic potential, Journal of Physics: Condensed Matter **33**, 405401 (2021).

[13] C. Verdi, F. Karsai, P. Liu, R. Jinnouchi, and G. Kresse, Thermal transport and phase transitions of zirconia by on-the-fly machine-learned interatomic potentials, npj Computational Materials **7**, 156 (2021).

[14] M. F. Langer, F. Knoop, C. Carbogno, M. Scheffler, and M. Rupp, Heat flux for semilocal machine-learning potentials, Phys. Rev. B **108**, L100302 (2023).

[15] G. C. Sosso, G. Miceli, S. Caravati, J. Behler, and M. Bernasconi, Neural network interatomic potential for the phase change material gete, Phys. Rev. B **85**, 174103 (2012).

[16] V. L. Deringer and G. Csányi, Machine learning based interatomic potential for amorphous carbon, Phys. Rev. B **95**, 094203 (2017).

[17] V. L. Deringer, N. Bernstein, A. P. Bartók, M. J. Cliffe, R. N. Kerber, L. E. Marbella, C. P. Grey, S. R. Elliott, and G. Csányi, Realistic atomistic structure of amorphous silicon from machine-learning-driven molecular dynamics, The Journal of Physical Chemistry Letters **9**, 2879 (2018), pMID: 29754489.

[18] G. Sivaraman, A. N. Krishnamoorthy, M. Baur, C. Holm, M. Stan, G. Csányi, C. Benmore, and Á. Vázquez-Mayagoitia, Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide, npj Computational Materials **6**, 104 (2020).

[19] W. Li, Y. Ando, E. Minamitani, and S. Watanabe, Study of Li atom diffusion in amorphous Li3PO4 with neural network potential, The Journal of Chemical Physics **147**, 214106 (2017).

[20] C. Wang, K. Aoyagi, P. Wisesa, and T. Mueller, Lithium ion conduction in cathode coating materials from on-the-fly machine learning, Chemistry of Materials **32**, 3741 (2020).

[21] Y. Shao, L. Knijff, F. M. Dietrich, K. Hermansson, and C. Zhang, Modelling bulk electrolytes and electrolyte interfaces with atomistic machine learning, Batteries & Supercaps **4**, 585 (2021).

[22] A. Hajibabaei and K. S. Kim, Universal machine learning interatomic potentials: Surveying solid electrolytes, The Journal of Physical Chemistry Letters **12**, 8115 (2021), pMID: 34410138.

[23] A. M. Goryaeva, J.-B. Maillet, and M.-C. Marinica, Towards better efficiency of interatomic linear machine learning potentials, Computational Materials Science **166**, 200 (2019).

[24] S. R. Xie, M. Rupp, and R. G. Hennig, Ultra-fast interpretable machine-learning potentials, npj Computational Materials **9**, 162 (2023).

[25] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nature Communications **13**, 2453 (2022).

[26] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csanyi, Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, in *Advances in Neural Information Processing Systems*, Vol. 35, edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Curran Associates, Inc., 2022) pp. 11423–11436.

[27] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nature Machine Intelligence , 1–11 (2023).

[28] H. Yu, Y. Zhong, L. Hong, C. Xu, W. Ren, X. Gong, and H. Xiang, Spin-dependent graph neural network potential for magnetic materials, Phys. Rev. B **109**, 144426 (2024).

[29] A. Grisafi and M. Ceriotti, Incorporating long-range physics in atomic-scale machine learning, The Journal of Chemical Physics **151**, 204105 (2019).

[30] L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, and W. E, A deep potential model with long-range electrostatic interactions, The Journal of Chemical Physics **156**, 124107 (2022).

[31] A. Gao and R. C. Remsing, Self-consistent determination of long-range electrostatics in neural network potentials, Nature Communications **13**, 1572 (2022).

[32] T. Jaffrelot Inizan, T. Plé, O. Adjoua, P. Ren, H. Gökcan, O. Isayev, L. Lagardère, and J.-P. Piquemal, Scalable hybrid deep neural networks/polarizable potentials biomolecular simulations including long-range effects, Chem. Sci. **14**, 5438 (2023).

[33] D. M. Anstine and O. Isayev, Machine learning interatomic potentials and long-range physics, The Journal of Physical Chemistry A **127**, 2417 (2023).

[34] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. D. Cubuk, Scaling deep learning for materials discovery, Nature **624**, 80 (2023).

[35] J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. **98**, 146401 (2007).

[36] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, Phys. Rev. Lett. **104**, 136403 (2010).

[37] H. Wang, L. Zhang, J. Han, and W. E, Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics, Computer Physics Communications **228**, 178 (2018).

[38] I. S. Novikov, K. Gubaev, E. V. Podryabinkin, and A. V. Shapeev, The mlip package: moment tensor potentials with mpi and active learning, Machine Learning: Science and Technology **2**, 025002 (2020).

[39] J. Mockus, On bayesian methods for seeking the extremum, in *Optimization Techniques* (1974).

[40] S. Dasgupta and D. Hsu, Hierarchical sampling for active learning, in *Proceedings of the 25th International Conference on Machine Learning*, ICML '08 (Association for Computing Machinery, New York, NY, USA, 2008) p. 208–215.

[41] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, Active learning of uniformly accurate interatomic potentials for materials simulation, Phys. Rev. Mater. **3**, 023804 (2019).

[42] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang, and W. E, Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models, Computer Physics Communications **253**, 107206 (2020).

[43] J. Carrete, H. Montes-Campos, R. Wanzenböck, E. Heid, and G. K. H. Madsen, Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning, The Journal of Chemical Physics **158**, 204801 (2023).

[44] Y. Xie, J. Vandermause, L. Sun, A. Cepellotti, and B. Kozinsky, Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene, npj Computational Materials **7**, 40 (2021).

[45] Y. Xie, J. Vandermause, S. Ramakers, N. H. Protik, A. Johansson, and B. Kozinsky, Uncertainty-aware molecular dynamics from bayesian active learning for phase transformations and thermal transport in sic, npj Computational Materials **9**, 36 (2023).

[46] A. Zhu, S. Batzner, A. Musaelian, and B. Kozinsky, Fast uncertainty estimates in deep learning interatomic potentials, The Journal of Chemical Physics **158**, 164111 (2023).

[47] C. van der Oord, M. Sachs, D. P. Kovács, C. Ortner, and G. Csányi, Hyperactive learning for data-driven interatomic potentials, npj Computational Materials **9**, 168 (2023).

[48] V. Zaverkin, D. Holzmüller, H. Christiansen, F. Errica, F. Alesiani, M. Takamoto, M. Niepert, and J. Kästner, Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials, Preprint on Research Square (2023).

[49] M. Kulichenko, K. Barros, N. Lubbers, Y. W. Li, R. Messerly, S. Tretiak, J. S. Smith, and B. Nebgen, Uncertainty-driven dynamics for active learning of interatomic potentials, Nature Computational Science **3**, 230 (2023).

[50] Z. Li, J. R. Kermode, and A. De Vita, Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces, Phys. Rev. Lett. **114**, 096405 (2015).

[51] E. V. Podryabinkin and A. V. Shapeev, Active learning of linearly parametrized interatomic potentials, Computational Materials Science **140**, 171 (2017).

[52] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science **2**, 718 (2022).

[53] J. D. Morrow, J. L. A. Gardner, and V. L. Deringer, How to validate machine-learned interatomic potentials, The Journal of Chemical Physics **158**, 121501 (2023).

[54] Y. Chen, Y. Ou, P. Zheng, Y. Huang, F. Ge, and P. O. Dral, Benchmark of general-purpose machine learning-based quantum mechanical method AIQM1 on reaction barrier heights, The Journal of Chemical Physics **158**, 074103 (2023).

[55] J. Riebesell, R. E. A. Goodall, P. Benner, Y. Chiang, B. Deng, A. A. Lee, A. Jain, and K. A. Persson, Mat-

bench discovery – a framework to evaluate machine learning crystal stability predictions (2024), arXiv:2308.14920 [cond-mat.mtrl-sci].

[56] M. S. Green, Markoff Random Processes and the Statistical Mechanics of Time-Dependent Phenomena. II. Irreversible Processes in Fluids, The Journal of Chemical Physics **22**, 398 (1954).

[57] R. Kubo, Statistical-mechanical theory of irreversible processes. i. general theory and simple applications to magnetic and conduction problems, Journal of the Physical Society of Japan **12**, 570 (1957).

[58] R. Kubo, M. Yokota, and S. Nakajima, Statistical-mechanical theory of irreversible processes. ii. response to thermal disturbance, Journal of the Physical Society of Japan **12**, 1203 (1957).

[59] F. Knoop, T. A. R. Purcell, M. Scheffler, and C. Carbogno, Anharmonicity in thermal insulators: An analysis from first principles, Phys. Rev. Lett. **130**, 236301 (2023).

[60] G. S. Jung, J. Y. Choi, and S. M. Lee, Active learning of neural network potentials for rare events, Digital Discovery **3**, 514 (2024).

[61] D. West and S. K. Estreicher, First-principles calculations of vibrational lifetimes and decay channels: Hydrogen-related modes in si, Phys. Rev. Lett. **96**, 115504 (2006).

[62] A. Castellano, F. m. c. Bottin, J. Bouchet, A. Levitt, and G. Stoltz, *ab* initio canonical sampling based on variational inference, Phys. Rev. B **106**, L161110 (2022).

[63] A. Takahashi, A. Seko, and I. Tanaka, Conceptual and practical bases for the high accuracy of machine learning interatomic potentials: Application to elemental titanium, Phys. Rev. Mater. **1**, 063801 (2017).

[64] 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. D. 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.

[65] See Supplemental Material at [URL will be inserted by publisher] for the discussion, with citations [27, 52, 59, 64, 80, 82], on the universal MLIP models for the rare events prediction in CuI, the $\mathcal{AL}$ result of $KCaF_3$, the learning process of CuI during the $\mathcal{AL}$, the absence of rare events in CuI during MLIP-MD, the other cases of fictitious minima ($AgGaS_2$, $InNaO_2$, CsBr, CsCl, LiBr, LiCl, LiI, and Na_2Te), the distribution of training data from aiMD in $AgGaSe_2$, the phonon lifetimes from aiMD and MLIP-MD of CuI and $AgGaSe_2$, and perturbative phonon predictions using MLIP models for $AgGaSe_2$.

[66] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, and A. R. Oganov, Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning, Phys. Rev. B **99**, 064114 (2019).

[67] L. Kahle and F. Zipoli, Quality of uncertainty estimates from neural network potential ensembles, Phys. Rev. E **105**, 015311 (2022).

[68] P. Mendes, P. Romano, and D. Garlan, Error-driven uncertainty aware training (2024), arXiv:2405.01205.

[69] S. Lahlou, M. Jain, H. Nekoei, V. I. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, and Y. Bengio, Deup: Direct epistemic uncertainty prediction (2023), arXiv:2102.08501.

[70] F. Knoop, T. A. R. Purcell, M. Scheffler, and C. Carbogno, Anharmonicity measure for materials, Phys. Rev. Mater. **4**, 083809 (2020).

[71] 10.17172/nomad/2021.11.11-1.

[72] V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter, and M. Scheffler, Ab initio molecular simulations with numeric atom-centered orbitals, Computer Physics Communications **180**, 2175 (2009).

[73] F. Knuth, C. Carbogno, V. Atalla, V. Blum, and M. Scheffler, All-electron formalism for total energy strain derivatives and stress tensor components for numeric atom-centered orbitals, Computer Physics Communications **190**, 33 (2015).

[74] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, and K. Burke, Restoring the density-gradient expansion for exchange in solids and surfaces, Phys. Rev. Lett. **100**, 136406 (2008).

[75] F. Knoop, T. A. r. Purcell, M. Scheffler, and C. Carbogno, Fhi-vibes: *abinitio* vibrational simulations, Journal of Open Source Software **5**, 2671 (2020).

[76] A. Togo, L. Chaput, T. Tadano, and I. Tanaka, Implementation strategies in phonopy and phono3py, J. Phys. Condens. Matter **35**, 353001 (2023).

[77] A. Togo, First-principles phonon calculations with phonopy and phono3py, J. Phys. Soc. Jpn. **92**, 012001 (2023).

[78] A. Togo, L. Chaput, and I. Tanaka, Distributions of phonon lifetimes in brillouin zones, Phys. Rev. B **91**, 094306 (2015).

[79] Available at: `https://keysongkang.github.io/ALmoMD/`.

[80] A. H. Larsen, J. J. 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. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, The atomic simulation environment—a python library for working with atoms, Journal of Physics: Condensed Matter **29**, 273002 (2017).

[81] B. Berne, G. Ciccotti, and D. Coker, *Classical And Quantum Dynamics In Condensed Phase Simulations: Proceedings Of The International School Of Physics* (World Scientific Publishing Company, 1998).

[82] L. R. Corrales, R. M. Van Ginhoven, J. Song, and H. Jónsson, Vacancy migration barrier energetics and pathways in silica, MRS Online Proceedings Library **538**, 317 (1998).

[83] G. Henkelman and H. Jónsson, Improved tangent esti-

mate in the nudged elastic band method for finding minimum energy paths and saddle points, The Journal of Chemical Physics **113**, 9978 (2000).

[84] H. A. Sturges, The choice of a class interval, Journal of the American Statistical Association **21**, 65 (1926).

[85] M. Simoncelli, N. Marzari, and F. Mauri, Wigner formulation of thermal transport in solids, Phys. Rev. X **12**, 041011 (2022).

[86] G. Caldarelli, M. Simoncelli, N. Marzari, F. Mauri, and L. Benfatto, Many-body green's function approach to lattice thermal transport, Phys. Rev. B **106**, 024312 (2022).

[87] A. Ioffe and A. Regel, Non-crystalline, amorphous, and liquid electronic semiconductors, Progress in semiconductors , 237 (1960).

[88] F. Knoop, M. Scheffler, and C. Carbogno, Ab initio green-kubo simulations of heat transport in solids: Method and implementation, Phys. Rev. B **107**, 224304 (2023).