Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials

Viktor Zaverkin, David Holzmüller, Henrik Christiansen, Federico Errica, Francesco Alesiani, Makoto Takamoto, Mathias Niepert, Johannes Kästner

npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01254-1

Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials

Viktor Zaverkin[1,2] ✉, David Holzmüller[3,4], Henrik Christiansen[1], Federico Errica[1], Francesco Alesiani[1], Makoto Takamoto[1], Mathias Niepert[1,5] & Johannes Kästner[2]

Efficiently creating a concise but comprehensive data set for training machine-learned interatomic potentials (MLIPs) is an under-explored problem. Active learning, which uses biased or unbiased molecular dynamics (MD) to generate candidate pools, aims to address this objective. Existing biased and unbiased MD-simulation methods, however, are prone to miss either rare events or extrapolative regions—areas of the configurational space where unreliable predictions are made. This work demonstrates that MD, when biased by the MLIP's energy uncertainty, simultaneously captures extrapolative regions and rare events, which is crucial for developing uniformly accurate MLIPs. Furthermore, exploiting automatic differentiation, we enhance bias-forces-driven MD with the concept of bias stress. We employ calibrated gradient-based uncertainties to yield MLIPs with similar or, sometimes, better accuracy than ensemble-based methods at a lower computational cost. Finally, we apply uncertainty-biased MD to alanine dipeptide and MIL-53(Al), generating MLIPs that represent both configurational spaces more accurately than models trained with conventional MD.

Computational techniques are invaluable for exploring complex configurational and compositional spaces of molecular and material systems. The accuracy and efficiency, however, depend on the chosen computational methods. Ab initio molecular dynamics (MD) simulations using density-functional theory (DFT) provide accurate results but are computationally demanding. Atomistic simulations with classical force fields offer a faster alternative but often lack accuracy. Thus, developing accurate and computationally efficient interatomic potentials is a key challenge successfully addressed by machine-learned interatomic potentials (MLIPs)[1–5]. An essential component of any MLIP is the accurate encoding of the atomic system by a local representation, which depends on configurational (atomic positions) and compositional (atomic types) degrees of freedom[6]. Recently, a wide range of MLIPs have been introduced, comprising linear and kernel-based models[7–10], Gaussian approximation[11,12], and neural network (NN) interatomic potentials[13–17], including graph NNs[18–24], all demonstrating remarkable success in atomistic simulations.

The effectiveness of MLIPs, however, crucially relies on training data sufficiently covering configurational and compositional spaces[25,26]. Without such training data, MLIPs cannot faithfully reproduce the underlying physics. An open challenge, therefore, is the generation of comprehensive training data sets for MLIPs, covering relevant configurational and compositional spaces and ensuring that resulting MLIPs are uniformly accurate across these spaces. This objective must be realized while reducing the number of expensive DFT evaluations, which provide reference energies, atomic forces, and stresses. This challenge is further complicated by the limited knowledge of physical conditions, such as temperature and pressure, at which configurational changes occur. Setting temperatures and pressures excessively high can result in atomic system degradation before exploring the relevant phase space.

To address this challenge, iterative active learning (AL) algorithms are used to improve the accuracy of MLIPs by providing an augmented data set[27–34]; see Fig. 1(a). They select the data most informative to the model, i.e., atomic configurations with high energy and force uncertainties, as estimated by the model. This data is drawn from configurational and compositional spaces explored during, e.g., MD simulations. Reference DFT energies, atomic forces, and stresses are evaluated for the selected configurations. Furthermore, energy and force uncertainties indicate the onset of extrapolative regions—regions where unreliable predictions are made—

[1]NEC Laboratories Europe GmbH, Kurfürsten-Anlage 36, 69115 Heidelberg, Germany. [2]Institute for Theoretical Chemistry, University of Stuttgart, Pfaffenwaldring 55, 70569 Stuttgart, Germany. [3]Institute for Stochastics and Applications, University of Stuttgart, Pfaffenwaldring 57, 70569 Stuttgart, Germany. [4]SIERRA, INRIA Paris, 2 rue Simone Iff, 75012 Paris, France. [5]Institute for Artificial Intelligence, University of Stuttgart, Universitätsstraße 32, 70569 Stuttgart, Germany.
✉e-mail: viktor.zaverkin@neclab.eu

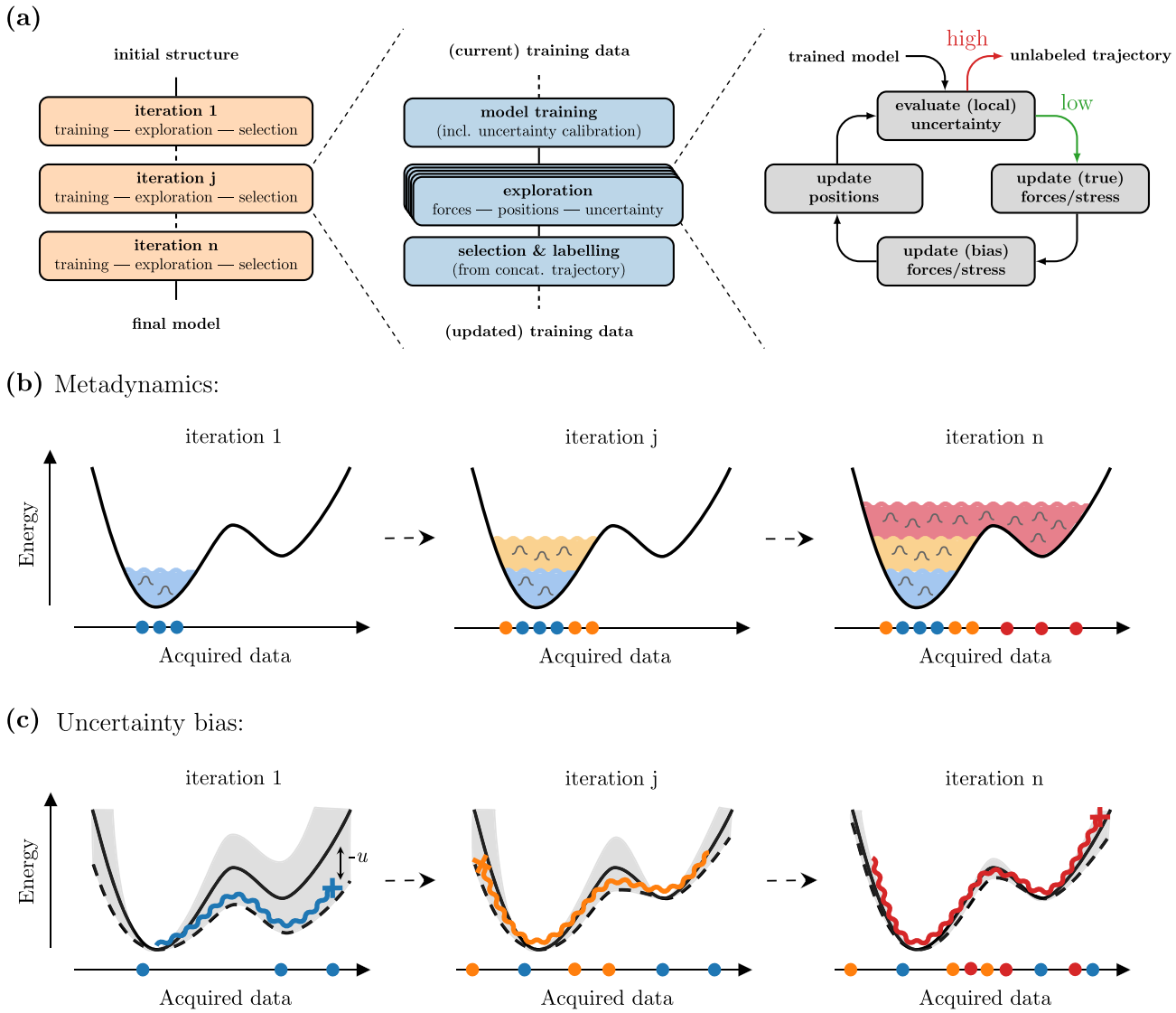

Fig. 1 | A schematic overview of an AL algorithm for MLIP training. Training structures are selected from data gathered during biased or unbiased MD simulations. **a** An AL experiment begins with training an MLIP in the first iteration using a small set of randomly perturbed initial configurations. The current MLIP is employed in each iteration to run parallel MD simulations. Each simulation continues until it reaches a predefined uncertainty threshold. Then, a batch of configurations is selected from all trajectories. Reference energies and forces of these samples are evaluated using a DFT solver, updating the training data set. The updated data set is employed for training the MLIP in the next iteration. **b** Adaptive biasing strategies like metadynamics enhance the exploration of the configurational space. In metadynamics, exploration along manually defined CVs is facilitated by adding Gaussian functions to a history-dependent bias (areas filled by blue, orange, and red colors). However, even for well-defined CVs, exploring the configurational space of interest may require long simulation times due to the diffusive motion along these CVs. **c** Uncertainty-biased MD aims to minimize uncertainty u (grey shaded area) related to the actual error, thereby facilitating the exploration of the configurational space. In uncertainty-biased MD, we subtract the MLIP's energy uncertainty from the predicted energy (continuous black line) and run MD simulations using the altered energy surface (dashed black line). Curved lines denote distinct MD trajectories. Unlike metadynamics, uncertainty-biased MD operates without defining CVs and drives MD simulations toward high uncertainty regions in each iteration.

prompting the termination of MD simulations and the evaluation of reference DFT values. In this AL setting, covering the configurational space and exploring extrapolative configurations might require running longer MD simulations and defining physical conditions for observing slow configurational changes (rare events).

Alternatively, enhanced sampling methods can significantly speed up the exploration of the configurational space by using adaptive biasing strategies such as metadynamics[35–41]; see Fig. 1(b). However, metadynamics requires manually selecting a few collective variables (CVs) that are assumed to describe the system. The limited number of CVs restricts exploration, as they might miss relevant transitions and parts of the configurational space. In contrast, MD simulations biased toward regions of high uncertainty can enhance the discovery of extrapolative configurations[42,43]. A related work utilizes uncertainty gradients for adversarial training of MLIPs[44,45]. To obtain MLIPs that are uniformly accurate across the relevant configurational space, however, simultaneous exploration of rare events and extrapolative configurations is necessary. The extent to which uncertainty-biased MD can achieve this objective remains an unexplored research area.

This work demonstrates the capability of uncertainty-biased MD to explore the configurational space, including fast exploration of rare events and extrapolative regions; see Fig. 1(c). We achieve this by exploring the CVs of alanine dipeptide—a widely used model for protein backbone structure. To assess the coverage of the CV space, we introduce a measure using a tree-based weighted recursive space partitioning. Furthermore, we extend

existing uncertainty-biased MD simulations by automatic differentiation (AD) and propose a biasing technique that utilizes bias stresses obtained by differentiating the model's uncertainty with respect to infinitesimal strain deformations. We assess the efficiency of the proposed technique by running MD simulations in isothermal-isobaric (*NpT*) statistical ensemble and exploring cell parameters of MIL-53(Al)—a flexible metal-organic framework (MOF) featuring closed- and large-pore stable states. Both benchmark systems are often used in studies assessing enhanced sampling and data generation methods[36,38,41,44].

A key ingredient of AL algorithms with dynamically generated candidate pools is a sensitive metric for detecting the onset of extrapolative regions. These regions are typically associated with large errors in MLIP predictions. However, MLIP uncertainties often underestimate actual errors[46,47], resulting in the exploration of unphysical regions, negatively affecting MLIP training. Thus, calibrated uncertainties are crucial for generating high-quality MLIPs with AL, which involves configurations explored during MLIP-based MD[47–49], but might be unnecessary in AL tasks that rely on relative uncertainties[50–52]. In our setting, we demonstrate that conformal prediction (CP) helps align the largest force error with its corresponding uncertainty value. This approach effectively makes MLIPs not underestimate force errors, which is important for preventing MD from exploring unphysical configurations. Thus, CP-based uncertainty calibration helps set reasonable uncertainty thresholds without limiting the exploration of the configurational space. In contrast, conventional approaches drive MD away from high-uncertainty regions, which can hinder exploration[53].

Contrary to existing work[42,43], which relies on ensembles of MLIPs for uncertainty quantification, we propose using ensemble-free uncertainties of NN-based MLIPs derived from gradient features[50–52]. These features can be interpreted as the sensitivity of a model's output to parameter changes. Recent studies demonstrate that gradient-based uncertainties perform comparably to ensemble-based counterparts in AL[51,52,54]. Furthermore, they yield the exact posterior in the case of linear models[9,10]. We demonstrate that gradient features can define uncertainties of total and atom-based properties, such as energy and atomic forces. To make gradient-based uncertainties computationally efficient, we employ the sketching technique[55] and reduce the dimensionality of gradient features. For many NN-based MLIPs, gradient-based approaches can significantly reduce the computational cost of uncertainty quantification and accelerate the time-consuming MD simulations compared to ensemble-based methods. However, the latter can be made computationally efficient, e.g., through parallelization or employing specific settings with non-trainable descriptors and gradient-free force uncertainties[45].

We further enhance configurational space exploration and improve the computational efficiency of AL by employing batch selection algorithms[51,52]. These algorithms simultaneously select multiple atomic configurations from trajectories generated during parallel MD simulations. Batch selection algorithms enforce the informativeness and diversity of the selected atomic structures. Thus, they ensure the construction of maximally diverse training data sets.

Results

Overview

In the following, we first demonstrate the necessity of uncertainty calibration on an example of MIL-53(Al) to constrain MD to physically reasonable regions of the configurational space. Then, we present two complementary analyses demonstrating the improved data efficiency of MLIPs obtained by our AL approach, developing MLIPs for alanine dipeptide and MIL-53(Al). Furthermore, we investigate how uncertainty-biased MD enhances the exploration of the configurational space, utilizing bias forces and stress. To benchmark our results, we draw a comparison with MD run at elevated temperatures and pressures as well as metadynamics simulations. The details on the ensemble-free uncertainties (distance- and posterior-based ones) and uncertainty-biased MD can be found in Methods.

Calibrating uncertainties with conformal prediction

Total and atom-based uncertainties are typically poorly calibrated[47], meaning that they often underestimate actual errors. The underestimation of atomic force errors is particularly dangerous when dynamically generating candidate pools, as it may result in exploring unphysical configurations with extremely large errors in predicted forces. These unphysical configurations often cause convergence issues in reference DFT calculations. Additionally, poor calibration complicates defining an appropriate uncertainty threshold for prompting the termination of MD simulations and the evaluation of reference DFT energies, atomic forces, and stresses. To address this issue, we utilize inductive CP, which computes a re-scaling factor based on predicted uncertainties and prediction errors on a calibration set. The confidence level $1-\alpha$ in CP is defined such that the probability of underestimating the error is at most α on data drawn from the same distribution as the calibration set. For more details, see Methods.

Figure 2 demonstrates the correlation of maximal atom-based uncertainties, $\max_i u_i$, with maximal atomic force RMSEs, $\max_i \sqrt{\frac{1}{3}\sum_{k=1}^{3}(\Delta F_{i,k})^2}$, for the MIL-53(Al) test data set from ref. 41 based on numerous first principles MD trajectories at 600 K. We chose maximal atomic force RMSE as our target metric to identify extrapolative regions due to its high sensitivity to unphysical local atomic environments. In MLIP-based atomistic simulations, we model it using maximal atom-based uncertainty. Employing quantiles or averages of atomic force RMSE could extend simulation time by reducing sensitivity to extreme values; however, exploring these alternatives is left for future work.

In Fig. 2, transparent hexbins represent uncertainties calibrated with a lower confidence ($\alpha = 0.5$; see Methods), while opaque ones depict those calibrated with a higher confidence ($\alpha = 0.05$). The presented uncertainties are derived from gradient features or an ensemble of three MLIPs and calibrated using CP with atomic force RMSEs[49]. For posterior- and distance-based uncertainties, which are unitless, the re-scaling with CP ensures that the resulting uncertainties are provided in correct units, i.e., eV $Å^{-1}$. Ensemble-based uncertainty quantification already provides correct units, which CP preserves. Equivalent results for alanine dipeptide, including the correlation between average uncertainties and average force RMSEs, can be found in the Supplementary Information.

Figure 2 (top) demonstrates results for MLIPs trained on 45 MIL-53(Al) configurations, while five samples were used for early stopping and uncertainty calibration. Figure 2 (bottom) shows the results for MLIPs trained and validated on 450 and 50 MIL-53(Al) configurations, respectively. In both experiments, the training and validation samples were selected from the data sets provided by ref. 41. The first 50 samples correspond to randomly perturbed structures, while the remaining 450 are generated using metadynamics combined with incremental learning[41]. The latter is an iterative algorithm that improves MLIPs by training on configurations generated sequentially over time, using the last frame of atomistic simulations.

We observe that uncertainties calibrated with a lower confidence level often underestimate actual errors. In this case, MD can explore unphysical regions before reaching the uncertainty threshold, especially in cases with a weak correlation between uncertainties and actual errors. By employing CP with higher confidence, we help align the largest prediction error with the corresponding uncertainty, thereby improving its ability to identify the onset of extrapolative regions. This alignment becomes apparent in Fig. 2, where CP shifts the hexbin points to be on or below the diagonal.

In Fig. 2 (top), we find that even training and calibrating models with a few randomly perturbed atomic configurations is sufficient for robust identification of unreliable predictions. This result is crucial as we rely on such data sets to initialize our AL experiments, eliminating the need for predefined data sets[42,43]. Furthermore, we observe that, for MIL-53(Al), calibrated uncertainties from model ensembles tend to overestimate the actual error to a greater extent than gradient-based approaches. While this may not be critical when exploring unphysical configurations, it can prematurely terminate MD simulations. This trend is consistent across all

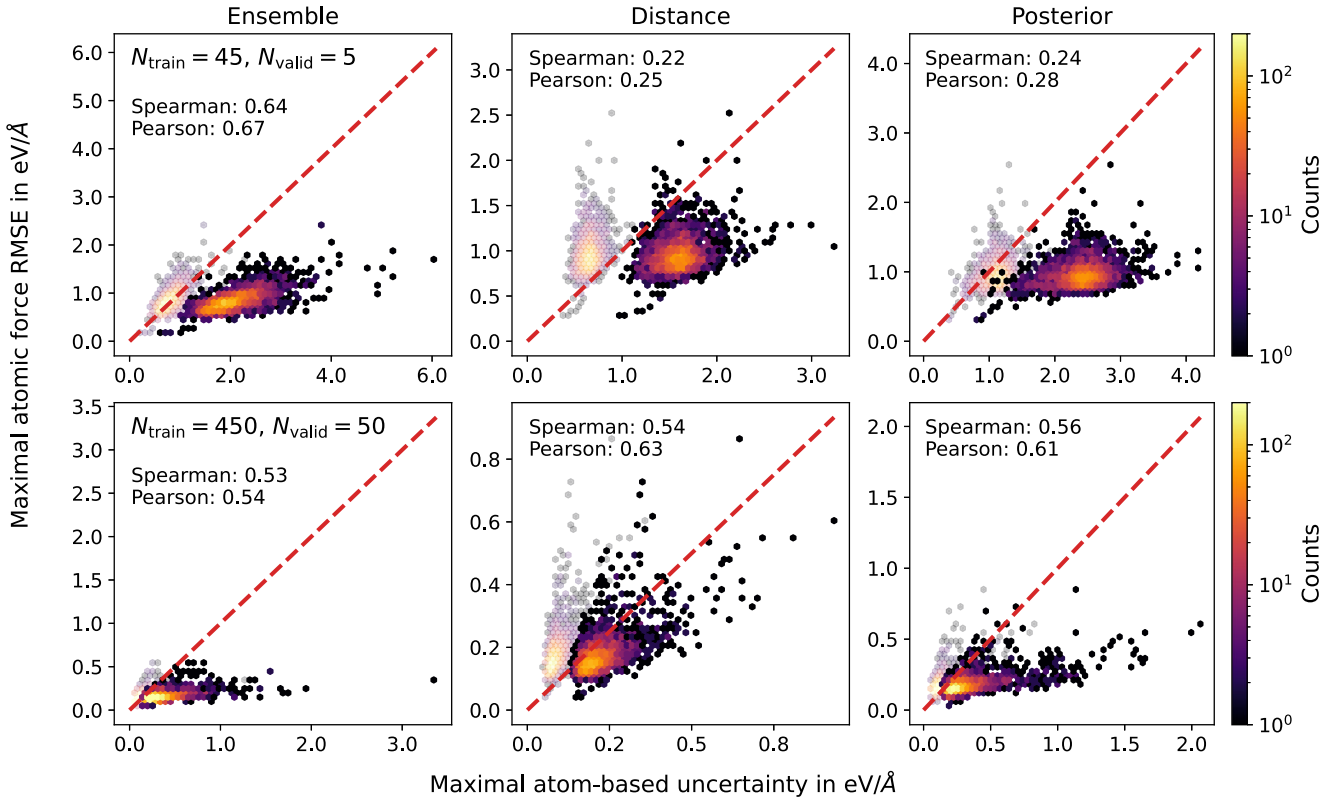

Fig. 2 | Correlation of maximal atom-based uncertainties with maximal atomic force RMSEs for MIL-53(Al). The results are presented for the test data set from ref. 41. All uncertainty quantification methods are calibrated using CP and atomic force RMSEs. The top row shows the results of MLIPs trained using 45 atomic configurations, while five are used for early stopping and uncertainty calibration. The bottom row shows the results obtained with 450 and 50 MIL-53(Al) configurations, respectively. The training and validation data are taken from ref. 41. Transparent hexbin points represent uncertainties calibrated with $\alpha = 0.5$ (low confidence; see Methods), while opaque ones denote uncertainties calibrated with $\alpha = 0.05$ (high confidence). Calibrating uncertainties with a high confidence level helps align the largest actual error with the corresponding uncertainty, shifting the hexbin points to or below the red diagonal line. This alignment is crucial for identifying unreliable predictions and prompting the termination of MD simulations.

training and calibration data sizes. Lastly, the results provided here and in the Supplementary Information demonstrate that all uncertainty methods perform comparably regarding Pearson and Spearman correlation coefficients.

Performance of bias-forces-driven active learning

Exploring the configurational space of complex molecular systems, particularly those with multiple stable states, is essential for developing accurate and robust MLIPs. We apply bias-forces-driven MD combined with AL to develop MLIPs for alanine dipeptide in vacuum. This dipeptide exhibits two stable conformers characterized by the backbone dihedral angles ϕ and ψ (see Fig. 3): the C_{7eq} state with $\phi \approx -1.5$ rad and $\psi \approx 1.19$ rad and the C_{ax} state with $\phi \approx 0.9$ rad and $\psi \approx -0.9$ rad[56]. We use unbiased MD as the baseline for generating candidate pools in two scenarios: AL with candidates selected from unbiased MD trajectories based on their uncertainty (and diversity) and candidates sampled from them at random. The performance of MLIPs is assessed employing the test data obtained from a long MD trajectory at 1200 K; see Methods. We employ the AMBER ff19SB force field for reference energy and force calculations[57], as implemented in the TorchMD package using PyTorch[58,59].

Figure 3 demonstrates the performance of MLIPs obtained for alanine dipeptide depending on the number of acquired configurations. Table 1 presents error metrics evaluated for MLIPs at the end of each experiment. Here, we provide results for the posterior-based uncertainty and uncertainty-biased MD at 300 K. The Supplementary Information presents equivalent results for other uncertainty methods and temperatures. Figure 3a presents the coverage of the CV space defined by ϕ and ψ, calculated using all MD trajectories up to the current AL step. We measure the coverage of the respective space by a tree-based weighted recursive space partitioning; see Methods. AL experiments combined with unbiased MD at 1200 K serve as the upper-performance limit for MLIPs in the case of alanine dipeptide, achieving the highest coverage of 0.97 after acquiring 512 configurations. Increasing temperature even further while using interatomic potentials, which allow for bond breaking and formation, may lead to the degradation of the molecule. Uncertainty-biased MD simulations at 300 K result in slightly lower coverage values, surpassing the coverages achieved by unbiased MD at 300 K and 600 K.

Furthermore, biased MD at 300 K outperforms unbiased dynamics at 1200 K, efficiently covering the CV space before acquiring ~ 200 configurations. This observation is attributed to the gradual increase in driving forces induced by the uncertainty bias, resulting in a more gradual distortion of the atomic structure. In contrast, high-temperature unbiased simulations perturb the system more strongly and rapidly enter extrapolative regions without exploring relevant configurational changes. Thus, high-temperature simulations may also cause the degradation of the investigated atomic systems, unlike uncertainty-biased dynamics applied at mild physical conditions.

Figure 3b, c present energy and force RMSEs evaluated on the alanine dipeptide test data set; see Methods. Consistent with the findings in Fig. 3a, AL approaches combined with biased MD at 300 K outperform their unbiased counterparts at 300 K and 600 K once they acquire ~ 100 configurations. Biased AL experiments achieve energy RMSE of 1.97 meV $atom^{-1}$, close to those observed in high-temperature MD simulations, surpassing others by a factor of more than 13. A similar trend is observed for force RMSE. Biased AL experiments achieve an RMSE of 0.071 eV $Å^{-1}$, outperforming their counterparts at 300 K and 600 K by factors of 2.1 and 1.6, respectively.

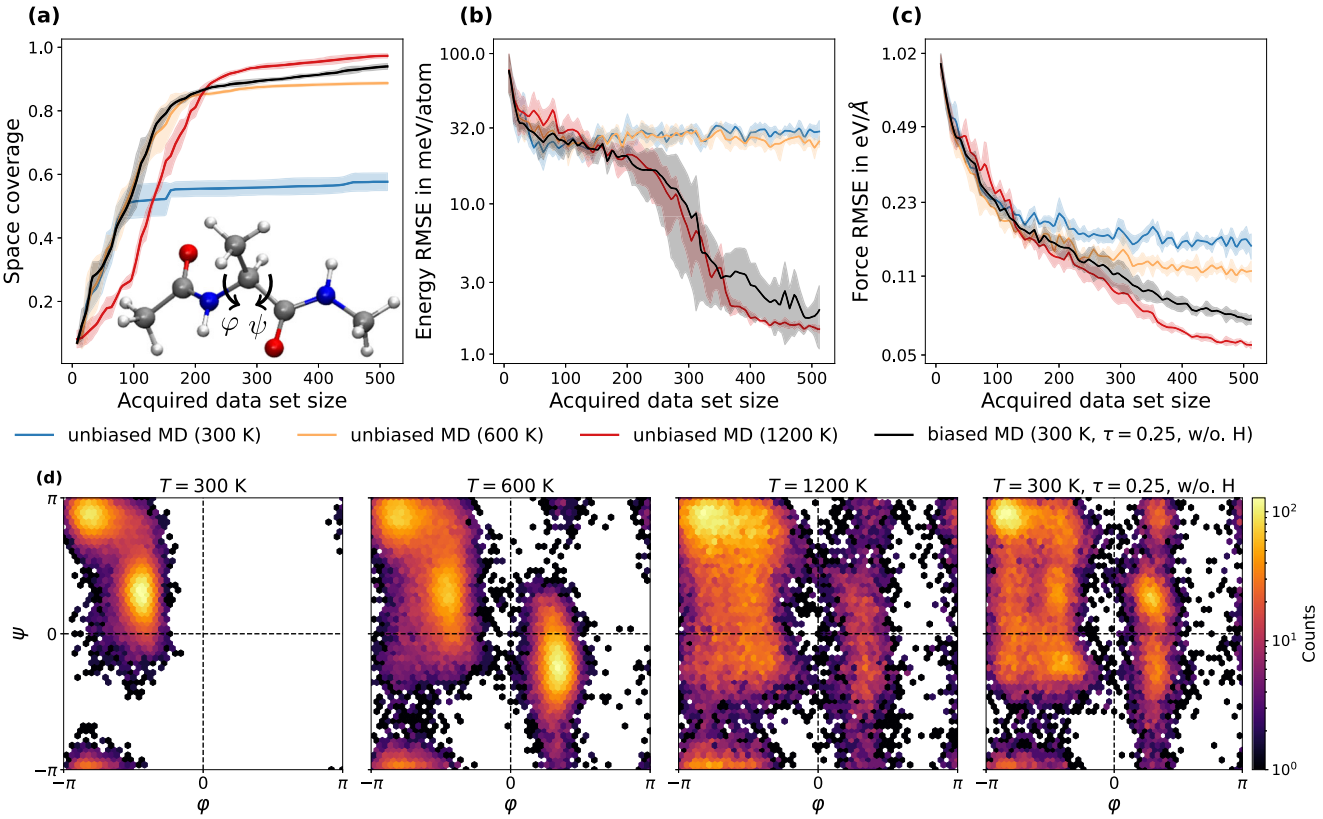

Fig. 3 | Comparison of AL approaches employing biased and unbiased MD simulations to generate the candidate pool of atomic configurations for alanine dipeptide. Results are provided for the posterior-based uncertainty quantification derived from sketched gradient features. Unlike unbiased MD simulations, which rely on atom-based uncertainties to terminate MD simulations, biased MD simulations use total and atom-based uncertainties to bias MD simulations and prompt their termination, respectively. We use three metrics to assess the performance of our AL approaches: (**a**) Coverage of the CV space; (**b**) Energy RMSE; and (**c**) Force RMSE. All RMSEs are evaluated on the alanine dipeptide test data set; see Methods. Shaded areas denote the standard deviation across five independent runs. The alanine dipeptide molecule, including its CVs, is shown as an inset in (**a**). The color code of the inset molecule is C grey, O red, N blue, and H white. **d** Ramachandran plots demonstrating the CV spaces explored by the four AL experiments. Biased MD simulations achieve exceptional performance, close to those of MD conducted at 1200 K, without knowledge of temperatures that accelerate transitions between stable states. The CV space covered by uncertainty-biased MD simulations at 300 K matches that of unbiased simulations at 1200 K, significantly outperforming the coverage achieved by unbiased MD at 300 K and 600 K.

These results demonstrate the efficiency of uncertainty-biased dynamics in exploring the configurational space and developing accurate and robust MLIPs. Moreover, generating training data that sufficiently covers the configurational space by combining AL with biased MD does not significantly increase the computational demand compared to conventional AL with unbiased MD; see the Supplementary Information. Lastly, MLIPs trained with candidates selected based on their uncertainty (and diversity) from biased and unbiased MD trajectories systematically outperform MLIPs trained with candidates selected at random; see Table 1.

Biased AL experiments achieve exceptional performance without knowledge of temperatures that accelerate transitions between stable states; see Fig. 3d. Identifying these temperatures requires running MD simulations at different conditions to explore the configurational space without degrading the atomic system. In contrast, given the mild physical conditions such as temperatures of 300 K and 600 K, biased MD simulations outperform their unbiased counterparts at 300 K and 600 K and achieve comparable performance to experiments at 1200 K for $\tau \lesssim 0.5$ and $0.2 \lesssim \tau \lesssim 0.4$, respectively. The available range of biasing strength values may be more restricted at more extreme conditions. Adding uncertainty bias to MD at 1200 K results in an even stronger system perturbation than during unbiased MD without yielding any improvement. For additional details, see the Supplementary Information.

Our results offer evidence of rare event exploration (the exploration of both stable states of alanine dipeptide) through uncertainty-biased dynamics. The following section will present a detailed analysis of the exploration rates. Additionally, we have identified how to further improve our biased MD simulations by making biasing strengths species dependent; see the Supplementary Information. The results presented in this section, achieved with a biasing strength of zero for hydrogen atoms, outperform settings where all atoms are biased equally, with improvements by a factor of 1.08 in coverage and 1.15 in force RMSE; see Table 1. Thus, a more sophisticated data-driven redistribution of biasing strengths can further enhance the performance of bias-forces-driven MD simulations. However, learning species-dependent biasing strengths necessitates defining a suitable loss function that promotes the fast exploration of phase space[60], which falls beyond the scope of this work.

Exploration rates for collective variables of alanine dipeptide

We have observed that uncertainty-biased MD simulations effectively explore the configurational space of alanine dipeptide, defined by its CVs. Figure 4 evaluates the extent to which the introduced bias forces in MD simulations accelerate their exploration. In Fig. 4a, we present the coverage of the CV space as a function of simulation time, i.e., of the effective number of MD steps. The figure demonstrates that uncertainty-biased AL experiments at 300 K outperform unbiased experiments at 300 K and 600 K. They achieve the same coverage in considerably shorter simulation times, thereby enhancing exploration rates by a factor of larger than two. At the same time, biased MD simulations yield results comparable to those obtained from unbiased MD simulations at 1200 K. Thus, uncertainty-biased MD explores configurational space at a similar rate to unbiased MD at 1200 K.

Table 1 | CV space coverage, atomic energy (E-) and atomic force (F-) RMSEs, as well as position (Pos.) and uncertainty (Unc.) auto-correlation times (ACTs) for alanine dipeptide experiments conducted with posterior-based uncertainties

Experiment	CV space cov.	E-RMSE	F-RMSE	Pos. ACT[a]	Unc. ACT[a]
random sel. (300 K)	0.58 ± 0.03	34.09 ± 6.29	0.191 ± 0.019	–	–
random sel. (600 K)	0.76 ± 0.04	31.44 ± 4.77	0.143 ± 0.015	–	–
random sel. (1200 K)	0.95 ± 0.01	19.83 ± 4.62	0.116 ± 0.017	–	–
unbiased MD (300 K)	0.58 ± 0.03	30.29 ± 5.47	0.149 ± 0.019	2.07 ± 0.11	327.11 ± 8.69
unbiased MD (600 K)	0.89 ± 0.00	26.03 ± 2.23	0.116 ± 0.012	1.23 ± 0.02	257.88 ± 22.01
unbiased MD (1200 K)	**0.97 ± 0.01**	**1.47 ± 0.09**	**0.055 ± 0.002**	0.74 ± 0.02	21.41 ± 4.91
biased MD (300 K, $\tau = 0.25$, w. H)	0.87 ± 0.02	5.09 ± 5.40	0.082 ± 0.016	2.08 ± 0.13	**19.38 ± 7.42**
biased MD (300 K, $\tau = 0.25$, w/o. H)	0.94 ± 0.01	1.97 ± 0.88	0.071 ± 0.003	**0.69 ± 0.04**	52.79 ± 19.40

[a]ACTs computed for experiments with the random selection (random sel.) strategy are excluded from the analysis because different approaches may introduce systematic biases, making the comparison unreliable.

E- and F-RMSEs are reported for MLIPs obtained at the end of each experiment, while CV space coverage and ACTs are computed using the entire trajectory obtained throughout the experiment. E-RMSE is given in meV atom^{-1}, while F-RMSE is in eV Å^{-1}. All E-RMSE and F-RMSE values are computed for the test data set obtained from a long MD trajectory at 1200 K; see Methods. ACTs are provided in ps. For biased MD, we compare two cases: one with (w.) biasing hydrogen atoms and one without (w/o.). We also compare biased and unbiased MD with experiments that involve the random selection (random sel.) strategy for acquiring training data. The best performance is highlighted in bold, and the second-best performance is underlined.

The exploration rates estimated from Fig. 4a provide an approximate measure of how uncertainty-biased dynamics accelerate the exploration of configurational space. To offer a more thorough assessment, we examine auto-correlation functions (ACFs) computed for both position and uncertainty spaces in Fig. 4b, c. Here, a faster decay corresponds to a faster exploration of the respective space. We compute ACFs using MD trajectories from all AL iterations. Additionally, we calculate the auto-correlation time (ACT) for each experiment. For the definition of ACF and ACT, see Methods. Table 1 presents ACTs for all AL experiments. Smaller ACTs correspond to a faster decay of ACFs, indicating a faster exploration of the respective spaces.

ACTs demonstrate that uncertainty-biased MD at 300 K explores position and uncertainty spaces two to six times faster than unbiased MD at 300 K and 600 K. Compared to unbiased MD at 1200 K, it achieves comparable exploration rates in the position space and rates lower by a factor of two for the uncertainty space. Biasing hydrogen atoms reduces the uncertainty ACT compared to experiments with zero hydrogen biasing strength but increases the position ACT by a factor of three. Thus, stronger atomic bond distortions, resulting in fast exploration of extrapolative regions, can explain a shorter uncertainty ACT of unbiased MD at 1200 K. While this effect can be unfavorable for promoting the exploration of rare events in biased MD, incorporating small, non-zero biasing strengths for hydrogen atoms may be necessary to ensure the robustness of MD simulations at elevated temperatures. Interestingly, we observe that uncertainty-biased MD explores both stable states in alanine dipeptide, even though 27 degrees of freedom (C, N, and O atoms) were effectively biased, demonstrating its remarkable efficiency.

To gain insight into the exploration of the CV space during AL, we refer to Fig. 4d, e, which illustrate the time evolution of the maximal atom-based uncertainty and the CV space coverage for selected AL iterations. Biased MD systematically explores configurations with higher uncertainty values than unbiased MD at 300 K and 600 K. Furthermore, bias forces drive the exploration of both stable states of alanine dipeptide and promote transitions between them, similar to higher temperatures in unbiased MD. Later AL iterations in Fig. 4d, e demonstrate that MD driven by bias forces reduces the uncertainty level uniformly across the configurational space. Thus, given the correlation between uncertainties and actual errors, uncertainty-biased MD generates MLIPs uniformly accurate across the configurational space.

Performance of bias-stress-driven active learning

Generating training data for bulk material systems with large unit cells and multiple stable states poses a significant challenge in developing MLIPs. Therefore, we assess the performance of the bias-stress-driven AL applied to MIL-53(Al), a flexible MOF that undergoes reversible, large-amplitude volume changes under external stimuli, such as temperature and pressure (see Fig. 5). MIL-53(Al) features two stable phases: the closed-pore state with a unit cell volume of $V \sim 830$ Å^3 and the large-pore state with $V \sim 1419$ Å^3. For reference energy, force, and stress calculations, we use the CP2K simulation package (version 2023.1)[61] and DFT at the PBE-D3(BJ) level[62,63]. Our baseline for generating candidate pools for AL involves unbiased MD and training data selected based on their uncertainty (and diversity) or at random. We also employ metadynamics[41], which uses an adaptive biasing strategy for cell parameters of MIL-53(Al), as a baseline. We assess the performance of MLIPs for MIL-53(Al) using the test data set presented by ref. 41.

Figure 5a–c demonstrate the performance of MLIPs developed for MIL-53(Al) depending on the number of acquired configurations. Table 2 presents error metrics evaluated for MLIPs at the end of each experiment. Here, we present results for the posterior-based uncertainty. The Supplementary Information presents equivalent results for other uncertainty methods and pressures. We observe that MLIPs trained with configurations generated using metadynamics outperform the others for data set sizes below ~ 200 samples. This difference in performance can be attributed to how perturbed configurations are generated and the differing experimental settings between incremental learning and AL applied here. Bias-stress-driven AL outperforms metadynamics-based experiments after acquiring ~ 200 configurations regarding force and stress RMSEs.

Metadynamics-based experiments achieve performance on par with unbiased AL experiments conducted at 0 MPa after they reach a data set size of ~ 200 configurations. For uncertainty-biased MD, the force RMSE improves by a factor of 1.14, and the stress RMSE improves by a factor of two compared to zero-pressure unbiased MD. Furthermore, AL experiments with biased MD simulations outperform unbiased MD simulations at 250 MPa regarding stress RMSE. Thus, bias-stress-driven MD generates a data set that better represents the relevant configurational space of flexible MOFs compared to MLIPs trained with conventional MD and metadynamics simulations. This improvement is achieved without significantly increasing the computational cost of data generation; see the Supplementary Information. Lastly, similar to the results obtained for alanine dipeptide, AL with a more advanced selection strategy outperforms experiments where training data is picked at random; see Table 2.

Figure 5d, e show the main advantage of biased MD simulations over unbiased and metadynamics-based approaches. While exploring the large-pore state less frequently than metadynamics-based counterparts, bias-stress-driven MD spans a broader range of volumes and uniformly reduces energy, force, and stress RMSEs across the entire volume space. Compared to zero-pressure unbiased MD simulations, it promotes the exploration of the large-pore state. However, this state can be modeled using atomic

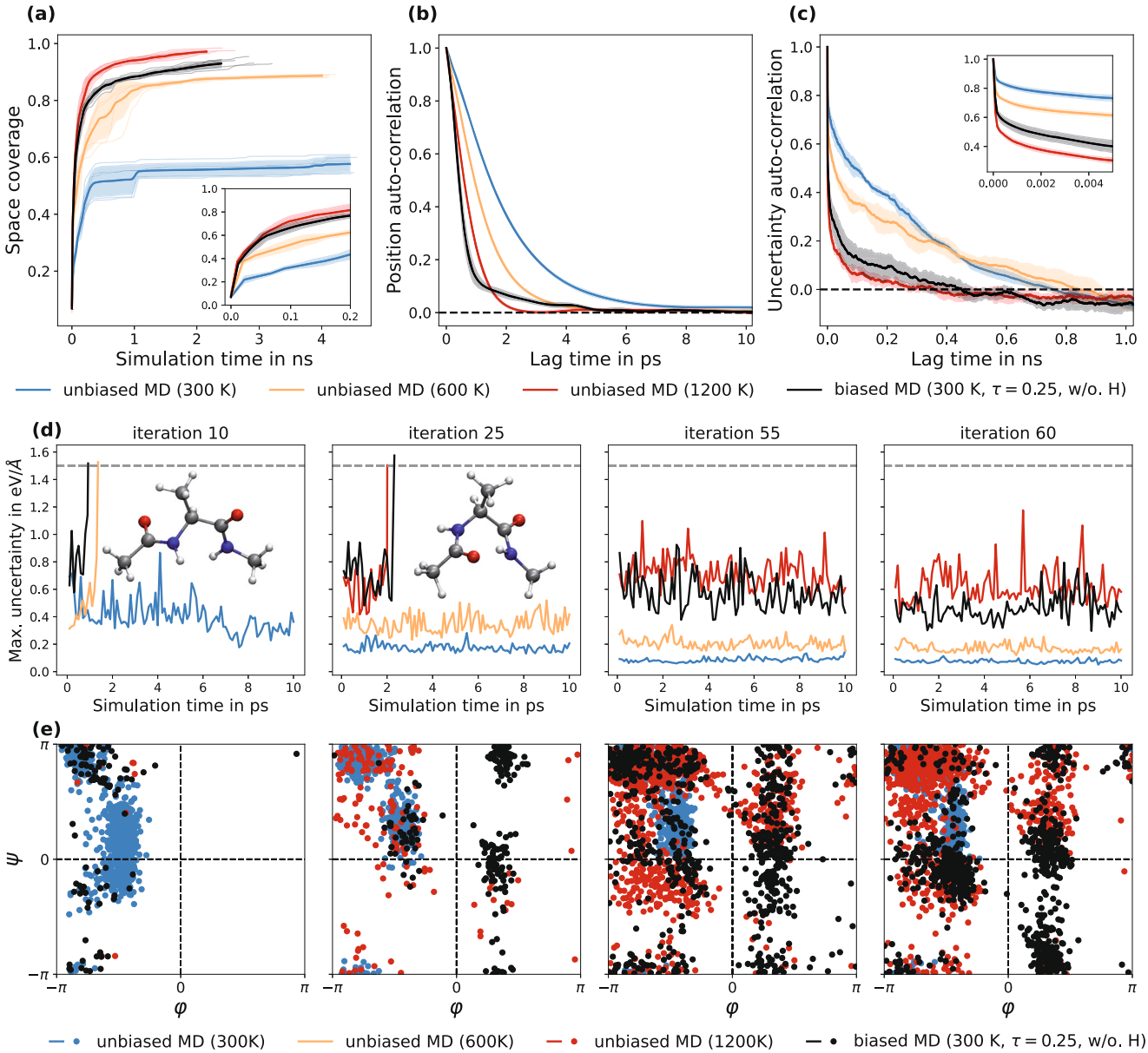

Fig. 4 | Evaluation of CV space exploration rates for biased and unbiased MD simulations of alanine dipeptide. Here, MD simulations generate candidate pools of atomic configurations for AL algorithms. Results are provided for the posterior-based uncertainty quantification derived from sketched gradient features. Unlike unbiased MD simulations, which rely on atom-based uncertainties to terminate MD simulations, biased MD simulations use total and atom-based uncertainties to bias MD simulations and prompt their termination, respectively. We use three metrics to asses the exploration rates: (**a**) Coverage of the CV space; (**b**) Auto-correlation functions of atomic positions; and (**c**) Auto-correlation functions of atom-based uncertainties. Shaded areas denote the standard deviation across five independent runs. **d** Time evolution of the maximal atom-based uncertainty within an AL iteration and the entire experiment. Time evolution is shown for one of the eight MD simulations. The dashed gray line represents the uncertainty threshold of 1.5 eV Å^{-1}. The insets show configurations that reached the uncertainty threshold for uncertainty-biased MD. **e** Ramachandran plots illustrate the exploration of the CV space over AL iterations and the entire experiment. Ramachandran plots are presented for unbiased MD simulations at 300 K and 1200 K and biased MD simulations at 300 K. Simulation time refers to the effective number of MD steps (× 0.5 fs) required to reach the final coverage, while lag time denotes the time interval between two successive MD frames. Biased MD simulations at 300 K exhibit at least two times higher exploration rates than their unbiased counterparts at 300 K and 600 K. Their exploration rates are comparable to those of unbiased MD simulations at 1200 K, with the advantage of gradually distorting the molecule, reducing the risk of its degradation.

environments from the closed-pore one. Thus bias stress does not excessively favor exploration of the former. Instead, it drives the dynamics more toward smaller volumes, for which all other approaches tend to predict energy, force, and stress values with larger errors. Note that, in Fig. 5e, we reduce the temperature to 300 K and initiate AL experiments with 256 configurations, each having a unit cell volume below 1200 Å^3 (drawn from the training data in ref. 41). Using a lower temperature and learning the configurational space around the closed-pore state is required to decrease the probability of MD simulations exploring the large-pore stable state of MIL-53(Al). In contrast, we found that using randomly perturbed atomic configurations can lead to underestimated energy barriers by MLIPs, thus facilitating the transition between both stable phases in initial AL iterations.

These results show that uncertainty-biased MD simulations aim to uniformly reduce errors across the relevant configurational space and promote the simultaneous exploration of extrapolative regions and

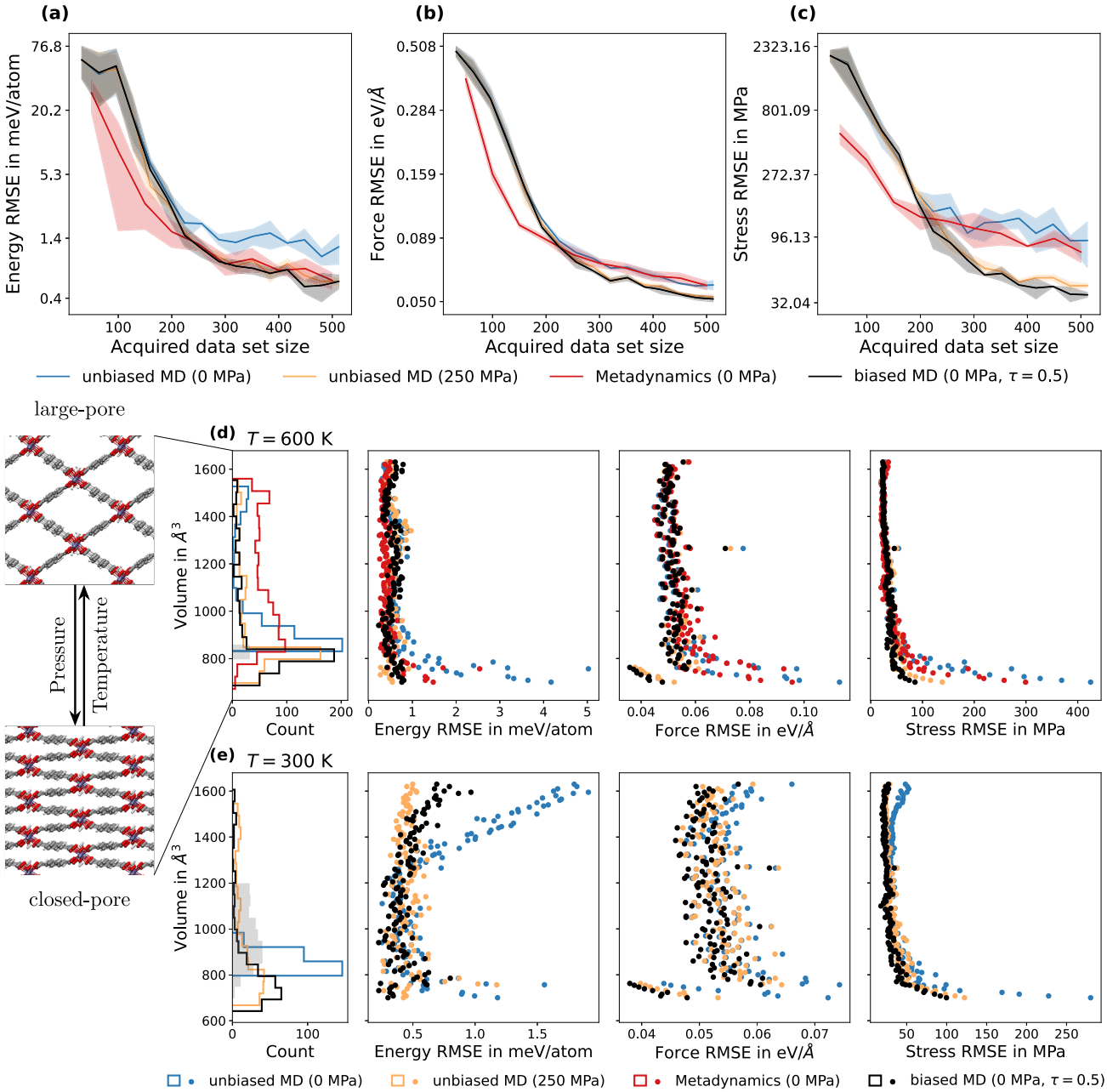

Fig. 5 | Comparison of AL approaches employing biased and unbiased MD simulations to generate the candidate pool of atomic configurations for MIL-53(Al). Results are provided for the posterior-based uncertainty quantification derived from sketched gradient features. Unlike unbiased MD simulations, which rely on atom-based uncertainties to terminate MD simulations, biased MD simulations use total and atom-based uncertainties to bias MD simulations and prompt their termination, respectively. We use three metrics to assess the performance of our AL approaches: (**a**) Energy RMSE; (**b**) Force RMSE; and (**c**) Stress RMSE. All RMSEs are evaluated on the MIL-53(Al) test data set[41]. Shaded areas denote the standard deviation across three independent runs, except for metadynamics. For it, shaded areas denote standard deviation across three randomly initialized MLIPs. **d** Volume distribution for atomic configurations acquired during MD at 600 K, along with volume-dependent energy, force, and stress RMSEs. **e** Volume distribution for configurations acquired during MD at 300 K, along with volume-dependent energy, force, and stress RMSEs. We employ a temperature of 300 K to reduce the probability of exploring the large-pore state of MIL-53(Al). Bias-stress-driven MD simulations outperform metadynamics-based simulations with adaptive biasing of the cell parameters. Metadynamics aims to cover the volume space uniformly. In contrast, uncertainty-biased MD generates training data sets that uniformly reduce energy, force, and stress RMSEs. Additionally, biased MD simulations enhance the exploration of closed- and large-pore states of MIL-53(Al) shown in the inset of (**d**).

transitions between stable states. Also, under selected physical conditions ($T = 600$ K and $p = 0$ MPa), the performance of our uncertainty-biased MD exhibits low sensitivity to stress biasing strength values for $\tau \geq 0.5$; see the Supplementary Information. Metadynamics, in contrast, may require longer simulation times to generate equivalent candidate pools as it focuses on generating configurations uniformly distributed in the CV space, which is unnecessary for developing MLIPs.

Exploration rates for cell parameters of MIL-53(Al)

Figure 6 assesses the extent to which uncertainty-biased (bias stress) MD simulations enhance the exploration of the extensive volume space of MIL-53(Al). In Fig. 6a, we observe a higher frequency of transitions between stable phases for biased MD simulations than for zero-pressure counterparts. Additionally, uncertainty-biased simulations favor the exploration of smaller MIL-53(Al) volumes, in line with the

Table 2 | Atomic energy (E-), atomic force (F-), and stress (S-) RMSEs, as well as position (Pos.) and uncertainty (Unc.) autocorrelation times (ACTs) for MIL-53(Al) experiments conducted with posterior-based uncertainties

Experiment	E-RMSE	F-RMSE	S-RMSE	Pos. ACT[a]	Unc. ACT[a]
$T = 600$ K					
random sel. (0 MPa)	1.62 ± 0.52	0.062 ± 0.002	145.3 ± 35.49	–	–
random sel. (250 MPa)	0.84 ± 0.09	0.057 ± 0.001	63.8 ± 15.82	–	–
unbiased MD (0 MPa)	1.17 ± 0.36	0.058 ± 0.002	90.81 ± 32.82	10.60 ± 9.54	88.05 ± 2.53
unbiased MD (250 MPa)	**0.57 ± 0.05**	0.052 ± 0.001	42.72 ± 1.37	**2.08 ± 0.58**	66.32 ± 2.02
Metadynamics (0 MPa)	0.58 ± 0.10	0.058 ± 0.002	74.83 ± 11.89	–	–
biased MD (0 MPa, $\tau = 0.5$)	**0.57 ± 0.08**	**0.051 ± 0.001**	**36.60 ± 1.46**	2.75 ± 0.46	**44.87 ± 14.08**
$T = 300$ K					
random sel. (0 MPa)	1.04 ± 0.26	0.058 ± 0.001	70.49 ± 6.61	–	–
random sel. (250 MPa)	0.58 ± 0.08	0.055 ± 0.002	52.19 ± 2.22	–	–
unbiased MD (0 MPa)	0.88 ± 0.20	0.056 ± 0.001	58.57 ± 5.94	3.45 ± 4.06	99.25 ± 10.34
unbiased MD (250 MPa)	**0.48 ± 0.01**	0.054 ± 0.000	39.88 ± 1.76	**1.86 ± 0.14**	54.56 ± 4.82
biased MD (0 MPa, $\tau = 0.5$)	0.49 ± 0.09	**0.052 ± 0.001**	**33.89 ± 3.06**	42.92 ± 14.18	**26.89 ± 8.94**

[a]ACTs computed for experiments with the random selection (random sel.) strategy are excluded from the analysis because different approaches may introduce systematic biases, making the comparison unreliable.

E-, F-, and S-RMSEs are reported for MLIPs obtained at the end of each experiment, while ACTs are computed using the entire trajectory sampled throughout the experiment. E-RMSE is given in meV atom^{-1}, F-RMSE in eV Å^{-1}, and S-RMSE in MPa. All E-RMSE, F-RMSE, and S-RMSE values are computed for the test data set obtained based on first principles MD trajectories at 600 K; see ref. 41. ACTs are provided in ps. We also compare biased and unbiased MD with experiments that involve the random selection (random sel.) strategy for acquiring training data. The best performance is highlighted in bold, and the second-best performance is underlined.

results shown in Fig. 5. Figure 6b, c present ACFs for position and uncertainty spaces, with estimated ACTs provided in Table 2. Here, a faster decay of ACFs corresponds to shorter ACTs and indicates a faster exploration of the respective space. These results indicate that bias-stress-driven MD is at least as efficient as high-pressure MD simulations in exploring both spaces. Figure 6d demonstrates the time evolution of energy, force, and stress RMSEs. It reveals that local atomic environments in the large-pore state are well represented by those in the closed-pore state, explaining the stronger preference for smaller volumes by biased MD; see Figs. 6a and 5d, e. This effect is evident from the low force and stress RMSEs in the early AL iterations for the large-pore state, even though this state has not been explored yet. Furthermore, uncertainty-biased MD simulations surpass the performance of their counterparts already in the early stages by aiming to reduce errors across the test volume space uniformly.

From these results and the findings in Fig. 5d, we conclude that bias-stress-driven MD significantly enhances the exploration of the relevant configurational space, including rare events (i.e., transitions between stable phases). However, in Table 2, we obtained longer ACTs for biased MD at 300 K compared to its unbiased counterparts, which contradicts our previous arguments. When examining the ACF shown in Fig. 7, it becomes evident that a stronger correlation in the position space results from the volume fluctuations induced in MIL-53(Al) by the bias stress. These fluctuations can be represented by a sine wave with additive random noise and a period twice the simulation's length; see Methods. This observation implies that bias stress induces correlated motions in the MIL-53(Al) system, causing it to expand and contract alternately for half of the simulation time. This phenomenon results in periodic exploration of small and large volumes within the configurational space.

In contrast to the conventional approaches, including the bias-forces-driven MD simulations, which aim for uncorrelated random-walk-like behavior of predetermined CVs to capture configurational changes, our method introduces correlated motion that explores the entire configurational space. Increasing the amplitude of random noise in the sine wave reduces the amplitude of these fluctuations in the ACF, similar to raising the temperature in an atomic system. This decrease in the amplitude explains why this effect is not observed in Fig. 6b.

Discussion

This work investigates an uncertainty-driven AL approach for data set generation, facilitating the development of high-quality MLIPs for chemically complex atomic systems. We employ uncertainty-biased MD simulations to generate candidate pools for AL algorithms. Our results show that applying uncertainty bias facilitates simultaneous exploration of extrapolative regions and rare events. Efficient exploration of both is crucial in constructing comprehensive training data sets, enabling the development of uniformly accurate MLIPs. In contrast, classical enhanced sampling techniques (e.g., metadynamics) or unbiased MD simulations at elevated temperatures and pressures often cannot simultaneously explore extrapolative regions and rare events. Enhanced sampling techniques were designed to ensure the reconstruction of the underlying Boltzmann distribution. However, this property is unnecessary for data set generation and may limit their effectiveness in this context.

The performance of enhanced sampling techniques depends on the manual definition of hyper-parameters, e.g., CVs for metadynamics. Setting them requires expert knowledge because the wrong choice can limit the range of explored configurations. Uncertainty-biased MD only needs to define an uncertainty threshold and biasing strength. Both parameters influence the exploration rate of configurational space without constraining the space that can be explored. Under milder conditions, uncertainty-biased MD simulations outperform their unbiased counterparts for a broad range of biasing strength values, making the latter's choice more accessible. Yet, the dependence of the performance on the biasing strength value becomes more noticeable under extreme conditions, sometimes with no improvement by adding uncertainty bias to MD. A similar behavior can also be expected for metadynamics simulations[64]. Additionally, employing species-dependent biasing strength can restrict biasing in sensitive configurational regions, e.g., biasing hydrogen atoms.

Identifying extreme conditions like high temperatures and pressures can also accelerate phase space exploration in unbiased MD. However, a wrong choice of temperature and pressure may result in unphysical force predictions and degradation of the atomic system. In contrast, uncertainty-biased MD, conducted under milder conditions, explores relevant phase space at rates comparable to those obtained under extreme conditions and reduces the risk of degrading the atomic

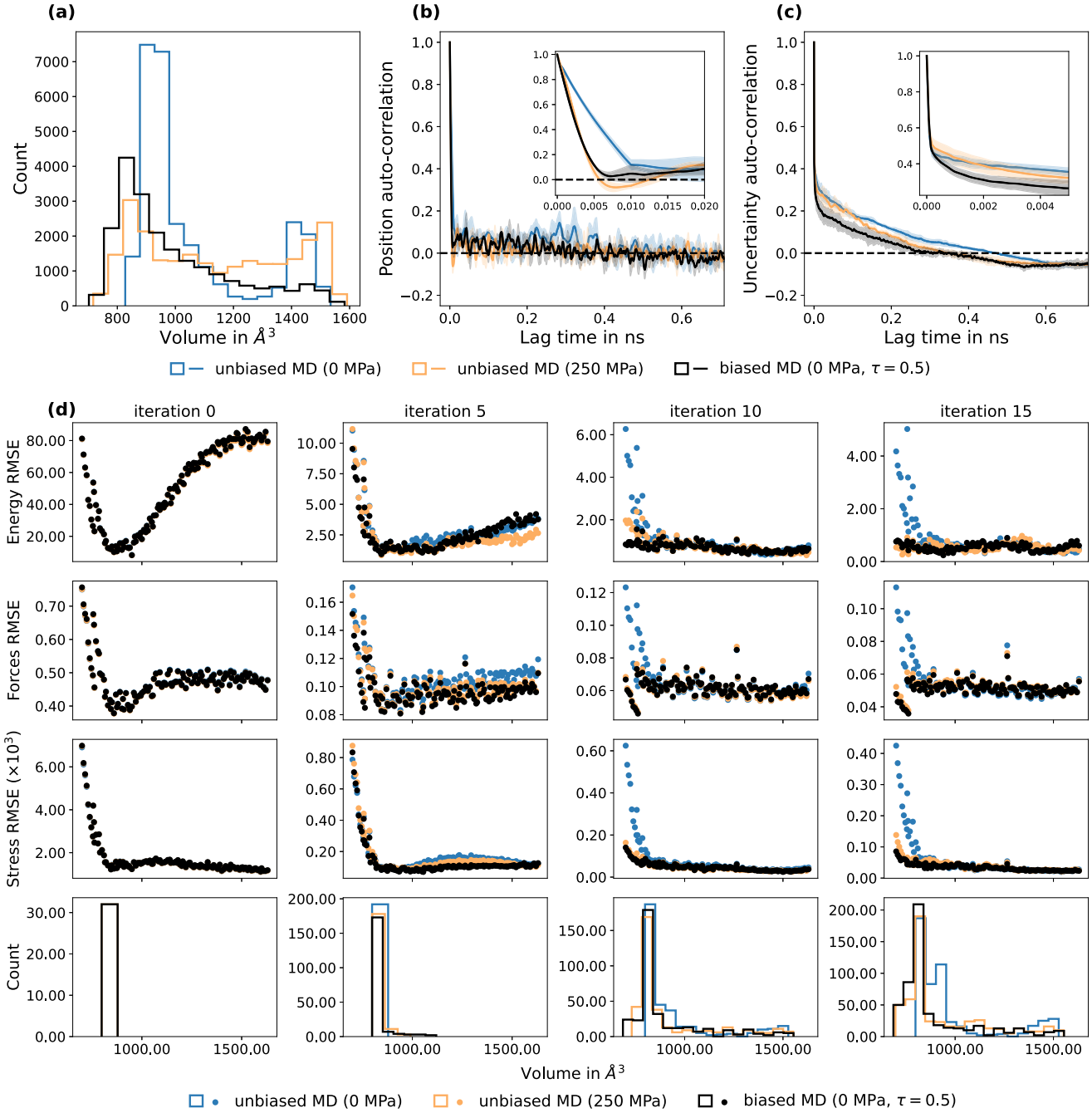

Fig. 6 | Evaluation of configurational space exploration rates for biased and unbiased MD simulations of MIL-53(Al). Here, MD simulations generate candidate pools of atomic configurations for AL algorithms. Results are provided for the posterior-based uncertainty quantification derived from sketched gradient features. Unlike unbiased MD simulations, which rely on atom-based uncertainties to terminate MD simulations, biased MD simulations use total and atom-based uncertainties to bias MD simulations and prompt their termination, respectively. We use three metrics to asses the exploration rates: (**a**) Volume distribution of configurations sampled throughout the experiment; (**b**) Auto-correlation functions for positions; and (**c**) Auto-correlation functions for atom-based uncertainties. Shaded areas denote the standard deviation across three independent runs. **d** Time evolution of the volume distribution of configurations acquired during training and of energy, force, and stress RMSEs evaluated on the test data set[41] depending on the unit cell volume. Bias-stress-driven MD simulations achieve exploration rates comparable to those of high-pressure unbiased MD simulations. They aim to reduce RMSEs uniformly across the entire volume space, even in the early stages of AL, surpassing the performance of unbiased simulations.

system. As mentioned, uncertainty-biased MD simulations outperform their unbiased counterparts for a broad range of biasing strength values in our setting. Furthermore, while evaluating uncertainty gradients increases the inference times by a factor of 1.4 to 1.7 compared to unbiased MD, applying uncertainty bias leads to, on average, shorter MD simulations. Thus, the difference in the computational cost between biased and unbiased MD is typically insignificant.

We compare uncertainty quantification methods, including the variance of an ensemble of MLIPs, and ensemble-free methods derived from sketched gradient features, focusing on configurational space exploration rates and generating uniformly accurate potentials; see the Supplementary Information. Overall, gradient-based approaches yield MLIPs with similar performance to those created using ensemble-based uncertainty while significantly reducing the computational cost of uncertainty quantification.

For MIL-53(Al), we find that ensemble-based uncertainties overestimate the force error more strongly than gradient-based approaches, resulting in earlier termination of MD simulations and potentially worse configurational space exploration. For alanine dipeptide, using an ensemble of MLIPs improves their robustness during MD simulations, facilitating CV space exploration. Therefore, improving the robustness of a single MLIP during an MD simulation is a promising research direction[65], combined with the proposed ensemble-free techniques.

While this study thoroughly investigates AL with uncertainty-biased MD for generating candidate pools, further research is still necessary. For example, one should analyze how well uncertainty-biased MD explores a configurational space with multiple stable states and how it identifies the respective slow modes using solely uncertainty bias. Also, assessing the uniform accuracy of resulting MLIPs and the enhanced exploration in higher-dimensional CV spaces remains challenging. Furthermore, the applicability of the proposed data generation approach to more complex molecular and material systems, such as biological polymers[66] and multicomponent alloys[5], is yet to be explored. Unlike MD, Monte Carlo simulations generally allow significant configurational changes, eliminating the need to explore intermediate transition paths. Combined with uncertainty bias, they might avoid exploring intermediate, low-uncertainty transition regions, improving the efficiency of uncertainty-driven data generation. Lastly, the extent to which MLIPs based on graph NNs can enhance the efficiency of the proposed data generation approach remains to be seen.

Methods

Machine-learned interatomic potentials

We define an atomic configuration, $S = \{\mathbf{r}_i, Z_i\}_{i=1}^{N_{\text{at}}}$, where $\mathbf{r}_i \in \mathbb{R}^3$ are Cartesian coordinates and $Z_i \in \mathbb{N}$ is the atomic number of atom i, with a total of N_{at} atoms. Our focus lies on interatomic NN potentials, which map an atomic configuration to a scalar energy E. The mapping is denoted as $f_{\boldsymbol{\theta}} : S \mapsto E \in \mathbb{R}$, where $\boldsymbol{\theta}$ denotes the trainable parameters. By assuming the locality of interatomic interactions, we decompose the total energy of the system into individual atomic contributions[13]

$$E(S, \boldsymbol{\theta}) = \sum_{i=1}^{N_{\text{at}}} E_i(S_i, \boldsymbol{\theta}), \tag{1}$$

where S_i is the local environment of atom i, defined by the cutoff radius r_c. The trainable parameters $\boldsymbol{\theta}$ are learned from atomic data sets containing atomic configurations and their energies, atomic forces, and stress tensors.

Gradient-based uncertainties

We quantify the uncertainty of a trained MLIP by expanding its energy per atom $E_{\text{at}} = E/N_{\text{at}}$ around the locally optimal parameters $\boldsymbol{\theta}^*$[50–52]

$$E_{\text{at}}(S, \boldsymbol{\theta}) \approx E_{\text{at}}(S, \boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \underbrace{\nabla_{\boldsymbol{\theta}} E_{\text{at}}(S, \boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}}_{=\boldsymbol{\phi}(S)}, \tag{2}$$

where S denotes an atomic configuration as defined in the previous section. Gradient features $\boldsymbol{\phi}(S) \in \mathbb{R}^{N_{\text{feat}}}$ can be interpreted as the sensitivity of the energy to small parameter perturbations. Here, N_{feat} is the number of trainable parameters of the MLIP. We employ the energy per atom E_{at} in Eq. (2), as it accounts for the extensive nature of the energy, whose value depends on the system size. This choice ensures that uncertainties defined using gradient features do not favor the selection of larger structures. Gradient features can also be expressed as the mean of their atomic contributions: $\boldsymbol{\phi} = \sum_{i=1}^{N_{\text{at}}} \boldsymbol{\phi}_i / N_{\text{at}}$. For atomic gradient features $\boldsymbol{\phi}_i$, using the energy per atom in Eq. (2) is unnecessary. Here, we use $\boldsymbol{\phi} = \boldsymbol{\phi}(S)$ and $\boldsymbol{\phi}_i = \boldsymbol{\phi}_i(S_i)$, with S_i denoting the local environment of an atom i, to simplify the notation. Thus, gradient features can be used to quantify uncertainties in total and atom-based properties of an atomic system, such as energy and atomic forces, respectively.

Particularly, we define the atom-based model's uncertainty (atomic forces) by employing squared distances between atomic gradient features

$$u_i^2 = \min_{\boldsymbol{\phi}_j \in \Phi_{\text{train}}} \|\boldsymbol{\phi}_i - \boldsymbol{\phi}_j\|_2^2. \tag{3}$$

Alternatively, we consider Bayesian linear regression in Eq. (2) and compute the posterior uncertainty as

$$u_i^2 = \lambda^2 \boldsymbol{\phi}_i^\top \left(\Phi_{\text{train}}^\top \Phi_{\text{train}} + \lambda^2 \mathbf{I}\right)^{-1} \boldsymbol{\phi}_i, \tag{4}$$

where λ is the regularization strength. Here, we define $\Phi_{\text{train}} = \boldsymbol{\phi}_j(\mathscr{X}_{\text{train}}) \in \mathbb{R}^{(N_{\text{at}} \cdot N_{\text{train}}) \times N_{\text{feat}}}$ with $\mathscr{X}_{\text{train}}$ denoting the local atomic environments of configurations in the training set of size N_{train}. In this work, we refer to our uncertainties as distance- and posterior-based uncertainties. Equivalent results can be obtained for total uncertainties (energy), employing gradient features $\boldsymbol{\phi} = \sum_{i=1}^{N_{\text{at}}} \boldsymbol{\phi}_i / N_{\text{at}}$ with $\Phi_{\text{train}} = \boldsymbol{\phi}(\mathscr{X}_{\text{train}}) \in \mathbb{R}^{N_{\text{train}} \times N_{\text{feat}}}$.

Calculating uncertainties using gradient features is computationally challenging, especially for the posterior-based approach, for which a single uncertainty evaluation scales as $\mathcal{O}(N_{\text{feat}}^2)$. Therefore, we employ the sketching technique[55] to reduce the dimensionality of gradient features, i.e., $\boldsymbol{\phi}_i^{\text{rp}} = \mathbf{U}\boldsymbol{\phi}_i \in \mathbb{R}^{N_{\text{rp}}}$ with N_{rp} and $\mathbf{U} \in \mathbb{R}^{N_{\text{rp}} \times N_{\text{feat}}}$ denoting the number of random projections and a random matrix, respectively[51,52]. In previous work[51], we have observed that uncertainties derived from sketched gradient features demonstrate a better correlation with RMSEs of related properties than those based on last-layer features[50,67,68]. More details on sketched gradient features can be found in the following sections. Atom-based uncertainties, defined by the distances between gradient features, scale linearly with both the system size and the number of training structures, i.e., as $\mathcal{O}(N_{\text{at}} N_{\text{train}})$. Consequently, they require an additional approximation to ensure computational efficiency. To address this, we employed the batch selection algorithm that maximizes distances within the training set, allowing us to identify the most representative subset of atomic gradient features; see the following sections.

Uncertainty-biased molecular dynamics

Following previous work[42,43], we define the biased energy as

$$E^{\text{biased}}(S, \boldsymbol{\theta}) = E(S, \boldsymbol{\theta}) - \tau u(S, \boldsymbol{\theta}), \tag{5}$$

where τ denotes the biasing strength. The negative sign ensures that negative uncertainty gradients with respect to atomic positions (bias forces) drive the system toward high uncertainty regions; see Fig. 1c. In this work, we use AD to compute bias forces acting on atom i, denoted as $-\nabla_{\mathbf{r}_i} u(S, \boldsymbol{\theta})$ with atomic positions $\mathbf{r}_i$. The total biased force on atom i reads

$$\mathbf{F}_i^{\text{biased}}(S, \boldsymbol{\theta}) = \mathbf{F}_i(S, \boldsymbol{\theta}) + \tau \nabla_{\mathbf{r}_i} u(S, \boldsymbol{\theta}). \tag{6}$$

These biased forces can be used for MD simulations in, e.g., canonical (NVT) statistical ensemble to bias the exploration of the configurational space.

In the case of bulk atomic systems, the configurational space often includes variations in cell parameters, which define the shape and size of the unit cell, necessitating enhanced exploration of them. For this purpose, we propose the concept of bias stress, defined by

$$\frac{1}{V} \nabla_{\boldsymbol{\epsilon}} u(S, \boldsymbol{\theta})|_{\boldsymbol{\epsilon}=\mathbf{0}},$$

with V denoting the volume of the periodic cell. This expression is motivated by the definition of the stress tensor[69]. Here, $u(S, \boldsymbol{\theta})$ denotes the uncertainty after a strain deformation of the bulk atomic system with the symmetric tensor $\boldsymbol{\epsilon} \in \mathbb{R}^{3\times3}$, i.e., $\tilde{\mathbf{r}} = (\mathbf{1} + \boldsymbol{\epsilon}) \cdot \mathbf{r}$. The calculation of the bias stress is

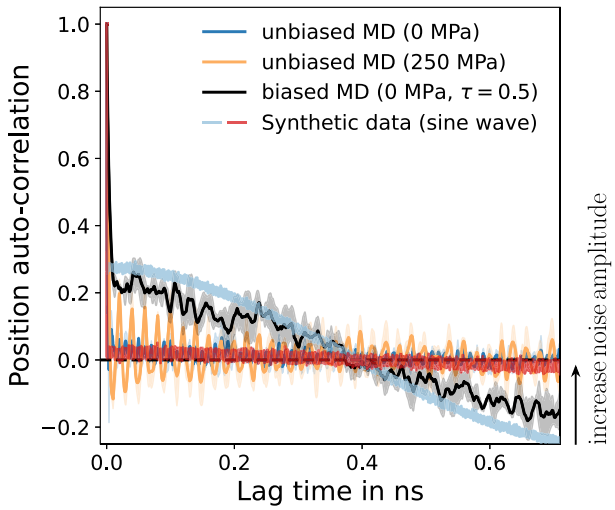

Fig. 7 | Position ACF obtained by running biased and unbiased MD simulations at 300 K for MIL-53(Al). Shaded areas denote the standard deviation across three independent runs. We employ a temperature of 300 K to reduce the probability of exploring the large-pore state of MIL-53(Al). The ACF exhibits strongly correlated motions attributed to volume fluctuations induced by the bias stress. These fluctuations can be modeled by a sine wave with a period twice the length of the simulation. The red line denotes a sine wave with a larger noise amplitude than the one denoted by the blue line.

straightforward with AD. The total biased stress reads

$$\boldsymbol{\sigma}^{\text{biased}}(S,\boldsymbol{\theta}) = \boldsymbol{\sigma}(S,\boldsymbol{\theta}) - \tau\frac{1}{V}\nabla_{\epsilon}u(S,\boldsymbol{\theta})|_{\epsilon=\mathbf{0}}. \tag{7}$$

The bias stress tensor in Eq. (7) effectively reduces the internal pressure in the bulk atomic system. We propose combining the bias stress tensor with MD simulations conducted in isothermal-isobaric (NpT) statistical ensemble to enhance the data-driven exploration of cell parameters and pressure-induced transitions in bulk materials.

Uncertainty gradients exhibit different magnitudes compared to energy gradients. Thus, re-scaling uncertainty gradients is necessary to ensure consistent driving toward uncertain regions. Building upon the approach introduced in ref. 43, we implement a re-scaling technique that monitors the magnitudes of both actual and bias forces (alternatively, actual and bias stresses) over N steps and then computes the ratio between them. To re-scale bias forces, we use the following expression

$$\tau_t = \tau_0 \times \frac{\sqrt{\sum_{n=0}^{N-1}\left\|\mathbf{F}_{t-n\Delta t}\right\|_2^2}}{\sqrt{\sum_{n=0}^{N-1}\left\|\nabla_{\mathbf{r}_i}u_{t-n\Delta t}\right\|_2^2}}. \tag{8}$$

An equivalent expression is applied for bias stresses.

The re-scaling of uncertainty gradients is reminiscent of the AdaGrad algorithm[70], which dynamically adjusts the learning rate (analogous to the biasing strength) based on historical gradients from previous iterations. While incorporating momentum through exponential moving averages can improve the AdaGrad approach, treating all past gradients with equal weight is essential within the context of this study. Our attempts to damp learning along directions with high curvature (high-frequency oscillations), similar to the Adam optimizer[71], did not yield improved performance. We further find that employing species-dependent biasing strengths for bias forces, $\tau \to \tau_{Z_i}$, with a particular emphasis on damping biasing of hydrogen atoms, improves the efficiency of biased MD simulations.

We employ biased MD simulation to generate a candidate pool for AL, as depicted in Fig. 1a. We employ multiple parallel MD simulations to enhance the exploration of the configurational space further and improve the computational efficiency of AL. We expect biased MD simulations to have relatively short auto-correlation times (ACTs) obtained from position and uncertainty auto-correlation functions (ACFs). Short ACTs imply that the generated candidates will be less correlated than those generated with unbiased MD simulations. However, we cannot guarantee the generation of uncorrelated samples with biased MD simulations throughout AL, particularly in later AL iterations when the uncertainty level is reduced. Therefore, we propose to use batch selection algorithms (see later sections) that select $N_{\text{batch}} > 1$ samples at once. These algorithms enforce the informativeness and diversity of the selected atomic configurations and the resulting training data set.

Gaussian moment neural network

This work uses the Gaussian moment neural network (GM-NN) approach for modeling interatomic interactions[16,17]. GM-NN employs an artificial NN to map a local atomic environment S_i to the atomic energy $E_i(S_i,\boldsymbol{\theta})$; see Eq. (1). It uses a fully-connected feed-forward NN with two hidden layers[16,17]

$$\begin{aligned} y_i \;=\; & 0.1\cdot\mathbf{b}^{(3)} + \tfrac{1}{\sqrt{d_2}}\mathbf{W}^{(3)}\phi\Big(0.1\cdot\mathbf{b}^{(2)} + \\ & \tfrac{1}{\sqrt{d_1}}\mathbf{W}^{(2)}\phi\Big(0.1\cdot\mathbf{b}^{(1)} + \tfrac{1}{\sqrt{d_0}}\mathbf{W}^{(1)}\mathbf{G}_i\Big)\Big), \end{aligned} \tag{9}$$

with $\mathbf{W}^{(l+1)} \in \mathbb{R}^{d_{l+1}\times d_l}$ and $\mathbf{b}^{(l+1)} \in \mathbb{R}^{d_{l+1}}$ representing the weights and biases of layer $l+1$. In this work, we employ a NN with $d_0 = 910$ input neurons (corresponding to the dimension of the input feature vector $\mathbf{G}_i = \mathbf{G}_i(S_i)$), $d_1 = d_2 = 512$ hidden neurons, and a single output neuron, $d_3 = 1$. The network's weights $\mathbf{W}^{(l+1)}$ are initialized by selecting entries from a normal distribution with zero mean and unit variance. The trainable bias vectors $\mathbf{b}^{(l+1)}$ are initialized to zero. To improve the accuracy and convergence of the GM-NN model, we implement a neural tangent parameterization (factors of 0.1 and $1/\sqrt{d_l}$)[72]. For the activation function ϕ, we use the Swish/SiLU function[73,74].

To aid the training process, we scale and shift the output of the NN

$$E_i(S_i,\boldsymbol{\theta}) = c\cdot(\rho_{Z_i}y_i + \mu_{Z_i}), \tag{10}$$

where the trainable shift parameters μ_{Z_i} are initialized by solving a linear regression problem, and the trainable scale parameters ρ_{Z_i} are initialized to one. The per-atom RMSE of the regression solution determines the constant c[17].

GM-NN models employ the Gaussian moment (GM) representation to encode the invariance of total energy with respect to translations, rotations, and permutations of the same species[16]. By computing pairwise distance vectors $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and then splitting them into radial and angular components, denoted as $r_{ij} = \|\mathbf{r}_{ij}\|_2$ and $\hat{\mathbf{r}}_{ij} = \mathbf{r}_{ij}/r_{ij}$, respectively, we obtain GMs as follows

$$\boldsymbol{\Psi}_{i,L,s} = \sum_{j\neq i} R_{Z_i,Z_j,s}(r_{ij},\boldsymbol{\beta})\hat{\mathbf{r}}_{ij}^{\otimes L}, \tag{11}$$

where $\hat{\mathbf{r}}_{ij}^{\otimes L} = \hat{\mathbf{r}}_{ij}\otimes\cdots\otimes\hat{\mathbf{r}}_{ij}$ is the L-fold outer product. The nonlinear radial functions $R_{Z_i,Z_j,s}(r_{ij},\boldsymbol{\beta})$ are defined as a sum of Gaussian functions $\Phi_{s'}(r_{ij})$ ($N_{\text{Gauss}} = 9$ for this work)[17]

$$R_{Z_i,Z_j,s}(r_{ij},\boldsymbol{\beta}) = \frac{1}{\sqrt{N_{\text{Gauss}}}}\sum_{s'=1}^{N_{\text{Gauss}}}\beta_{Z_i,Z_j,s,s'}\Phi_{s'}(r_{ij}). \tag{12}$$

The factor $1/\sqrt{N_{\text{Gauss}}}$ impacts the effective learning rate inspired by neural tangent parameterization[72]. The radial functions are centered at equidistantly spaced grid points ranging from $r_{\min} = 0.5$ Å to r_c, set to 5.0 Å and 6.0 Å for alanine dipeptide and MIL-53(Al), respectively. The radial functions are re-scaled by a cosine cutoff function[13], to ensure a smooth

dependence on the number of atoms within the cutoff sphere. Chemical information is embedded in the GM representation through trainable parameters $\beta_{Z_i,Z_j,s,s'}$, with the index s iterating over the number of independent radial basis functions ($N_{\text{basis}} = 7$ for this work).

Features invariant to rotations, $\mathbf{G}_i$, are obtained by computing full tensor contractions of tensors defined in Eq. (11), e.g.[16,17],

$$G_{i,s_1,s_2,s_3} = (\boldsymbol{\Psi}_{i,1,s_1})_a (\boldsymbol{\Psi}_{i,1,s_2})_b (\boldsymbol{\Psi}_{i,2,s_3})_{a,b}, \tag{13}$$

where we use Einstein notation, i.e., the right-hand side is summed over $a, b \in \{1, 2, 3\}$. Specific full tensor contractions are defined by using generating graphs[75]. In a practical implementation, we compute all GMs at once and reduce the number of invariant features based on the permutational symmetries of the respective graphs.

All parameters $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}, \boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\mu}\}$ of the NN are optimized by minimizing the combined squared loss on training data $\mathscr{D}_{\text{train}} = \left(\mathscr{X}_{\text{train}}, \mathscr{Y}_{\text{train}}\right)$, with $\mathscr{X}_{\text{train}} = \{S^{(k)}\}_{k=1}^{N_{\text{train}}}$ and $\mathscr{Y}_{\text{train}} = \{E_k^{\text{ref}}, \{\mathbf{F}_{i,k}^{\text{ref}}\}_{i=1}^{N_{\text{at}}}, \boldsymbol{\sigma}_k^{\text{ref}}\}_{k=1}^{N_{\text{train}}}$,

$$\begin{aligned}\mathscr{L}\left(\boldsymbol{\theta}, \mathcal{D}_{\text{train}}\right) = \sum_{k=1}^{N_{\text{train}}} \Bigg[& C_{\text{e}} \left\| E_k^{\text{ref}} - E(S^{(k)}, \boldsymbol{\theta}) \right\|_2^2 \\ & + C_{\text{f}} \sum_{i=1}^{N_{\text{at}}^{(k)}} \left\| \mathbf{F}_{i,k}^{\text{ref}} - \mathbf{F}_i\left(S^{(k)}, \boldsymbol{\theta}\right) \right\|_2^2 \\ & + C_{\text{s}} \left\| V_k \boldsymbol{\sigma}_k^{\text{ref}} - V_k \boldsymbol{\sigma}\left(S^{(k)}, \boldsymbol{\theta}\right) \right\|_2^2 \Bigg].\end{aligned} \tag{14}$$

We have chosen $C_{\text{e}} = 1.0$, $C_{\text{f}} = 4.0$ Å^2, and $C_{\text{s}} = 0.01$ to balance the relative contributions of energies, forces, and stresses, respectively.

Using AD, we compute atomic forces as negative gradients of total energy with respect to atomic coordinates

$$\mathbf{F}_i\left(S^{(k)}, \boldsymbol{\theta}\right) = -\nabla_{\mathbf{r}_i} E\left(S^{(k)}, \boldsymbol{\theta}\right). \tag{15}$$

Furthermore, we use AD to compute stress tensor, defined by[69]

$$\boldsymbol{\sigma}\left(S^{(k)}, \boldsymbol{\theta}\right) = \frac{1}{V_k} \nabla_{\boldsymbol{\epsilon}} E\left(S^{(k)}, \boldsymbol{\theta}\right)\Big|_{\boldsymbol{\epsilon}=\mathbf{0}}, \tag{16}$$

where $E\left(S^{(k)}, \boldsymbol{\theta}\right)$ is total energy after a strain deformation with symmetric tensor $\boldsymbol{\epsilon} \in \mathbb{R}^{3\times3}$, i.e., $\tilde{\mathbf{r}} = (\mathbf{1} + \boldsymbol{\epsilon}) \cdot \mathbf{r}$. As the stress tensor is symmetric, we use only its upper triangular part in the loss function. Here, V_k is the volume of the periodic cell.

We employ the Adam optimizer[71] to minimize the loss function. The respective parameters of the optimizer are $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-7}$. Usually, we work with a mini-batch of 32 molecules. However, smaller mini-batches were used in the initial AL iterations because the training data sizes were less than 32. The layer-wise learning rates are decayed linearly. The initial values are set to 0.03 for the parameters of the fully connected layers, 0.02 for the trainable representation, as well as 0.05 and 0.001 for the scale and shift parameters of atomic energies, respectively. The training is performed for 1000 training epochs. To prevent overfitting during training, we employ the early stopping technique[76]. All models are trained using PyTorch[59].

Sketched gradient features

We obtain atomic gradient features by computing gradients of Eq. (1) with respect to the parameters of the fully connected layers in Eq. (9). Particularly, we make use of the product structure of atomic gradient features. To obtain the latter, we re-write the network in Eq. (9) as follows

$$\begin{aligned}\mathbf{z}_i^{(l+1)} &= \tilde{\mathbf{W}}^{(l+1)} \tilde{\mathbf{x}}_i^{(l)} \in \mathbb{R}^{d_{l+1}}, \\ \tilde{\mathbf{W}}^{(l+1)} &= \left(\mathbf{W}^{(l+1)}, \mathbf{b}^{(l+1)}\right) \in \mathbb{R}^{d_{l+1} \times (d_l+1)}, \\ \tilde{\mathbf{x}}_i^{(l)} &= \left(\frac{1}{\sqrt{d_l}} \mathbf{x}_i^{(l)}, 0.1\right)^{\top} \in \mathbb{R}^{d_l+1},\end{aligned} \tag{17}$$

where $\mathbf{z}^{(l)}$ and $\mathbf{x}^{(l)}$ denote the pre- and post-activation vectors of layer l. Thus, atomic gradient features read

$$\begin{aligned}\phi_i(S_i) &= \left(\frac{\partial \mathbf{z}_i^{(L)}}{\partial \tilde{\mathbf{W}}^{(1)}}, \cdots, \frac{\partial \mathbf{z}_i^{(L)}}{\partial \tilde{\mathbf{W}}^{(L)}}\right) \\ &= \left(\frac{\partial \mathbf{z}_i^{(L)}}{\partial \mathbf{z}_i^{(1)}} \otimes \tilde{\mathbf{x}}_i^{(0)}, \cdots, \frac{\partial \mathbf{z}_i^{(L)}}{\partial \mathbf{z}_i^{(L)}} \otimes \tilde{\mathbf{x}}_i^{(L-1)}\right).\end{aligned} \tag{18}$$

To make the calculation of gradient features computationally tractable, we employ the random projections (sketching) technique[55], as proposed in refs. 51,52. For atomic gradient features $\phi_i(S_i) \in \mathbf{R}^{N_{\text{feat}}}$ and a random matrix $\mathbf{U} \in \mathbb{R}^{N_{\text{rp}} \times N_{\text{feat}}}$—with N_{feat} and N_{rp} denoting the number of atomic features and random projections, respectively—we can define randomly projected atomic gradient features as

$$\phi_i^{\text{rp}}(S_i) = \mathbf{U}\phi_i(S_i) \in \mathbb{R}^{N_{\text{rp}}}. \tag{19}$$

While a Gaussian sketch could be employed, where the elements of $\mathbf{U}$ are drawn from standard normal distributions, we use a tensor sketching approach that is more runtime and memory efficient[52]. Specifically, denoting the element-wise or Hadamard product as $\odot$, we compute

$$\phi_i^{\text{rp}}(S_i) = \sum_{l=1}^{L} \left(\mathbf{U}_{\text{out}}^{(l)} \phi_{i,\text{out}}^{(l)}(S_i)\right) \odot \left(\mathbf{U}_{\text{in}}^{(l-1)} \phi_{i,\text{in}}^{(l-1)}(S_i)\right), \tag{20}$$

with $\phi_{i,\text{out}}^{(l)}(S_i) = \partial \mathbf{z}_i^{(L)} / \partial \mathbf{z}_i^{(l)}$ and $\phi_{i,\text{in}}^{(l)}(S_i) = \tilde{\mathbf{x}}_i^{(l)}$. All entries of $\mathbf{U}_{\text{in}}^{(l)}$ and $\mathbf{U}_{\text{out}}^{(l)}$ are sampled independently from a standard normal distribution.

For atom-based uncertainties, we can directly use the sketched atomic gradient features. For (total) uncertainties per atom, we need to work with a mean $\phi(S) = \sum_{i=1}^{N_{\text{at}}} \phi_i(S_i)/N_{\text{at}}$. Thus, we use that the individual projections (rows of Eq. (20)) are linear in the features and obtain for the (total) gradient features[51]

$$\phi^{\text{rp}}(S) = \frac{1}{N_{\text{at}}} \sum_{i=1}^{N_{\text{at}}} \sum_{l=1}^{L} \left(\mathbf{U}_{\text{out}}^{(l)} \phi_{i,\text{out}}^{(l)}(S_i)\right) \odot \left(\mathbf{U}_{\text{in}}^{(l-1)} \phi_{i,\text{in}}^{(l-1)}(S_i)\right), \tag{21}$$

given that all of the individual random projections use the same random matrices.

Ensemble-based uncertainty quantification

The variance of the predictions of individual models in an ensemble of MLIPs can be used to quantify their uncertainty. Thus, we define the variance of predicted energy as

$$u^2 = \frac{1}{M} \sum_{j=1}^{M} \|E_j - \bar{E}\|_2^2, \tag{22}$$

where M is the number of models in the ensemble. The variance of atomic forces reads

$$u_i^2 = \frac{1}{3M} \sum_{j=1}^{M} \|\mathbf{F}_{i,j} - \bar{\mathbf{F}}_i\|_2^2, \tag{23}$$

Here, $\bar{E}$ and $\bar{\mathbf{F}}_i$ denote the arithmetic mean of the predictions from individual models. Our experiments demonstrated that $M = 3$ is sufficient to obtain good performance. Using larger ensembles would make the ensemble-based uncertainty quantification even more computationally inefficient than gradient-based alternatives.

Batch selection methods

The simplest batch selection method is based on querying points only by their uncertainty values. Specifically, given the already selected structures $\mathcal{X}_{\text{batch}}$ from an unlabeled pool $\mathcal{X}_{\text{pool}}$ we select the next point by

$$S = \underset{S \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}}{\arg\max} \; u(S), \tag{24}$$

until $N_{\text{batch}} > 1$ structures are selected. In this work, we use this selection method combined with ensemble-based uncertainties.

For the posterior-based uncertainty, we can constrain the diversity of the selected batch by using the posterior covariance between structures

$$\text{Cov}(S, S') = \lambda^2 \phi(S)^\top \left(\Phi_{\text{train}}^\top \Phi_{\text{train}} + \lambda^2 \mathbf{I}\right)^{-1} \phi(S'), \tag{25}$$

with $\Phi_{\text{train}} = \phi\left(\mathcal{X}_{\text{train}}\right)$. The corresponding method greedily selects structures, i.e., one structure per iteration, such that the determinant of the covariance matrix is maximized[51,52,77]

$$S = \underset{S \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}}{\arg\max} \; \det\left[\text{Cov}\left(\mathcal{X}_{\text{batch}} \cup \{S\}, \mathcal{X}_{\text{batch}} \cup \{S\}\right)\right]. \tag{26}$$

For the distance-based uncertainty, we ensure the diversity of the acquired batch by greedily selecting structures with a maximum distance to all previously selected and training data points. The respective selection method reads[51,52,78]

$$S = \underset{S \in \mathcal{X}_{\text{pool}} \setminus \mathcal{X}_{\text{batch}}}{\text{argmax}} \; \min_{S' \in \mathcal{X}_{\text{train}} \cup \mathcal{X}_{\text{batch}}} \left\| \phi(S) - \phi(S') \right\|_2^2. \tag{27}$$

We also applied this batch selection method to define the most representative subset of atomic gradient features when calculating atom-based uncertainty using feature space distances.

Lastly, to compare the performance of uncertainty-based data generation approaches with conventional random sampling from an ab initio MD, we employ a random selection strategy combined with posterior-based uncertainty to terminate MD simulations. We define random selection as

$$S \sim \mathcal{U}\left(\mathcal{X}_{\text{pool}}\right), \tag{28}$$

where $\mathcal{U}$ is the uniform distribution over $\mathcal{X}_{\text{pool}}$.

Conformal prediction

Conformal prediction methods offer distribution-free uncertainty quantification with guaranteed finite sample coverage[49,79–82], thus ensuring calibration. Finite sample coverage can be defined as

$$\mathbb{P}\{y_{\text{test}} \in C(x_{\text{test}})\} \geq 1 - \alpha. \tag{29}$$

Here, $(x_{\text{test}}, y_{\text{test}})$ are the newly observed data, while C defines the prediction set based on previous observations $\{(x_k, y_k)\}_{k=1}^{N_{\text{calibr}}}$. The user determines the hyper-parameter α and defines the desired confidence level. CP methods guarantee that the prediction set contains the true label with a probability of almost $1 - \alpha$.

We employ inductive CP, which comprises the following steps[49,79]: (1) A subset of calibration data, sized N_{calibr}, is selected, and the corresponding errors are computed on this subset. For atomic forces, we employ RMSEs $\Delta\mathbf{F}_i^2 = \frac{1}{3}\left\|\mathbf{F}_i - \mathbf{F}_i^{\text{ref}}\right\|_2^2$, while for total energies the respective energy absolute errors per atom, $\Delta e = |E - E^{\text{ref}}|/N_{\text{at}}$, are used. (2) The uncertainty $u(S)$ is calculated for this subset of data. (3) The ratio $\Delta e/u(S)$ or $\Delta\mathbf{F}_i/u\left(S_i\right)$ is computed. (4) Utilizing quantile regression, the $(1-\alpha)\left(N_{\text{calibr}}+1\right)/N_{\text{calibr}}$-th quantile, denoted as s, is determined. (5) This s value is applied to new observations, resulting in the re-scaled and calibrated uncertainty, $\tilde{u} = s \cdot u$.

Coverage of collective variable space

To measure how well different methods explore the (bounded) space of interest, we implement a tree-based weighted recursive partitioning of a d-dimensional Euclidean space, which is reminiscent of quadtrees[83] and matrix-based octrees[84] but allows to choose how many times n to split each dimension. Thus, the variety of the tree is $k = n^d$. Each node of this complete k-ary tree encodes a generalized hypercube of d dimensions, where each side length depends on the boundaries of the original space. The root node represents the full bounded space. A tree of height L has total number of partitions equal to $(k^{L+1} - 1)/(k - 1)$, and each level ℓ has k^ℓ nodes. The hyper-parameters we choose in this paper are $n = 2, d = 2$ (for the CVs ϕ and ψ of alanine dipeptide), and $L = 5$, for a total of 1365 partitions of the space of interest.

Our proposed surface coverage metric uses this data structure as a proxy to capture how many space partitions a method can explore in the least amount of iterations. At the same time, we need to penalize methods that get stuck in a region of the space, exploring partitions of smaller volumes, that is, those represented by nodes at deeper levels in the tree. For this reason, each node at level ℓ is associated with a reward (or weight) of $1/k^\ell$, so each level of the tree has a cumulative reward of 1. The optimal strategy would be to perform a breadth-first search of the nodes of this tree, which translates into observing the largest partitions of unobserved space first. In addition, partitions that are revisited by the methods give no additional reward, so there is no gain in getting stuck in a certain partition. We visually represent the idea of the algorithm in the Supplementary Information for the simple case of $d = 2$.

Auto-correlation analysis

We evaluate the performance of uncertainty-biased MD simulations by investigating the auto-correlation between subsequent time frames of the MD trajectory. The auto-correlation function (ACF) is defined as[85]

$$A_{\mathcal{O}}(k) = \frac{\langle \mathcal{O}_i \mathcal{O}_{i+k} \rangle - \langle \mathcal{O}_i \rangle^2}{\langle \mathcal{O}_i^2 \rangle - \langle \mathcal{O}_i \rangle^2}, \tag{30}$$

where $\langle \cdots \rangle$ denotes the thermodynamic expectation value, k is the lag time, and $\mathcal{O}$ is an observable, e.g., atomic positions or atom-based uncertainties. From ACF, we can calculate the auto-correlation time (ACT) for an MD trajectory of length N

$$\text{ACT}_{\mathcal{O}} = \frac{1}{2} + \sum_{k=1}^{N} A_{\mathcal{O}}(k)\left(1 - \frac{k}{N}\right). \tag{31}$$

ACT is related to effective sample size (ESS) by

$$\text{ESS}_{\mathcal{O}} = \frac{N}{2 \cdot \text{ACT}_{\mathcal{O}}}. \tag{32}$$

In this work, we calculate ESS as implemented in TensorFlow[86] and use it to estimate the ACT.

Test data set for alanine dipeptide

The test data set for alanine dipeptide comprises 2000 configurations randomly selected from an MD trajectory at 1200 K. This trajectory was generated within the ASE simulation package[87] by running an MD simulation in the canonical (*NVT*) statistical ensemble using the Langevin thermostat. We have used a time step of 0.5 fs and a total simulation time of 1 ns. The

AMBER ff19SB force field has provided forces[57], as implemented in the TorchMD package using PyTorch[58,59]. The data set effectively covers the relevant configurational space of alanine dipeptide, representing an upper boundary in exploring its collective variables (CVs).

MLIP learning details for alanine dipeptide

Each AL experiment starts with training an MLIP with eight alanine dipeptide configurations randomly perturbed from its initial configuration in the C_{7eq} state. Trained MLIPs are then used to run eight parallel MD simulations, initialized from the initial configuration or configurations selected in later iterations. Each MD simulation runs until reaching an empirically defined uncertainty threshold of 1.5 eV Å^{-1}. A lower threshold value may result in slower CV space exploration, while a larger one would lead to the exploration of unphysical configurations. The maximum data set size, comprising training and validation data, is limited to 512 configurations. The Supplementary Information presents the scaling of the presented AL experiments to larger data set sizes, acquiring data sets of 1024 samples. Biased (bias-forces-driven) and unbiased MD simulations are performed using the canonical (*NVT*) statistical ensemble within the ASE simulation package[87]. Unbiased MD simulations are run with the Langevin thermostat at temperatures of 300 K, 600 K, and 1200 K, whereas biased simulations are performed at a constant temperature of 300 K. We have chosen an integration time step of 0.5 fs and set a maximum of 20,000 steps for an MD simulation. A biasing strength of $\tau = 0.25$ was also chosen for biased AL experiments. In reference calculations, we employ a force threshold of 20 eV Å^{-1} to exclude unphysical structures, potentially expected at high biasing strengths (equivalently, a smaller integration time step could be used). All AL experiments have been repeated five times.

Reference DFT calculations for MIL-53(Al)

DFT calculations for MIL-53(Al) were performed using the CP2K simulation package (version 2023.1)[61]. To ensure consistency with incremental learning experiments[41], we employed the PBE functional[62] with Grimme D3 dispersion correction[63]. A hybrid basis set, combining TZVP Gaussian basis functions and plane waves, was employed[88]. GTH pseudopotentials were used to smoothen the electron density near the nuclei[89]. To ensure the convergence of force and stress calculations, a plane wave cutoff energy of 1000 Ry was selected.

MLIP learning details for MIL-53(Al)

In each AL experiment, we start with 32 MIL-53(Al) configurations randomly perturbed around its closed-pore state, with 90% reserved for training. Trained MLIPs are then used to perform 32 parallel MD simulations, each running until it reaches an uncertainty threshold of 1.0 eV Å^{-1}. The maximum data set size is limited to 512 configurations, comprising training and validation data. The Supplementary Information presents the scaling of the presented AL experiments to larger data set sizes, acquiring data sets of 1024 samples. Both biased (bias-stress-driven) and unbiased MD simulations use the isothermal-isobaric form of the Nosé–Hoover dynamics[90,91]. Unbiased MD simulations are carried out at 600 K and 0 MPa, as well as ± 250 MPa (half of the simulations each), while biased simulations are performed at 600 K and 0 MPa. The characteristic time scales of the thermostat and barostat are set to 0.1 ps and 1 ps, respectively. We have chosen an integration time step of 0.5 fs and set a maximum of 20,000 MD steps for an MD simulation. A stress-biasing strength of $\tau = 0.5$ is used in biased AL experiments. In reference calculations, we employ a force threshold of 20 eV Å^{-1} to exclude strongly distorted structures. We use the data set from ref. 41 as a metadynamics-generated baseline and select the first 500 sequentially generated configurations. All AL experiments are repeated three times, except for metadynamics, which was run once[41]. For metadynamics, we train three MLIPs initialized using different random seeds.

Random perturbation of atomic configurations

We obtain randomly perturbed atomic configurations by adding atomic shifts, denoted as $\boldsymbol{\delta}_i$, to the original atomic positions $\mathbf{r}_i$

$$\tilde{\mathbf{r}}_i = \mathbf{r}_i + \boldsymbol{\delta}_i. \tag{33}$$

The components of $\boldsymbol{\delta}_i$ are sampled independently from a uniform distribution: for alanine dipeptide, the range is between −0.02 Å and 0.02 Å, and for MIL-53(Al), it is between −0.08 Å and 0.08 Å. Additionally, for MIL-53(Al), we introduce random perturbations to its periodic cell $\mathbf{B}$ using a strain deformation $\boldsymbol{\epsilon} = \left(\mathbf{A} + \mathbf{A}^{\top}\right)/2$, where the components of $\mathbf{A}$ are sampled independently from a uniform distribution between −0.02 and 0.02. This transformation can be expressed as

$$\tilde{\mathbf{B}} = \mathbf{B}(\mathbf{I} + 2\boldsymbol{\epsilon})^{1/2}. \tag{34}$$

The shifted atomic positions are re-scaled according to

$$\tilde{\tilde{\mathbf{r}}}_i = (\mathbf{I} + 2\boldsymbol{\epsilon})^{1/2}\tilde{\mathbf{r}}_i. \tag{35}$$

Sine wave with additive random noise

We model large-amplitude volume fluctuations in MIL-53(Al) induced by the bias stress using a sine wave with period T_0 and additive random noise $N(t)$

$$A \sin\left(\frac{2\pi t}{T_0}\right) + BN(t),$$

where A and B denote the sine wave's amplitude and random noise, respectively. In this work, $N(t) \sim \mathcal{N}(0, 1)$ represents random noise following a normal distribution with zero mean and unit variance. We chose $A = 1.0$ and $B = 0.5$ for the blue line in Fig. 7. For the red line, we increase the noise amplitude to $B = 2.0$. To represent the volume fluctuations induced in MIL-53(Al) (see Fig. 7), a sine wave with the period twice the length of the MD simulation, i.e., $T_0 = 3.2$ ns is required.

Data availability

The data sets generated during this study are available in the Zenodo repository: https://doi.org/10.5281/zenodo.10776838. The MIL-53(Al) test data set is available at https://doi.org/10.5281/zenodo.6359970 (ref. 41).

Code availability

The source code for this study is available on GitHub and can be accessed via this link: https://github.com/nec-research/alebrew.

Received: 31 October 2023; Accepted: 22 March 2024;
Published online: 29 April 2024

References

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
2. Smith, J. S. et al. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, 134 (2020).
3. Chanussot, L. et al. Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catal.* **11**, 6059–6072 (2021).
4. Xie, Y. et al. Uncertainty-aware molecular dynamics from Bayesian active learning for phase transformations and thermal transport in SiC. *npj Comput. Mater.* **9**, 36 (2023).
5. Gubaev, K. et al. Performance of two complementary machine-learned potentials in modelling chemically complex systems. *npj Comput. Mater.* **9**, 129 (2023).

6. Langer, M. F., Goeßmann, A. & Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *npj Comput. Mater.* **8**, 41 (2022).
7. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
8. Faber, F. A., Christensen, A. S., Huang, B. & von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **148**, 241717 (2018).
9. Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
10. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
11. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
12. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
13. Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
14. Artrith, N. & Urban, A. An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. *Comput. Mater. Sci.* **114**, 135–150 (2016).
15. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
16. Zaverkin, V. & Kästner, J. Gaussian Moments as Physically Inspired Molecular Descriptors for Accurate and Scalable Machine Learning Potentials. *J. Chem. Theory Comput.* **16**, 5410–5421 (2020).
17. Zaverkin, V., Holzmüller, D., Steinwart, I. & Kästner, J. Fast and Sample-Efficient Interatomic Neural Network Potentials for Molecules and Materials Based on Gaussian Moments. *J. Chem. Theory Comput.* **17**, 6658–6670 (2021).
18. Schütt, K. et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **30**, 992–1002 (2017).
19. Schütt, K. T., Unke, O. T. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Int. Conf. Mach. Learn.* **139**, 9377–9388 (2021).
20. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
21. Batatia, I., Kovacs, D. P., Simm, G. N. C., Ortner, C. & Csanyi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *Adv. Neural Inf. Process. Syst.* **35**, 11423–11436 (2022).
22. Gasteiger, J., Becker, F. & Günnemann, S. GemNet: Universal Directional Graph Neural Networks for Molecules. *Adv. Neural Inf. Process. Syst.* **34**, 6790–6802 (2021).
23. Liao, Y.-L., Wood, B., Das, A. & Smidt, T. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. *Int. Conf. Learn. Represent*. https://arxiv.org/abs/2306.12059 (2024).
24. Passaro, S. & Zitnick, C. L. Reducing SO(3) Convolutions to SO(2) for Efficient Equivariant GNNs. *Int. Conf. Mach. Learn.* **202**, 27420–27438 (2023).
25. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761 (2021).
26. Unke, O. T. et al. Machine Learning Force Fields. *Chem. Rev.* **121**, 10142–10186 (2021).
27. Li, Z., Kermode, J. R. & De Vita, A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
28. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
29. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
30. Zhang, L., Lin, D.-Y., Wang, H., Car, R. & E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
31. Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 20 (2020).
32. Shuaibi, M., Sivakumar, S., Chen, R. Q. & Ulissi, Z. W. Enabling robust offline active learning for machine learning potentials using simple physics-based priors. *Mach. Learn.: Sci. Technol.* **2**, 025007 (2021).
33. Briganti, V. & Lunghi, A. Efficient generation of stable linear machine-learning force fields with uncertainty-aware active learning. *Mach. Learn.: Sci. Technol.* **4**, 035005 (2023).
34. Wang, X. et al. Generalization of Graph-Based Active Learning Relaxation Strategies Across Materials. *Mach. Learn.: Sci. Technol.* https://doi.org/10.1088/2632-2153/ad37f0 (2024).
35. Huber, T., Torda, A. E. & van Gunsteren, W. F. Local elevation: A method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aid. Mol. Des.* **8**, 695–708 (1994).
36. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA* **99**, 12562–12566 (2002).
37. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020603 (2008).
38. Demuynck, R. et al. Efficient Construction of Free Energy Profiles of Breathing Metal-Organic Frameworks Using Advanced Molecular Dynamics Simulations. *J. Chem. Theory Comput.* **13**, 5861–5873 (2017).
39. Yoo, D., Jung, J., Jeong, W. & Han, S. Metadynamics sampling in atomic environment space for collecting training data for machine learning potentials. *npj Comput. Mater.* **7**, 131 (2021).
40. Yang, M., Bonati, L., Polino, D. & Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catal. Today* **387**, 143–149 (2022).
41. Vandenhaute, S., Cools-Ceuppens, M., DeKeyser, S., Verstraelen, T. & Van Speybroeck, V. Machine learning potentials for metal-organic frameworks using an incremental learning approach. *npj Comput. Mater.* **9**, 19 (2023).
42. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
43. van der Oord, C., Sachs, M., Kovács, D. P., Ortner, C. & Csányi, G. Hyperactive learning for data-driven interatomic potentials. *npj Comput. Mater.* **9**, 168 (2023).
44. Schwalbe-Koda, D., Tan, A. R. & Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun.* **12**, 5104 (2021).
45. Carrete, J., Montes-Campos, H., Wanzenböck, R., Heid, E. & Madsen, G. K. H. Deep ensembles vs committees for uncertainty estimation in neural-network force fields: Comparison and application to active learning. *J. Chem. Phys.* **158**, 204801 (2023).
46. Kuleshov, V., Fenner, N. & Ermon, S. Accurate Uncertainties for Deep Learning Using Calibrated Regression. *Int. Conf. Mach. Learn.* **80**, 2796–2804 (2018).
47. Pernot, P. The long road to calibrated prediction uncertainty in computational chemistry. *J. Chem. Phys.* **156**, 114109 (2022).

48. Tran, K. et al. Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn.: Sci. Technol.* **1**, 025006 (2020).
49. Hu, Y., Musielewicz, J., Ulissi, Z. W. & Medford, A. J. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Mach. Learn.: Sci. Technol.* **3**, 045028 (2022).
50. Zaverkin, V. & Kästner, J. Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design. *Mach. Learn.: Sci. Technol.* **2**, 035009 (2021).
51. Zaverkin, V., Holzmüller, D., Steinwart, I. & Kästner, J. Exploring chemical and conformational spaces by batch mode deep active learning. *Digital Discovery* **1**, 605–620 (2022).
52. Holzmüller, D., Zaverkin, V., Kästner, J. & Steinwart, I. A framework and benchmark for deep batch active learning for regression. *J. Mach. Learn. Res.* **24**, 1–81 (2023).
53. Schran, C., Brezina, K. & Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **153**, 104105 (2020).
54. Kirsch, A. Black-Box Batch Active Learning for Regression. *Transact. Mach. Learn. Res*. https://arxiv.org/abs/2302.08981 (2023).
55. Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.* **10**, 1–157 (2014).
56. Bolhuis, P. G., Dellago, C. & Chandler, D. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci. USA* **97**, 5877–5882 (2000).
57. Tian, C. et al. ff19sb: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.* **16**, 528–552 (2020).
58. Doerr, S. et al. Torchmd: A deep learning framework for molecular simulations. *J. Chem. Theory Comput.* **17**, 2355–2363 (2021).
59. Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
60. Christiansen, H., Errica, F. & Alesiani, F. Self-tuning Hamiltonian Monte Carlo for accelerated sampling. *J. Chem. Phys.* **159**, 234109 (2023).
61. Kühne, T. D. et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 194103 (2020).
62. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
63. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **132**, 154104 (2010).
64. Nagyfalusi, B., Udvardi, L. & Szunyogh, L. First principles and metadynamics study of the spin-reorientation transition in Fe/Au(001) films. *J. Phys. Conf. Ser.* **903**, 012016 (2017).
65. Ibayashi, H. et al. Allegro-legato: Scalable, fast, and robust neural-network quantum molecular dynamics via sharpness-aware minimization. *High Perform. Comput*. https://doi.org/10.1007/978-3-031-32041-5_12 (2023).
66. Zhao, J., Kennedy, S. D. & Turner, D. H. Nuclear Magnetic Resonance Spectra and AMBER OL3 and ROC-RNA Simulations of UCUCGU Reveal Force Field Strengths and Weaknesses for Single-Stranded RNA. *J. Chem. Theory Comput.* **18**, 1241–1254 (2022).
67. Janet, J. P., Duan, C., Yang, T., Nandy, A. & Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **10**, 7913–7922 (2019).
68. Zhu, A., Batzner, S., Musaelian, A. & Kozinsky, B. Fast uncertainty estimates in deep learning interatomic potentials. *J. Chem. Phys.* **158**, 164111 (2023).
69. Knuth, F., Carbogno, C., Atalla, V., Blum, V. & Scheffler, M. All-electron formalism for total energy strain derivatives and stress tensor components for numeric atom-centered orbitals. *Comput. Phys. Comm.* **190**, 33–50 (2015).
70. Duchi, J., Hazan, E. & Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011).
71. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *Int. Conf. Learn. Represent*. https://arxiv.org/abs/1412.6980 (2015).
72. Jacot, A., Gabriel, F. & Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *Adv. Neural Inf. Process. Syst.* **31**, 8580–8589 (2018).
73. Elfwing, S., Uchibe, E. & Doya, K. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning. *Neural Netw.* **107**, 3–11 (2018).
74. Ramachandran, P., Zoph, B. & Le, Q. V. Searching for activation functions. *Int. Conf. Learn. Represent*. https://arxiv.org/abs/1710.05941 (2018).
75. Suk, T. & Flusser, J. Tensor method for constructing 3D moment invariants. In *Computer Analysis of Images and Patterns* (eds Real, P. et al.), 212–219 (Springer, 2011).
76. Prechelt, L. Early stopping—but when? In *Neural Networks: Tricks of the Trade* (eds. Montavon, G. et al.), 53–67 (Springer, 2012).
77. Kirsch, A., Van Amersfoort, J. & Gal, Y. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning. *Adv. Neural Inf. Process. Syst.* **32**, 7026–7037 (2019).
78. Sener, O. & Savarese, S. Active learning for convolutional neural networks: A core-set approach. *Int. Conf. Learn. Represent*. https://arxiv.org/abs/1708.00489 (2018).
79. Vovk, V., Gammerman, A. & Shafer, G. Algorithmic learning in a random world (Springer, 2005).
80. Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J. & Wasserman, L. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* **113**, 1094–1111 (2018).
81. Romano, Y., Patterson, E. & Candés, E. J. Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**, 3543–3553 (2019).
82. Angelopoulos, A. N. & Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **16**, 494–591 (2023).
83. Finkel, R. A. & Bentley, J. L. Quad trees a data structure for retrieval on composite keys. *Acta Inform.* **4**, 1–9 (1974).
84. Meagher, D. J. Octree Encoding: A New Technique for the Representation, Manipulation and Display of Arbitrary 3-D Objects by Computer (Electrical and Systems Engineering Department Rensseiaer Polytechnic, 1980).
85. Janke, W. Monte Carlo Simulations in Statistical Physics – From Basic Principles to Advanced Applications. In *Order, Disorder and Criticality*, 93–166 (World Scientific, 2013).
86. Dillon, J. V. et al. Tensorflow distributions. Preprint at https://arxiv.org/abs/1711.10604 (2017).
87. Larsen, A. H. et al. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
88. Lippert, G., Hutter, J. & Parrinello, M. A hybrid Gaussian and plane wave density functional scheme. *Mol. Phys.* **92**, 477–487 (1997).
89. Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **54**, 1703–1710 (1996).
90. Melchionna, S., Ciccotti, G. & Holian, B. L. Hoover NPT dynamics for systems varying in shape and size. *Mol. Phys.* **78**, 533–544 (1993).
91. Melchionna, S. Constrained systems and statistical distribution. *Phys. Rev. E* **61**, 6165–6170 (2000).

Acknowledgements

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 – 390740016. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting David Holzmüller.

Author contributions

All authors designed the project, discussed the results, and wrote the manuscript. Viktor Zaverkin performed the calculations.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01254-1.

Correspondence and requests for materials should be addressed to Viktor Zaverkin.