npj | computational materials

Article

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

https://doi.org/10.1038/s41524-025-01703-5

Learning atomic forces from uncertainty-calibrated adversarial attacks

Henrique Musseli Cezar[1], Tilmann Bodenstein[1], Henrik Andersen Sveinsson[2], Morten Ledum[1], Simen Reine[1] & Sigbjørn Løland Bore[1] ✉

Adversarial approaches, which intentionally challenge machine learning models by generating difficult examples, are increasingly being adopted to improve machine learning interatomic potentials (MLIPs). While already providing great practical value, little is known about the actual prediction errors of MLIPs on adversarial structures and whether these errors can be controlled. We propose the Calibrated Adversarial Geometry Optimization (CAGO) algorithm to discover adversarial structures with user-assigned errors. Through uncertainty calibration, the estimated uncertainty of MLIPs is unified with real errors. By performing geometry optimization for calibrated uncertainty, we reach adversarial structures with the user-assigned target MLIP prediction error. Integrating with active learning pipelines, we benchmark CAGO, demonstrating stable MLIPs that systematically converge structural, dynamical, and thermodynamical properties for liquid water and water adsorption in a metal-organic framework within only hundreds of training structures, where previously many thousands were typically required.

By representing the potential energy surface of atoms with neural networks, machine learning interatomic potentials (MLIPs) can be trained to predict the outcomes of costly quantum mechanical calculations in milliseconds instead of hours. With pioneering MLIPs, such as Behler-Parinello neural networks, structural and thermodynamic properties are already well captured[1]. The latest equivariant message passing neural network approaches, such as NequIP[2], Allegro[3], and MACE[4] have further significantly reduced prediction errors by nearly an order of magnitude. These approaches typically achieve training and test set errors far lower than typical errors associated with the details of the underlying quantum mechanical calculations they are trained on, for example, basis set truncation and functional choice in the case of Density Functional Theory (DFT). Although MLIPs rely on the inductive bias of short-range interactions, their accuracy makes these methods, in principle, capable of accurately representing the energy landscape of chemically complex systems.

Parameterization of reliable MLIPs remains a challenge, often requiring significant human time and trial-and-error experimentation. The challenge in developing a reliable MLIP presents a chicken-and-egg scenario. On the one hand, the ultimate goal of MLIPs is to explore the unknown, reaching beyond the capabilities of ab initio MD, extending both to longer timescales and to larger length scales. On the other hand, the phase space covered by the training set is limited by the computational cost of ab initio calculations. As a result, all applications rely to varying degrees on the MLIP's ability to generalize to structures outside the training set. This ability to generalize is not just a property of the MLIP but an interaction between the MLIP and the training set. This dependency can become very problematic, as highlighted in recent studies[5,6] which demonstrate that MLIPs, in some situations, struggle to provide stable dynamics, accurate sampling, or reproduce the underlying physics.

Since current MLIP approaches achieve very good training set and validation set accuracy, this problem ultimately arises due to shifting features (molecular structures in our case) from training to production. When the MLIP is used in practice, the molecular structures are different from the structures in the training data. This phenomenon is known in a broader statistics context as a covariate shift[7]. Active learning approaches have become the go-to solution for reducing covariate shifts in structures that typically occur during molecular dynamics (MD) sampling with MLIP-based models. The central idea is to use the MLIP to extend the training set by sampling structures that the MLIP encounters during production[8]. In practice, this is typically achieved by performing iterations of MD sampling with the MLIP, adding new structures until the MLIP model reaches specific convergence criteria, such as simulation stability or the accurate reproduction of structural and thermodynamic properties of reference ab initio simulations.

To avoid unnecessary costly reference calculations on structures already well represented in the training set, active learning procedures typically employ uncertainty quantification to select structures that enhance the training data, i.e., those with high prediction uncertainty. There are

[1]Hylleraas Centre for Quantum Molecular Sciences and Department of Chemistry, University of Oslo, PO Box 1033 Blindern, 0315 Oslo, Norway. [2]The Njord Centre, Department of Physics, University of Oslo, PO Box 1048 Blindern, 0316 Oslo, Norway. ✉e-mail: s.l.bore@kjemi.uio.no

various approaches to uncertainty estimation, including Gaussian Process regression[9,10], dropout in neural networks[11], and committees of MLIPs[12]. While the first two methods rely on specific MLIP architectures, the committee approach is broadly applicable, only requiring computing the variance of a property between multiple models trained on different training sets or seeds, for example, through bootstrapping. However, it is not without drawbacks: The members of the MLIP committee can potentially all agree on an incorrect answer. Furthermore, training and running multiple models adds to the computational cost. Despite these issues, the committee approach remains widely used due to its simplicity and ease of implementation, with many well-established active learning softwares using it[13–16].

For uncertainty estimation to effectively extend the selection of new training set structures, it is essential to offer a diverse pool of candidate structures that expand the phase space of the existing training set. The typical strategy involves MD sampling with MLIPs across various thermodynamic conditions. The inclusion of structures that represent rare events through enhanced sampling techniques is also increasingly being recognized as important[16,17]. Although MLIP-based sampling with MD is a simple approach suitable for automated active learning pipelines to develop models, it has some weaknesses. These include prolonged correlation times, leading to high inter-structure correlation, and the inherent Boltzmann bias toward low-energy structures. Therefore, when the standard MLIP sampling approach fails, there are numerous pragmatic solutions to target more diverse structures, such as sampling at high temperatures and pressures, normal-mode sampling, or perturbing structures via random atomic displacements[14,18]. While these methods offer structures contrasting those of standard MD simulations, they often result in high-energy structures with atomic overlap. The result is, therefore, an undesirable compromise between including structures with high prediction uncertainty and structures with significant forces from atomic overlap. The latter can reduce the overall accuracy of the MLIP[19]. In addition, such approaches tend to localize the prediction uncertainty to a few atoms, which is problematic for large molecular assemblies. It stands to reason that a few carefully optimized structures can offer the same learning content as many highly correlated structures.

In machine learning for image classification problems, adversarial approaches provide a systematic framework for building more robust models by training against examples that aim to trick the model[20]. This approach has also been applied to the active learning of MLIPs. To the best of our knowledge, this idea was first introduced by Cubuk et al.[21] to move atoms toward high prediction uncertainty for the potential energy. A similar idea was also applied in ref. 22 using a bias towards high uncertainty with metadynamics simulations. Such approaches are algorithmically elegant by requiring only a standard MLIP single point force calculation. However, energy uncertainty leaves $3N+6$ labels of learning content associated with gradients and virials up to chance. In this regard, the Bombarelli Group extended the approach of Cubuk to include prediction uncertainty of forces[23]. Using a differentiable MLIP architecture, they performed an adversarial active learning approach capable of discovering structures with high force uncertainty, reaching robust models for challenging systems including zeolite and alanine molecules. This work was also recently extended to non-periodic system reference calculations[24], and force uncertainty-driven dynamics[25].

While such adversarial approaches have demonstrated great potential in active learning, very little attention has so far been put towards quantifying the errors of MLIPs on the adversarial structures. As such, a fundamental question remains: To what extent can we control the actual errors of the MLIP on the adversarial structures? To answer this question, we have developed the Calibrated Adversarial Geometry Optimization (CAGO) algorithm, which aims to discover new adversarial structures with target force errors preassigned by the user (Fig. 1). To unify estimated prediction uncertainties with real errors, we perform uncertainty calibration. To control the MLIP prediction errors on adversarial structures, we optimize structures to reach moderate target errors. These structures are within the range of validity of the uncertainty calibration while being challenging structures for the MLIPs, from which they can learn. To demonstrate the usefulness of our approach, we integrate CAGO into an active learning framework, as shown in Fig. 1, enabling us to learn liquid-water dynamics and water adsorption in metal-organic frameworks from small datasets.

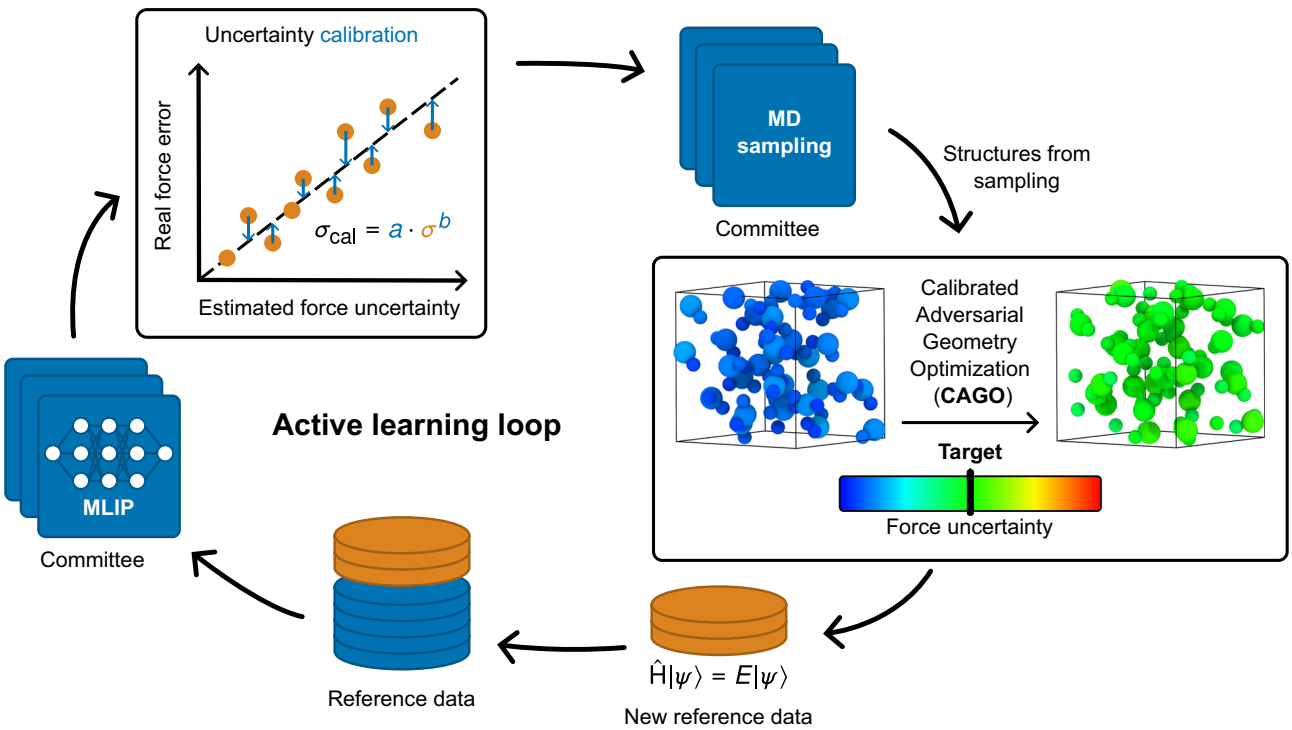

Fig. 1 | Active learning loop with calibrated adversarial geometry optimization (CAGO). Uncertainty calibration, the determination of calibration parameters, is performed using training data and a committee of MLIPs. The CAGO algorithm is applied to the MD-sampled structures from the current iteration of the MLIPs to obtain new structures. These structures are subsequently used in reference ab initio calculations and added to the training set.

Results

Uncertainty quantification and calibration

MLIPs predict quantities y, such as forces, energies, and virials, depending on the structure $\mathbf{x}$, and we would like to estimate the root mean square error σ_{rmse} of the predictions $\hat{y}$ with respect to a ground truth reference y_{ref}:

$$\sigma^2_{\text{rmse}}(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} |\hat{y}^m - y_{\text{ref}}|^2, \tag{1}$$

where m denotes one of the M models. In the committee approach, this error is estimated from the standard deviation of predictions:

$$\hat{\sigma}^2 = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{y}^m - \bar{y}\right)^2, \tag{2}$$

where $\bar{y}$ denotes the committee mean. In practice, the committee uncertainty estimates $\hat{\sigma}$ typically underestimates the actual prediction error σ_{rmse}. To achieve a statistically correct uncertainty estimate, we employ uncertainty calibration, following the procedure introduced in ref. 26, where the uncertainty estimate σ is considered well-calibrated when the ratio

$$r = \frac{\hat{y} - y_{\text{ref}}}{\sigma} \tag{3}$$

is normally distributed with bias zero, and the standard deviation is equal to 1. Many different uncertainty calibration schemes have been proposed[27,28]. Here, we opt for a power law calibration strategy[27]:

$$\sigma_{\text{cal}} = a \cdot \hat{\sigma}^b, \tag{4}$$

where a and b are determined by optimizing the negative log-likelihood that $\hat{y} - y_{\text{ref}}$ was drawn from a normal distribution with zero bias and standard deviation in equation (4) over structures $\mathbf{x}$[26]:

$$a,\ b = \arg\min_{a',\,b'} \sum_{\mathbf{x}} \left[2\pi + \ln\left(a'\hat{\sigma}^{b'}\right)^2 + \frac{|\hat{y}(\mathbf{x}) - y_{\text{ref}(\mathbf{x})}|^2}{\left(a'\hat{\sigma}^{b'}\right)^2} \right]. \tag{5}$$

Adversarial structures

Like in ref. 23, we create adversarial attacks by optimizing the structure $\mathbf{x}$ according to a fitness function $\mathcal{L}$. However, instead of maximizing the committee uncertainty estimate $\hat{\sigma}$, we optimize calibrated prediction uncertainties σ_{cal} towards the error target δ:

$$\mathcal{L}(\sigma_{\text{cal}}) = \left(\sigma_{\text{cal}}(\mathbf{x}) - \delta\right)^2. \tag{6}$$

By targeting a specific prediction uncertainty δ, we aim to push the predictions $\hat{y}$ outside the MLIPs comfort zone, where the error is considerably higher than the training set error. This ensures that the resulting structures contain information that expands the training set while at the same time maintaining the consistency between real errors and estimated errors.

Biasing adversarial structures

In addition to the prediction uncertainty of adversarial structures, it can be desirable to bias the adversarial structural properties toward a target value for y_{bias}, e.g., certain pressures or force magnitudes less than a specified threshold. This can be achieved by supplementing the fitness function by:

$$\mathcal{L}_{\text{bias}}(\bar{y}) = l_{\text{bias}}\left(\bar{y}(\mathbf{x}) - y_{\text{bias}}\right)^2, \tag{7}$$

where l_{bias} is a prefactor determining how strictly the bias is applied.

Force-based calibrated adversarial geometry optimization

Throughout the rest of this paper, we will consider force-based adversarial geometry optimization, with the molecular structure $\mathbf{x}$ being determined by the optimization problem:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \sum_{i=1}^{N_{\text{atoms}}} \left((\hat{\sigma}_{\mathbf{F}_i}(\mathbf{x}) - \delta)^2 + l_{\text{bias}} |\bar{F}_i(\mathbf{x})|^2 \right), \tag{8}$$

where N_{atoms} is the number of atoms, $\hat{\sigma}_{\mathbf{F}_i}$ is given by

$$\hat{\sigma}_{\mathbf{F}_i} = \sqrt{\hat{\sigma}^2_{|F_{ix}|} + \hat{\sigma}^2_{|F_{iy}|} + \hat{\sigma}^2_{|F_{iz}|}}, \tag{9}$$

and $|\bar{F}_i(\mathbf{x})|$ denotes the norm of the average force on atom i.

Calibrated adversarial geometry optimization to target error

We start by considering the error of force predictions on liquid water from a committee of 20 MLIPs trained on a synthetic training set of DNN@MB-pol[29] MD trajectories (see Methods IV E for more details). Figure 2a compares the ratio of the distribution error and the estimated calibrated/uncalibrated uncertainties against the normal distribution from ideal error calibration[26]. While both curves show a bell curve, the wider shape of the force error–uncalibrated uncertainty ratio shows that the uncalibrated

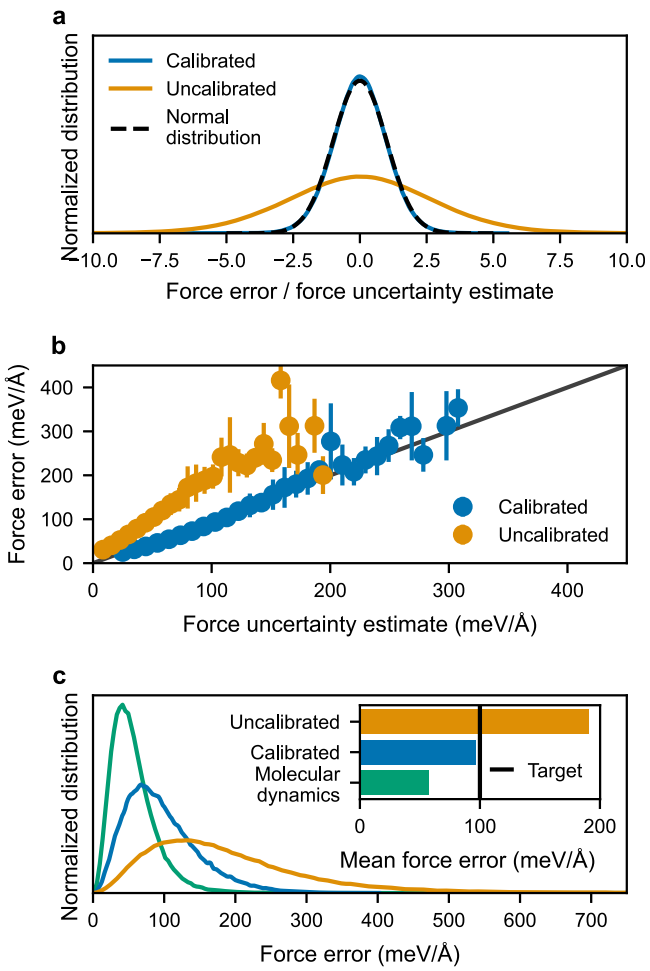

Fig. 2 | Uncertainty calibration and error statistics on adversarial structures. **a** Comparison of error statistics of forces predicted by a committee of models divided by estimated prediction uncertainty against normal distribution. **b** Force errors for a committee of models vs. estimated uncertainty. **c** Force error statistics for structures from MD, CAGO with calibrated uncertainty, and CAGO without uncertainty-calibration, with associated estimated mean errors in the inset figure.

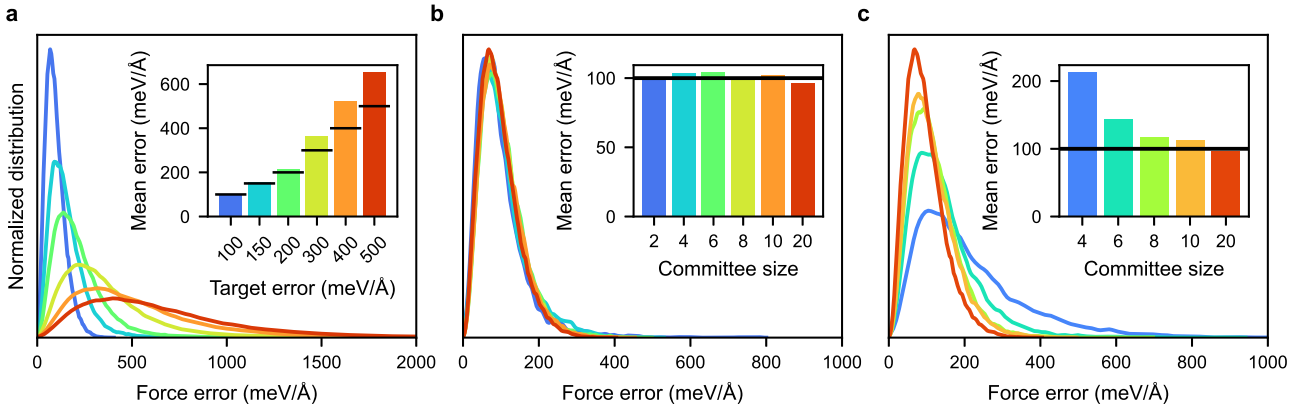

Fig. 3 | Benchmark on the effect of CAGO algorithm hyperparameters. The panels show the distribution of MLIP force errors on adversarial structures, with corresponding mean errors in the inset of each panel. Line legends and colors are specified by ticks in inset figures. **a** Force error statistics for different error targets. **b** Force error statistics for adversarial structures optimized for a target 100 meV/Å uncertainty with different committee sizes, using uncertainty calibration from a committee of 20 MLPs. **c** Force error statistics for adversarial structures optimized for a target 100 meV/Å uncertainty, considering uncertainty calibration performed with its respective committee size.

uncertainty significantly underestimates the true uncertainty. In contrast, the calibrated uncertainty leads to an almost perfect agreement. In the Supplementary Fig. 2, we show that this is also the case for energy and virial predictions. The near-perfect error statistics for the force errors of the calibrated uncertainty is reflected in the one-to-one correlation between the mean force error and the force uncertainty estimate in Fig. 2b. For high force errors, the correlations are good, albeit more noisy than in the low-force-errors regime. High-error samples are inherently less frequent than low-error samples. This means the uncertainty calibration has access to less data for high-error samples and is therefore expected to be less accurate in this case than in the low-error regime. As such, the force error versus the force uncertainty can act as a heuristic to determine the validity range for calibrated uncertainty estimation, in this case, up to about ~ 200 meV/Å.

Next, we perform CAGO to solve equation (8) without a bias-term by geometry optimization using numerical finite-difference gradients for 40 structures (see Fig. 1, all structures reaching the target force error indicated in green). Figure 2c reports the corresponding error statistics, i.e., individual model forces versus reference DNN@MB-pol forces. The error statistics for adversarial structures for CAGO with calibrated uncertainties lead to a mean error close to the target error, whereas the uncalibrated case has errors about twice as high as the target. The potential usefulness of such calibrated adversarial structures can be understood when compared to ordinary MD structures, which have about half the error and would, therefore, not provide the same level of learning content if added to the training set.

The CAGO algorithm has several hyperparameters that can be adjusted in accordance with user goals and are benchmarked in Fig. 3. Figure 3a reports the error statistics for different error targets δ in equation (8). By tuning this parameter, the real error of adversarial structures can be controlled. A clear trend here is that CAGO is more accurate when targets have moderate values, in this case, up to about 200 meV/Å, while for higher values, the real errors are higher than the target errors. This is in line with our observations from Fig. 2b, where above 200 meV/Å the uncertainty calibration is less accurate.

When performing CAGO, using a small committee of models is desirable due to the linear increase in computational cost with committee size. Interestingly, Fig. 3b shows that the error statistics are not very sensitive to committee size, with a few models sufficing to reach close to the target error. On the contrary, Fig. 3c shows how committee size is critical for achieving reliable calibration, where adversarial structures from small committees calibrated with the same number of models do not hit their target. In this case, a committee of around 10 models is necessary to reach the target error, but this may be system- and MLIP-architecture-dependent. This indicates that using many models to calibrate the uncertainties (Fig. 3c) and performing CAGO using a few models (Fig. 3b) is a good heuristics for performing CAGO reliably and efficiently. CAGO can also be performed with bias terms while maintaining the target error, as demonstrated by our benchmark reported in Supplementary Fig. 3.

Learning liquid water from a single structure

An MLIP is only as good as its reference data. Therefore, creating a training set with as high-quality references as possible is desirable. However, higher quality is generally associated with higher computational cost. In particular, density-corrected DFT (DC-DFT) and density-corrected R^{2}SCAN (DC-R^{2}SCAN) achieve excellent correspondence in energies with respect to coupled cluster methods[30,31]. However, DC-DFT and DC-R^{2}SCAN are rather expensive compared to ordinary functionals, making the standard route of starting from a training set of ab initio MD trajectories prohibitively expensive. It is therefore important to have a method that can start the active learning process with as few training structures as possible. Taking this to the extreme, here we use active learning to learn an MLIP for DC-R^{2}SCAN starting from only a single structure of liquid water (Fig. 4a). Figure 4 reports our benchmark on the performance of MLIP iterations during CAGO-based active learning for Allegro with error target 100 meV/Å (Allegro CAGO 100 meV/Å), and, for DeePMD, with targets 100 meV/Å and 200 meV/Å (CAGO 100 meV/Å and 200 meV/Å), standard active learning selecting maximum uncertainty structures from MD (max. uncertainty), and sampling random structures from MD at 500 K (random 500 K), please refer to Methods IV F for additional details. Note that for the first two iterations of CAGO-based active learning, we do CAGO from the first starting geometry, and not from MD sampled geometries (see Methods IV F for additional details).

Figure 4 b reports the percentage of stable models. At around 80 training-set structures, all Allegro models become stable, while the CAGO-based DeePMD models achieve stability at around 220 or 350 for the 200 meV/Å and 100 meV/Å targets, respectively. The slower convergence of DeePMD is in line with past research, which shows that MLIPs based on equivariant layers are not only more accurate but also more data-efficient[2,3], with similar equivariant NequIP also being benchmarked to be more stable than DeePMD[6]. However, DeePMD, with our CAGO structures, performs rather well compared to its data-hungry reputation[2]. To put this into context, ref. 6 used 10,000 structures and achieved stability for an average of 0.247 nanoseconds under *NVT* conditions with a simple water force-field. In contrast, we achieve 100% stability for two nanoseconds under the more challenging *NPT* conditions for a complex many-body ab initio method. Interestingly, active learning using more standard approaches (max. uncertainty and random 500 K) achieves stability quicker than CAGO-based active learning. This could be due to the sampling yielding higher uncertainty structures than CAGO, which targets moderate force errors.

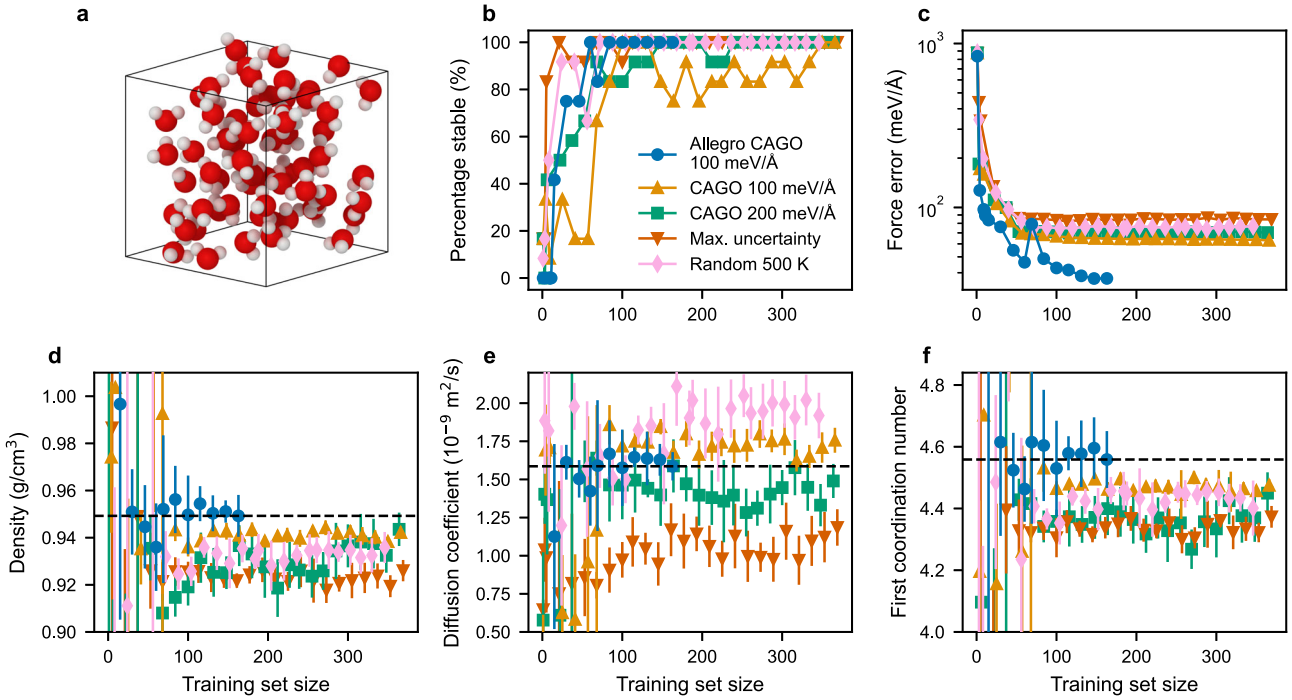

Fig. 4 | Benchmark of convergence of liquid-water properties for CAGO-based and standard active learning starting out from a single structure. We report data for the Allegro MLIP with target error 100 meV/Å and for DeePMD with target error 100 meV/Å, 200 meV/Å, as well as picking maximum uncertainty structure from MD and random structure sampling from 500 K MD simulations (see Methods IV F for more details). Each point in every graph has been computed from 12 models using the subset of stable MLIPs, with three replica MD simulations lasting 2 ns each, for a total of 540 Allegro and 3744 DeepMD simulations. The error bars are the standard deviation between the different committee members, and the horizontal lines correspond to the average final value of the property for Allegro CAGO. **a** The liquid water box structure that was used to start the active learning training. Oxygen is represented in red, with hydrogen in white. **b** Percentage of stable MLIPs in accordance with the stability criteria in Methods IV F. **c** Average force error of MLIPs on the liquid water structures. **d** Mean liquid water density. **e** Self-diffusion coefficient. **f** First coordination number (integral of the first radial distribution function peak).

This would be in line with the previously observed trend of higher target errors leading to more stable models.

To benchmark how well CAGO-based active learning performs for energetics, we report in Fig. 4c the average force error for 100 undistorted liquid water MD structures. In agreement with the literature[3], the Allegro force errors are considerably smaller than those of DeePMD, with Allegro CAGO 100 meV/Å converging to 37 meV/Å, and DeePMD CAGO 100 meV/Å and 200 meV/Å converging to 63 meV/Å and 70 meV/Å, respectively. Similar force error magnitudes were also reported in ref. 2. It is noteworthy that while CAGO distorts the structures from MD during training, CAGO improves force errors of MLIPs on undistorted liquid MD structures. Although stability for DeePMD is achieved with fewer training structures for standard active learning approaches, the force errors are higher: max. uncertainty gives an error of 85 meV/Å, (approximately 35% higher than CAGO 100 meV/Å), and random 500 K gives an error of 77 meV/Å (approximately 22% higher than CAGO 100 meV/Å).

Next, in Fig. 4d–f, we benchmark the thermodynamical, dynamical, and structural properties of liquid water, respectively. As with the stability, for all properties, Allegro converges around 80 structures, with DeePMD following closely. Overall, DeePMD exhibits larger variability for these properties. This is consistent with past benchmark studies on MB-pol-based DeePMD, which, even for huge training sets, struggle to achieve the correct density[5,32], whereas Allegro gets it spot on[33]. Nonetheless, we only observe variability within a few percent, and CAGO-based DeepMD models are within one standard deviation of the final Allegro iteration. For DeePMD, comparing against more standard active learning approaches (max. uncertainty and random 500 K), CAGO with 100 meV/Å performs best, followed by CAGO with 200 meV/Å, followed closely by random 500 K, and max. uncertainty. This order of accuracy is the same as the order of accuracy for the energetics as reported in Fig. 4c (see Supplementary Table 3 for the numbers of the final models).

Learning water-adsorption of a metal-organic framework

Water in nano-porous materials constitutes an extra challenging system due to the combination of a vast configurational space associated with water as a guest molecule and the potential for the material to catalyze chemical transformations, such as proton hopping. In Fig. 5, we report a benchmark of CAGO-based active learning with DC-R^{2}SCAN for water adsorption inside UiO-66, a zirconium-based metal-organic framework (MOF) with a very large surface area as well as high thermal stability[34]. We start the training from 11 structures with different water content (see Supplementary Note 2). For all benchmarked properties, we see a systematic convergence with training set data just after ~ 250 training set structures. This is a higher number than for liquid water, but this system also involves two additional chemical species with a higher level of chemical variability. To put this into context, a similar study of zeolite-OSADA pair with adversarial active learning started from a total of 17 492 structures and achieved a 97% stability rate after adding 573 adversarial structures[23]. In another impressive study, a training set of 400 000 energies and forces from adversarial active learning, using gas-phase calculations, was used to develop an MLIP for silica with reactive water[24].

Discussion

In this paper, we presented the CAGO algorithm to systematically generate adversarial molecular structures with user-assigned MLIP prediction errors. Through our benchmarks of the different hyperparameters of CAGO, we establish suitable minimums. In particular, we found it important to use many models, about 10 models in our case, to obtain a good error calibration, while the CAGO algorithm itself works well even for small committees. For the true error to converge to the target error, moderate error targets should be used. This is in line with our error calibration analysis, where we saw that error calibration is less robust for the high-error regime, where calibration data is scarce.

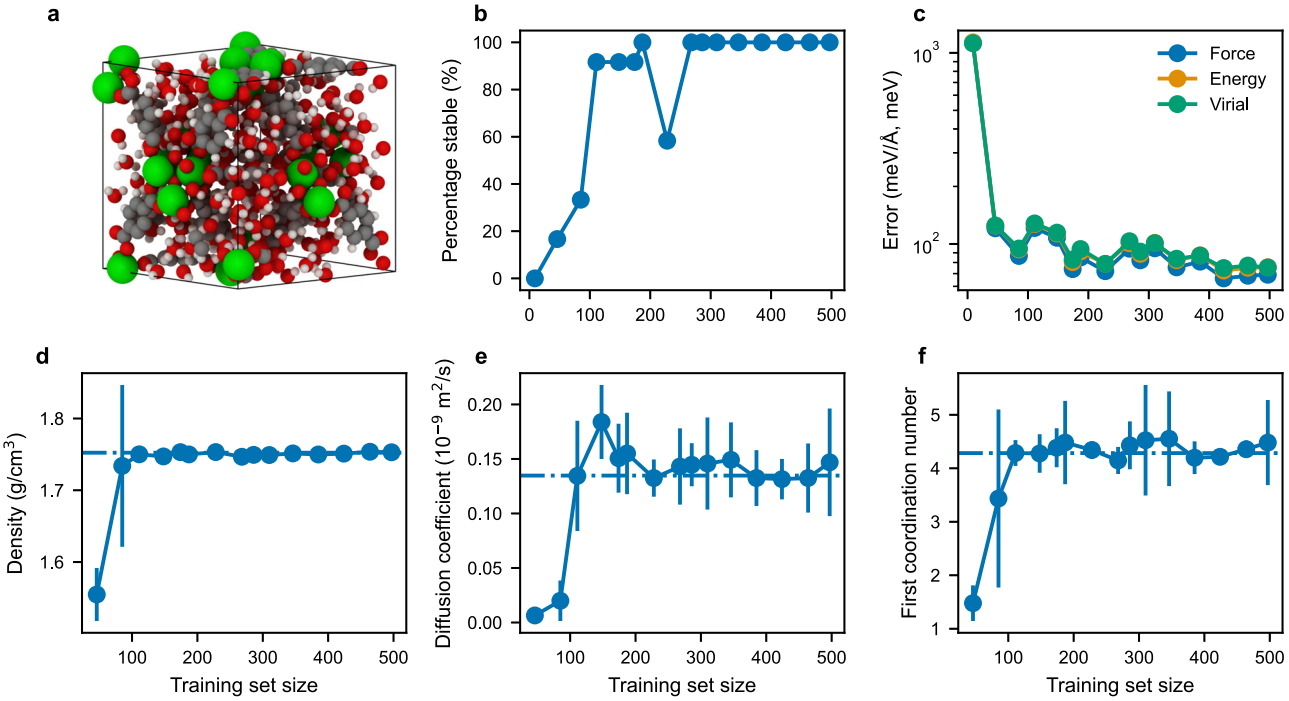

Fig. 5 | Benchmark of convergence of properties for water-adsorption in MOF for CAGO-based active learning for DC-R^{2}SCAN, with Allegro MLIPs. When presented, the error bars are the standard deviation of the measurements of each committee member at that iteration. Each point in the graphs has been computed from 12 models using the subset of stable MLIPs, with three replica simulations each, for a total of 576 two nanoseconds MD simulations. The error bars and horizontal lines hold the same meaning as in Fig. 4. **a** Water MOF system used in this benchmark. The color coding for the atoms is red, white, gray, and green for oxygen, hydrogen, carbon, and zirconium, respectively. **b** Percentage MLIPs fulfilling stability criteria in Methods IV F. **c** Average force, energy, and virials error on MOF/water structures. **d** Mean mass density. **e** Water diffusion coefficient. **f** First oxygen-oxygen coordination number.

These insights enabled us to perform active learning using minimal datasets with simple active learning protocols. In particular, we learned liquid water from a single structure of water for both Allegro and DeePMD and only 11 for water adsorption in the UiO-66 metal-organic framework. Although CAGO seemingly introduces an extra step of complexity into the active learning scheme, it ultimately simplifies the rest of the workflow. With CAGO, fewer manual adjustments and tricks are needed, such as low/high-temperature sampling or reducing the size of MD steps to achieve accurate models. Moreover, we have used orders of magnitude fewer structures than is typically reported in other works on adversarial active learning[23,24]. A key problem with current MLIPs is their typical dependence on large and approximate but cheap DFT datasets. CAGO-based active learning offers a promising solution by only requiring a few structures, allowing highly accurate but costly coupled-cluster and quantum Monte Carlo reference calculations to be used as training data.

Methods

CAGO algorithm and implementation

The CAGO algorithm works much like ordinary molecular geometry/structure optimization, but instead of minimizing the energy, CAGO minimizes the fitness function in equation (8), to obtain a prescribed uncertainty for the MLIPs. In particular, we use the L-BFGS algorithm, as implemented in the SciPy library[35], to optimize equation (8) with respect to scaled coordinates and cell vectors. The gradients are computed numerically by a two-point finite difference operator. While computing the hessian analytically is possible and most likely faster for differentiable MLIP architectures, as done in ref. 23, numerical derivatives have certain advantages. First, single-point calculations are fast with MLIPs, making the $3N+6+1$ calculations needed for numerical gradient calculations feasible. Second, these single-point calculations can be prepared for batch-based calculations, rather than being performed one by one, which is efficient with GPUs. Third, the calculations are trivially parallelizable, making linear scaling by adding more GPUs possible. Fourth, Hessians are generally not highly optimized with PyTorch and incur significant overheads. Finally and most importantly, while single-point calculations are ubiquitous among MLIPs, Hessians are rarely available out of the box and, therefore, not amenable to general active learning pipelines, forcing a reliance on specific MLIP architectures that may not be state-of-the-art.

Machine learning interatomic potentials

We tested CAGO on two different MLIPs with different architectures: *Allegro* and *DeepMD*. For Allegro models[3], we used equivariant E(3) products up to $L_{max} = 2$ in the tensor layers, with two interaction layers, including three tensor product layers of 128 neurons each. Before the interaction layer, a feature layer with 16 input features was used. Following the interaction layer block, we used three latent layers, each containing 128 neurons. A polynomial cutoff of 6 Å was used with a trainable Bessel basis set of 8 basis functions to form the feature descriptors. We used a radial cutoff of 6 Å for building the neighbor lists. Our loss function used forces, energy, and stresses, with coefficients 1, 5, and 100, respectively. These settings are similar to those previously used to generate water MLIPs with Allegro[33].

For DeePMD models[36], we used the se_e2_a architecture and 25, 50, 100 neurons for the hidden embedding layer, with the submatrix of the embedding matrix using 16 neurons. Similarly to Allegro, a distance cutoff of 6 Å was used, but with a smoothing region of 0.5 Å. The potential is represented by a fully connected deep neural network with three layers of 240 neurons each. These settings have also been used before in the study of different water phases[5,29].

Input files detailing the setup of MLIPs are available in the data repository associated with this manuscript.

Reference calculations for training and test set data

For benchmarking CAGO in II E, we ran reference caclulations using the DNN@MB-pol model developed in refs. 29,37 based on the MB-pol water model[38,39], using the se_e2_a DeePMD MLIP architecture[36].

To benchmark CAGO as part of active learning, we employed the DC-R^{2}SCAN method. To calculate the DC-R^{2}SCAN reference energies, forces, and virials for the MLIPs, we used the recent implementation for density-corrected DFT in CP2K[40], which has in a series of papers demonstrated great accuracy with benchmarks against coupled-cluster levels of theory[30,31]. The atomic core electrons were described using Goedecker-Teter-Hutter (GTH) pseudopotentials, while valence electron molecular orbitals were expanded in triple-zeta double-polarized basis sets (TZV2P) optimized for the SCAN functional. The kinetic energy cutoff for the plane-wave expansion of the density was set to 2500 Ry, which was required to get a numerical accuracy of ~ 2 meV/Å, and ~ 0.05 kBar for forces and stress-tensor trace, respectively. The truncated Coulomb operator with a cutoff radius between 5.0 and 5.5 Å was used depending on the system, corresponding to approximately half the length of the smallest edge of the simulation cell. To overcome the expense of the Hartree-Fock calculation, the ADMM approximation was used with an optimized valence basis set (BASIS_ADMM_UZH). The Schwarz integral screening threshold was set to 10^{-6} atomic units, starting from a converged PBE calculation. All the DC-R^{2}SCAN calculations in this work were carried out with CP2K, with associated inputs and outputs stored in the data repository associated with this manuscript.

CAGO-based active learning procedure and implementation

The CAGO-based active learning procedure is an iterative process depicted in Fig. 1. Our implementation begins with an initial set of reference data, reserving 10% of this data for a test set. The remaining 90% reference data is partitioned into an 80% training set and a 10% validation set through bootstrapping for each machine learning interatomic potential (MLIP) committee member, as described in ref. 26. After training the committee of MLIPs, their uncertainty is calibrated using the test set. Subsequently, MD simulations are conducted with the MLIP committee. From MD simulations, a subset of structures are randomly sampled to create adversarial structures with CAGO. To reduce the risk for failed reference calculations, a filter is applied to these structures, such as an upper threshold for the mean force magnitude of the MLIP committee. Reference calculations are then performed on the filtered adversarial structures before undergoing a final filter, checking for convergence and avoiding unphysical structures dominated by sterics before incorporating them into the reference data set, completing one iteration of the active learning cycle. This entire active learning loop is implemented in the Hylleraas Software Platform[41], which interfaces with various MLIPs and quantum chemistry softwares, enabling the use of heterogeneous computing environments on high-performance computer clusters.

Computational procedure for calibrated adversarial geometry optimization to target error

For benchmarking the CAGO algorithm in section II E, we targeted the DNN@MB-pol water model as specified in Methods IV C. First, we trained a committee of 20 DeePMD MLIPs on a training set of MD trajectories. Second, using these 20 models and a separate test set, we performed uncertainty calibration in accordance with equation (5), the results of which are reported in Fig. 2. Third, sampling 40 structures from the test set, we perform CAGO with various hyperparameters, such as different error targets and committee sizes. The errors we report are root mean square errors of individual models against DNN@MB-pol. For more details on the training and test set, we refer to Supplementary Note 1.

Computational procedure for learning liquid water from a single structure

For the active learning workflow, we used committees with 12 MLIPs to learn liquid water from a single structure of 64 water molecules, with details on initial structures reported in Supplementary Note 2. In each of the first two active learning iterations with CAGO, we sampled five structures from the training set and optimized the structures with CAGO using random subsets of 3 committee members. For all subsequent iterations, MD sampling with LAMMPS[42] was performed before CAGO, and in the case of pure water, 20 samples were extracted and taken for CAGO considering all 12 committee members. Due to the low number of structures in the first iterations of active learning, the MLIPs tend to be unstable; we, therefore, performed *NVT* simulations at 300 K to generate new structures for CAGO for iterations 3 and 4. These simulations were run for 50 ps using the velocity-Verlet integrator with a 0.25 fs timestep and Nosé-Hoover thermostat chain with 3 thermostats and a 0.5 ps relaxation time. From iteration 5 and onwards, we performed *NPT* simulations at 300 K and 1 bar, sampling for 100 ps, using the same settings for the thermostat part, and a Nosé-Hoover barostat chain of 3 barostats with 1 ps relaxation time. Structures derived from CAGO with maximum forces (mean by committee members) higher than 40 eV/Å or maximum force from reference calculations higher than 30 eV/Å, or minimum distance lower than 0.75 Å, or with convergence warnings from CP2K, were filtered out. Note that the DeePMD MLIPs used slightly varying settings for the training throughout the active learning loop to avoid overfitting. In the first two iterations, we trained the models for 50 000 steps to avoid overfitting with small training sets. As more structures were added, we increased the number of training steps to 150 000 for the next two active learning iterations, and from iteration 5 and onwards, we used 400 000 training steps. For active learning with maximum uncertainty and elevated temperature, MD-sampled structures were used for all iterations. 500 K was used for elevated temperature active learning. For maximum uncertainty, we picked structures based on the highest force uncertainty on any atom in the structure. For elevated temperature active learning, we sampled structures at random. In both cases we filter high-force structures as in the CAGO-based active learning.

For the benchmark of MLIPs along active learning iterations in Fig. 4, we ran 3 independent simulations starting from different initial conditions in a simulation box with 256 water molecules. We first performed two thermalizations for each system, one in *NVT* (25 ps) and another in the *NPT* (100 ps) ensemble, using the Berendsen thermostat and barostat. Finally, we ran a 2 ns *NPT* simulation using the Nosé-Hoover thermostat and barostat. These simulations used a 0.5 fs timestep, and all other settings were the same as the ones described above for the sampling phase of active learning.

We use two criteria to define the stability of the MLIPs during MD simulations. First, we check every 100 timesteps that the minimal distance between atoms in the system is never below 0.4 Å. Second, we monitor the system density, verifying that it does not diverge to values lower than 0.25 times or higher than 2 times the initial density. Furthermore, we ran multiple replicas, considering the MLIP to be stable only if all trajectories of the replicas meet our stability criteria.

Computational procedure for learning water-adsorption of a metal-organic framework

The water adsorption in the UiO-66 simulations was performed using similar workflow settings as for liquid water in Methods, with a few modifications as follows. In particular, for this active learning workflow, we considered multiple systems, including liquid water and 10 structures of UiO-66 with differing water content (see Supplementary Note 2). Therefore, in this workflow, we started from an initial reference data set of 11 structures and extended the reference data set for each individual system. In each of the first two active learning iterations, we sampled five structures from the training set and optimized the structures with CAGO and random subsets of 3 committee members for each of the systems. For the rest of the iterations, we performed *NPT* MD sampling for each system with all committee members, followed by CAGO on five structures with four committee members for each of the 11 systems. We used a timestep of 0.5 fs and ran the simulations for 100 ps during the active learning sampling phase. The benchmark of MLIPs along active learning iterations in Fig. 5 were performed using the system presented in Fig. 5a, which is UiO-66 with a water weight percentage of 45, with the same MD protocol as in Methods.

Data availability

The dataset and scripts used to produce the data in this study are publicly available via the Norwegian it National e-Infrastructure for Research Data} (NIRD) at https://doi.org/10.11582/2025.00018.

Code availability

The code implementing the active learning loop using CAGO is publicly available at GitLab: https://gitlab.com/hylleraasplatform/hyal.

Received: 7 March 2025; Accepted: 13 June 2025;
Published online: 01 July 2025

References

1. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
2. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
3. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
4. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S. et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (Curran Associates, Inc., 2022).
5. Zhai, Y., Caruso, A., Bore, S. L., Luo, Z. & Paesani, F. A "short blanket" dilemma for a state-of-the-art neural network potential for water: Reproducing experimental properties or the physics of the underlying many-body interactions? *J. Chem. Phys.* **158**, 084111 (2023).
6. Fu, X. et al. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research* (2023).
7. Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *J. Stat. Plan. Inference* **90**, 227–244 (2000).
8. Yang, Y., Zhang, S., Ranasinghe, K. D., Isayev, O. & Roitberg, A. E. Machine learning of reactive potentials. *Annu. Rev. Phys. Chem.* **75**, 371–395 (2024).
9. Vandermause, J. et al. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 20 (2020).
10. 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).
11. Wen, M. & Tadmor, E. B. Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Comput. Mater.* **6**, 124 (2020).
12. Krogh, A. & Vedelsby, J. Neural network ensembles, cross validation, and active learning. In Tesauro, G., Touretzky, D. & Leen, T. (eds.) *Advances in Neural Information Processing Systems*, vol. 7 (MIT Press, 1994).
13. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
14. Zhang, Y. et al. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Computer Phys. Commun.* **253**, 107206 (2020).
15. Schran, C. et al. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. USA* **118**, e2110077118 (2021).
16. 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).
17. 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).
18. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148** (2018).
19. Zeng, J., Cao, L., Xu, M., Zhu, T. & Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **11**, 5713 (2020).
20. Goodfellow, I. *Deep learning*, vol. 196 (MIT press, 2016).
21. Cubuk, E. D. & Schoenholz, S. S. Adversarial Forces of Physical Models. *3rd NeurIPS workshop on Machine Learning and the Physical Sciences* (2020).
22. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
23. Schwalbe-Koda, D., Tan, A. R. & Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun.* **12**, 5104 (2021).
24. Roy, S., Dürholt, J. P., Asche, T. S., Zipoli, F. & Gómez-Bombarelli, R. Learning a reactive potential for silica-water through uncertainty attribution. *Nat. Commun.* **15**, 6030 (2024).
25. Zaverkin, V. et al. Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials. *npj Comput. Mater.* **10**, 83 (2024).
26. Palmer, G. et al. Calibration after bootstrap for accurate uncertainty quantification in regression models. *npj Comput. Mater.* **8**, 115 (2022).
27. Musil, F., Willatt, M. J., Langovoy, M. A. & Ceriotti, M. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).
28. Imbalzano, G. et al. Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **154**, 074102 (2021).
29. Bore, S. L. & Paesani, F. Realistic phase diagram of water from "first principles" data-driven quantum simulations. *Nat. Commun.* **14**, 3349 (2023).
30. Dasgupta, S., Lambros, E., Perdew, J. P. & Paesani, F. Elevating density functional theory to chemical accuracy for water simulations through a density-corrected many-body formalism. *Nat. Commun.* **12**, 6359 (2021).
31. Dasgupta, S., Cassone, G. & Paesani, F. Nuclear quantum effects and the Grotthuss mechanism dictate the pH of liquid water (2024).
32. Zhai, Y., Rashmi, R., Palos, E. & Paesani, F. Many-body interactions and deep neural network potentials for water. *J. Chem. Phys.* **160** (2024).
33. Maxson, T. & Szilvási, T. Transferable water potentials using equivariant neural networks. *J. Phys. Chem. Lett.* **15**, 3740–3747 (2024).
34. Cavka, J. H. et al. A new zirconium inorganic building brick forming metal organic frameworks with exceptional stability. *J. Am. Chem. Soc.* **130**, 13850–13851 (2008).
35. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
36. Zeng, J. et al. DeePMD-kit v2: A software package for deep potential models. *J. Chem. Phys.* **159**, 054801 (2023).
37. Sciortino, F., Zhai, Y., Bore, S. & Paesani, F. Constraints on the location of the liquid–liquid critical point in water. *Nature Physics* 1–6 (2025).
38. Babin, V., Leforestier, C. & Paesani, F. Development of a "first principles" water potential with flexible monomers: Dimer potential energy surface, vrt spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **9**, 5395–5403 (2013).
39. Babin, V., Medders, G. R. & Paesani, F. Development of a "first principles" water potential with flexible monomers. ii: Trimer potential energy surface, third virial coefficient, and small clusters. *J. Chem. Theory Comput.* **10**, 1599–1607 (2014).
40. Belleflamme, F. & Hutter, J. Radicals in aqueous solution: assessment of density-corrected SCAN functional. *Phys. Chem. Chem. Phys.* **25**, 20817–20836 (2023).
41. *Hylleraas Software Platform***hyal**, https://gitlab.com/hylleraasplatform (2025).
42. Thompson, A. P. et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Computer Phys. Commun.* **271**, 108171 (2022).

Acknowledgements
The work was supported by the Research Council of Norway through the Centre of Excellence Hylleraas Centre for Quantum Molecular Sciences (Grant 262695) and the Young Researcher Talent grants 344993 and 354100. We acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call (Grants EHPC-REG-2023R02-088, EHPC-REG-2023R03-146). Support was also received from the Centre for Advanced Study in Oslo, Norway, which funded and hosted the SLB Young CAS Fellow research project during the academic year of 23/24 and 24/25. Part of the simulations were performed on resources provided by Sigma2 — the Norwegian National Infrastructure for High-Performance Computing and Data Storage (grant numbers NN4654K and NS4654K).

Author contributions
H.M.C., T.B., H.A.S., M.L., S.R. and S.L.B. contributed to the methods and software that are behind the presented results. S.L.B. and H.M.C. ran simulations and analyzed data. H.M.C., T.B., H.A.S., M.L., S.R. and S.L.B. contributed to the writing of the manuscript.

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/s41524-025-01703-5.

Correspondence and requests for materials should be addressed to Sigbjørn Løland Bore.