Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncertainty Driven Dynamics for Active Learning of Interatomic Potentials

Maksim Kulichenko (mbkulichenko@gmail.com)
Los Alamos National Laboratory https://orcid.org/0000-0002-6194-3008
Kipton Barros
Los Alamos National Laboratory
Nicholas Lubbers
Los Alamos National Laboratory
Ying Wai Li
Los Alamos National Laboratory
Richard Messerly
Los Alamos National Laboratory
Sergei Tretiak
Los Alamos National Laboratory https://orcid.org/0000-0001-5547-3647
Justin Smith
Los Alamos National Laboratory
Benjamin Nebgen
Los Alamos National Laboratory

Article

Keywords:

Posted Date: October 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2109927/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Uncertainty Driven Dynamics for Active Learning of Interatomic Potentials.

Maksim Kulichenko[1*], Kipton Barros[1,2], Nicholas Lubbers[3], Ying Wai Li[3], Richard Messerly[1], Sergei Tretiak[1,2,4], Justin S. Smith[1,5,*],Benjamin Nebgen[1,*]

[1] Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
[2] Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
[3] Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
[4] Center for Integrated Nanotechnologies, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
[5] Nvidia Corporation, Santa Clara, CA 9505 , United States

[*] Corresponding authors: maxim@lanl.gov, jusmith@nvidia.com, bnebgen@lanl.gov

Abstract

Machine learning (ML) models, if trained to datasets of high-fidelity quantum simulations, produce accurate and efficient interatomic potentials. Active learning (AL) is a powerful tool to iteratively generate diverse datasets. In this approach, the ML model provides an uncertainty estimate along with its prediction for each new atomic configuration. If the uncertainty estimate passes a certain threshold, then the configuration is included in the dataset. A key challenge in this process is locating structures for which the model lacks underlying training data. Here, we develop a strategy to more rapidly discover configurations that meaningfully augment the training dataset. The approach, uncertainty driven dynamics for active learning (UDD-AL), modifies the potential energy surface used in molecular dynamics simulations to favor regions of configuration space for which there is large model uncertainty. Performance of UDD-AL is demonstrated for two challenging AL tasks: sampling the conformational space of glycine and sampling the promotion of proton transfer in acetylacetone. The method is shown to efficiently explore chemically relevant configuration

space, which may be inaccessible using regular dynamical sampling at target temperature conditions.

Main

Machine learning (ML) is a firmly established approach in chemical science that demonstrates great promise for the acceleration of physical simulations. A particular strength of ML models is a robust representation of potential energy surfaces of molecular and materials systems, when trained to large and diverse datasets of high-fidelity quantum chemistry simulations. For example, ML based potentials[1–11] approach *ab initio*[12,13] or density functional theory (DFT)[14] levels of accuracy at a computational cost near that of classical force fields.[15–17] Over the recent years, various ML-models — such as neural networks (NNs),[18–24] Gaussian approximation potentials,[25] spectral neighbor analysis potentials,[26] moment tensor potentials,[27] symmetric gradient domain machine learning[28,29] — have demonstrated remarkable success in the field atomic-scale discovery.

No matter how sophisticated the ML model architecture, however, the quality and diversity of the training data remains crucial to ultimate model accuracy since ML models are known to extrapolate poorly to unseen data. Therefore, training sets for ML potentials need to span as much phase (structural) space as possible to perform meaningful simulations. Additionally, the training set needs to be as diverse as possible to avoid overfitting towards excessively represented training data (e.g., near-equilibrium configurations in MD trajectories).

Entropy-maximization techniques[30,31] help to partially overcome these problems by maximizing the structural diversity of a data set. When acquiring new data, these methods are focused on the structural dissimilarity compared to the existing data. However, these methods usually require training of a separate Gaussian process model and significantly rely on the structural representation in latent space. There are other opportunities for improvement as well. Active learning (AL)[32,33] attempts to expand the dataset in areas where the ML model is most

uncertain, which leads to more rapid model improvement. Another feature of AL is that it can employ physically meaningful dynamical trajectories for the sampling of configurations. In the present work, we illustrate how to keep these benefits of AL, while accelerating the rate of new data collection.

AL[32,33,34] aims to iteratively collect diverse training datasets addressing any weaknesses identified in an ML model prediction. For this, it is necessary to estimate uncertainty for a model's predictions.[35–45] A well-established practical strategy for AL with NN potentials is Query by Committee[46] (QBC); here, the estimate of uncertainty is the disagreement between a collection of models within an ensemble. Typically, there are 5 to 10 NNs in an ensemble, and these share the same architecture and hyperparameters but, crucially, use a different initial randomization of the model parameters prior to training, and different splits of the training/validation data. It is empirically observed that the variance of the ensemble predictions correlates strongly with actual prediction error,[36] suggesting that the prediction task requires extrapolation beyond the range of the training data. In the QBC strategy, if this ensemble variance is observed to be large, then the training set will be augmented with new quantum simulation data.

AL estimates uncertainty in properties predicted for structures generated by an underlying sampler at each iteration. Molecular dynamics (MD) is the most popular method for sampling chemically meaningful potential energy surfaces. However, MD is susceptible to trapping in near-minimum conformations and only rarely enters chemically important regions such as transition states, which are key data for reactive simulations of chemical processes. In general, capturing thermodynamically rare events is a challenging task for any sampler. For example, metadynamics[47–50] is an effective method of potential energy surface exploration, which operates on the concept of collective variables (CVs). However, CVs require manual selection, and their number is limited in practice. The user needs to “look” at each specific molecule to determine the desired structural parameters to be scanned. Therefore, this approach is not suitable for automatic sampling.

Here, following the idea of QBC and ensemble uncertainty, we propose a new AL sampling algorithm biased towards regions of high uncertainty - Uncertainty Driven Dynamics (UDD). Due to model random initializations and the stochastic nature of training, regions of chemical space with low ensemble uncertainty will typically arise when similar regions are prevalent in the current training dataset, such that every member of the ensemble is making an accurate inference. Thus, biasing molecular dynamics in the direction of high ensemble uncertainty encourages the dynamics to visit new configurations, which are relevant for improving the diversity of the training set.

One can regard the uncertainty-based bias potential as similar to metadynamics in the sense that the sampling trajectories are pushed towards less-visited configurational regions. Here, however, CVs need not be defined. We show that within MD-based AL data acquisition, UDD helps substantially reduce the MD simulation time required to enter the high uncertainty region. Most importantly, the proposed approach enables efficient conformational and configurational sampling at low-temperature (low-T) conditions, making this approach essential for temperature sensitive molecules. UDD assists in sampling the chemically relevant subspace of high-energy space which contains important data such as transition states.

The value of the proposed approach is demonstrated in two test cases. First, UDD-AL is used for conformational sampling of a glycine molecule. We find that the bias potential technique generates a diverse dataset covering both low and high energy regions. This contrasts with high-temperature (high-T) MD-AL, which tends to skip over low energy regions. Second, in tests with acetylacetone at low-T conditions, the bias potential is observed to encourage the sampling of the phase space relevant to a proton transfer. Here we find that, in contrast with regular high-T MD, the bias potential technique encourages the reactive transition with very little distortion to the distribution of other degrees of freedom in the system.

Results

Uncertainty driven dynamics for active learning (UDD-AL)

Before introducing UDD-AL, let us first set the context by reviewing the related method of metadynamics. Here the CVs $s(q)$ are user-defined structural parameters being scanned by external Gaussian bias potentials. Usually, $3N-6$ dimensional atomic coordinates q of the simulated system are mapped to CVs $s(q)$. The corresponding energy function is defined as

$$E_{\text{metadynamics}}(s,t) = \sum_{k\tau<t} W(k\tau)\exp[-\sum_{i=1}^{N_{\text{CV}}} \frac{1}{2b_i^2}(s_i - s_i(q(k\tau)))^2], \quad (1)$$

where b_i is the width of the Gaussian function for the i^{th} collective variable, $W(k\tau)$ is the height of the Gaussian at the simulation time $t = k\tau$, which is constant in the case of standard metadynamics, τ is the deposition rate of the Gaussian functions, and N_{CV} is the number of CVs. During simulations, more Gaussians are added, thus discouraging the system to go back to its previous steps.

Like metadynamics, UDD-AL method modifies the physical energy by adding a bias potential E_{bias}. Here, however, E_{bias} will be defined in terms of the model uncertainty rather than CVs. Such uncertainty estimates can be used to assist in the sampling of atomistic data.[44] In the QBC approach, an ensemble of NN-potentials is trained, and the level of agreement between the NN predictions serves as the estimate of overall model uncertainty. The QBC uncertainty estimate is proportional to the variance in the energy (σ_E^2)

$$\sigma_E^2 = \frac{1}{2}\sum_{\iota}^{N_{\text{M}}}(\widehat{E_\iota} - \hat{E})^2, \quad (2)$$

where $\widehat{E_\iota}$ is the energy predicted by an ensemble member, $\hat{E}$ is its ensemble average, and N_{M} is the number of ensemble members, i.e., NN-potentials. Here, ensembles of ANI potentials are prepared using an 8-fold cross validation split of the data set which yields $N_{\text{M}} = 8$ ensemble members (see Methods). In previous applications of QBC-based AL, new data is collected when the uncertainty estimator

$$\rho = \sqrt{2/N_M N_A}\,\sigma_E \quad (3)$$

exceeds a threshold, where N_{A} is the number of atoms in a configuration.

We seek to construct a bias energy $E_{\mathrm{bias}}(\sigma_E^2)$ that favors configurations with larger uncertainties $\rho \propto \sigma_E$; such configurations are expected to correlate with regions that are underrepresented in the training data. A reasonable choice is the Gaussian function,

$$E_{\mathrm{bias}}(\sigma_E^2) = A\left[\exp\left(-\frac{\sigma_E^2}{N_{\mathrm{M}}N_A B^2}\right) - 1\right]. \quad (4)$$

The magnitude *A* and width *B* of the biasing should be selected empirically. The bias potential goes to zero in the absence of uncertainty, $E_{\mathrm{bias}}(0) = 0$. Configurations with large uncertainty, $\rho \gg B$, are favored by a bias energy of magnitude $E_{\mathrm{bias}} \approx -A$. Forces derived from the bias potential are strongest when the uncertainty ρ is of the same order as the parameter B.

The combined potential $\hat{E} + E_{\mathrm{bias}}$ is used to define an uncertainty driven dynamics (UDD). In applications to AL the overall strategy will be denoted UDD-AL. The schematic workflow of UDD-AL is depicted in Fig. 1a. It should be compared to the usual MD based approach (MD-AL), which does not incorporate the E_{bias} term.

The optimal parameters *A* and *B* will be context dependent. For example, our applications to glycine and acetylacetone tests discussed below suggest that the bias magnitude *A* should be at least of the order of the energy barriers of interest. The bias width, B, should also be carefully selected. If B is either too small or too large, the bias forces could become negligibly small for the typical uncertainties ρ that are observed in the AL sampling.

Effective bias forces on an atom at position r can be calculated using the chain rule,

$$-\frac{\partial}{\partial r}E_{\mathrm{bias}}(\sigma_E^2) = -E_{\mathrm{bias}}(\sigma_E^2)'\frac{\partial}{\partial r}\sigma_E^2 \quad (5)$$

and

$$-\frac{\partial}{\partial r}\sigma_E^2 = -\sum_i^M(\hat{E}_i - \hat{E})\frac{\partial}{\partial r}(\hat{E}_i - \hat{E}) = \sum_i^M(\hat{E}_i - \hat{E})(\hat{f}_i - \hat{f}) \quad (6)$$

where $\widehat{f}_i$ denotes the force vector predicted by an ensemble member and $\hat{f}$ is the ensemble-averaged prediction. Since σ_E^2, $\widehat{E}_i$, and $\widehat{f}_i$ are calculated at each MD step in any AL approach, including the bias potential has a negligible effect on the computational time.

Glycine conformational space sampling

The glycine molecule is shown in Fig. 1a. We are interested in sampling the conformational space without bond-breaking events using various AL protocols. Dihedral rotations of $–NH_2$ and –COOH groups correspond to 2.5-3.5 kcal/mol barriers. Our numerical tests have shown that the bias magnitude *A* approximately five times higher than the average barrier of interest provides the best results for the glycine test. It helps keep the bias essentially high at higher uncertainty values. It also helps overcome possible barrier bottlenecks caused by possible overestimation of barrier heights by ML model. The geometries near global energy minimum (GM) of a glycine are already present in the initial training set, and each sampling MD simulation starts with this kind of structure. Depending on the AL iteration, near-GM structures have ρ of 0.024±0.005$[\text{kcal} \times \text{mol}^{-1} \times (\sqrt{N_A})^{-1}]$ (σ_E =0.15±0.03 kcal/mol). Thus, we selected *A* equal to 15.4 kcal/mol which corresponds to the bias potential value of ~15.0 kcal/mol at near-GM values of ρ (with $B = 0.12$ kcal/mol). We further increase *A* by 15% at 140ps time step if the uncertainty criterion is not met at this simulation stage. Our tests also show that best results are achieved for B being close to near-GM uncertainty σ_E. Here we use *B* = 0.12 kcal/mol for glycine.

We next compare the two active learning approaches – UDD-AL and MD-AL – for the task of collecting a dataset of the glycine conformational space. Each AL iteration performs 16 MD simulations with 200ps time limit and 1fs step (see Methods). An ensemble of NN-potentials for the first AL iteration is trained on the initial data set of 125 conformers that span near-equilibrium structures of the glycine global minimum. At each subsequent AL iteration, the MD simulation employs an ensemble of ANI-type NN-potentials (see Methods and Ref. 51), trained on the initial

data and data accumulated on all previous AL iterations. The starting geometries for MD simulations and the initial training set contain only near-equilibrium geometries of a glycine global minimum (Extended Data Fig. 1). Stated differently, NNs have no initial information about higher energy conformers, and MD simulations have to reach them from the bottom of the potential energy surface. Each MD simulation is terminated when the system meets the uncertainty selection criteria ρ of 0.35 $[\text{kcal} \times \text{mol}^{-1} \times (\sqrt{N_a})^{-1}]$ (see Methods). If the MD simulation reaches the time limit, then the structure from the trajectory with the highest uncertainty is selected. Density Functional Theory (DFT) reference data (see Methods) is then computed for final conformations and added to the training set for the next iteration of the AL process.

Figure 1b shows the average MD simulation time required to meet the uncertainty criterion in different AL approaches with respect to AL iteration. The MD-AL at low-T conditions (350K) reaches the MD time limit at ~20^{th} iteration which is continued until the final AL iteration (Fig. 1b, blue line). This means that the specified uncertainty criterion is almost never met, and the sampler returns the geometry of maximum available uncertainty from the MD trajectory. The uncertainty bias potential is introduced in low-T (350K) MD simulations (orange line). We do not activate the bias potential at earlier AL iterations for two reasons. First, the low-T MD-AL (Fig. 1b, blue line) does not reach the MD simulation time limit up to 15-20^{th} iteration. Thus, at this stage, the regular MD-AL manages to acquire new data satisfying the uncertainty criteria. Second, the NN potential might be unstable or not smooth at earlier iterations due to the lack of data. Therefore, the bias is activated at the 15^{th} iteration to avoid moving systems toward unphysical configurations. When the uncertainty bias potential is on (UDD-AL regime), it reduces the number of MD steps needed to meet the desired uncertainty. Moreover, the MD time limit plateau is still not reached till the final iteration.

Perhaps the most common way to accelerate sampling of high-energy states is to run high-T MD. Thus, to illustrate the difference between the bias potential and a simple temperature

increase, we also compare the low-T 350K UDD-AL with the high-T MD-ALs at 600K and 1000K simulation conditions (Fig. 1b). Like in case of UDD-AL sampling, the temperature is increased at the 15th iteration. 600K MD-AL approaches the MD time limit at ~40th iteration while UDD-AL reaches the time limit plateau by the end of AL procedure. Thus, 600K MD-AL does not perform as well as UDD-AL in terms of simulation time. On the other hand, the 1000K MD-AL (cyan line) exhibits a faster sampling, and the average MD time does not exceed 35ps even at the last AL iteration.

Energy ranges sampled by each AL approach are depicted in Fig.1c. Expectedly, high-T MD-AL data (green and cyan histograms) span wider energy ranges compared to low-T 350K MD-AL (blue histogram). What is more interesting is that the energy distribution of data from the 350K UDD-AL (orange histogram) is very similar to the shape of 600K MD-AL data; an advantage of UDD-AL is that many fewer MD steps are required to collect these samples.

After completing the entire AL procedure, the final models are trained on 1280 glycine conformers collected during the entire AL procedure (+125 conformers in the initial training set). In order to access the accuracy of the four models, we use a test set comprising 50,000 glycine structures from a 50ns MD simulation via the ANI-1ccx potential run at 400K with a 0.5fs time step (see Methods).[19] As depicted in Fig. 1d, all models perform reasonably well with root-mean-square deviations (RMSEs) less than 0.3 kcal/mol. However, the RMSE of the model trained on low-T 350K MD-AL data is slightly yet systematically lower than RMSEs of the rest of the models with ~0.11 kcal/mol difference. This is likely because MD simulations tend to oscillate in near-equilibrium positions, which is why this test set is dominated by low-energy geometries. Low-T 350K MD-AL, in turn, densely spans a narrow low-energy range, which might explain a slightly better performance of this model on a test set derived from a 50ns MD. Indeed, normalized histograms in Fig.1e point out that this seemingly large 50ns test set spans the energy region closest (but even lower) to the one covered by low-T 350K MD-AL data.

Note that the 50ns test set is a MD trajectory with no AL involved while MD-AL data comprises structures with high uncertainties. High uncertainty usually corresponds to a higher energy due to poor sampling in normal low-T MD. In other words, the 50ns MD test set is biased towards near-equilibrium oscillations, but MD-AL data are selectively augmented with higher energy isomers. Similar RMSE shift is observed for low energy rotations of –COOH and $–NH_2$ functional groups (Extended Data Fig.2). Therefore, the various models should also be tested on higher energy pathways.

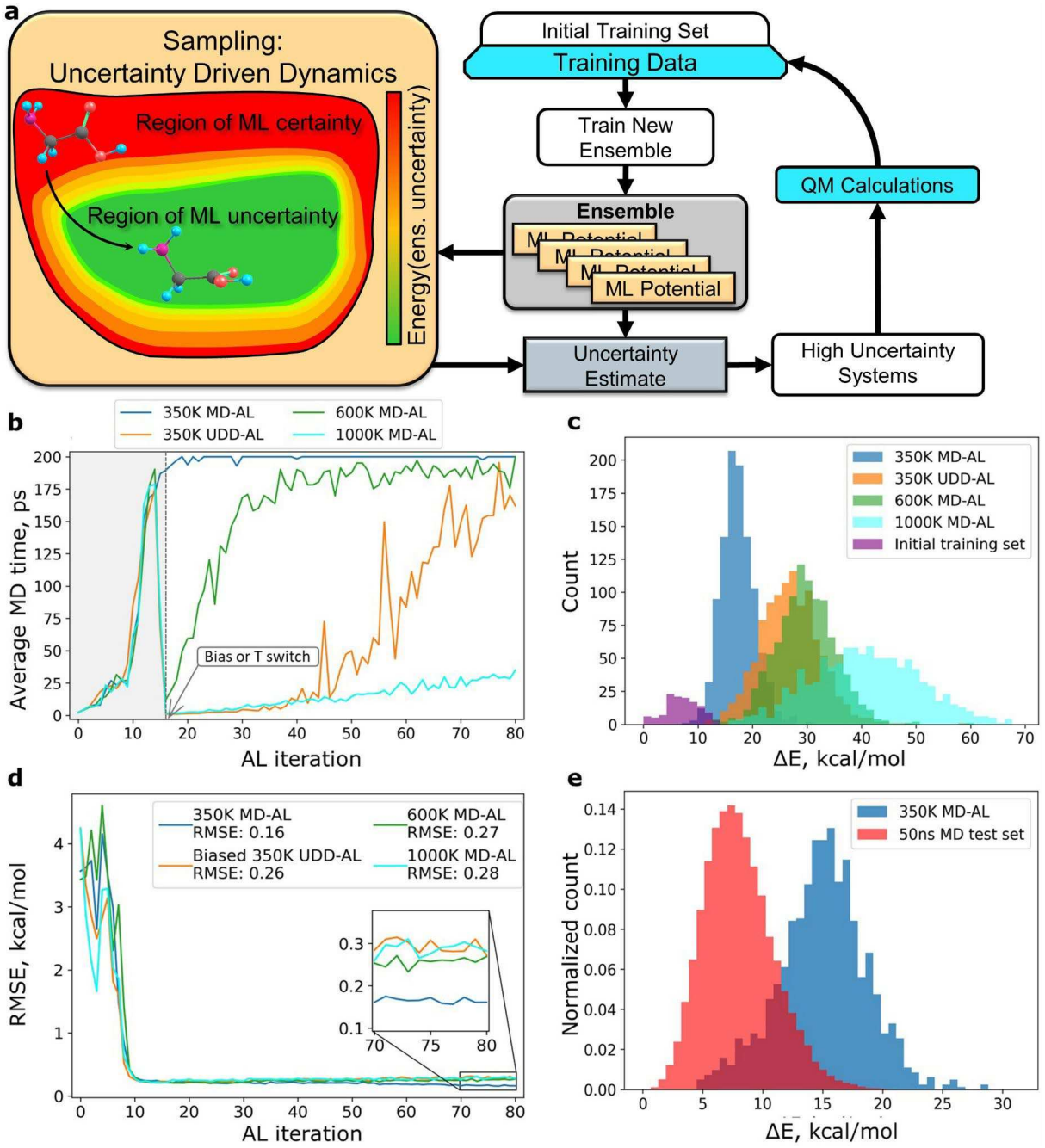


Fig.1 | Comparison of UDD-AL and MD-AL approaches for a glycine test case. a Schematic representation of UDD-AL workflow. **b** Average MD time required to meet the uncertainty criterion vs. AL iteration for four different MD simulation types: 350K MD-AL (blue), 600K MD-AL (green), 1000K MD-AL (cyan), and 350K UDD-AL (orange). **c** Energy distribution histograms of four datasets sampled by 350K MD-AL (blue), 600K AL MD-AL (green), 1000K MD-AL (cyan), and 350K UDD-AL (orange). Data from iterations 0-14 is omitted because the bias is off, or the temperature is not increased at this stage. **d**

Comparison of potential energy RMSE obtained on the 50ns test set vs. AL iteration (i.e., training set size, 16 new glycine conformations per iteration). The legend shows RMSEs for models trained on data from the entire AL procedure. **e** Normalized energy distribution histograms of 50ns test set (red) and training set sampled by 350K MD-AL (blue). Lines in **b** and **d**, are averaged over three ensembles, each trained on data from an independent AL procedure.

To illustrate what types of chemical processes appear in the system, and how each sampling method covers them, we next visualize the glycine conformational space using dimensionality reduction. We project samples to a 2D plane using the Uniform Manifold Approximation and Projection (UMAP)[52] technique, where each conformation is characterized by the 672-D vector of activations (concatenated atomic environment vectors) in the first layer of an independent pre-trained ANI-1x model.[19] Figure 2a shows the 50ns MD dataset spanning four low-lying glycine conformers represented by four regions in the 2D space. Torsional conformations, N-H bond scan, and C=O bond scan are depicted in Fig.2b-c to illustrate structural profiles in 2D space.

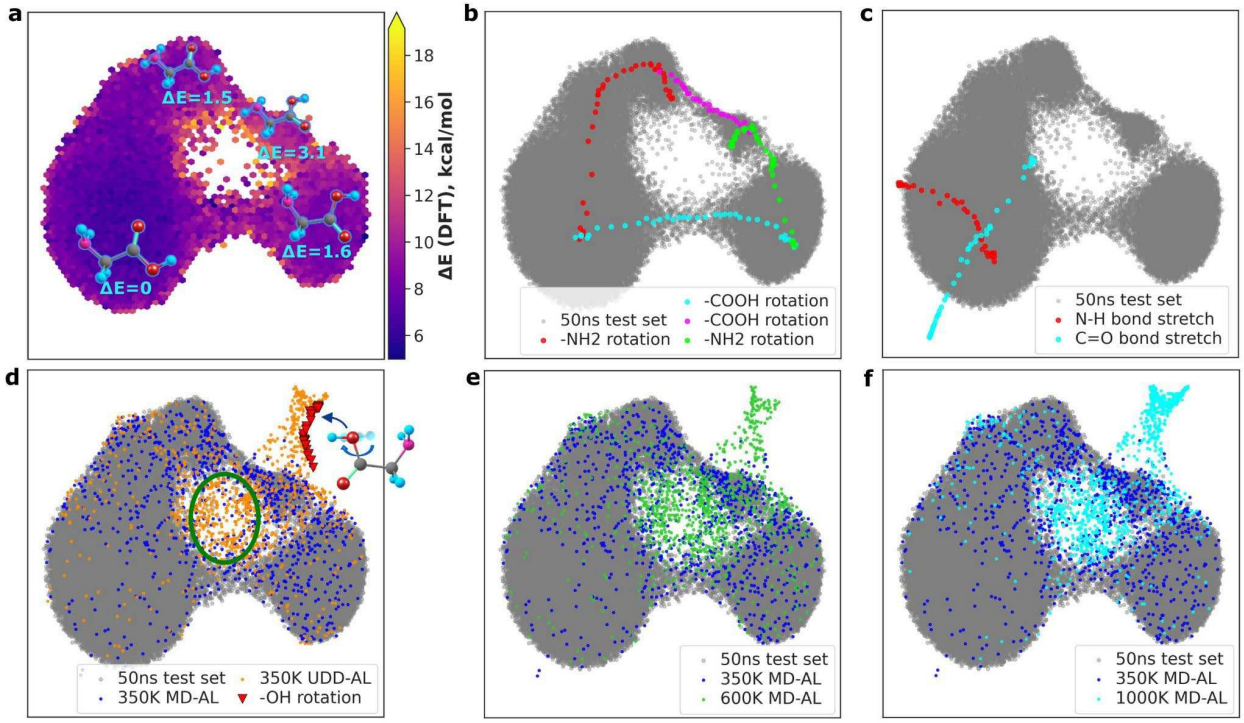


Fig. 2 | 2D representation of glycine conformational space processed by UMAP dimensionality reduction technique. a 50ns test set. Heat-map represents the relative DFT energy. Glycine insets denote the corresponding conformational region. In panels **b-d**, data sets and scans are placed over the 50ns test set (gray). **b** Conformational paths through -COOH (cyan and purple) and -NH_2 (red and green) rotations. **c** N-H (red) and C=O (cyan) bond length scans. **d** Comparison of training sets sampled by 350K MD-AL (blue) and 350K UDD-AL (orange). Green circle denotes the inner high-energy region. Red triangles denote the scan of -OH rotation around the C-O bond. **e** Comparison of training sets sampled by 350K MD-AL (blue) and 600K MD-AL (green). **f** Comparison of training sets sampled by 350K MD-AL (blue) and 1000K MD-AL (cyan).

Figure 2d-f depicts the four datasets visualized over the 50ns test set. Figure 2d provides a visual comparison of data sampled by MD-AL and UDD-AL at 350K. Both datasets cover the 50ns MD data reasonably well. However, there are three key differences. First, a high-energy configurational space (points inside the green circle) is more densely sampled in the UDD-AL dataset: there are 289 points in this non-equilibrium region in the UDD-AL sampling, compared to

105 points in 350K MD-AL. Second, UDD-AL encountered a new conformational path in the top right corner of the Fig.2d which was not accessed by 350K MD-AL. This region corresponds to the rotation of the –OH group around the C-O bond which is a distinct conformational transition and a high-energy profile with a 15 kcal/mol barrier.

Figure 2e presents a visual comparison of the sampling performance of MD-ALs at 350K and 600K. The 2D representation of 600K data is quite similar with the one of 350 UDD-AL data: there are 290 data points inside the inner circled region and a good coverage of –OH rotation region. As expected, MD-AL at the extreme temperature of 1000K (Fig.2f) samples the inner high-energy region even more densely (394 data points) as well as the –OH rotation region. This, however, comes at a cost. The low energy region in the lower left of Fig.2f clearly demonstrates lack of sampling. This is the primary deficiency of using high-T MD: as temperature increases, the system spends less time near low energy regions, since in these regions the kinetic energy is typically the greatest. Thus, it will be possible to ‘skip over’ regions of high stability thus resulting in a poor data coverage of near equilibrium region. On the other hand, UDD-AL sampling does not run this risk by sufficiently sampling any relevant region.

Thus, Figure 2 indicates that the UDD-AL is a good, balanced way of sampling chemical space, reaching most of the high energy points achieved with 1000K MD sampling without losing data density in low energy regions. However, alone, it is not clear that biased sampling presents advantages over unbiased high-T sampling; UDD-AL appears to sample similar configurations to 600K MD-AL. Therefore, we performed additional tests of high-energy pathways that illuminate the differences between 600K MD-AL and 350K UDD-AL. The discussion on high energy profiles - angle and bond scans - can be found in Supplementary Information. The overall trend is that the 350K UDD-AL model exhibits much better accuracy than the model trained on low-T 350K MD-AL data. When comparing the UDD-AL with 600K and 1000K MD-AL models, the former results in a better or, at least, comparable accuracy.

We also provide an overall assessment of the performance of the sampling strategy by cross-testing the associated models on the data from all sampling strategies in Supplementary information. Supplementary Table 2 summarizes RMSEs of the four models on the test sets accumulated by each AL sampler: 350K MD-AL, 350K UDD-AL, 600K MD-AL, and 1000K MD-AL. When testing models on data sets that are not generated by the same corresponding sampler, the UDD-AL model outperforms all other models. A detailed discussion on models' cross testing can be found in Supplementary Information.

Ultimately, when looking at a variety of bond rotations and stretches, the most accurate energy profile changes depending on the energy range of the specific scan. Low energy profiles tend to be modeled better by the low-T dataset while higher energy scans are accessed better by the high-T dataset. However, the UDD-AL sampling method yields a model that performs well on a wide range of energy profiles, while also maintaining a low error on each sampling method's held out test set (Table 2 in Supplementary Information). This difference suggests that UDD-AL is able to avoid the higher-energy and less chemically relevant structural distortions, which are typical at very high temperatures. Meanwhile, chemically relevant data present in the UDD-AL data set enables efficient extrapolation to higher-energy structures present in 1000K MD-AL data. As can be seen in Fig. 3, the shapes of interatomic distance distributions in UDD-AL closely mimic sharp distributions in low-T 350K MD-AL, although having a larger standard deviation. This deviation, however, is lower than in 600K and 1000K MD-AL data sets that span a wider distance range. This, in turn, further suggests that the UDD sampler tends to avoid random distortions found in high-T regimes.

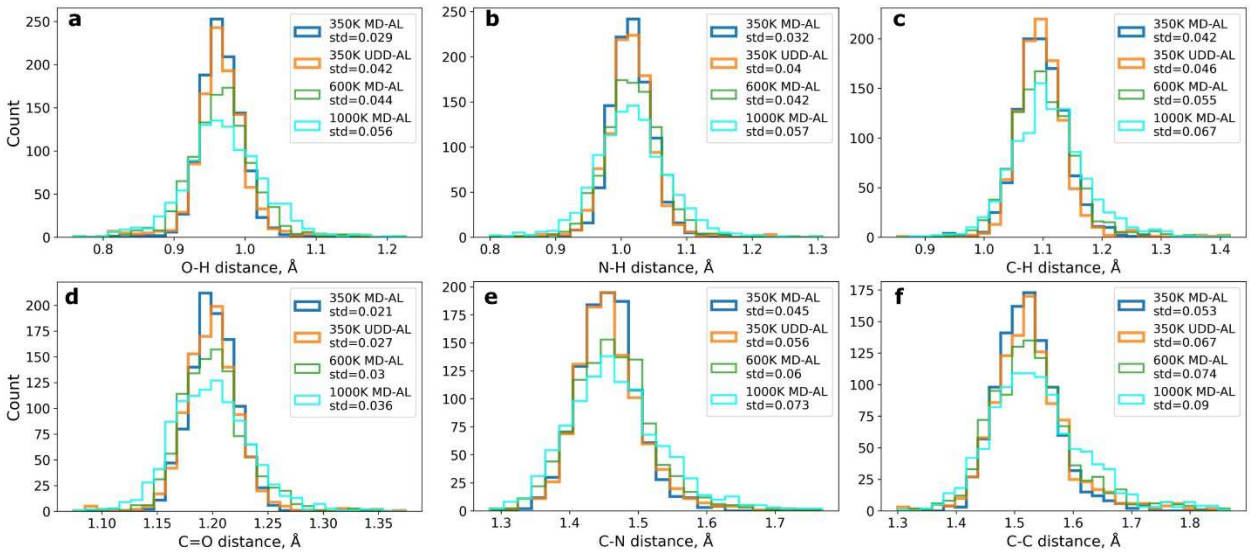


Fig. 3 | Glycine interatomic distance distributions in MD-AL and UDD-AL data sets. Each subplot shows a comparison of bond length distributions in 350K MD-AL (blue), 350K UDD-AL (orange), 600K MD-AL (green), and 1000K MD-AL (cyan) data sets. Each subplot also lists the bond length standard deviation, in the legend. **a** O-H. **b** N-H. **c** C-H. **d** C=O, **e** C-N. **f** C-C.

Since all models used the same hyperparameters, it is possible that each model could perform better if individual hyperparameter searches are carried out. Perhaps, data sets that cover a broader chemical space need more learnable parameters to be flexible enough to fit the effectively larger degrees of freedom they are being trained to. This will be a subject of future studies.

Proton transfer in acetylacetone

We further examine the UDD performance for sampling of a reactive pathway in an acetylacetone enol tautomer depicted in Fig. 4a. Particularly, we are interested in the proton transfer between the two oxygen atoms, considering the proton position as a free variable. Instead of using AL techniques, here we use an ensemble of pre-trained ANI-1x interatomic potentials,[19] which were not trained on bond-breaking reactions, and analyze trajectories from UDD and MD simulations. ANI-1x was trained on wB97x DFT level of theory which yields a 4.7 kcal/mol barrier. However, ANI-1x significantly overestimates the barrier giving the value of 6.3 kcal/mol. This value

is then selected as a bias magnitude *A*. Uncertainty values *ρ* of near-equilibrium acetylacetone structures within ANI-1x model are higher by the order of magnitude than the ones produced by the newly trained model for glycine. Therefore, we set a higher value of bias width $B = 0.45$ kcal/mol, being an empirically adjusted parameter.

Figure 4b shows log-normalized uncertainty *ρ* of the acetylacetone system with respect to the position of the proton. The dark region (near x=0 Å and y=-1.5 Å) demonstrates that there is a high-uncertainty region between oxygen atoms which corresponds to a proton transfer transition state. Figure 4c depicts the log-normalized relative potential energy (ANI-1x) of the system with respect to the position of the free proton. The dark region (near x=-0.5 Å and y=-1.5 Å) indicates that the lowest energy corresponds to a proton position near the oxygen atom. This is an expected result since the hydrogen bound to the oxygen atom is the most stable geometry. However, as shown in Fig. 4d, the energy minimum can be shifted to the central position between the oxygen atoms (dark region near x=0 Å) when the bias potential is applied. For illustrative purposes, here we use a high value of bias magnitude *A*=56.0 kcal/mol. In practice, we use *A*=6.3 kcal/mol for the MD simulation discussed below. The effect of different *A* values on the total energy landscape is depicted in Extended Data Fig. 3.

We further analyze results from 0.5ns trajectories obtained using UDD and regular MD simulation techniques. No proton transfer occurs during the regular 350K MD simulation. Meanwhile, the uncertainty bias can direct the proton toward a high-uncertainty region between two oxygen atoms - 90 proton transitions observed in 350K UDD simulation. Finally, unbiased high-T 620K MD results in 48 proton transitions. Although at a lower rate compared to UDD, increased temperature also facilitates the proton transfer. Time traces of the two O-H distances can be further found in Extended Data Fig. 4.

A key difference between low-T UDD and regular high-T MDs can be found when analyzing oscillations of interatomic distances. Certainly, high-T conditions will affect the entire molecule causing larger distance deviations compared to low-T conditions. Indeed, the overall

spread of O-H distances is comparable using 350K MD (Extended Data Fig. 4), but far wider at 620K MD, even in segments of the trajectory without proton transfer. Further analysis of C-H distances in the molecule, shown in Fig. 4e-f, confirms this phenomenon. Figure 4e shows C-H distance distributions in the methyl group in 350K MD, 350K UDD, and 620K MD simulations. 620K MD exhibits higher deviations from the equilibrium C-H bond length compared to low-T UDD. The standard deviation of the methyl C-H distance in 620K MD is 0.04 Å against 0.03 Å in 350K UDD. Notably, the low-T 350K MD trajectory has the standard deviation of 0.03 Å, same as in low-T UDD. The same picture holds for the central C-H bond shown in Fig. 4f.

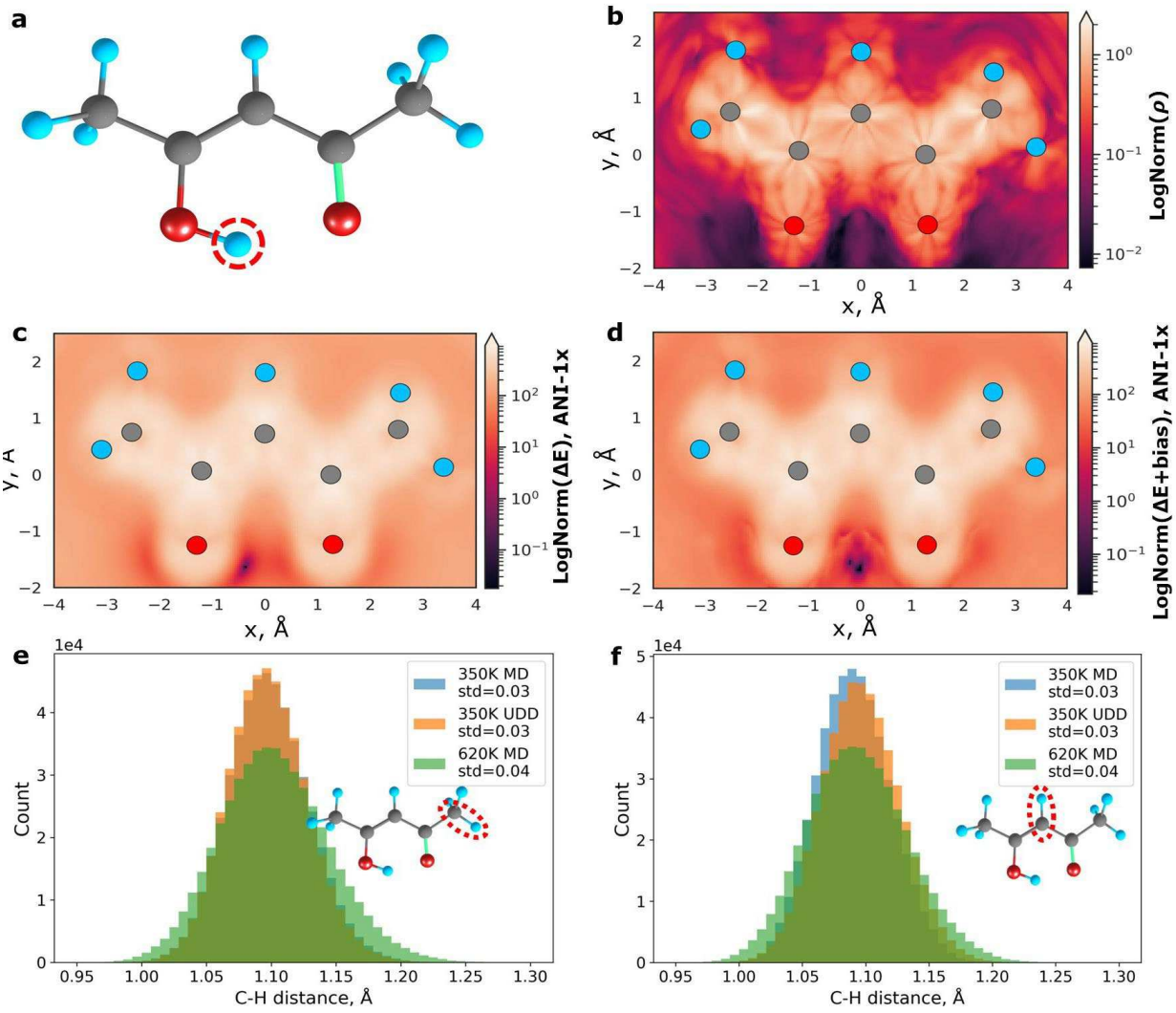


Fig. 4 | Ensemble uncertainty and UDD in acetylacetone. a Acetylacetone molecule. Red circle denotes a hydrogen atom involved in a proton transfer. **b** Log-normalized map of disagreement ρ of ANI-1x model ensemble with respect to a position of a circled hydrogen. **c** Log-normalized map of physical energy. **d** Log-normalized map of summed physical and bias energy. **e** and **f** show a comparison of C-H bond length distributions in the methyl and central groups, respectively, for 350K MD (blue), 350K UDD (orange), and 620K MD simulations. Red ellipse denotes a bond under consideration. Each subplot also lists the bond length standard deviation from the equilibrium distance per legend.

These observations confirm that although high-T sampling promotes the activation of reactive pathways, it has a global effect on all degrees of freedom in the system, whereas the

UDD technique allows to sample the reactive pathway without significant changes to equilibrium distributions of other degrees of freedom. This is likely due to the composition of ANI-1x training data, which has information on non-equilibrium extended bonds. However, a hydrogen which is equidistant between two oxygen atoms is not a commonly encountered configurational data. Thus, the UDD potential promotes sampling of this specific region.

Discussion

In this work, we present a MD-based active learning (AL) algorithm assisted by the uncertainty-based bias potential. The algorithm is abbreviated as UDD-AL (uncertainty driven dynamics for active learning) and is compared to MD-based AL (MD-AL) in common use. We show that an uncertainty metric provided by an ensemble of NN-potentials can be used to construct the bias potential. The resulting energy term encourages the system to move toward underrepresented chemical regions, accelerating and improving sampling of high-energy regions.

The ANI model, trained on glycine conformational data from low-T UDD-AL, is shown to properly reproduce conformational pathways not accessible by low-T MD-AL. The resulting ANI model exhibits high accuracy for important high-energy conformational profiles. Importantly, while this accuracy is better than or comparable to the results for models trained on high-T MD-AL data, a low-T UDD-AL training set spans an energy range narrower than that in high-T sets. However, a model trained on low-T UDD-AL data exhibits excellent accuracy on data sampled at unbiased high-T conditions which is not the case in the reverse test. This, in turn, suggests that the UDD-AL approach avoids oversampling the extreme structural distortions that are common at very high temperatures.

The test case of acetylacetone shows that bias potential applied to a pre-trained ANI-1x model promotes proton transfer in the enol tautomer keeping oscillations in the rest of the bond distances identical to low-T conditions. Opposite situation is observed for the unbiased high-T MD where interatomic distances exhibit larger deviations from equilibrium positions with a lower rate

of a proton transfer. The analysis of interatomic distances clearly indicates an advantage of the bias potential over the high-T approach: bias potential facilitates sampling of important underrepresented chemical data without random structural distortions caused by high-T conditions. This feature can be used for efficient sampling of conformational and/or configurational space of temperature sensitive systems.

Our tests indicate that the bias potential can facilitate sampling of high energy chemical space without sacrificing the sampling of low energy configurations. This means that UDD will produce robust datasets that are applicable to both lower energy, near global minimum data and high energy chemical space which usually corresponds to important reactive structural data such as transition states and intermediates.

As the results show, uncertainty-based bias potential is a promising technique for sampling rare events while being relatively faithful to the physical equilibrium distribution. The uncertainty driven dynamics is similar to metadynamics[47–50] in its use a bias potential but avoids the need to manually select collective variables or to identify basins of attraction. In a way, it defines the best CV for the purpose of AL: training a more general and robust ML potential. The approach requires selection of two parameters: the bias magnitude and width; developing a method which can tune these algorithmically would be a productive future activity. Additionally, the algorithm for automatic selection of uncertainty criteria could improve the sampling efficiency.

Methods

Active learning

For glycine simulations, we use ANI deep learning model[51] to generate ensembles of NN potentials prepared using an 8-fold cross validation split of the data set. The empirical value 0.23 $[\text{kcal} \times \text{mol}^{-1} \times (\sqrt{N_A})^{-1}]$ for the uncertainty selection criteria ρ, equation (3), provided in the original work on active learning for organic molecules[36] turned out to be too low for the purposes

of training on one chemical system. It causes unnecessarily dense sampling of glycine conformational space which, in turn, hinders the MD simulation to reach higher energy regions. Therefore, we use a higher value of 0.35 $[\text{kcal} \times \text{mol}^{-1} \times (\sqrt{N_A})^{-1}]$ for this test case. Overall, automatic selection of uncertainty criteria is a nontrivial question which deserves a separate discussion and goes beyond the scope of this work. Each MD simulation is terminated when the system meets the uncertainty selection criteria ρ.

The initial training set consists of 125 glycine geometries that span near-equilibrium structures of the glycine global minimum. This data is acquired from a separate 5ps MD trajectory at 350K with a 0.5 fs timestep, initialized from the glycine global minimum. Every 80th MD step is included in the initial data training set. The MD simulation for the initial training set is carried out using the pre-trained ANI-1x potential.[51]

MD Simulations

In all discussed cases, the Atomic Simulation Environment (ASE) Langevin thermostat was used to maintain temperature with a friction coefficient of 0.01 a.u. Each AL iteration performs 16 MD simulations with 1fs time step and a 200 000 steps limit (200ps). At each AL iteration, the MD is driven by an ensemble of ANI-type ML potentials, trained on initial data and data accumulated on previous AL iterations. The NN-based MD is interfaced with ASE code.[53] The final data set has 1280 data points sampled in AL procedure + 125 data points from the initial data set.

The set of seed geometries for MD simulations comprises 25 structures that correspond to near-equilibrium geometries of a glycine global minimum (GM). These are selected as the first 25 structures from the initial training set. AL sampler randomly selects one of them for each MD initialization. Energies and forces of new conformers are calculated using the WB97X-D/cc-pVTZ[54,55] level of theory as implemented in PSI4 code.[56]

NN architecture

Parameters for the atomic environment vector[51] (a numerical vector used to encode the atomic local environment in ANI) used during the AL process were constant. 32 evenly spaced shifting parameters are used for the radial part of the vector with 4.6 Å cutoff radius and a total of 8 radial and 8 angular shifting parameters are used for the angular part with 3.5 Å cutoff radius. With four atom types, this gives 768 elements in the descriptor. The first atom-centered function is shifted to 0.8 Å from the atomic center. The ANI potential used in this work contains three hidden layers and has the following architecture: 768:32:16:8:1, each number describing the number of neurons at each subsequent layer in the network.

The ANI potential used in this work contains three hidden layers and has the following architecture: 768:32:16:8:1. Gaussian activation functions are used in hidden layers and linear activation in the final layer.

Acknowledgements

K. B., N. L., R. M., S. T., and B. N. acknowledge support from the US DOE, Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under Triad National Security, LLC ("Triad") contract Grant 89233218CNA000001 (FWP: LANLE3F2). M. K. and Y. W. L. acknowledge support from the Los Alamos National Laboratory (LANL) Directed Research and Development funds. This work was performed in part at the Center for Nonlinear Studies and the Center for Integrated Nanotechnology, a US Department of Energy (DOE) and Office of Basic Energy Sciences user facility. This research used resources provided by the LANL Institutional Computing Program, which is supported by the US DOE National Nuclear Security Administration under Contract 89233218CNA000001. We also acknowledge the CCS-7 Darwin cluster at LANL for additional computing resources.

Data Availability

Source data are provided with this Paper.

Code Availability

Two implementations of the ANI neural network architecture are available online: TorchANI (https://github.com/aiqm/torchani) and NeuroChem (https://github.com/atomistic-ml/neurochem).

References

1. Kulichenko, M. *et al.* The Rise of Neural Networks for Materials and Chemical Dynamics. *J. Phys. Chem. Lett.* **12**, 6227–6243 (2021).
2. Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
3. Zubatiuk, T. & Isayev, O. Development of Multimodal Machine Learning Potentials: Toward a Physics-Aware Artificial Intelligence. *Acc. Chem. Res.* **54**, 1575–1585 (2021).
4. Westermayr, J. & Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **121**, 9873–9926 (2021).
5. Unke, O. T. *et al.* Machine Learning Force Fields. *Chem. Rev.* **121**, 10142–10186 (2021).
6. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761 (2021).
7. von Lilienfeld, O. A., Müller, K.-R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat Rev Chem* **4**, 347–358 (2020).
8. Zhai, H. & Alexandrova, A. N. Ensemble-Average Representation of Pt Clusters in Conditions of Catalysis Accessed through GPU Accelerated Deep Neural Network Fitting Global Optimization. *J. Chem. Theory Comput.* **12**, 6213–6226 (2016).
9. Keith, J. A. *et al.* Combining Machine Learning and Computational Chemistry for Predictive

Insights Into Chemical Systems. *Chem. Rev.* **121**, 9816–9872 (2021).

10. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

11. Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *J. Phys. Chem. Lett.* **11**, 9656–9658 (2020).

12. Purvis, G. D. & Bartlett, R. J. A full coupled-cluster singles and doubles model: The inclusion of disconnected triples. *J. Chem. Phys.* **76**, 1910–1918 (1982).

13. Krishnan, R. & Pople, J. A. Approximate fourth-order perturbation theory of the electron correlation energy. *International Journal of Quantum Chemistry* **14**, 91–100 (1978).

14. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

15. Hornak, V. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics* **65**, 712–725 (2006).

16. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* **31**, 671–690 (2010).

17. Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *npj Comput Mater* **2**, 1–14 (2016).

18. Behler, J. & Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).

19. 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).

20. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, 241715 (2018).

21. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable

multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science Advances* **5**, eaav6490.

22. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet – A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).

23. Batzner, S. *et al.* E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat Commun* **13**, 2453 (2022).

24. Thölke, P. & De Fabritiis, G. TorchMD-NET: Equivariant Transformers for Neural Network based Molecular Potentials. Preprint at https://doi.org/10.48550/arXiv.2202.02541 (2022).

25. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *International Journal of Quantum Chemistry* **115**, 1051–1057 (2015).

26. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics* **285**, 316–330 (2015).

27. Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. The MLIP package: moment tensor potentials with MPI and active learning. *Mach. Learn.: Sci. Technol.* **2**, 025002 (2021).

28. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat Commun* **9**, 3887 (2018).

29. Chmiela, S., Sauceda, H. E., Poltavsky, I., Müller, K.-R. & Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Computer Physics Communications* **240**, 38–45 (2019).

30. Karabin, M. & Perez, D. An entropy-maximization approach to automated training set generation for interatomic potentials. *J. Chem. Phys.* **153**, 094110 (2020).

31. Allotey, J., Butler, K. T. & Thiyagalingam, J. Entropy-based Active Learning of Graph

Neural Network Surrogate Models for Materials Properties. Preprint at https://doi.org/10.48550/arXiv.2108.02077 (2021).

32. Gal, Y., Islam, R. & Ghahramani, Z. Deep Bayesian Active Learning with Image Data. Preprint at https://doi.org/10.48550/arXiv.1703.02910 (2017).

33. Hernández-Lobato, J. M., Requeima, J., Pyzer-Knapp, E. O. & Aspuru-Guzik, A. Parallel and Distributed Thompson Sampling for Large-scale Accelerated Exploration of Chemical Space. Preprint at https://doi.org/10.48550/arXiv.1706.01825 (2017).

34. Zaverkin, V., Holzmüller, D., Steinwart, I. & Kästner, J. Exploring chemical and conformational spaces by batch mode deep active learning. *Digital Discovery*, Advance Article (2022).

35. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **20**, 458–465 (2015).

36. 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).

37. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).

38. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science* **140**, 171–180 (2017).

39. Browning, N. J., Ramakrishnan, R., von Lilienfeld, O. A. & Roethlisberger, U. Genetic Optimization of Training Sets for Improved Machine Learning Models of Molecular Properties. *J. Phys. Chem. Lett.* **8**, 1351–1359 (2017).

40. Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **146**, 244108 (2017).

41. Peterson, A. A., Christensen, R. & Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Phys. Chem. Chem. Phys.* **19**, 10978–10985 (2017).

42. Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. Machine learning of molecular properties: locality and active learning. *The Journal of Chemical Physics* **148**, 241727 (2018).

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

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. Uncertainty estimation for molecular dynamics and sampling: The Journal of Chemical Physics: Vol 154, No 7. https://aip.scitation.org/doi/10.1063/5.0036522.

46. Seung, H. S., Opper, M. & Sompolinsky, H. Query by committee. in *Proceedings of the fifth annual workshop on Computational learning theory* 287–294 (Association for Computing Machinery, 1992). doi:10.1145/130385.130417.

47. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562–12566 (2002).

48. Laio, A. & Gervasio, F. L. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.* **71**, 126601 (2008).

49. Sutto, L., Marsili, S. & Gervasio, F. L. New advances in metadynamics. *WIREs Computational Molecular Science* **2**, 771–779 (2012).

50. Valsson, O., Tiwary, P. & Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry* **67**, 159–184 (2016).

51. 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).

52. Sainburg, T., McInnes, L. & Gentner, T. Q. Parametric UMAP embeddings for representation and semi-supervised learning. Preprint at

https://doi.org/10.48550/arXiv.2009.12981 (2021).

53. Larsen, A. H. *et al.* The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002 (2017).

54. Chai, J.-D. & Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **10**, 6615–6620 (2008).

55. Dunning, T. H. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).

56. Smith, D. G. A. *et al.* PSI4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **152**, 184108 (2020).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SI.docx
- ExtendedDataFig1.jpg
- ExtendedDataFig2.jpg
- ExtendedDataFig3.jpg
- ExtendedDataFig4.jpg