The Journal of Chemical Physics

RESEARCH ARTICLE | APRIL 27 2023

Fast uncertainty estimates in deep learning interatomic potentials

Special Collection: Machine Learning Hits Molecular Simulations

Albert Zhu ; Simon Batzner ; Albert Musaelian ; Boris Kozinsky

J. Chem. Phys. 158, 164111 (2023)
https://doi.org/10.1063/5.0136574

AIP Publishing

The Journal of Chemical Physics ARTICLE scitation.org/journal/jcp

Fast uncertainty estimates in deep learning interatomic potentials

Cite as: J. Chem. Phys. 158, 164111 (2023); doi: 10.1063/5.0136574
Submitted: 27 November 2022 • Accepted: 10 April 2023 •
Published Online: 27 April 2023

Albert Zhu,[1] Simon Batzner,[1,a)] Albert Musaelian,[1] and Boris Kozinsky[1,2,a)]

AFFILIATIONS

[1] Harvard University, Cambridge, Massachusetts 02138, USA
[2] Robert Bosch Research and Technology Center, Cambridge, Massachusetts 02472, USA

Note: This paper is part of the JCP Special Topic on Machine Learning Hits Molecular Simulations.
[a)]**Authors to whom correspondence should be addressed:** batzner@g.harvard.edu and bkoz@seas.harvard.edu

ABSTRACT

Deep learning has emerged as a promising paradigm to give access to highly accurate predictions of molecular and material properties. A common short-coming shared by current approaches, however, is that neural networks only give point estimates of their predictions and do not come with predictive uncertainties associated with these estimates. Existing uncertainty quantification efforts have primarily leveraged the standard deviation of predictions across an ensemble of independently trained neural networks. This incurs a large computational overhead in both training and prediction, resulting in order-of-magnitude more expensive predictions. Here, we propose a method to estimate the predictive uncertainty based on a *single* neural network without the need for an ensemble. This allows us to obtain uncertainty estimates with virtually no additional computational overhead over standard training and inference. We demonstrate that the quality of the uncertainty estimates matches those obtained from deep ensembles. We further examine the uncertainty estimates of our methods and deep ensembles across the configuration space of our test system and compare the uncertainties to the potential energy surface. Finally, we study the efficacy of the method in an active learning setting and find the results to match an ensemble-based strategy at order-of-magnitude reduced computational cost.

INTRODUCTION

Over the past decade, the construction of high-dimensional potential energy surfaces (PES) based on machine learning (ML) has become a promising avenue to enable linear-scaling and computationally efficient molecular simulations that retain the quantum chemical accuracy of their training data.[1–23] A large variety of methods have been proposed to regress energies and forces obtained from *ab initio* calculation as a function of atomic positions and chemical species, including kernel-based approaches,[4,8,20] linear models,[5,6] and neural networks (NNs).[3,7,12–14] Among these, deep NNs, in particular, have shown remarkable accuracy and fast progress in their predictive accuracy.[13,14,24–26] The high accuracy of NN-based approaches, however, comes at a cost: common to all existing neural approaches is that they provide only point estimates of their predictions instead of the full predictive distribution. This differs from Bayesian methods such as Gaussian Processes, which inherently come with a measure of predictive uncertainty. Uncertainties have been shown to be of tremendous value in ML-driven molecular simulations.[17,18,20,21,27] In particular, uncertainties have been used to bootstrap simulations without the need for a training set via an active learning loop.[20] In such an approach, the model's uncertainty is assessed at every integration step: if the uncertainty is low, the model's predictions are used. If, instead, the uncertainty exceeds a certain threshold, high-accuracy quantum mechanical calculations, such as density functional theory (DFT), are invoked, the new data point is added to the dataset, and the model is re-trained. Provided the uncertainty measure used is of high fidelity, such an approach can greatly enhance the robustness, reliability, and ease-of-use of ML-driven atomistic simulations. In this work, we present a novel, computationally inexpensive method to obtain uncertainty measures in deep learning interatomic potentials, evaluate its performance compared to existing approaches, and demonstrate that it produces order-of-magnitude faster uncertainty estimates.

Related work

Given the high impact reliable uncertainty estimates would have on the usefulness of NN-based machine learning interatomic potentials (MLIPs), great effort has gone into the development of

techniques that enhance the point estimate predictions of NNs with a measure of predictive uncertainty.[28–39] The most widely used approach among these is an *ensemble* of NNs, all trained on the same data that differ in their initial weights and perhaps other training or model hyperparameters. The mean of the predictions of all constituent networks is used as the ensemble's prediction, and the standard deviation of the predictions is used as a measure of uncertainty. Intuitively, if a structure seen at test time falls within the input domain that the NNs are confident about, their predictions should agree. In contrast, if a test structure lies outside of the training distribution, the networks' predictions should differ, resulting in a higher standard deviation. This method has been widely used since the first generation of NN interatomic potentials.[28–30] Systematic analysis and improvements are desired in this direction, e.g., to avoid situations where all models share the same bias not captured in the training data or their functional form, resulting in models sharing confident but erroneous predictions. Recent work has explored the over-confidence of NN-based ensembles and the correlation between the test set error with the true predictive error.[40] While larger ensembles can provide more statistics, the need to train and evaluate all constituent networks (often $N \geq 10$) incurs significantly greater computational expenses, lowering inference speed and limiting practical applications of ensemble NN models for molecular dynamics or Monte Carlo sampling calculations. To alleviate this, a series of single-model methods have been proposed.[35,36,39] However, these methods often either still require multiple evaluations at inference time or have not been demonstrated to work in applications of molecular dynamics, where force uncertainty is the key objective.[36] Finally, two papers that came to our attention during the final preparation of this manuscript[39,41] propose a further single-model approach to uncertainty quantification; however, the primary version of their approach only obtains an uncertainty estimate for an entire structure. In contrast, our approach yields uncertainties on each individual atom's force predictions, which affords spatially resolved analysis and active learning strategies that may be more informative and useful in simulations of large and heterogeneous systems. While these works[39,41] introduce a second version that can produce an uncertainty estimate for each atomic force component, it incurs high computational costs due to requiring the computation of many different gradients.

METHODS

Ensembles of neural networks

We investigate uncertainty quantification in Neural Equivariant Interatomic Potentials (NequIP), an E(3)-equivariant neural network for learning interatomic potentials that achieves state-of-the-art accuracy on a challenging and diverse set of molecules and materials at remarkable data efficiency in comparison to other MLIPs.[13] To establish a baseline for our proposed uncertainty quantification approach, we train two sets of ensembles, each consisting of ten NequIP neural networks: a "traditional" ensemble consisting of networks differing solely in their weight initialization and the order in which mini-batches are sampled during training, and a "diverse" ensemble consisting of three networks from the traditional ensemble and seven additional networks, each with different hyperparameters (listed in the supplementary material, Table 1). Other types of ensembles based on subsampling or bootstrapping of the training set also exist; however, we choose not to analyze these in this work because a bootstrapped ensemble may consist of individual models that miss certain high-energy and important configurations in their subsampled training set, leading to uncertainty that simply originates from subsampling the training set. Furthermore, the high cost of training the ensemble still remains. To demonstrate the robustness of our methods and conclusions to the width/capacity of the networks, we train all networks with a hidden feature dimension of $f = 32$ in one setting and $f = 16$ in another setting.

At run time, the force predictions of an ensemble, denoted $\bar{F}$, are calculated as the mean of the predictions of individual models in the ensemble, component-wise,

$$\bar{F} = \left(\bar{F}_x, \bar{F}_y, \bar{F}_z\right), \tag{1}$$

where $\bar{F}_\alpha$ denotes the mean of the α-component of the predicted forces of all constituent models. To evaluate the model's fidelity, we calculate the per-atom root mean square error (RMSE), ϵ, of the ensemble's predicted force as

$$\epsilon = \sqrt{\frac{1}{3}\left(\left(\bar{F}_x - F_x\right)^2 + \left(\bar{F}_y - F_y\right)^2 + \left(\bar{F}_z - F_z\right)^2\right)} \tag{2}$$

and the force RMSE, $\bar{\epsilon}$, over all N atoms in the test set as

$$\bar{\epsilon} = \sqrt{\frac{1}{3N}\sum_{\alpha\in x,y,z}\left(\sum_{i=1}^{N}\left(\bar{F}_{i,\alpha} - F_{i,\alpha}\right)^2\right)}, \tag{3}$$

where $\bar{F}_{i,\alpha}$ and $F_{i,\alpha}$ denote the predicted and true α-component of the force on atom i, respectively.

To obtain an uncertainty estimate, we calculate the standard deviation of a predicted force, σ, over the constituent networks. Because we are primarily interested in predictive uncertainties for molecular dynamics simulations, we investigate the uncertainty in the force components as opposed to the energies, since the forces determine the dynamics of the system. We thus calculate σ as the square root of the mean of component-wise variances of the predicted forces,

$$\sigma = \sqrt{\frac{1}{3J}\sum_{\alpha\in x,y,z}\left(\sum_{j}\left(\hat{F}_{j,\alpha} - \bar{F}_\alpha\right)^2\right)}, \tag{4}$$

where J is the number of constituent models (we use $J = 10$) and $\hat{F}_{j,\alpha}$ denotes the α-component of network j's predicted force.

Gaussian mixture model

The aim of this work is to understand whether Gaussian mixture models (GMM), trained on a network's learned features, may provide a faster and more memory-efficient approach to uncertainty quantification in MLIPs. A GMM is a probabilistic model used in many applications—including speaker verification,[42] language identification,[43] and computer vision[44]—due to its ability to represent a large class of data distributions,[45] which motivates us to investigate

its capability of modeling a NequIP network's learned features. A GMM models a data distribution as a weighted sum of M Gaussians,

$$p(x|\theta) = \sum_{m=1}^{M} w_m \mathcal{N}(x|\mu_m, \Sigma_m), \tag{5}$$

where x is a D-dimensional, continuous-valued vector, w_m is the weight of the mth Gaussian with the constraint $\sum_{m=1}^{M} w_m = 1$, and $\mathcal{N}(x|\mu_m, \Sigma_m)$ are the D-variate Gaussian densities,

$$\mathcal{N}(x|\mu_m, \Sigma_m) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} e^{-\frac{1}{2}(x-\mu_m)^T \Sigma_m^{-1}(x-\mu_m)}, \tag{6}$$

with D-dimensional mean vector $\mu_m \in \mathbb{R}^D$ and $D \times D$-dimensional covariance matrix Σ_m.[45] The parameters of the complete GMM are collected into θ,

$$\theta = \{w_m, \mu_m, \Sigma_m\}, \quad m \in [1, M]. \tag{7}$$

To construct the GMM, we first train and evaluate a NequIP model on the training set to access the per-atom final projected scalar features extracted immediately before the linear projection down into the per-atom energy prediction (not the latent features themselves). These final features are half the respective latent feature widths $f = 16$ ($x \in \mathbb{R}^8$) and $f = 32$ ($x \in \mathbb{R}^{16}$) used in our experiments, but in practice, users can define the final feature dimension and independently set the latent feature dimension. We then fit a GMM to model the distribution of these feature vectors as evaluated on the training set, denoted X, using the expectation maximization (EM) algorithm with each initial μ_m determined by k-means clustering.[45,46] We fit the GMM using a full covariance matrix for each Gaussian, meaning that each Σ_m is full rank and not shared between Gaussians.[45] We select the number of Gaussians using the Bayesian Information Criterion (BIC). To then estimate the uncertainty of a trained NequIP model on a test data point, we run a forward pass through NequIP to extract the final layer features for the atoms of that test structure. Subsequently, we evaluate the fitted GMM on the feature vector for each test atom x, obtaining a negative log-likelihood $\mathrm{NLL}(x|X)$,

$$\mathrm{NLL}(x|X) = -\log\left(\sum_{m=1}^{M} w_m \mathcal{N}(x|\mu_m, \Sigma_m)\right). \tag{8}$$

A higher $\mathrm{NLL}(x|X)$ indicates higher uncertainty. Since the GMM is computationally light-weight, almost all computational burden lies in the evaluation of the NequIP features, which now occurs once instead of J times in the ensembles.

RESULTS AND DISCUSSION

Dataset

We conduct our experiments on the 3-(benzyloxy)pyridin-2-amine (3BPA) transferability benchmark.[23] 3BPA (Fig. 1) is a flexible drug-like molecule whose configurational diversity is largely determined by the three dihedral angles α, β, and γ and is explored more fully at higher temperatures, making it a challenging test case for MLIPs.

In the first setting, we use three datasets of 3BPA structures sampled at three temperatures: 300, 600, and 1200 K. We train

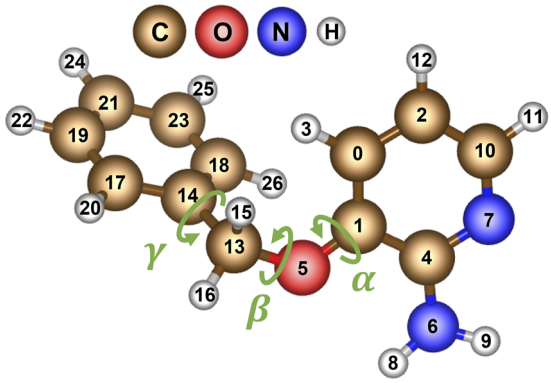

FIG. 1. 3D model of the 3BPA molecule with atomic indices and α, β, and γ dihedral angles depicted.

on structures sampled at 300 K (once with 50 structures, $D^{300}_{\mathrm{train},50}$, once with 100 structures, $D^{300}_{\mathrm{train},100}$). We set aside a pool of additional training data at 300 K to select from (D^{300}_{pool}). Finally, we have three test sets of structures at each temperature, i.e., D^{300}_{test}, D^{600}_{test}, and D^{1200}_{test}, along with three additional datasets ($D_{\beta=120°}$, $D_{\beta=150°}$, and $D_{\beta=180°}$), each consisting of structures with a fixed β dihedral angle and uniformly sampled α and γ angles.

In a second setting, we combine all data from the three temperatures and similarly split the combined data into training sets (of size 50, $D^{\mathrm{mixed}}_{\mathrm{train},50}$, and size 100, $D^{\mathrm{mixed}}_{\mathrm{train},100}$), a pool of additional training data to sample from ($D^{\mathrm{mixed}}_{\mathrm{pool}}$), and a test set ($D^{\mathrm{mixed}}_{\mathrm{test}}$).

Uncertainty quantification

In Fig. 2, we plot the uncertainty estimates of the ensembles and the GMM against the measured per-atom RMSE ϵ for models with hidden feature dimension $f = 32$, trained on $D^{300}_{\mathrm{train},100}$ and evaluated on D^{1200}_{test}. Plots for the evaluations on D^{300}_{test} and D^{600}_{test} and for the mixed-temperature setting show similar results (supplementary material, Fig. 4). Likewise, plots for models with hidden feature dimension $f = 16$ and models trained on $D^{300}_{\mathrm{train},50}$ and $D^{\mathrm{mixed}}_{\mathrm{train},50}$ generally demonstrate the same results (supplementary material, Figs. 1–3). We note that these uncertainty estimates in their current form cannot be interpreted as error estimates. Previous works have calibrated their error estimates to improve their accuracy,[32,47] and the conversion of our uncertainty estimates into error estimates and subsequent calibration is a promising direction for future work. We observe in Fig. 2 that the traditional ensemble achieves the lowest $\bar{\epsilon}$ while the single model used for fitting the GMM has the highest $\bar{\epsilon}$. This result is expected as the average ensemble prediction has been observed to be more accurate than the prediction of a single model.[48] The traditional ensemble generally performs better than the diverse ensemble since the diverse ensemble contains simpler models on average. The distribution of σ in the diverse ensemble is generally shifted toward higher values than the traditional ensemble, likely due to the differences in network architecture resulting in more diverse predictions. As expected, supplementary material, Fig. 4 shows that the distribution of ϵ shifts toward higher errors as the temperature of the test set increases from 300 to 1200 K.

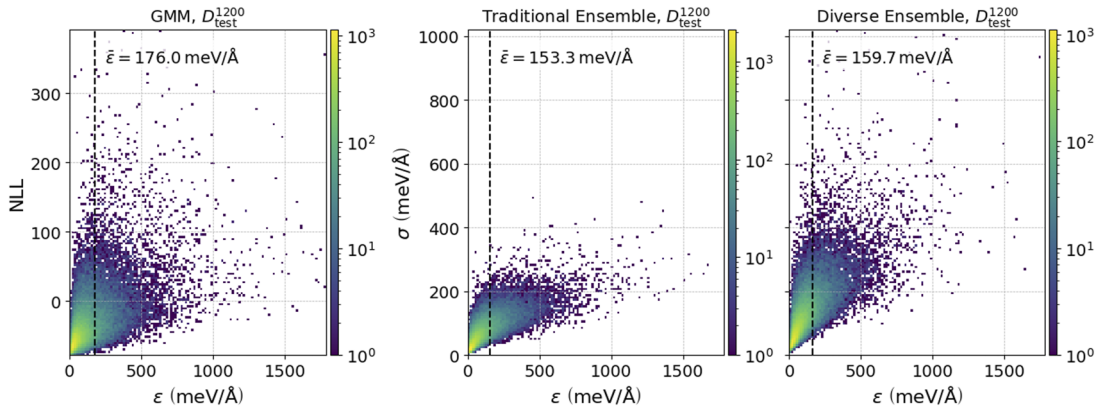

FIG. 2. Plots of the uncertainty metric (σ for the ensembles and NLL for the GMM) vs ϵ for models with hidden feature dimension $f = 32$, trained on $D_{\text{train},100}^{300}$ and evaluated on all atoms of all configurations in D_{test}^{1200}. Each point (ϵ_i, U_i) in a given plot represents the model's force RMSE and predictive uncertainty, respectively, on a single atom i. The color bar represents the number of points within each bin. The vertical dashed line in each plot marks the average force RMSE $\bar{\epsilon}$ over all atoms [see Eq. (3)].

More notably, the distribution of σ of all methods shifts toward larger positive values with increasing temperature, demonstrating the ability of all methods to detect high-energy, out-of-distribution configurations. Overall, we observe a modest correlation between the uncertainty metric and ϵ for each of the approaches, with the GMM's correlation similar to those of the ensembles. Most importantly, a GMM evaluated on a *single network* has similar predictive power of the uncertainty as an ensemble of ten networks, providing a way to reduce the computational cost of uncertainty quantification by an order of magnitude while maintaining the state-of-the-art performance of ensemble-based uncertainties.

To further quantify the quality of these uncertainty metrics, we establish certain criteria that a good uncertainty metric should meet. Since uncertainty estimates are often used to identify high-error structures for which to invoke first principles calculations, we set some uncertainty cutoff U_{cutoff} for capturing such structures. In particular, we classify all configurations with uncertainty $U > U_{\text{cutoff}}$ as "high-error" and all configurations with uncertainty $U \leq U_{\text{cutoff}}$ as "low error." A good uncertainty metric should simultaneously classify a large proportion of configurations with $\epsilon > \epsilon_{\text{cutoff}}$ as high-error to avoid missing configurations with high true ϵ while classifying a small proportion of configurations with true $\epsilon \leq \epsilon_{\text{cutoff}}$ as high-error configurations to avoid redundant calls to DFT calculations. In other words, a good uncertainty metric should achieve a high true positive rate (TPR) and a high positive predictive value (PPV),[40]

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{9}$$

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{10}$$

where TP (true positives) is the number of configurations with $U > U_{\text{cutoff}}$ and $\epsilon > \epsilon_{\text{cutoff}}$, FN (false negatives) is the number of configurations with $U \leq U_{\text{cutoff}}$ and $\epsilon > \epsilon_{\text{cutoff}}$, and FP (false positives) is the number of configurations with $U > U_{\text{cutoff}}$ and $\epsilon \leq \epsilon_{\text{cutoff}}$. Thus, to quantify the quality of each method's uncertainty estimates, we calculate their TPR and PPV (as done in Ref. 40) for a range of U_{cutoff} and ϵ_{cutoff}.

Figure 3 shows the TPR and PPV of the uncertainty metrics for ranges of U_{cutoff} and ϵ_{cutoff} on all atoms in D_{test}^{1200}, along with the product of the TPR and PPV (TPR × PPV). In each plot, ϵ_{cutoff} ranges from 0 meV/Å to 99th percentile of the ϵ's of the corresponding method on all atoms, and U_{cutoff} ranges from the 1st to 99th percentile of that method's uncertainty estimates. Ranges are chosen to minimize the impacts of outliers on the TPR and PPV. Supplementary material Figs. 5–8 show plots for the other test sets, training sets, and model hyperparameters and yield similar findings.

We observe that for a given ϵ_{cutoff} in all approaches, lowering U_{cutoff} captures more high true-error points but incurs more false positives, as evidenced by an increase in the TPR but a decrease in the PPV. The GMM achieves TPR ≈ 1 for slightly smaller but comparable ranges of U_{cutoff} and ϵ_{cutoff} compared to both ensemble types. All approaches achieve PPV ≈ 1 for similar ranges of U_{cutoff} and ϵ_{cutoff}, with the GMM's PPV decaying slightly more slowly than that of the ensembles as ϵ_{cutoff} increases and U_{cutoff} decreases. Notably, the similarity in the TPR and PPV profiles between the two ensemble types indicates that diversifying an ensemble increases σ for each data point but does not necessarily improve the quality of the uncertainty estimate. Overall, we conclude that for all methods, it is difficult to simultaneously achieve TPR ≈ 1 and PPV ≈ 1 unless we set a U_{cutoff} around the 30th percentile of all uncertainty estimates of a given method, marking a majority of configurations for recalculation with DFT. Most importantly, the GMM produces uncertainty estimates comparable in quality to both ensemble types at a much lower computational cost.

Uncertainty landscapes

To further investigate the similarity between the uncertainty metrics obtained by the ensembles and the GMM, we create "uncertainty landscapes" in which we plot the uncertainty of each

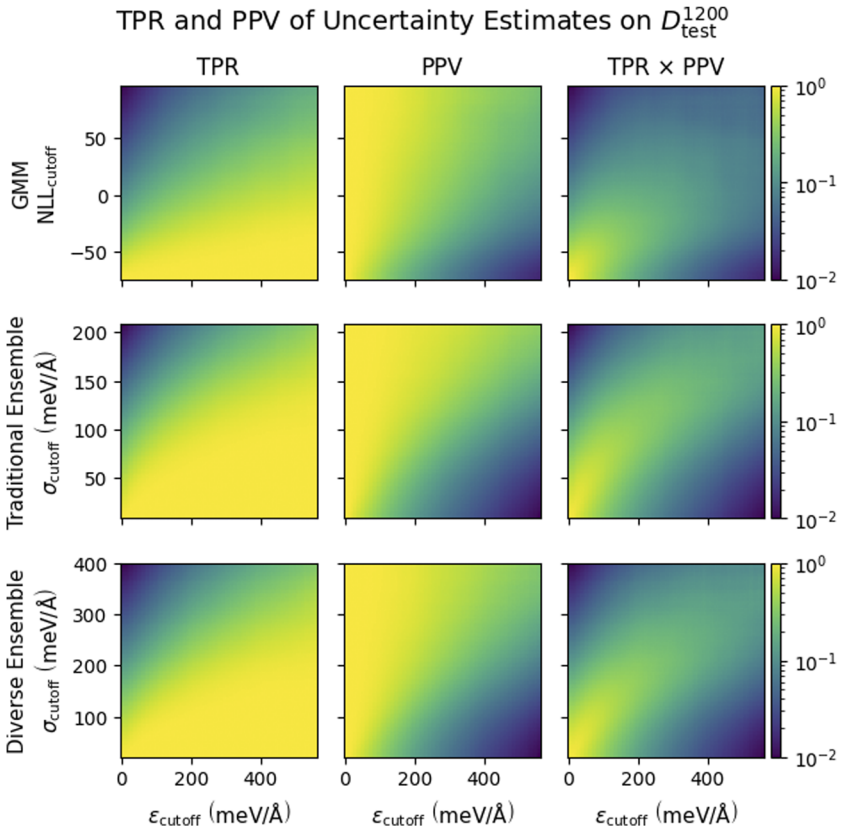

FIG. 3. TPR and PPV profiles of uncertainty estimates on all atoms in D_{test}^{1200}.

method for configurations of fixed β and varying α and γ, similar to a potential energy landscape. We evaluate the ensembles and a single NequIP model on $D_{\beta=120^\circ}$, $D_{\beta=150^\circ}$, and $D_{\beta=180^\circ}$ and obtain force uncertainties per atom per configuration as usual. For the ensembles, we define a single aggregate uncertainty value for a molecular structure as the square root of the sum of all 27 atomic force variances. For the GMM, we sum all 27 atomic force NLL's for each structure to obtain a single uncertainty value. Finally, we normalize these aggregate molecular uncertainties for each method to be between 0 and 1 for a clearer comparison. For example, if the uncertainty over all configurations ranges from U_{low} to U_{high} for a particular method, a configuration with uncertainty U would have a normalized uncertainty of

$$\frac{U - U_{low}}{U_{high} - U_{low}}. \tag{11}$$

We note that a user of these uncertainty quantification approaches would not have to generate such test sets and uncertainty landscapes; this analysis simply further validates our approaches.

The top row of Fig. 4 shows the uncertainty landscape of the GMM, and both ensemble types with models with $f = 32$ trained on $D_{train,100}^{300}$ and evaluated on $D_{\beta=120^\circ}$ (supplementary material Figs. 9–12 show results for the $D_{\beta=150^\circ}$ and $D_{\beta=180^\circ}$ test sets, models with $f = 16$, and models trained on $D_{train,50}^{300}$, which demonstrate comparable findings). All three uncertainty landscapes are very similar in that they generally label the same configurations with relatively high uncertainty. For example, all methods label configurations with $\alpha \leq 60^\circ$, $\alpha \approx 180^\circ$, and $(\alpha \approx 240^\circ, \gamma \approx 10^\circ)$ with high relative uncertainty. Moreover, the uncertainty landscapes of all methods closely resemble their corresponding potential energy landscapes and the reference DFT potential energy landscape (bottom row of Fig. 4).[23] These similarities further demonstrate that a single NequIP model and GMM evaluation can achieve uncertainty estimates comparable to those of an ensemble at a greatly reduced computational cost. Note that the aggregation of uncertainty values assumes independence between atomic forces within a structure. However, we cannot know the true correlation of forces within a structure as they may vary between systems. Furthermore, this assumption does not affect our analysis because in supplementary material Figs. 25–28 where we plot similar landscapes of the standard deviation of the ensembles' total energy predictions, we observe that the total energy standard deviation landscapes closely resemble the aggregated uncertainty and potential energy landscapes. Because the total energy standard deviation landscapes do not make any assumptions of independence, the assumption of independence between the atomic forces does not affect our conclusion.

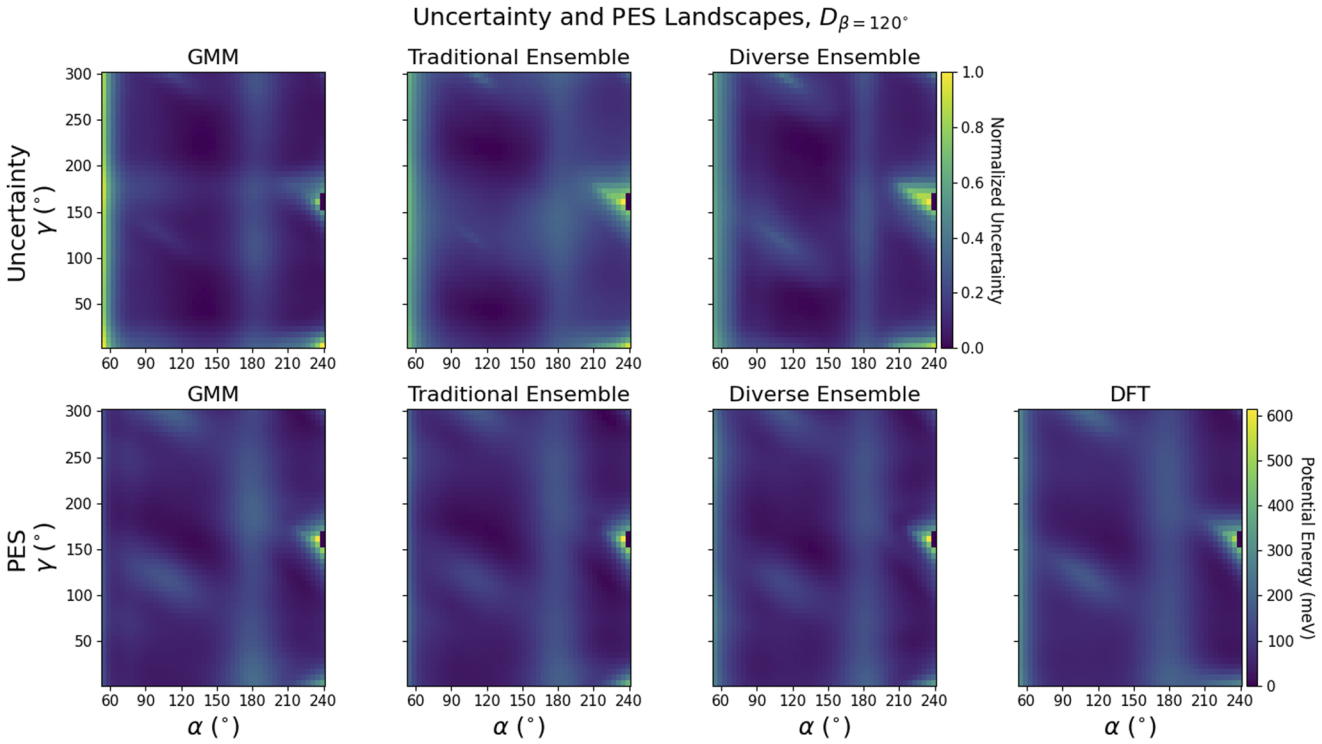

FIG. 4. Top row: Uncertainty landscape of the GMM and ensembles for $D_{\beta=120^\circ}$. Color scale represents normalized summed force uncertainty over all atoms in a configuration. Bottom row: Potential energy surface (PES) landscape of the GMM, ensembles, and reference DFT calculations for $D_{\beta=120^\circ}$. The color scale represents energy relative to the minimum energy over all configurations in units of meV, with yellow indicating higher energy. Purple space at $\alpha \approx 240°$, $\gamma \approx 160°$ indicates missing data.

Active learning

Active learning is a procedure in which a model explores a data distribution and iteratively chooses new data to label and add to its training set to augment it with maximally informative samples. Active learning has been of great use, in particular, in building training sets for machine learning interatomic potentials.[17,18,20,21,27,30,39,41,49] Crucially, in such a setting, one requires a method to decide which data points to label. Using uncertainty estimates from the GMM and the ensembles, we conduct experiments in which we select new data from a set of hold-out examples based on the model's estimated uncertainty on those data points. This uncertainty-based selection should ideally result in a training set that improves the model's generalization error more than by adding an identical number of randomly chosen data points.

To measure the effectiveness of the different uncertainty estimates, we perform one round of this active learning task (which we will refer to as "active learning") using a single model with GMM uncertainties and compare the results to active learning with uncertainties from both ensemble types. We evaluate a trained ensemble or single NequIP model on a pool of additional training data, obtaining an uncertainty $U_k = \sigma_k$ for each atom k in each molecular structure. Let k^* be the atom within a 3BPA molecule with the highest uncertainty, i.e.,

$$k^* = \arg\max_k \{U_k\}. \tag{12}$$

We select the N_{train} structures with the highest U_{k^*} to add the original training set of size N_{train}, creating a new training set $D_{\text{train,active}}$. We also randomly select N_{train} to add to the original training set, creating another training set $D_{\text{train, random}}$. We re-train all models with the same parameters from scratch on each of these two new training sets and evaluate them on various test sets. For further details, see the Active learning section in the supplementary material.

Assuming that atoms with the highest ϵ (i.e., outliers) are labeled with the highest uncertainties, we ideally expect our active learning procedure to improve generalization error on those atoms the most. Thus, to compare the effectiveness of active learning based on the three uncertainty metrics, we consider not only $\bar{\epsilon}$ [Eq. (3)] over all atoms in a test set but also the distribution of ϵ, the average force RMSE of all ϵ above 200 meV/Å ($\bar{\epsilon}_{200}$), and the maximum force RMSE ($\epsilon_{\max}$) between the three approaches.

Figure 5 shows the results of evaluating the ensembles and single NequIP model on $D_{\beta=120^\circ}$ after one round of active learning, with all models having hidden feature dimension $f = 16$, initially trained on $D^{300}_{\text{train,50}}$, and with new data sampled from D^{300}_{pool}. We observe that active learning improves $\bar{\epsilon}$ on $D_{\beta=120^\circ}$ by around 10 meV/Å for the three methods, compared to the random selection baseline. The improvement is even greater on $D_{\beta=150^\circ}$ and $D_{\beta=180^\circ}$, reaching nearly 30 meV/Å on $D_{\beta=180^\circ}$ using the GMM (supplementary material, Fig. 13). Similar to Fig. 2, we observe that the ensembles generally achieve lower error than a single model irrespective of the sampling scheme due to the fact that the average ensemble

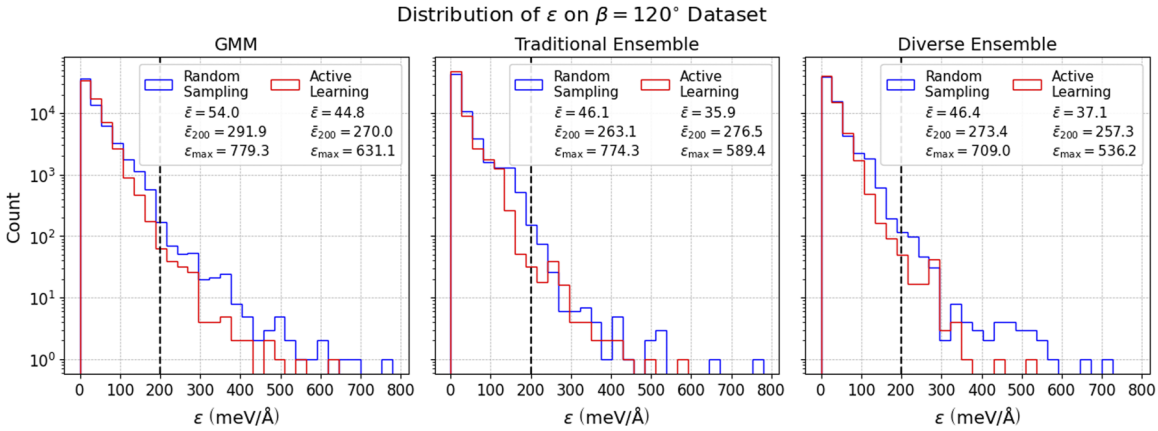

FIG. 5. Distribution of ϵ of the GMM and ensembles on $D_{\beta=120°}$ for models with hidden feature dimension $f = 16$ and trained on $D^{300}_{\text{train},50}$. A vertical dashed line in each plot marks the 200 meV/Å cutoff for determining the RMSE of outliers.

TABLE I. Improvement of $\bar{\epsilon}$, $\bar{\epsilon}_{200}$, and ϵ_{max} with active learning over random sampling for each of the three methods on $D_{\beta=120°}$, $D_{\beta=150°}$, and $D_{\beta=180°}$ (for models with hidden feature dimension $f = 16$ and initially trained on $D^{300}_{\text{train},50}$).

		Absolute improvement (meV/Å)			Percentage improvement (%)		
		GMM	Traditional	Diverse	GMM	Traditional	Diverse
$D_{\beta=120°}$	$\bar{\epsilon}$	9.2	10.2	9.3	17.0	22.1	20.0
	$\bar{\epsilon}_{200}$	21.9	−13.4	16.1	7.5	−5.1	5.9
	ϵ_{max}	148.2	184.9	172.8	19.0	23.9	24.3
$D_{\beta=150°}$	$\bar{\epsilon}$	15.6	9.5	7.8	26.0	25.5	17.9
	$\bar{\epsilon}_{200}$	118.8	41.6	90.0	28.6	14.0	26.0
	ϵ_{max}	290.8	114.1	232.1	40.9	41.9	41.9
$D_{\beta=180°}$	$\bar{\epsilon}$	29.6	15.2	23.4	39.9	32.1	39.0
	$\bar{\epsilon}_{200}$	140.4	70.3	134.6	27.5	18.5	30.2
	ϵ_{max}	451.9	169.0	537.4	44.3	27.3	54.9

prediction is typically more accurate than the prediction of a single model.[48]

Additionally, we observe that active learning reduces $\bar{\epsilon}_{200}$ and ϵ_{max} by a much larger amount compared to the reduction in $\bar{\epsilon}$, as reflected by the presence of fewer high-error data points in the distribution of ϵ resulting from active learning. On $D_{\beta=150°}$ and $D_{\beta=180°}$, $\bar{\epsilon}_{200}$ is even reduced by over 100 meV/Å with active learning, as seen in the supplementary material, Fig. 13. Table I summarizes the absolute and percentage improvements of active learning over random sampling for all three methods on these performance metrics on $D_{\beta=120°}$, $D_{\beta=150°}$, and $D_{\beta=180°}$. From these results, we conclude that the improvement in $\bar{\epsilon}$ is significant for all methods, and in particular, using GMM uncertainties for active learning yields improvements in overall error and outlier generalization error comparable to those achieved with ensemble-based active learning at a significantly lower computational cost.

We note that assessing the effectiveness of uncertainty quantification and active learning approaches requires configurations of sufficiently high rarity. This ensures that the model is not already fully trained to accurately predict all points in the test set. For instance, active learning achieved minimal improvement over random sampling for all methods when testing on D^{300}_{test}, D^{600}_{test}, D^{1200}_{test}, and $D^{\text{mixed}}_{\text{test}}$ (supplementary material, Figs. 17–20). D^{300}_{test} and $D^{\text{mixed}}_{\text{test}}$ contain structures from the same distribution as their respective training sets, rendering them "easier" test sets that the model marginally improves on with additional training data. Similarly, while D^{600}_{test} and D^{1200}_{test} contain more high-energy structures than $D^{300}_{\text{train},50}$ and $D^{300}_{\text{train},100}$, their distributions still overlap significantly because structures are Boltzmann-sampled at each temperature. In comparison, $D_{\beta=120°}$, $D_{\beta=150°}$, and $D_{\beta=180°}$ contain structures spanning a much wider range of α and γ angles for a fixed β.[23] Furthermore, compared to active learning with models initially trained on $D^{300}_{\text{train},50}$, active learning with models initially trained on $D^{300}_{\text{train},100}$ generally achieves smaller improvement over random sampling (supplementary material, Figs. 15 and 16). Similarly, for models initially trained on $D^{300}_{\text{train},100}$, active learning using models with $f = 32$

generally achieves slightly lower improvement over random sampling compared to models with $f = 16$ (compare supplementary material, Figs. 15 and 16), which may also be due to the fact that models with a higher latent feature dimension typically perform slightly better before active learning (compare supplementary material, Tables 3 and 4). In summary, we emphasize that for developing and benchmarking uncertainty quantification and active learning approaches, one must establish that active learning methods are justified and statistically distinguishable in performance from random sampling.

CONCLUSION

Algorithmic advances in fast and accurate uncertainty quantification for deep neural network interatomic potentials are needed to enable robust large-scale uncertainty-aware simulations. While an efficient method that can give access to predictive uncertainties in deep learning interatomic potentials has been a long-standing goal, so far it has not been achieved, and current methods still rely on leveraging an ensemble of networks, thereby incurring a massive computational overhead. Here—by training a probabilistic model on the feature space of the neural network—we show that it is possible to retain the accuracy of ensemble uncertainty estimates with a single neural network evaluation, resulting in large computational savings in training and inference. In particular, we show that a Gaussian Mixture Model trained on NequIP features produces uncertainty estimates of similar quality to deep ensembles while requiring only a single model evaluation, resulting in a significant reduction of computational cost in both training and inference. While we find that the GMM models are competitive with ensembles, which are currently the state of the art methodology, significant improvement is desired for reliable predictions of quantitative uncertainties in both types of approaches. One future direction is to compare the methods explored here with rigorous Bayesian inference techniques, both in neural networks and kernel-based learning models.

SUPPLEMENTARY MATERIAL

The supplementary material contains additional experiments, as referenced in the main text.

ACKNOWLEDGMENTS

We thank Yu Xie and Julia Yang for their helpful discussions.

Work at Harvard University was supported by Bosch Research, the U.S. Department of Energy, Office of Basic Energy Sciences, under Award No. DE-SC0022199 and by the NSF through the Harvard University Materials Research Science and Engineering Center under Grant No. DMR-2011754. A.M. is supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Computational Science Graduate Fellowship under Award No. DE-SC0021110. A.Z. was supported by the Harvard College Research Program. The authors acknowledge computing resources provided by the Harvard University FAS Division of Science Research Computing Group.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

A.Z. trained the NequIP networks, conducted all uncertainty experiments, active learning experiments, and analysis of models, and wrote the first version of the manuscript. S.B. also trained some of the NequIP networks, proposed to use a GMM on final layer features, and contributed to the first version of the manuscript. A.M. contributed to the software development for ensembles and GMM-based evaluations. All authors discussed the results and designed the experiments. B.K. supervised and guided the project from conception to design of experiments, implementation, theory, as well as analysis of data. All authors contributed to the manuscript.

Albert Zhu: Conceptualization (equal); Data curation (lead), Investigation (lead); Methodology (lead); Software (lead); Validation (lead); Visualization (lead); Writing – original draft (lead); Writing – review & editing (lead). **Simon Lutz Batzner**: Conceptualization (equal); Data curation (supporting); Investigation (supporting); Methodology (supporting); Software (supporting); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal). **Albert Musaelian**: Conceptualization (equal); Investigation (supporting); Methodology (supporting); Software (supporting); Supervision (equal); Writing – original draft (supporting); Writing – review & editing (supporting). **Boris Kozinsky**: Conceptualization (equal); Funding acquisition (lead); Investigation (supporting); Methodology (supporting); Project administration (lead); Resources (lead); Supervision (lead), Writing – review & editing (supporting).

DATA AVAILABILITY

The dataset of structures for the 3BPA molecule is publicly available at https://pubs.acs.org/doi/full/10.1021/acs.jctc.1c00647. An open-source software implementation of NequIP is available at https://github.com/mir-group/nequip.

REFERENCES

[1] T. B. Blank, S. D. Brown, A. W. Calhoun, and D. J. Doren, "Neural network models of potential energy surfaces," J. Chem. Phys. **103**, 4129–4137 (1995).
[2] C. M. Handley, G. I. Hawe, D. B. Kell, and P. L. A. Popelier, "Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning," Phys. Chem. Chem. Phys. **11**, 6365–6376 (2009).
[3] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[4] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).
[5] A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, "Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials," J. Comput. Phys. **285**, 316–330 (2015).
[6] A. V. Shapeev, "Moment tensor potentials: A class of systematically improvable interatomic potentials," Multiscale Model. Simul. **14**, 1153–1173 (2016).
[7] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "SchNet—A deep learning architecture for molecules and materials," J. Chem. Phys. **148**, 241722 (2018).

[8]S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," Nat. Commun. **9**, 3887 (2018).
[9]O. T. Unke and M. Meuwly, "PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges," J. Chem. Theory Comput. **15**, 3678–3693 (2019).
[10]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).
[11]A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, "FCHL revisited: Faster and more accurate quantum machine learning," J. Chem. Phys. **152**, 044107 (2020).
[12]J. Klicpera, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in International Conference on Learning Representations 2020, 26 April–1 May (2020).
[13]S. Batzner *et al.*, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nat. Commun. **13**, 2453 (2022).
[14]A. Musaelian *et al.*, "Learning local equivariant representations for large-scale atomistic dynamics," Nat. Commun. **14**, 579 (2023).
[15]J. P. Mailoa *et al.*, "A fast neural network approach for direct covariant forces prediction in complex multi-element extended systems," Nat. Mach. Intell. **1**, 471–479 (2019).
[16]C. W. Park *et al.*, "Accurate and scalable multi-element graph neural network force field and molecular dynamics with direct force architecture," arXiv:2007.14444 (2020).
[17]Y. Xie, J. Vandermause, L. Sun, A. Cepellotti, and B. Kozinsky, "Bayesian force fields from active learning for simulation of inter-dimensional transformation of stanene," npj Comput. Mater. **7**, 40 (2021).
[18]Y. Xie *et al.*, "Uncertainty-aware molecular dynamics from Bayesian active learning for phase transformations and thermal transport in SiC," npj Comput. Mater. **9**, 36 (2023).
[19]L. Zhang, J. Han, H. Wang, R. Car, and W. E, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," Phys. Rev. Lett. **120**, 143001 (2018).
[20]J. Vandermause *et al.*, "On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events," npj Comput. Mater. **6**, 20 (2020).
[21]J. Vandermause, Y. Xie, J. S. Lim, C. Owen, and B. Kozinsky, "Active learning of reactive Bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt," Nat. Commun. **13**, 5183 (2022).
[22]B. Anderson, T. S. Hy, and R. Kondor, "Cormorant: Covariant molecular neural networks," Adv. Neural Inf. Process. Syst. **32**, 14537–14546 (2019).
[23]D. P. Kovács *et al.*, "Linear atomic cluster expansion force fields for organic molecules: Beyond RMSE," J. Chem. Theory Comput. **17**, 7696–7711 (2021).
[24]K. Schütt *et al.*, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," Adv. Neural Inf. Process. Syst. **30**, 991–1001 (2017).
[25]Z. Qiao *et al.*, "Informing geometric deep learning with electronic interactions to accelerate quantum chemistry," Proc. Natl. Acad. Sci. **119**, e2205221119 (2022).
[26]J. Klicpera, F. Becker, and S. Günnemann, "GemNet: Universal directional graph neural networks for molecules," Adv. Neural Inf. Process. Syst. **34**, 6790–6802 (2021).
[27]A. Johansson *et al.*, "Micron-scale heterogeneous catalysis with Bayesian force fields from first principles and active learning," arXiv:2204.12573 (2022).
[28]J. Behler, "Constructing high-dimensional neural network potentials: A tutorial review," Int. J. Quantum Chem. **115**, 1032–1050 (2015).
[29]C. Schran, K. Brezina, and O. Marsalek, "Committee neural network potentials control generalization errors and enable active learning," J. Chem. Phys. **153**, 104105 (2020).
[30]L. Zhang, D.-Y. Lin, H. Wang, R. Car, and W. E, "Active learning of uniformly accurate interatomic potentials for materials simulation," Phys. Rev. Mater. **3**, 023804 (2019).
[31]K. Tran *et al.*, "Methods for comparing uncertainty quantifications for material property predictions," Mach. Learn.: Sci. Technol. **1**, 025006 (2020).
[32]L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, and C. W. Coley, "Uncertainty quantification using neural networks for molecular property prediction," J. Chem. Inf. Model. **60**, 3770–3780 (2020).
[33]Y. Hu, J. Musielewicz, Z. Ulissi, and A. J. Medford, "Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials," Mach. Learn.: Sci. Technol. **3**, 045028 (2022).
[34]A. P. Soleimany *et al.*, "Evidential deep learning for guided molecular property prediction and discovery," ACS Cent. Sci. **7**, 1356–1367 (2021).
[35]M. Wen and E. B. Tadmor, "Uncertainty quantification in molecular simulations with dropout neural network potentials," npj Comput. Mater. **6**, 124 (2020).
[36]J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, "A quantitative uncertainty metric controls error in neural network-driven chemical discovery," Chem. Sci. **10**, 7913–7922 (2019).
[37]B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," Adv. Neural Inf. Process. Syst. **30**, 6402–6413 (2017).
[38]Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning* (PMLR, 2016), pp. 1050–1059.
[39]V. Zaverkin and J. Kästner, "Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design," Mach. Learn.: Sci. Technol. **2**, 035009 (2021).
[40]L. Kahle and F. Zipoli, "Quality of uncertainty estimates from neural network potential ensembles," Phys. Rev. E **105**, 015311 (2022).
[41]V. Zaverkin, D. Holzmüller, I. Steinwart, and J. Kästner, "Exploring chemical and conformational spaces by batch mode deep active learning," Digital Discovery **1**, 605–620 (2022).
[42]D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Process. **10**, 19–41 (2000).
[43]P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using Gaussian mixture model tokenization," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, 2002), Vol. 1, pp. I–757.
[44]Z.-K. Huang and K.-W. Chau, "A new image thresholding method based on Gaussian mixture model," Appl. Math. Comput. **205**, 899–907 (2008).
[45]D. A. Reynolds, "Gaussian mixture models," Encycl. Biom. **741**, 659–663 (2009).
[46]F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res. **12**, 2825–2830 (2011).
[47]G. Palmer *et al.*, "Calibration after bootstrap for accurate uncertainty quantification in regression models," npj Comput. Mater. **8**, 115 (2022).
[48]T. G. Dietterich, "Ensemble methods in machine learning," in *International Workshop on Multiple Classifier Systems* (Springer, 2000), pp. 1–15.
[49]E. V. Podryabinkin and A. V. Shapeev, "Active learning of linearly parametrized interatomic potentials," Comput. Mater. Sci. **140**, 171–180 (2017).