Spatially resolved uncertainties for machine learning potentials

Esther Heid,[a] Johannes Schörghuber, Ralf Wanzenböck, and Georg K. H. Madsen
Institute of Materials Chemistry, TU Wien, A-1060 Vienna, Austria

Machine learning potentials have become an essential tool for atomistic simulations, yielding results close to ab-initio simulations at a fraction of computational cost. With recent improvements on the achievable accuracies, the focus has now shifted on the dataset composition itself. The reliable identification of erroneously predicted configurations to extend a given dataset is therefore of high priority. Yet, uncertainty estimation techniques have achieved mixed results for machine learning potentials. Consequently, a general and versatile method to correlate energy or atomic force uncertainties with the model error has remained elusive to date. In the current work, we show that epistemic uncertainty cannot correlate with model error by definition, but can be aggregated over groups of atoms to yield a strong correlation. We demonstrate that our method correctly estimates prediction errors both globally per structure, and locally resolved per atom. The direct correlation of local uncertainty and local error is used to design an active learning framework based on identifying local sub-regions of a large simulation cell, and performing ab-initio calculations only for the sub-region subsequently. We successfully utilize this method to perform active learning in the low-data regime for liquid water.

I. INTRODUCTION

In recent years, machine learning potentials have gained importance as data-driven energy and force predictors for atomistic simulations, achieving an accuracy close to ab-initio results while offering a considerable speedup. The underlying model architectures have improved from neural networks built on simple invariant encodings of the atom environment[1–3] to elaborate equivariant, possibly multi-body interactions based on graph neural networks[4–7] or graph transformers[8] allowing both more precise and more data-efficient models.

With excellent models at hand, the focus has now shifted to the quality and quantity of the underlying data. New datasets such as the Open Catalyst 2020 and 2022 datasets[9,10] for adsorbates on surfaces, ANI-1x[11] for organic molecules, or Transition1x[12] for simple organic reactions have emerged, and can be used to train a baseline model, which can subsequently be fine-tuned to a system of interest. Since ab-initio calculations are costly it is nevertheless essential to develop clever data generation strategies, i.e. to identify where a model fails so that new ab-initio calculations can be issued and added to the training data. The quantification of the estimated error in a prediction is furthermore essential for decision-making processes. For atomistic simulations, the force uncertainty is key to determine whether the simulation is exploring structures that the machine learning potential is confident about.[13,14]

The identification of high-error predictions is still an open challenge in machine learning across many fields of research, and viable approaches depend on the details of the dataset, task, and model architecture. The error in a model prediction can be dissected into aleatoric (irreducible by addition of data) and epistemic (reducible by addition of data) contributions.[15–19] The aleatoric contribution stems from noise in the input data or missing input features, and can be learned by the model itself using heteroscedastic loss models (mean variance estimation) and variations thereof,[20,21] or can be estimated posthoc.[22,23] The epistemic contribution is associated with a limited knowledge of the model which can be further dissected into error from model variance and model bias.[19] It is usually approximated by training a committee of models, varying the model initialization seed, hyperparameters, architecture, or training data, and monitoring their disagreement on a prediction to obtain a proxy for the error from model variance.[24–26] Other techniques furthermore obtain a combined measure for aleatoric and epistemic error.[27,28] Yet, especially epistemic error is notoriously difficult to model, since the above approaches only capture error from variance, but not from model bias. Yet, model bias can be the major source of error in a model especially for small datasets and difficult-to-learn targets, so that the epistemic uncertainty obtained from committees often underestimates the model error.[19,29]

In the field of machine learning potentials, uncertainty estimation techniques have achieved mixed results.[21,30–33] Here, aleatoric uncertainty is usually negligible, since there is a direct, learnable relation between the input (the atomic numbers and coordinates) and the target (the ab-initio energies and forces) if different spin states and magnetic states are not taken into account. However, there is no direct correlation between the epistemic uncertainty of a single data point obtained from the standard deviation of committee predictions and the absolute error. This behavior has been reported for a wide variety of datasets and model architectures,[30,32,34–36] but its origin and possible remedies have not been identified yet. Heuristic approaches to average force uncertainties over structures as a proxy for model error have achieved success in active learning settings,[31,37] but are missing a theoretical framework and explanation of why and when such an approach is recommendable. Moreover, as simulations based on machine learning potentials are moving

[a] Electronic mail: esther.heid@tuwien.ac.at

towards trillions of atoms[38] and ab-initio calculations become infeasible for the full system, even a perfect estimator of the overall error is insufficient. For active learning it is therefore essential to trace down the overall error and uncertainty to smaller regions within a system, which can then be isolated. While recent approaches to active learning have attributed the uncertainty on a per atom basis,[39–41] there is no guarantee that the obtained atomic uncertainties actually correlate with the model error. In fact, recent studies reported that there is no direct correlation between the uncertainty of atomic forces and the actual error in the force prediction of that atom.[30,31]

In summary, the unresolved challenges of accurate global and local uncertainty estimates largely hinder the development of efficient active learning cycles for machine learning potentials. Yet, the field has a major advantage over other prediction tasks: Instead of learning to predict a single quantity (one target value per data point), a machine learning potential always predicts a molecular/full-structure energy, as well as atomic forces for each atom in each spatial direction. In the current study, we discuss why approaches based on aggregating uncertainties over all atoms in a data point give reliable uncertainty estimates, providing a theoretical framework for previous work.[31,37] We then take the concept a step further and present a new, model-agnostic, simple, and fast method to obtain spatially resolved uncertainties that correlate with the actual error for all atoms within a data point. We detail the theoretical basis of our approach and demonstrate it on diverse systems, namely organic reactions in gas phase, perovskite structures, and liquid water, where we find that global and spatially resolved local model errors can be predicted quantitatively. We then show how our approach enables a reliable identification of spatially resolved high-error regions using simple committee standard deviations and demonstrate the capabilities of these local uncertainties for active learning.

II. METHODS

A. Data sets and models

The equivariant message-passing neural network MACE[4] was used as provided, with hyperparameters as indicated in the following paragraphs. In all cases, training was performed using the AMSGrad optimizer with hyperparameters and learning rate schedules given by the defaults set in the MACE package. All models were constructed using two layers with 128 channels for even and odd parity features and a maximum order $l_{\max} = 3$ of spherical harmonics. Eight Bessel basis functions and a polynomial cutoff function of order $p = 5$ were used for generating the radial features, which are passed into MLPs with three layers of 64 nodes each and SiLU serving as the non-linear transfer function. The final readout function generating the atomic energies is given by a MLP with a single layer of 16 hidden features. In the Supporting Information, we furthermore report results with smaller MACE models (differing in the number of channels and cutoff radius), as well as invariant NeuralIL[2,31] models.

Transition1x[12] was downloaded as provided. From the roughly 10M data points of 10,073 reactions, we only kept the last, converged reaction pathways, where each pathway is made up of 10 images of the NEB search, resulting in 100k data points. The dataset was then split into a training and validation fraction made up of structures of the first or last two images (index 1, 2, 9, 10) in the NEB search (40k data points), corresponding to the equilibrated reactants, products, and configurations close to these equilibrated structures. All other image indices were put into the test set (60k data points). This split allows us to explore the correlation of epistemic uncertainty with the absolute error for regions the model has never seen, namely non-equilibrium configurations along diverse reaction paths. Since the image indices are known for all data points, we can evaluate the correlation as a function of the distance to the training image indices, and therefore explore mild (index 3, 8) to strong (index 5, 6) out-of-distribution examples.

MACE models for Transition1x were trained with a cutoff of 5 Å for a maximum of 1400 epochs with an early stopping patience of 50, with a force weight of 100.0 and an energy weight of 1.0. Then, 100 further epochs without early stopping were conducted with a force weight of 100.0 and an energy weight of 1000.0.

$SrTiO_3(110)$-4×1 structures were obtained from a subset of structures originally published in Ref. 42, which were then re-evaluated via VASP version 6.2.0[43] single-point evaluations with the r^{2}SCAN functional.[44] The energy cutoff was set to 440 eV and the width of Gaussian smearing to 0.02 eV. The final data set contained 889 unique structures, which were split randomly into training, validation, and test sets with 554, 237, and 98 data points, respectively.

MACE models for $SrTiO_3$ were trained with a cutoff of 4 Å for a maximum of 1200 epochs with an early stopping patience of 50, where the force and energy weights corresponded to 100.0 and 1.0, respectively. Subsequently, 300 epochs without early stopping were conducted at force and energy weights of 100.0 and 1000.0, respectively.

Finally, 1,593 structures of 64 water molecules originally calculated at the revPBE0-D3 level of theory and periodic boundary conditions were taken from Cheng et al.[45] as provided. Since the set contains five structures with duplicate atomic positions these were removed from the dataset, resulting in 1,588 unique structures. The energies and forces for these structures were recomputed at the RPBE-D3[46,47] level of theory using VASP version 6.4.2.[43] The hard PAW potential setups provided by the VASP package were used, with the cutoff energy set to 850 eV, the width for Gaussian smearing set to 0.05 eV and solely the Γ-point of the Brillouin zone being sampled. Following the results reported in Ref. 48, D3 corrections have been computed with the zero damping scheme.

The dataset was randomly split into training, validation and test sets with ratios 80:10:10.

MACE models for liquid water were all computed with a cutoff radius of 4 Å. Training was initially run with a force weight of 100.0 and an energy weight of 1.0 for 800 epochs without early stopping. Subsequently, the energy and force weights were adjusted to 1000.0 and 100.0 respectively and 200 more epochs were performed.

All datasets as used in this study are available at 10.5281/zenodo.11086346.

B. Active learning for a molecular dynamics simulation of water

Three independent active learning cycles were run for 15 iterations, one based on identifying new structures using local uncertainties, one based on atomic uncertainties, and one based on sampling randomly. All start from a model trained on an 80:20 split of the 50 highest energy structures contained in the water dataset described above. At each iteration, a 20 ps molecular dynamics simulation of a system containing 128 water molecules at a fixed density of $0.86\,\mathrm{g\,cm^{-3}}$ was run with a timestep of 0.5 fs in the NVT ensemble at 350 K using a Nosé-Hoover thermostat. LAMMPS version 2023.3.28[49] was used with MACE as the engine for energy and force evaluations. The given density lies below the experimental density of water, as expected for the RPBE-D3 level of theory.[48]

The trajectories obtained from the MD simulations were used to generate new training data to improve the models: At each active learning step, the first 10 ps of the trajectory were divided into ten evenly sized segments to guarantee new structures were sampled at different intervals of the simulation. For the active learning cycle based on randomly selecting new data, a random snapshot was selected from each of the ten segments. For the local uncertainty-based runs, snapshots were selected from the subsets by calculating the locally aggregated uncertainties (as detailed in the Results section) for each atom based on a committee of five MACE models using an aggregation $r_{\mathrm{cut}} = 4\,\text{Å}$ and choosing the frames featuring the environments with the highest uncertainties. For the atomic uncertainties, frames and atoms were selected based on the highest atomic force uncertainty (as detailed in the Results section). Boxes containing 64 water molecules were subsequently cut out from the snapshots: A central oxygen atom and the 63 closest oxygen atoms to it were selected, and the water molecules determined by selecting the two hydrogen atoms closest to each oxygen respectively. The box length for the new, smaller configuration is given by $l_{\mathrm{new}} = \left(\frac{1}{2}V_{\mathrm{init}}\right)^{\frac{1}{3}}$ and the originally selected center oxygen atom was placed in the center of a cubic box with this length. For the active learning cycle based on local uncertainties, the central atom was given by the oxygen atom featuring the maximum local uncertainty in the snapshot. For the run based on random selection, the index of the central oxygen atom was randomly generated. To avoid introducing high energies and forces due to cutting a periodic box from a larger initial configuration, the atoms close to the border of the new box were relaxed using the MACE model at the current active learning iteration based on the following procedure: All atoms within a distance of $0.8\,l_{\mathrm{new}}/2$ of the central oxygen atoms were kept fixed. The box was then padded by 2 Å in each direction. Five BFGS iterations were performed to relax the positions of the free atoms and the box size decreased by 0.2 Å in each direction. This procedure was repeated until the original box size was recovered, to allow the border regions to relax to physically meaningful structures. Energies and forces for the configurations obtained by this procedure were then calculated using the DFT setup described above. A new active learning iteration was then started by adding the new configurations to the dataset from the previous iteration and retraining MACE models from randomly initialized model parameters based on the new dataset.

We furthermore generated ten additional data points per AL cycle for both the random and local uncertainty-based selection following the exact same procedure as described above, utilizing the trajectory from 10-20 ps. This data was used as independent test set, totalling 360 data points.

III. RESULTS

A. Benchmark model

In this subsection, we detail the general method developed in this study, and verify it by the Monte Carlo simulation of a simple system featuring model variance as its only error source. In the following, we then discuss real data sets where model bias can play a significant role.

Due to the large flexibility of neural networks, training a model multiple times from different starting configurations yields slightly different predictions which can be assumed to approximately follow a Gaussian distribution. The following results will therefore be derived under the assumption that the committee predictions follow such a distribution. We first consider a committee of N_C independently trained models with model variance as the only source of error. The committee will provide N_C predictions $\hat{y}^l$ (with $l = 0, 1, ..., N_C$), with mean $\hat{y}$ and the committee standard deviation

$$s = \sqrt{\frac{1}{N_C - 1}\sum_{l}^{N_C}(\hat{y}^l - \hat{y})^2} \qquad (1)$$

computed as the unbiased estimator of the population standard deviation. In the following, we compute s as defined above, but note that for a low number of committee members, a correction to the underestimation of the true population standard deviation by the committee standard deviation can be applied.[50] The mean $\hat{y}$ is subsequently used as overall model prediction and s as an estimator of the model uncertainty. Repeating this experiment (i.e. training N_C models and averaging their predictions $\hat{y}^l$ to $\hat{y}$) multiple times would reveal that $\hat{y}$ is distributed around the target with a standard deviation of $s_{\hat{y}} = s/\sqrt{N_C}$. Choosing the target $y = 0$ we can, without loss of generality, then obtain the expectation value of the absolute error via

$$\langle|\hat{y}|\rangle = 2\int_0^\infty \hat{y}\frac{1}{s_{\hat{y}}\sqrt{2\pi}}e^{-\frac{\hat{y}^2}{2s_{\hat{y}}^2}}\,\mathrm{d}\hat{y} = \frac{2s_{\hat{y}}}{\sqrt{2\pi}} = \frac{2s}{\sqrt{2\pi N_C}}. \quad (2)$$

While Eq. (2) shows that the absolute error, on average, is related to the committee standard deviation by a factor $\alpha = 2/\sqrt{2\pi N_C}$, this is only true for a large number of repeated experiments of committee training and predictions. For a single committee prediction $\hat{y}$, for example predicting the molecular energy or the force of a single atom, the model error $|\hat{y}|$ cannot be directly correlated with the uncertainty obtained from the committee standard deviation s, because the prediction corresponds to a single random draw from the underlying distribution. At the same time a large number of repeated committee trainings is prohibitively expensive in practice, so that the correlation of the absolute error and the committee uncertainty through Eq. (2) might seem inapplicable to machine learning. Alternatively, the conversion factor from Eq. (2) can also be recovered by averaging over targets within a dataset of N_i data points, instead of averaging over multiple experiments of model retraining and prediction for a single target. In this case, the mean absolute error over all targets y_i is related to the mean committee standard deviation by a factor $2/\sqrt{2\pi N_C}$:

$$\frac{1}{N_i}\sum|\hat{y}_i - y_i| \simeq \frac{2}{\sqrt{2\pi N_C}}\cdot\frac{1}{N_i}\sum s_i. \quad (3)$$

This is frequently used to compute, e.g., error calibration curves.[36,51] Nevertheless, averaging over data points is undesirable, since we aim for a direct correlation of error and uncertainty for a single data point.

Machine learning potentials, however, do not only yield a prediction of a single energy value per configuration, but also forces for each atom. Thus, each data point i in a dataset of size N_i consists of $3N_j$ force components, where N_j is the number of atoms. As a result the model predictions $\hat{f}_{ij}^k$ aim to reproduce the target forces f_{ij}^k with k denoting the spatial direction. The predicted forces $\hat{f}_{ij}^k$ are obtained as a committee average over N_C committee predictions

$$\hat{f}_{ij}^k = \frac{1}{N_C}\sum_l^{N_C}\hat{f}_{ij}^{kl}, \quad (4)$$

resulting in a prediction error of

$$e_{ij}^k = |\hat{f}_{ij}^k - f_{ij}^k| \quad (5)$$

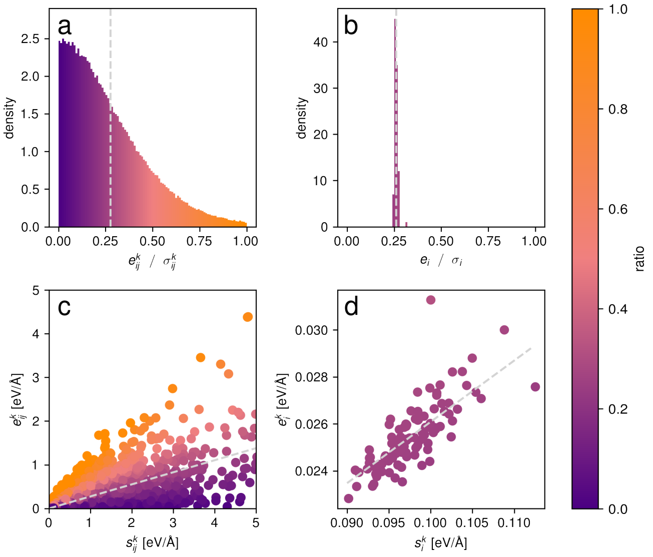

Figure 1. Histogram of the ratio of absolute error and uncertainty for Monte Carlo committee predictions featuring model variance as the only error source for a) individual atoms of all data points, and b) aggregated values within each data point. The gray dashed line indicates the average over all data points. The individual combinations of errors and uncertainties are furthermore shown c) per atom or d) per data point, colored according to the error/uncertainty ratio. The gray dashed line indicate the fit obtained via a mean α. The data were obtained with the parameters $N_C = 10$, $N_i = 100$, $N_j = 1000$ and ground-truth uncertainties drawn from an inverse-Gamma distribution.

and a committee standard deviation of

$$s_{ij}^k = \sqrt{\frac{1}{N_C - 1}\sum_l^{N_C}(\hat{f}_{ij}^{kl} - \hat{f}_{ij}^k)^2}. \quad (6)$$

Similar to the prediction of a single target discussed above, e_{ij}^k and s_{ij}^k cannot be correlated per atom, because again a single committee prediction corresponds to a single draw from a distribution centered around the target with a width determined by $s_{ij}^k/\sqrt{N_C}$. However, we can utilize force predictions averaged over sets of atoms to correlate variance error with uncertainty. Namely, by averaging over all directions k and atoms j, we arrive at a per-structure mean absolute error of

$$e_i = \frac{1}{N_j}\sum_j^{N_j}\frac{1}{3}\sum_{x,y,z}e_{ij}^k \quad (7)$$

and an average standard deviation of

$$s_i = \frac{1}{N_j}\sum_j^{N_j}\frac{1}{3}\sum_{x,y,z}s_{ij}^k. \quad (8)$$

In the following we will use Monte Carlo simulation to corroborate that e_i and s_i are strongly correlated whereas

e_{ij}^k and s_{ij}^k are not. For a dataset of N_i structures consisting of N_j atoms each, we draw σ_{ij}^k from an inverse-Gamma distribution as proposed in Ref. 28, so that each combination of i, j, and k gets assigned its own ground-truth uncertainty. We then draw from normal distributions with $\mu = 0$ and $\sigma = \sigma_{ij}^k$ to obtain N_C model predictions of a committee for each combination of i, j, and k. Subsequently, we compute the committee standard deviation s_{ij}^k and absolute error e_{ij}^k as well as the averaged values s_i and e_i within a data point, Eqs. (7) and (8), from the committee predictions. The simulation thus corresponds to a simple artificial system, where an otherwise perfect model only features variance error. Fig. 1 depicts histograms of the error-uncertainty ratio for the individual and aggregated case, as well as scatter plots of the errors versus uncertainties. In the Supporting Information, we furthermore report error-uncertainty ratios as a function of the size of the committee. For $N_C = 10$, the conversion factor between error and uncertainty from Eq. (3) is $2/\sqrt{2\pi N_C} = 0.25$. Clearly, both the individual and aggregated versions converge to a ratio of 0.25 averaged over all data points (gray dashed lines). However, as expected[30,32,34–36] the individual error-uncertainty ratios are broadly distributed so that there is no linear dependency between error and uncertainty for the individual data points, as visible in Fig. 1c. Fig. 1a and Fig. 1c furthermore depict that the distribution of the predicted values around the target lead to more predictions being observed towards the center of the distribution (small error, dark purple) than its outskirts (orange). Therefore, many individual uncertainty values obtained from the committee standard deviation significantly underestimate the actual model error, so that the individual committee uncertainties cannot be used to identify high-error atoms. The relation between atomic uncertainties and errors is therefore asymmetric, with large errors occurring only for large uncertainties, and small uncertainties permitting only small errors, but not the other way around.[34] In contrast, the ratio of averaged absolute errors over the uncertainties show a narrow distribution as visible in Fig. 1b and Fig. 1d, so that we can actually estimate the prediction error from the uncertainty by multiplication with 0.25, corroborating Eq. (3). Note that the conversion factor for this artificial system being lower than 1.0 results from the model being unbiased, which is usually not the case for real atomic systems.[30,31]

Monte Carlo simulations were performed for different N_C, N_i, N_j, as well as different distributions for σ_{ij}^k, namely uniformly random distributions within an interval, normal distributions, and even constant values, and observe the same behavior across all systems: The individual ratios are centered around $2/\sqrt{2\pi N_C}$ but are distributed too broadly to show a meaningful correlation between the absolute errors and uncertainties, while the ratio of the respective aggregated values are distributed in a narrow peak yielding a strong correlation.

We can therefore conclude that for models where the variance is the main source of error, the mean force error over a structure or molecule can be easily predicted by multiplication of the mean committee standard deviation by $2/\sqrt{2\pi N_C}$. For cases where model variance is the only source of error, we have thus identified a method to correlate the error and uncertainties within a single data point, by aggregating over the errors and uncertainties of all atoms within that data point. In the following, we explore whether this method is also applicable to models featuring bias errors of different magnitudes.

B. Real systems with mixed error sources

In addition to model variance, nearly all machine learning models also suffer from model bias, which can stem from fundamental shortcomings of the model, too little data, or ill-chosen features, amongst others.[29] Uncertainty estimates obtained from an ensemble of models only capture variance errors. Since model bias, as introduced due to the aforementioned reasons, can be the dominant contribution, ensemble uncertainties usually underestimate the actual error.[30,31] We therefore examine whether our approach also holds up for machine learning potentials trained on real data sets.

Since we aggregate force uncertainties over atoms, the number of atoms per data point is important and we chose three data sets examining a large variety of system sizes and configurations. 1) Transition1x[12] features nudged elastic band (NEB) searches for a wide range of organic reactions in gas phase with 7 to 23 H, C, N, and O atoms, and thus resembles rather small systems without periodic boundary conditions. From each reaction, we only used the last, converged reaction pathways made up of ten images each, where indices 1, 2, 9, and 10 were used as training and validation set. The test set can thus be split into different indices corresponding to a different level of extrapolation from the training configurations. 2) Surface reconstructions of crystalline $SrTiO_3$ were taken from Ref. 42, and feature 136 Sr, Ti, and O atoms per data point, thus resembling large systems with periodic boundary conditions. 3) A liquid water data set from Ref. 45 features 192 H and O atoms per data point, and thus resembles a large, homogeneous system with periodic boundary conditions that lends itself to dissecting the overall simulation cell into smaller sub-regions. On each dataset, we trained a five-membered ensemble of the equivariant message-passing neural network MACE.[4] See Methods for further details.

Fig. 2 depicts direct comparisons between individual (first column) and aggregated (second column) absolute errors versus uncertainties, as well as sparsification plots (third column) and the distribution of the ratios between absolute error and uncertainty α (fourth column). Further uncertainty metrics are reported in the Supporting Information. The sparsification curves are obtained by ordering the test data points by either the absolute error or the uncertainty, and then obtaining the mean absolute error over a fraction of the data points. Without remov-

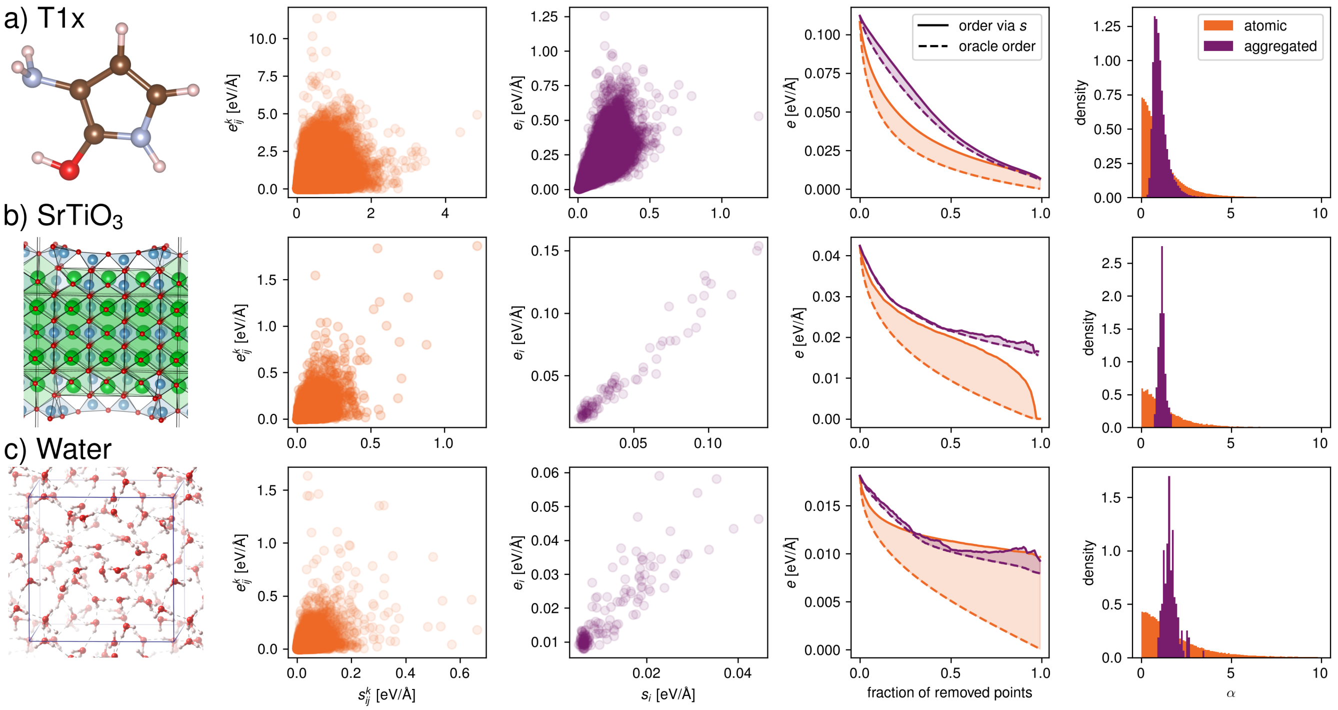

Figure 2. Relation between absolute error and uncertainty for the Transition1x dataset (a, top row), the SrTiO$_3$ dataset (b, middle row), and the liquid water dataset (c, bottom row). For each dataset, individual absolute errors versus uncertainties are depicted in the first column, aggregated absolute errors versus uncertainties in the second column, sparsification curves in the third column, and the distribution of the proportionality constant α in the fourth column.

ing any data points (i.e. the fraction of removed data points equals zero), this yields simply the mean absolute error averaged over all test data points, all atoms and all directions. By iteratively removing test data (highest values of error or uncertainty are removed first), the mean absolute error over the remaining data is therefore lowered if the ordering corresponds to the absolute error (oracle order). If the model uncertainties show the exact same order as the absolute errors, the area between the oracle curve and the sparsification curve is zero. A large area therefore indicates that the uncertainties are not ordered according the absolute error.

For all datasets, we observe large areas in the sparsification plots for the individual force uncertainties showing that they do not correlate with the absolute error, and cannot be used to order the test set according to the estimated error. In fact, as we have explored in the previous subsection, the single individual force uncertainties and errors cannot correlate for mathematical reasons. To obtain a direct correlation, uncertainties and errors have to be averaged over sets of atoms, here the full molecule or structure, to make use of Eq. (3). Although the data sets have a different number of data points, and a different number of atoms per data point, they all show a large improvement in correlation between error and uncertainty upon aggregation over all atoms in a data point. In all cases, an ordering of the predicted values according to their uncertainty corresponds nearly perfectly to the oracle ordering according to the true prediction error, and the distribution of α is narrow. Note that the distribution of α is not centered around the expected value of $2/\sqrt{2\pi N_C}$ anymore, since all models feature differing amounts of model bias, so that the error from model variance is only a small part of the overall prediction error. Moreover, the model error may not be Gaussian anymore, causing further deviations from Eq. (3). However, this only affects the magnitude of α (since a change in distribution changes its first moment), but does not impede the quality of correlation between the aggregated errors and uncertainties, as long as the error is still symmetric around zero, and the model still performs adequately for the data of interest (see Supporting Information for cases with low model performance).

Fig. 3a and Fig. 3b furthermore depict heatmaps of the individual and molecular errors versus uncertainties for the Transition1x dataset, since the actual functional dependence is hard to read from Fig. 2 due to the large number of test data points. Here, it becomes obvious that even for this difficult extrapolation task, the aggregation successfully leads to a highly correlated molecular error and uncertainty. The Transition1x dataset is especially interesting to research the correlation between error and uncertainty, since we can resolve it with respect to the NEB image index. Since indices 1, 2, 9, and 10 were used in the training and validation sets, we can plot the test set performance and uncertainty met-

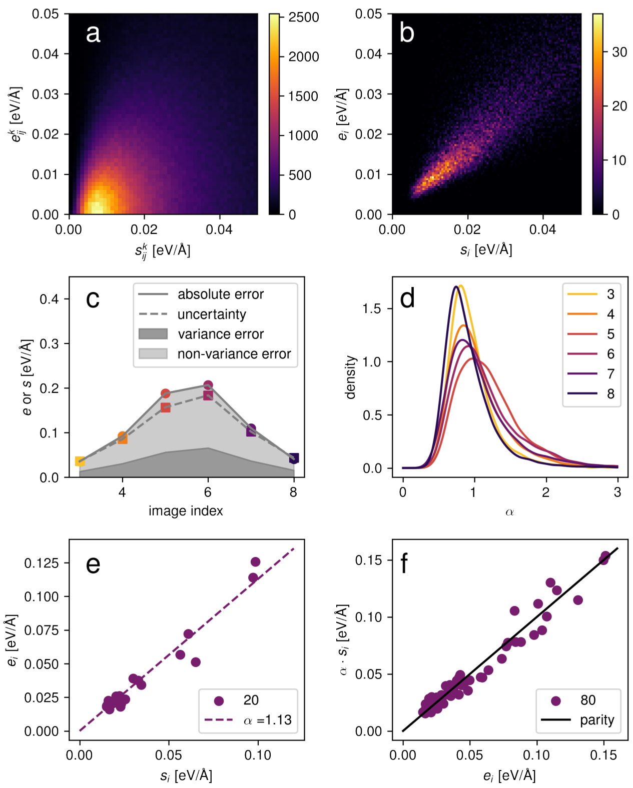

Figure 3. a) Replot of the individual errors vs uncertainties as heatmap for the Transition1x dataset. b) Replot of the molecular error vs uncertainty as heatmap for the Transition1x dataset. c) Mean absolute error and mean uncertainty over all data points and all atoms as a function of the reaction path image index for Transition1x. d) For each index, the distribution of α is plotted for the aggregated force uncertainties and errors. e) Fit of α obtained from 20 % of the $SrTiO_3$ test set. f) Prediction error obtained via the model uncertainty and the fitted value of α versus the true error for the remaining 80 % of the test set.

rics over the indices 3, 4, 5, 6, 7, and 8. The indices 5 and 6 are the most dissimilar to the training set, which has never seen any transition state structures, but only (close-to) equilibrium structures. We therefore expect the model to have a much larger model bias for the indices 5 and 6, than the indices 3 and 8. Fig. 3c depicts the mean absolute errors and mean uncertainty over all data points as a function of the index. We find that the model performs worst for indices 5 and 6, as expected. The model furthermore identifies these predictions as the most uncertain, featuring a high aggregated committee standard deviation. Further, the model is able to identify a higher model bias (extrapolation error), Fig. 3d, where the distribution of α shifts to higher values, the more the model encounters data points far away from the training set configurations. The values change from 0.93 for index 8 to 1.21 for index 5, as opposed to a pure-variance model with $\alpha = 2/\sqrt{2\pi N_C} = 0.36$ (all values given in the Supporting Information). Since Eq. (3) directly outputs the expected amount of variance error for a given dataset and model architecture, we can furthermore compute the fraction of variance and non-variance error of the overall error. The variance error is shaded in dark gray in Fig. 3c, as well as the non-variance error in light gray, which both visibly increase for structures far away from equilibrium. Together, the overall uncertainty strongly correlates with the actual error. The good correlation of aggregated uncertainties with aggregated errors for models with a significant amount of bias error may seem surprising, since technically the relation in Eq. (3) relation only holds for errors stemming from variance. Recently, the bias error was found to be correlated with the variance error when changing the dataset size or model architecture.[29] Here, we also find that the presence of model bias only changes the value of α and broadens its distribution slightly, but preserves the correlation. In fact, we find that the bias and variance error are correlated for machine learning potentials (the variance error also increased for index 5 and 6), but the amount of bias error differs between the test sets with different degrees of extrapolation. To further illustrate the behavior of α, we also trained less accurate versions of the presented MACE models and models based on the NeuralIL architecture.[2] The results are compiled in the Supplementary Information, together with correlation coefficients, mean α and model bias percentages for all model types. Overall, the aggregated force uncertainties reliably identify sets of data points with a higher model bias, here extrapolation error due to missing training data, which is an important prerequisite for successful active learning cycles.

Finally, we observe that although α never corresponds to the variance-only case, its true value can easily be obtained from a small test set, and subsequently be used to transform between the model uncertainty (committee standard deviation) and the prediction error. Fig. 3e depicts how α is fitted from 20 % of the test data of the $SrTiO_3$ system, which is then subsequently used to obtain the predicted error for the remaining 80 % of the test data, Fig. 3f. The predicted error is simply αs_i, and can approximate the true error up to a mean absolute deviation of only 0.005 eV Å^{-1} for each single data point(for a model with a force mean absolute error of 0.042 eV Å^{-1}), thus providing an accurate, quantitative proxy of the error without the need for a separate calibration step. This enables an easy and reliable identification of erroneous model predictions for single data points, as opposed to a full dataset. In the Supporting Information, we furthermore report predicted errors for the Transition1x dataset resolved per NEB index.

C. Spatially resolved uncertainty

For large systems, we can take the aggregation concept one step further, as we obtain sufficient statistics even when aggregating only over part of a large system instead of all atoms in the system. To obtain a spatially

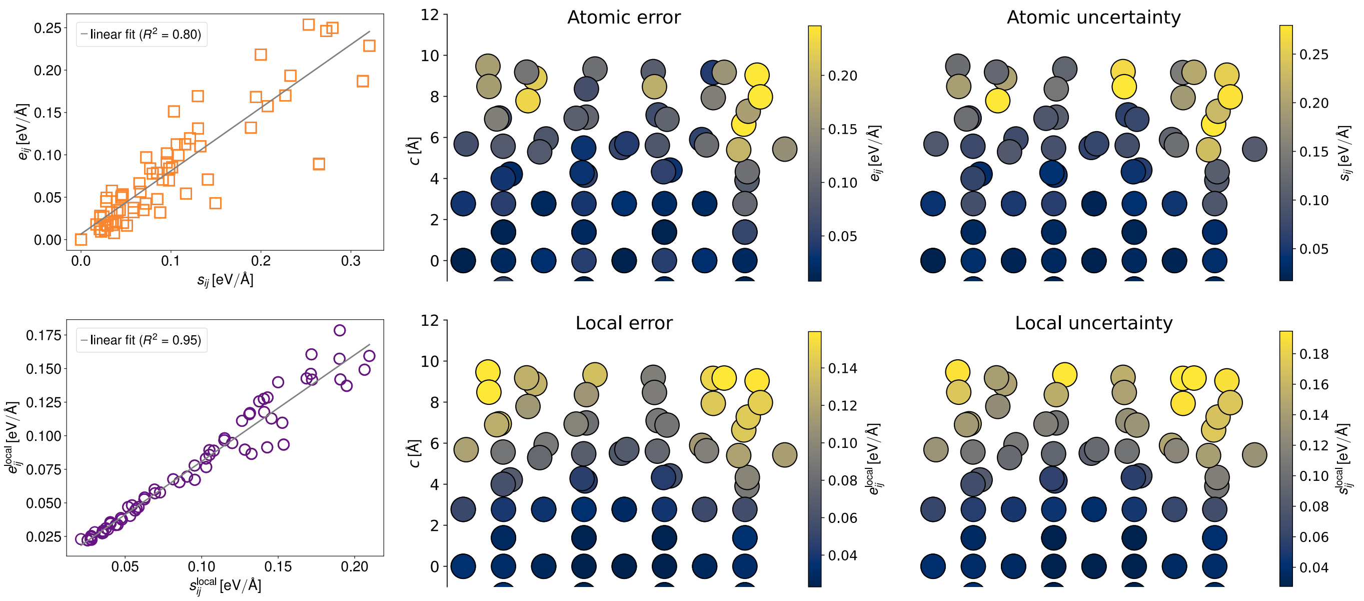

Figure 4. The two left-most panels show parity plots comparing uncertainty estimates to errors, atomic (top) and local (bottom), respectively. Spatially resolved errors and uncertainties for a data point from the $SrTiO_3$ dataset (side view) are illustrated in the other four panels. Top: The atomic error and committee uncertainty obtained by summing over the three spatial directions for each atom. Bottom: The local error and uncertainty, Eqs. (9) and (10), aggregated up to $r_{\text{cut}} = 4\,\text{Å}$.

resolved error and uncertainty, we aggregate locally over N_n neighboring atoms n around atom j (including j) located within a cutoff r_{cut} via

$$e^{\text{local}}_{ij} = \frac{1}{N_n}\sum_{n}^{N_n} \frac{1}{3}\sum_{x,y,z} e^k_{in}. \tag{9}$$

Similarly, the uncertainties can be averaged over all neighboring atoms as

$$s^{\text{local}}_{ij} = \frac{1}{N_n}\sum_{n}^{N_n} \frac{1}{3}\sum_{x,y,z} s^k_{in}. \tag{10}$$

Instead of a single absolute error and uncertainty per data point, we therefore obtain N_j local absolute errors and uncertainties for each data point, each centered around an atom j.

Spatially resolved errors and uncertainties for the $SrTiO_3$ surface are shown in Fig. 4, where a lighter color corresponds to a larger error or uncertainty. It is clear that mere atomic errors and uncertainties are not correlated, and the uncertainties provide lots of false positives for the expected error. This is in agreement with earlier findings where the aggregated uncertainty of the entire system was used for an active learning procedure, because the atomic uncertainties failed to correlate with the error.[31] In contrast, the local errors and uncertainties correlate very strongly, so that the local uncertainty can be used to reliably identify high-error regions.

Fig. 5a depicts the local mean absolute error versus the local uncertainty for $r_{\text{cut}} = 5\,\text{Å}$ for the liquid water dataset. In contrast to the individual force component errors and uncertainties in Fig. 2c which feature little correlation, the local errors and uncertainties are correlated. Furthermore, the maximum local uncertainty,

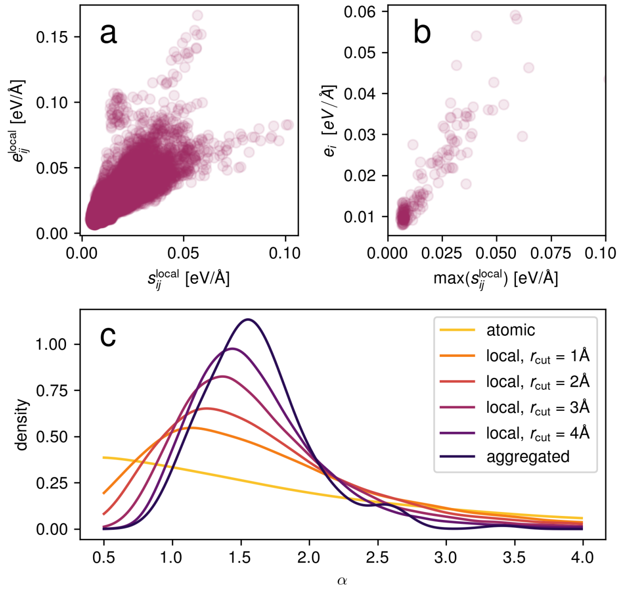

Figure 5. Local errors for the for the liquid water dataset. a) Local mean absolute error versus local uncertainty for $r_{\text{cut}} = 5\,\text{Å}$. b) Global mean absolute error for all atoms in a data point over the maximum of local uncertainties. c) Ratio of local error versus local uncertainty for different cutoff radii.

Fig. 5b, correlates with the overall mean absolute error averaged over all atoms in the data point, so that the locally aggregated force uncertainties can both be used to select high-error local substructures within a data point, but also high-error data points. Fig. 5c shows the distribution of the ratio of local error versus local uncertainty for different cutoff radii. Again it is seen that the atomic uncertainties are uncorrelated with the corresponding absolute error. In a molecular system like water, one might intuitively aggregate over individual water molecules. Fig. 5c shows that even such a highly localized aggregation, $r_{\mathrm{cut}} = 1\,\text{Å}$, leads to a certain degree of correlation. For larger cutoff radii, the correlation becomes stronger and the locally aggregated uncertainty and error naturally converge to the globally aggregated quantities eventually. The choice of cutoff radius thus presents a trade-off between the quality of the correlation and the locality of the obtained values.

In the Supporting Information, we report detailed uncertainty metrics comparing atomic, local, and per-structure errors and uncertainties, and find excellent correlation between the spatially resolved local errors and uncertainties developed in this study. The local uncertainties can be directly used to estimate local errors for the Transition1x, $SrTiO_3$, and water data sets, and feature high correlation coefficients, Spearman's rank coefficients, as well as excellent error-based calibration. We furthermore note that the local uncertainties do not require a recalibration, but only a simple fit of α to obtain direct error estimates via $\alpha s_{ij}^{\mathrm{local}}$.

In addition, the strong correlation between local error and uncertainty makes the obtained uncertainties sharp, tight, and disperse. Sharpness quantifies the mean prediction uncertainty, where a method yielding high uncertainties across all predictions (low sharpness) is undesirable. Tightness refers to the small-scale reliability of uncertainty predictions with respect to reference values, and is high for a perfect calibration.[34] Dispersion describes the spread of observed uncertainties, where a model predicting the same uncertainty across all predictions (low dispersion) is undesirable.[52] For all data sets and models reported, we observe sharp, tight, and disperse local uncertainty predictions due to their direct, strong correlation with error.

D. Active learning using spatially resolved uncertainty

With the spatially resolved uncertainties established as proxies for the local prediction error, we in the following explore their use in an active learning scenario. The concept of local uncertainties is most helpful for large systems, where they would enable an active learning loop involving cutting out a fragment of the system for ab-initio calculations, instead of recomputing the full system. Molecular dynamics (MD) simulations pose an important application, where a large simulation cell prohibits the addition of new training data of the full system.

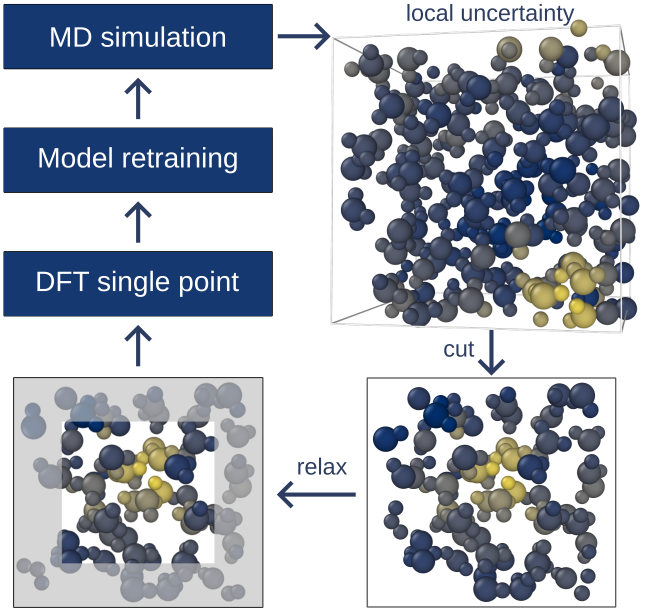

Figure 6. Schematic depiction of an active learning cycle, where an MD simulation with a machine learning potential generates a trajectory. Frames with high local uncertainties, indicated by lighter, yellow colors, are selected and cut into smaller boxes. The edge regions (gray background) are relaxed using the model and finally the energies and forces for the small boxes are then obtained ab-initio, and added to the training set, on which a new model is trained.

Here, we can make use of the direct correlation of local uncertainty with local error to design an active learning framework based on local subregions where a model prediction is identified to fail. Similar approaches have been proposed before, but suffer from selection based on atomic uncertainties,[35,39–41] which we show to not correlate with the actual error. Further, the scheme proposed in Ref. 39 only makes use of the forces on the central atom, thus not making use of a large amount of reference data. As detailed in the Methods section, we start with only 50 high energy water structures largely irrelevant to the density and temperature of interest. The active learning cycle is schematically depicted in Fig. 6. The procedure consists of iteratively selecting sub-regions of the large simulation cell exhibiting large local uncertainty and constructing new small cells for which ab-initio calculations are performed. Thereby, new training data is only added for the relevant subregions. Fig. 7 depicts the performance of the uncertainty-based models on test sets also constructed from MD simulation (see Methods for details). For all test sets the error falls off rapidly as data from the local substructures are added to the training set, with each active-learning cycle adding only ten data points. As comparison, we furthermore randomly selected frames and regions to cut out boxes, as well as based on the atomic, non-aggregated uncertainty. Averaged over all test data points, the accuracy of predicting forces does not differ significantly between the

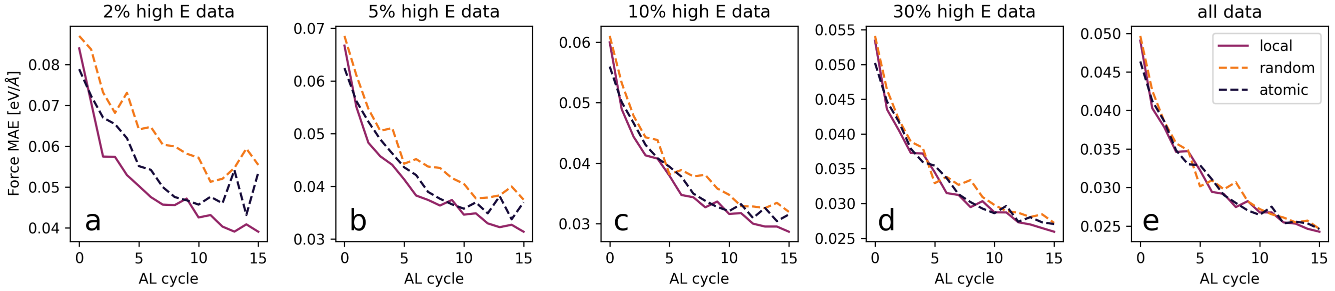

Figure 7. Model performance (mean absolute errors of the forces) as a function of the active learning cycle with each active-learning cycle adding ten data points of new ab-initio calculations. The active learning is performed via a random (orange dashed lines), atomic uncertainty (blue dashed lines) or local uncertainty-based (purple full lines) sampling strategy. The different panels show different subsets of the test data, sorted by their potential energy.

three approaches, because with the little amount of training data available, the addition of any data is helpful to the model. However, the uncertainty-sampling method largely outperforms random and atom-uncertainty sampling for high-energy structures even after only a few active learning cycles. This is especially important for molecular dynamics simulations, where even infrequent wrong predictions of the forces can cause the trajectory to deviate from an ab-initio trajectory significantly, or can cause the simulation to crash. We therefore largely favor a model not only able to predict typical configurations well, but also extrapolate outside of that realm. The identification of ill-predicted environments via the local uncertainty therefore enables an efficient, fast, and effective way to collect new training data in an active learning scheme for molecular dynamics simulation and yields stable models better able to extrapolate.

IV. CONCLUSION

We have shown that the epistemic uncertainty of a single prediction obtained via a committee standard deviation cannot be directly correlated to the absolute error, reiterating the well-known asymmetric relation between error and uncertainty.[34] For machine learning potentials, this holds true both for energy and atomic force predictions. Building on these results, we have developed an approach to nevertheless use force uncertainties to identify high-error data points by aggregating force uncertainties and errors over groups of atoms. For variance-dominated models, a strong correlation between aggregated uncertainties and errors can be proven mathematically. For machine learning potentials, we further find that the approach holds for models containing a substantial amount of bias. For such models, we observe that the presence of bias amplifies the proportionality factor between aggregated uncertainties and errors compared to the variance-only case. Since the proportionality factor needs to be obtained from a fit to a validation set for any real-world scenario anyways, this does not impede the reliability of the proposed method. The aggregated force uncertainty is directly correlated both with the aggregated absolute error over that group of atoms but also the total absolute error of the model prediction for various datasets. Our approach confidently identifies high-error data points for systems with low and high numbers of atoms, and is applicable to periodic and non-periodic systems in gas, liquid, and solid phase alike. We furthermore demonstrated that an aggregation must not necessarily include all atoms of a structure, but can be restricted to neighboring atoms around an atom of interest. Locally aggregated uncertainties can then be applied to identify high-error local substructures, and thus resolve absolute errors on an atomic scale. The benefits of locally aggregated uncertainties were showcased for an active learning study, indicating that data selection via spatially resolved uncertainties allows for a data-efficient and fast training of accurate machine learning potentials. We therefore envision this workflow to be very powerful for active learning frameworks in low-data regimes for large, demanding systems, such as multi-phase systems.

We furthermore note that the generality of our approach allows for application to a variety of systems, and have already utilized spatially resolved uncertainties as developed in this study for surface reconstructions,[53] interfaces and defected structures (results not yet published). These type of configurations consist of substructures featuring higher prediction errors than the rest of the structure, which is well resolved by the demonstrated approach. We therefore envision the current study to spark new applications within machine learning potentials, especially active learning cycles, across a large variety of systems.

DATA AND SOFTWARE AVAILABILITY

The Transition1x, $SrTiO_3$, and water datasets including all data generated during the active learning loops for water are available on Zenodo at 10.5281/zenodo.11086346, together with scripts to calculate locally aggregated uncertainties and cut and relax water boxes for the active learning study, as well as a Jupyter notebook for the Monte Carlo experiment.

ACKNOWLEDGEMENT

This research was funded in part by the Austrian Science Fund (FWF) 10.55776/COE5 and 10.55776/F81. For open access purposes, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission.

AUTHOR CONTRIBUTIONS

E.H. conceived the project. E.H., J.S, R.W. and G. K. H. M. designed the experiments. E.H., J.S and R.W. trained and evaluated models, analyzed the data and wrote the underlying computer code. G.K.H.M. supervised the research. All authors contributed to the manuscript writing.

REFERENCES

[1] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).
[2] H. Montes-Campos, J. Carrete, S. Bichelmaier, L. M. Varela, and G. K. H. Madsen, J. Chem. Inf. Model. **62**, 88 (2021).
[3] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, J. Chem. Phys. **148**, 241722 (2018).
[4] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, Adv. Neural. Inf. Process. Syst. **35**, 11423 (2022).
[5] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, Nat. Commun. **13**, 2453 (2022).
[6] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, and B. Kozinsky, Nat. Commun. **14**, 579 (2023).
[7] S. Passaro and C. L. Zitnick, arXiv preprint arXiv:2302.03655 (2023).
[8] Y.-L. Liao, B. Wood, A. Das, and T. Smidt, arXiv preprint arXiv:2306.12059 (2023).
[9] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, *et al.*, Acs Catal. **11**, 6059 (2021).
[10] R. Tran, J. Lan, M. Shuaibi, B. M. Wood, S. Goyal, A. Das, J. Heras-Domingo, A. Kolluru, A. Rizvi, N. Shoghi, *et al.*, ACS Catal. **13**, 3066 (2023).
[11] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, J. Chem. Phys. **148**, 241733 (2018).
[12] M. Schreiner, A. Bhowmik, T. Vegge, J. Busk, and O. Winther, Sci. Data **9**, 779 (2022).
[13] J. P. Janet, C. Duan, T. Yang, A. Nandy, and H. J. Kulik, Chem. Sci. **10**, 7913 (2019).
[14] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, and B. Kozinsky, Nat. Commun. **13**, 5183 (2022).
[15] S. C. Hora, Reliab. Eng. Syst. Saf. **54**, 217 (1996).
[16] A. Der Kiureghian and O. Ditlevsen, Struct. Saf. **31**, 105 (2009).
[17] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, and E. Hüllermeier, Inf. Sci. **255**, 16 (2014).
[18] A. Kendall and Y. Gal, Adv. Neural Inf. Process. Syst. **30** (2017).
[19] E. Hüllermeier and W. Waegeman, Mach. Learn. **110**, 457 (2021).
[20] D. A. Nix and A. S. Weigend, in *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, Vol. 1 (IEEE, 1994) pp. 55–60.
[21] P. B. Jørgensen, J. Busk, O. Winther, and M. N. Schmidt, arXiv preprint arXiv:2312.04174 (2023).
[22] G. Scalia, C. A. Grambow, B. Pernici, Y.-P. Li, and W. H. Green, J. Chem. Inf. Model. **60**, 2697 (2020).
[23] N. Zhan and J. R. Kitchin, AIChE J. **68**, e17516 (2022).
[24] B. Efron, in *Breakthroughs in statistics* (Springer, 1992) pp. 569–593.
[25] Y. Gal and Z. Ghahramani, in *international conference on machine learning* (PMLR, 2016) pp. 1050–1059.
[26] I. Cortes-Ciriano and A. Bender, J. Chem. Inf. Model. **59**, 3330 (2019).
[27] B. Lakshminarayanan, A. Pritzel, and C. Blundell, in *Advances in Neural Information Processing Systems*, Vol. 30 (2017) pp. 6402–6413.
[28] A. P. Soleimany, A. Amini, S. Goldman, D. Rus, S. N. Bhatia, and C. W. Coley, ACS Cent. Sci. **7**, 1356 (2021).
[29] E. Heid, C. J. McGill, F. H. Vermeire, and W. H. Green, J. Chem. Inf. Model. **63**, 4012 (2023).
[30] L. Kahle and F. Zipoli, Phys. Rev. E **105**, 015311 (2022).
[31] J. Carrete, H. Montes-Campos, R. Wanzenböck, E. Heid, and G. K. H. Madsen, J. Chem. Phys. **158**, 204801 (2023).
[32] A. Zhu, S. Batzner, A. Musaelian, and B. Kozinsky, J. Chem. Phys. **158**, 164111 (2023).
[33] M. Kellner and M. Ceriotti, Machine Learning: Science and Technology **5**, 035006 (2024).
[34] P. Pernot, J. Chem. Phys. **157**, 144103 (2022).
[35] Y. Lysogorskiy, A. Bochkarev, M. Mrovec, and R. Drautz, Phys. Rev. Mater. **7**, 043801 (2023).
[36] M. H. Rasmussen, C. Duan, H. J. Kulik, and J. H. Jensen, J. Cheminform. **15**, 1 (2023).
[37] D. Schwalbe-Koda, A. R. Tan, and R. Gómez-Bombarelli, Nat. Commun. **12**, 5104 (2021).
[38] A. Johansson, Y. Xie, C. J. Owen, J. S. Lim, L. Sun, J. Vandermause, and B. Kozinsky, arXiv preprint arXiv:2204.12573 (2022).
[39] C. Zeng, X. Chen, and A. A. Peterson, J. Chem. Phys. **156**, 064104 (2022).
[40] S. Roy, J. P. Dürholt, T. S. Asche, F. Zipoli, and R. Gómez-Bombarelli, Nat. Commun. **15**, 6030 (2024).
[41] L. C. Erhard, J. Rohrer, K. Albe, and V. L. Deringer, Nat. Commun. **15**, 1927 (2024).
[42] R. Wanzenböck, M. Arrigoni, S. Bichelmaier, F. Buchner, J. Carrete, and G. K. H. Madsen, Digit. Discov. **1**, 703 (2022).
[43] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169 (1996).
[44] J. W. Furness, A. D. Kaplan, J. Ning, J. P. Perdew, and J. Sun, J. Phys. Chem. Lett. **11**, 8208– 8215 (2020).
[45] B. Cheng, E. A. Engel, J. Behler, C. Dellago, and M. Ceriotti, Proc. Natl. Acad. Sci. U.S.A. **116**, 1110 (2019).
[46] B. Hammer, Phys. Rev. B **59**, 7413 (1999).
[47] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, J. Chem. Phys. **132**, 154104 (2010).
[48] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, PNAS **113**, 8368 (2016).
[49] A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in 't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott, and S. J. Plimpton, Comput. Phys. Commun. **271**, 108171 (2022).
[50] R. M. Brugger, Am. Stat. **23**, 32 (1969).
[51] D. Levi, L. Gispan, N. Giladi, and E. Fetaya, Sensors **22**, 5540 (2022).
[52] K. Tran, W. Neiswanger, J. Yoon, Q. Zhang, E. Xing, and Z. W. Ulissi, Mach. Learn.: Sci. Technol. **1**, 025006 (2020).
[53] R. Wanzenböck, E. Heid, M. Riva, G. Franceschi, A. M. Imre, J. Carrete, U. Diebold, and G. K. H. Madsen, chemRxiv 10.26434/chemrxiv-2024-9l6jc-v2 (2024).