Review

Jin Dai, Santosh Adhikari and Mingjian Wen*

Uncertainty quantification and propagation in atomistic machine learning

https://doi.org/10.1515/revce-2024-0028
Received May 3, 2024; accepted December 4, 2024;
published online December 30, 2024

Abstract: Machine learning (ML) offers promising new approaches to tackle complex problems and has been increasingly adopted in chemical and materials sciences. In general, ML models employ generic mathematical functions and attempt to learn essential physics and chemistry from large amounts of data. The reliability of predictions, however, is often not guaranteed, particularly for out-of-distribution data, due to the limited physical or chemical principles in the functional form. Therefore, it is critical to quantify the uncertainty in ML predictions and understand its propagation to downstream chemical and materials applications. This review examines existing uncertainty quantification (UQ) and uncertainty propagation (UP) methods for atomistic ML under the framework of probabilistic modeling. We first categorize the UQ methods and explain the similarities and differences among them. Following this, performance metrics for evaluating their accuracy, precision, calibration, and efficiency are presented, along with techniques for recalibration. These metrics are then applied to survey existing UQ benchmark studies that use molecular and materials datasets. Furthermore, we discuss UP methods to propagate uncertainty in widely used materials and chemical simulation techniques, such as molecular dynamics and microkinetic modeling. We conclude with remarks on the challenges and opportunities of UQ and UP in atomistic ML.

Keywords: uncertainty quantification; machine learning; model calibration; reliability; molecular simulation

1 Introduction

Since the breakthrough in image recognition using deep neural networks (NNs) back in 2012 (Krizhevsky et al. 2012), machine learning (ML) approaches have been increasingly leveraged to study complex chemical and materials systems. They have achieved remarkable successes, from designing catalysts (Back et al. 2019; Zahrt et al. 2019), to discovering functional materials (Axelrod et al. 2022; Rao et al. 2022) and studying protein folding (Baek et al. 2021; Jumper et al. 2021), to name a few. The ML approaches applied in these tasks take advantage of a wide range of techniques, but they share a common core idea: modeling molecules, materials, and chemical reactions at the atomic scale and looking for patterns and trends in atomic data, which we refer to as *atomistic machine learning*.

One of the most impactful developments in atomistic ML for chemical and materials science is creating interatomic potentials (i.e., force fields) to model the interactions between atoms. This goes from early endeavors that model individual molecular/material systems (e.g., the feed-forward NN potential (Behler 2021; Behler and Parrinello 2007) and Gaussian approximation potential [GAP] (Bartók et al. 2010; Deringer et al. 2021)) to more recent efforts to build universal potentials for the entire periodic table (e.g., M3GNet (Chen and Ong 2022), CHGNet (Deng et al. 2023) and MACE-MP (Batatia et al. 2024)). While interatomic potentials remain an active focus of atomistic ML, the field has moved beyond and expanded to encompass the prediction of arbitrarily complicated molecular, materials, and reaction properties (Ceriotti 2022; Fedik et al. 2022). These include molecular dipole moments (Gastegger et al. 2021; Unke and Meuwly 2019), bond strength (St. John et al. 2020; Wen et al. 2021), high-rank material tensors (Pakornchote et al. 2023; Wen et al. 2024), neutron, X-ray, and vibrational spectroscopies (Chen et al. 2021; Schienbein 2023), and reaction rates and yields (Heid and Green 2022; Wen et al. 2023), among others.

Despite these successes, the reliability of atomistic ML models remains a significant concern (Heid et al. 2023; Peterson et al. 2017; Tavazza et al. 2021). The complex and high-dimensional nature of chemical and materials systems,

***Corresponding author: Mingjian Wen**, Department of Chemical and Biomolecular Engineering, University of Houston, Houston, TX, 77204, USA, E-mail: mjwen@uh.edu. https://orcid.org/0000-0003-0013-575X
Jin Dai and Santosh Adhikari, Department of Chemical and Biomolecular Engineering, University of Houston, Houston, TX, 77204, USA, E-mail: jdai9@central.uh.edu (J. Dai). https://orcid.org/0009-0000-6614-6852 (J. Dai). https://orcid.org/0000-0003-0551-4919 (S. Adhikari)

coupled with the limited availability of high-quality training data, can lead to models that are prone to overfitting, generalization errors, and poor transferability (Abdar et al. 2021; Gawlikowski et al. 2023; Psaros et al. 2023; Wen et al. 2022). To address these challenges, it is imperative to quantify the uncertainty associated with model predictions, providing a measure of confidence in the results and helping to identify areas where the model may be unreliable. The uncertainties can be broadly categorized into two types: aleatoric uncertainty and epistemic uncertainty (Abdar et al. 2021; Hüllermeier and Waegeman 2021). Aleatoric uncertainty, also known as data uncertainty, arises from the inherent and irreducible noise in the data used for model development. In atomistic ML, aleatoric uncertainty can come from, for example, the inexact exchange-correlation functional of the density function theory (DFT) employed to generate the training data (Henkel and Mollenhauer 2021; Lejaeghere et al. 2016; Ruiz et al. 2005; Wellendorff et al. 2012). Epistemic uncertainty, also referred to as model uncertainty, accounts for limitations of our knowledge or assumptions about a model. Factors such as model architecture, model parameters, and hyperparameter selection can all contribute to epistemic uncertainty.

Incorporating uncertainty quantification (UQ) into the ML model development process is a significant step forward; transmitting the uncertainty from the atomistic ML models to downstream tasks, known as uncertainty propagation (UP), is equally crucial (Abdi et al. 2024; Honarmandi and Arróyave 2020; Honarmandi et al. 2019; Wang and Sheen 2015). Atomistic ML models are typically trained to predict fundamental physical and chemical properties, such as the forces on atoms and chemical reaction rates. These properties are then used as inputs to other modeling techniques (analytical or numerical), such as molecular dynamics (MD) and microkinetic simulations, to obtain a final quantity of interest (QoI). The accurate and efficient propagation of uncertainty through the model chain to the QoI is essential to make informed decisions and assess the reliability of the final predictions. Therefore, UQ and UP should be considered together when developing and applying atomistic ML models for chemical and materials applications.

Uncertainty analysis has been a crucial topic in chemical engineering and materials science. Classical methods such as polynomial chaos (Wiener 1938) have long been applied to study the uncertainty associated with the key parameters and outcomes in these domains (Phenix et al. 1998; Reagan et al. 2005; Villegas et al. 2012). While uncertainty analysis traditionally serves to acknowledge model imperfections and has been primarily used for post-hoc analysis, with the emergence of ML approaches, the landscape has become increasingly rich and diverse. For example, uncertainty has become routinely used in active learning to select new atomic structures to enrich the dataset and subsequently retrain ML models. This has been demonstrated in both structure–property models (Gubaev et al. 2018; Liu et al. 2022; Tian et al. 2021) and MLIPs (Schwalbe-Koda et al. 2021; van der Oord et al. 2023; Zaverkin et al. 2024), among others. Such proactive usage of uncertainty can greatly enhance the robustness of ML models and, additionally, it is data efficient.

This abundance of choices in ML-based uncertainty analysis, however, has created a significant challenge in decision-making for practitioners. Several critical questions arise: What are the fundamental similarities and differences between these UQ methods? What constitutes good UQ methods, and how do their strengths and weaknesses compare in the context of atomistic ML? Can the uncertainty obtained from an ML estimator be propagated to downstream chemical and materials applications, and if so, how? Without clear answers to these questions, navigating the field of UQ and UP in atomistic ML can be a daunting task. It often leads to confusion, such as misinterpretation of the meaning and implications of the uncertainty, and difficulty in selecting an appropriate method for a given task.

In this work, we provide a comprehensive review of selected, representative UQ and UP methods for atomistic ML, aiming to answer the above questions. We assume the readers have a basic understanding of ML, but no prior knowledge of UQ and UP is required. We anticipate that this review will equip readers with a certain degree of certainty in navigating the space of uncertainty. The paper is structured as follows. First, we present a primer on probabilistic modeling in Section 2, setting the stage for the coming sections. Next, in Section 3, we review and categorize the selected UQ methods, leveraging the concepts introduced in Section 2. In Section 4, we discuss ways to evaluate the quality of the UQ methods from four different perspectives: accuracy, precision, calibration, and efficiency. Additionally, in Section 5, we explore UP techniques in widely used chemical and materials simulation techniques, using MD and microkinetic modeling as examples. While this work focuses on uncertainty in ML models, data used in chemical and materials science inherently contains uncertainties in its values; we briefly discuss recent efforts to quantify and propagate data uncertainty in Section 5.4. Finally, we conclude by summarizing the current challenges of UQ and UP for atomistic ML and outlining opportunities for future research. A summary of the most commonly used symbols is listed in Table 1, and a list of the abbreviations is provided in Appendix A.

Table 1: Notation: overview of the most commonly used symbols.

Symbol	Explanation	Alternative
θ	Model parameters	
x	Model input	
$\hat{y}$	Model output	$\hat{y} = f(x;\theta)$
X	Input set	$X = \{x_i\}_{i=1}^{N}$
Y	Output set	$Y = \{y_i\}_{i=1}^{N}$
D	Dataset	$D = (X, Y)$
$p(\theta)$	Prior	
$p(Y\|X, \theta)$	Likelihood	$p(D\|\theta)$
$p(Y\|X)$	Marginal likelihood	$p(D)$
$p(\theta\|X, Y)$	Posterior	$p(\theta\|D)$
$\mathcal{N}(\mu, \sigma^2)$	Gaussian distribution	
δ	Uncertainty	
$\mathbb{E}_\theta[\cdot]$	Expectation w.r.t. θ	

2 Primer on probabilistic modeling

In many atomistic ML problems, the goal is to obtain a regression model, $y = f(x; \theta)$, which maps the input x (e.g., a chemical reaction) to the corresponding output y (e.g., the reaction rate), where θ denotes all model parameters and can be determined from an observed dataset $D = \{(x_i, y_i)\}_{i=1}^{N} = (X, Y)$ consisting of N data points. In addition, we are interested in quantifying the uncertainty in the predicted outputs y.

This section provides a brief review of the basic concepts in probabilistic modeling, laying the foundation for the discussion of uncertainty in subsequent sections. We first introduce *Bayesian inference* to obtain the predictive distribution which expresses the uncertainty about the prediction y for each input x. Next, we discuss the frequentist *maximum likelihood estimation* of model parameters θ and its link to the widely-used least-squares minimization. Finally, we introduce *evidence approximation*, a framework integrating frequentist estimates into the Bayesian approach to find approximate solutions. Readers well-versed in probability theory may choose to skip this section and refer back to it as needed while reading the later sections.

2.1 Distribution functions

A *probability density function* (PDF), denoted as $p(\theta)$, is used to describe the probability distribution of a continuous random variable Θ. The PDF represents the relative likelihood of the random variable taking on a specific value. A *cumulative distribution function* (CDF), denoted as $F(\theta)$, describes the probability that a random variable Θ takes a value less than or equal to θ. The CDF can be obtained from the PDF via: $F(\theta) = P(\Theta \le \theta) = \int_{-\infty}^{\theta} p(t)\mathrm{d}t$. A *quantile function* (QF) maps probabilities to values of the random variable. It is defined as the inverse of the CDF, $Q(p) = F^{-1}(p)$, that is, for a given probability p, the quantile function returns the value θ such that its probability is less than or equal to an input probability value, i.e., $P(\Theta \le \theta) = p$. As a concrete example, Figure 1 presents the PDF, CDF, and QF of a one-dimensional standard Gaussian distribution $\mathcal{N}(0,1)$ with a mean of 0 and standard deviation of 1.

2.2 Bayesian inference

In the Bayesian view, probability provides a quantification of uncertainty (Gelman et al. 2013). Given an observed dataset D, we are interested in obtaining the conditional probability $p(y^*|x^*, D)$ of the output y^* for a new input x^*. From it, a statistical measure of the uncertainty in y^* such as the variance can then be obtained. Given a parametric model, $y = f(x; \theta)$, this can be achieved in two steps.

First, obtaining the *posterior distribution* over model parameters θ using Bayes' theorem (Gelman et al. 2013):

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}. \tag{1}$$

The *prior distribution* $p(\theta)$ represents our prior information as to which parameters θ are likely to have generated the outputs before observing any data, based on previous knowledge, experience, or physical limitations. The effects of the observed data D come from the *likelihood function*, $p(D|\theta)$, which quantifies the plausibility of D for different realizations of θ. The denominator,

$$p(D) = \int p(D|\theta)p(\theta)\,\mathrm{d}\theta, \tag{2}$$

is called the *marginal likelihood*, also known as the *evidence* in the context of Bayesian statistics, which ensures that the posterior is a proper probability and thus integrates into one. It represents the likelihood of the observed data D, considering all possible values of the parameter θ weighted by their prior distribution. Marginalization means evaluating this equation to obtain the marginal likelihood. Bayes' theorem converts the prior probability over model parameters into the posterior probability by incorporating the evidence provided by the observed data.

Second, with the posterior over θ, we can obtain the *predictive distribution* for a new data point (x^*, y^*) as

$$p(y^*|x^*, D) = \int p(y^*|x^*, \theta)p(\theta|D)\,\mathrm{d}\theta, \tag{3}$$

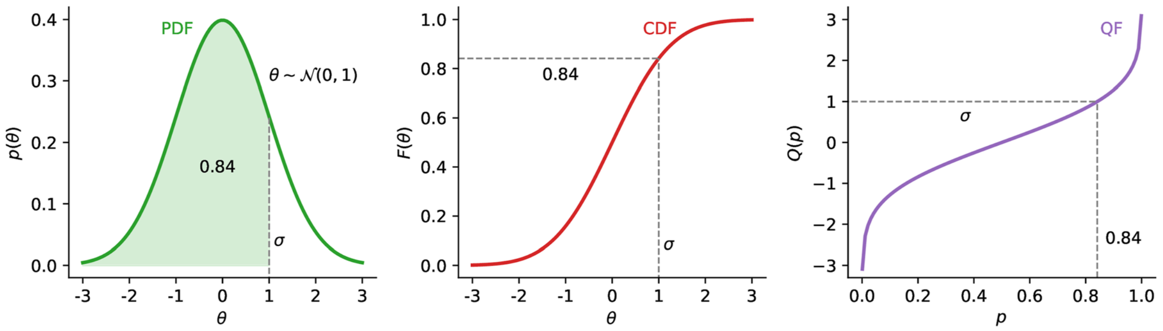

Figure 1: One-dimensional Gaussian distribution with a mean of $\mu = 0$ and standard deviation $\sigma = 1$. The probability density function (PDF), cumulative probability function (CDF), and quantile function (QF). The QF is the inverse of the CDF, as can be seen by switching the horizontal and vertical axes.

where $p(y^*|x^*, \theta)$ is the likelihood given the new data point. It is related to the likelihood function for a dataset $p(D|\theta)$ in Eq. (1), and further discussion on this will be provided in Section 2.3. From the predictive distribution, we can readily obtain, e.g., the mean as the final prediction and the variance as a point estimate of the uncertainty.

Although theoretically sound, a major practical limiting factor of the full Bayesian approach lies in the complexity of evaluating the posterior. In particular, the marginal likelihood in Eq. (2) can only be analytically evaluated for simple models like linear regression. Numerical techniques such as sampling methods and variational inference have to be undertaken to evaluate the predictive distribution for more complicated models (Bishop 2006).

One can sample the posterior distribution using Monte Carlo (MC) methods, such as Markov chain Monte Carlo (MCMC), and then obtain the predictive mean and uncertainty (Neal 1993, 2003). Sampling methods are accurate, flexible, and can be applied to a wide range of models. However, they are still computationally intensive because a large number of samples might be needed for convergence; thus, they are mainly used for small-scale problems (Bishop 2006).

Alternatively, the variational inference approach tackles the challenge by employing another distribution $q(\theta)$ to approximate the true posterior $p(\theta|D)$, and then using $q(\theta)$ to evaluate the predictive distribution in Eq. (3). The approximate distribution $q(\theta)$ is typically much simpler, and it is optimized to resemble the true posterior, e.g., by minimizing the Kullback–Leibler divergence (Kullback and Leibler 1951; MacKay 1992b) between $q(\theta)$ and $p(\theta|D)$. Although not exact, variational inference offers a computationally efficient approach to evaluate the predictive distribution. The MC dropout method to obtain uncertainty in NN models proposed by Gal (2016) adopts this approach, and it will be further discussed in Section 3.1.1.

2.3 Maximum likelihood

Maximum likelihood estimation (MLE) is a frequentist approach to estimate the optimal parameters θ of a model $y = f(x;\ \theta)$. It is equivalent to the least-squares parameter optimization technique widely used in science and engineering. MLE provides a point estimate of the parameters and thus predictive uncertainty cannot be directly quantified from a model trained using MLE alone. However, MLE serves as a fundamental concept in parameter estimation and forms the basis of many other UQ methods.

In MLE, we focus on the likelihood function $p(D|\theta)$, and do not care about the prior and posterior (see Eq. (1)). We assume that the observed output y is given by the model prediction $f(x;\ \theta)$ with an additive error ϵ, i.e., $y = f(x;\ \theta) + \epsilon$. A Gaussian distribution with zero mean is a reasonable choice for the error, $p(\epsilon) = \mathcal{N}(\epsilon|0, \sigma^2)$, where σ^2 is the variance. This is equivalent to the Gaussian distribution in which y is regarded as the random variable with the model prediction $\hat{y} = f(x;\theta)$ as its mean and σ^2 as the variance (Bishop 2006):

$$p(y|x,\theta) = \mathcal{N}(y|\hat{y},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right). \tag{4}$$

In other words, this is the likelihood of θ for a single data point (x, y). Now consider the observed dataset D where each data point is drawn independently from the distribution in Eq. (4). We can obtain the likelihood function for the dataset as the product of the likelihood for each data point:

$$p(D|\theta) = p(Y|X,\theta) = \prod_{i=1}^{N} p(y_i|x_i,\theta). \tag{5}$$

The optimal model parameters can thus be obtained by maximizing Eq. (5) with respect to (w.r.t.) θ, which is equivalent to minimizing the negative log-likelihood (NLL):

$$\mathrm{NLL} = -\log p(D|\theta) = \frac{1}{2}\sum_{i=1}^{N}\left(\log(2\pi\sigma^2) + \frac{(y_i-\hat{y}_i)^2}{\sigma^2}\right), \tag{6}$$

because the logarithm function is monotonically increasing. This transformation converts a product of probabilities into a sum of log probabilities, which is often more convenient to optimize.

We note that in MLE, the variance σ^2 is modeled as a single constant, albeit unknown. Consequently, minimizing the NLL is equivalent to the familiar least-squares minimization w.r.t. θ using the loss:

$$L(\theta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{7}$$

Recall that the model parameters θ are implicitly indicated in the model prediction $\hat{y} = f(x;\theta)$. With the optimal point estimate of the parameters, θ_{opt}, we can then get model prediction as $y^* = f(x^*; \theta_{\text{opt}})$ for any new input x^*. Again, we note that no information on uncertainty can be directly obtained from this approach only.

2.4 Evidence approximation

While the full Bayesian approach can produce predictive uncertainty, it can be computationally demanding to obtain. On the other hand, it is straightforward to get a point estimate of the optimal model parameters using MLE, but the uncertainty cannot be quantified. The *evidence approximation* approach (Gull 1989; MacKay 1992a), also known as the *empirical Bayes* (Bernardo and Smith 2009), *generalized maximum likelihood* (Wahba 1985), or *type 2 maximum likelihood* (Berger 1985), lies between the two extremes.

Evidence approximation is a method that looks at how well a model fits the data overall. It focuses on the marginal likelihood in Eq. (2), which provides the *model evidence* of observing the data marginalized over the parameters, meaning that it calculates the probability of seeing the observed data under all possible parameter values of the model. In evidence approximation, the prior $p(\theta)$ in Eq. (2) is further parameterized using a set of hyperparameters ξ, becoming $p(\theta|\xi)$. Consequently, the marginal likelihood is (Bishop 2006):

$$p(D|\xi) = \int p(D|\theta,\xi)p(\theta|\xi)\,\mathrm{d}\theta. \tag{8}$$

The introduction of the hyperparameters ξ increases the model's capacity, meaning that the model has increased flexibility to capture the underlying structure present in the data. In practice, the prior distribution $p(\theta|\xi)$ is often chosen to be conjugate to the likelihood $p(D|\theta, \xi)$. In other words, $p(\theta|\xi)$ is specifically selected to match the form of the likelihood $p(D|\theta, \xi)$ such that the integration in Eq. (8) has a closed-form solution, thereby simplifying the process to obtain $p(D|\xi)$. For example, if the likelihood is the Binomial distribution, a conjugate prior for it is the Beta distribution, and then Eq. (8) can be analytically evaluated to obtain the model evidence, which is a Beta distribution as well (DeGroot and Schervish 2012).

In a full Bayesian setting, after obtaining $p(D|\xi)$, one would also marginalize over the hyperparameters ξ to obtain $p(D)$ and then perform Bayesian inference. However, this marginalization typically does not have an analytical solution. Instead, in evidence approximation, we get a point estimate of ξ by maximizing the model evidence $p(D|\xi)$ w.r.t. ξ. Then, the model evidence, prior, and likelihood are all evaluated using the optimal hyperparameters ξ_{opt} to perform Bayesian inference using Eqs. (1) and (3).

Alternatively, one can obtain the predictive uncertainty directly from $p(\theta|\xi_{\text{opt}})$, provided that the prior is chosen to be of a specific form, e.g., as a high-order distribution on top of the likelihood. This is the *evidential regression* approach recently proposed by Amini et al. (2020). In Section 3.1.3, we will further discuss this approach and explain how to get the uncertainty.

3 Uncertainty quantification

A large number of UQ methods have been developed for ML models. Here, we discuss several selected, representative ones for atomistic ML for chemical and materials applications; in particular, we focus on UQ methods for NN models. We classify them into three categories (Figure 2) mainly based on the model construction strategy: probabilistic approach, ensemble approach, and feature space distance approach. A probabilistic approach models some distribution discussed in Section 2 and derives uncertainty from it. An ensemble approach builds multiple models and obtains the variance in the predictions as the uncertainty. A feature space distance approach measures some "distance" of a data point to the model training data and regards the distance as the uncertainty. Although each method is placed under a single category in Figure 2, some can belong to different categories. For example, MC dropout can also be regarded as an ensemble approach. We note that there are other ways to categorize the UQ methods, such as the one based on model utilization strategy (Gawlikowski et al. 2023).

In this section, we discuss the UQ methods, examining how a model is trained and how uncertainty is obtained. In addition, we provide example applications for chemical and materials' problems. A summary of the UQ methods is provided in Table 2.

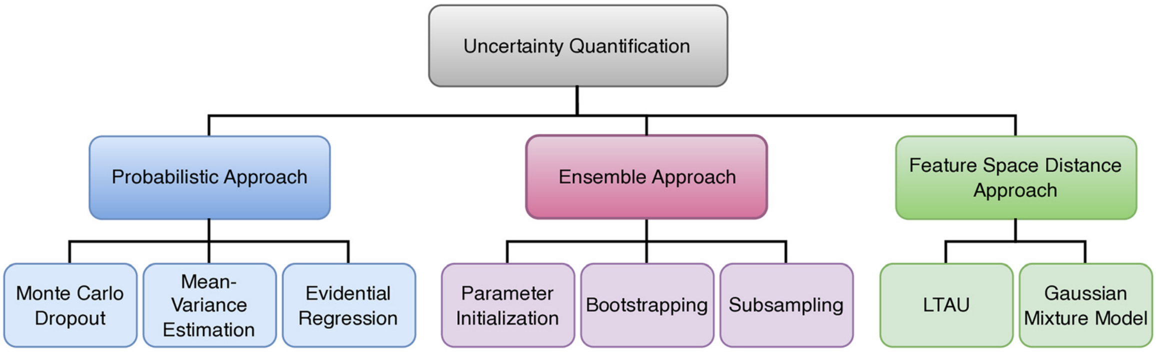

Figure 2: Categorization of uncertainty quantification methods in atomistic machine learning. LTAU: loss trajectory analysis for uncertainty.

Table 2: Summary of the UQ methods. Category denotes the class to which a method belongs. For the probabilistic methods, we provide the specific type of probabilistic approach the method belongs to. UQ measure denotes the quantity being used as the uncertainty. Efficiency means the number of models that need to be optimized in the training stage and the number of model evaluations that need to be performed to obtain the uncertainty in the inference stage. A method with a check mark indicates that it can be used to conduct sampling-based UP.

	Category	UQ measure	Efficiency (training/inference)	Sampling-based UP
Monte Carlo dropout[a]	Bayesian	Variance	One/Multiple	✓
Mean-variance estimation[b]	Maximum likelihood	Variance	One/One	
Evidential regression[c]	Evidence approximation	Variance	One/One	
Parameter initialization[d]	Ensemble	Variance	Multiple/Multiple	✓
Bootstrapping[e]	Ensemble	Variance	Multiple/Multiple	✓
Subsampling[e]	Ensemble	Variance	Multiple/Multiple	✓
LTAU[f]	Feature space distance	Error ratio	One/One	
Gaussian mixture model[g]	Feature space distance	NLL	Two/One	

[a]Gal and Ghahramani (2016), [b]Nix and Weigend (1994), [c]Amini et al. (2020), [d]Lakshminarayanan et al. (2017), [e]Hastie et al. (2009), [f]Vita et al. (2024), [g]Bishop (2006). LTAU: loss trajectory analysis for uncertainty; NLL: negative log-likelihood.

3.1 Probabilistic approach

3.1.1 Monte Carlo dropout

The dropout technique was originally proposed by Srivastava et al. (2014) as a regularization technique to alleviate overfitting in deep NN models. It was adapted by Gal and Ghahramani (2016) to approximate the Bayesian approach for UQ mentioned in Section 2.2. Their method, known as MC dropout, can be theoretically viewed as sampling from a Bernoulli prior distribution over the weights of the NN, and then taking advantage of the variational inference technique to approximate the posterior distribution. Practically, dropout is used at both training and inference time, allowing the model to estimate uncertainty by considering multiple predictions from different subsets of the network.

The model can be trained by minimizing a loss between its predictions and the corresponding reference values to obtain the optimal NN parameters. At each training step, dropout randomly sets the outputs of a fraction of the nodes in an NN to zero (e.g., the dashed nodes in Figure 3a), effectively creating an ensemble of thinned sub-networks. Once trained, we use the model in a similar way. Multiple forward passes are performed through the NN, each with different nodes being dropped, to obtain multiple predictions. The predictions are then averaged to obtain the final prediction and their variance is computed as the predictive uncertainty. In this sense, MC dropout can also be thought of as a frequentist ensemble approach to be discussed in Section 3.2.

It is important to note that MC dropout is an approximation and may not capture the full posterior distribution over the NN's weights. Nevertheless, it has gained popularity as a practical technique in atomistic ML. For example, Wen and Tadmor (2020) have developed a dropout NN model for carbon allotropes and shown that the obtained uncertainty can reliably distinguish diamonds from graphene and graphite.

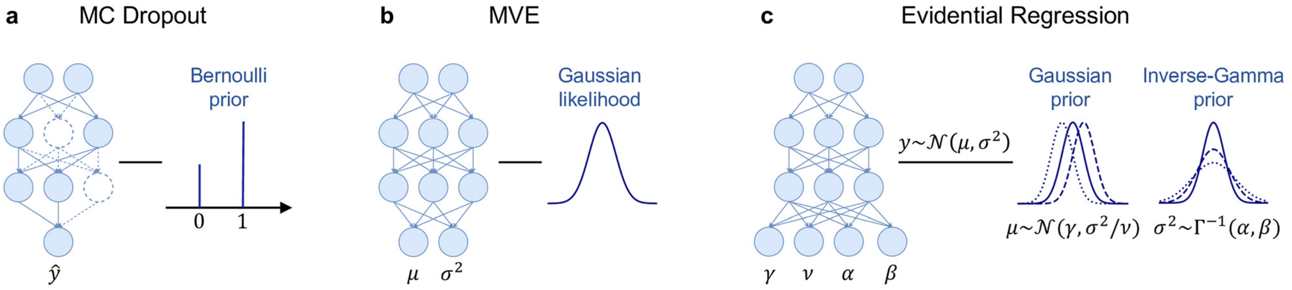
a
MC Dropout
Bernoulli prior
0
1
$\hat{y}$
b
MVE
Gaussian likelihood
μ σ^2
c
Evidential Regression
$y \sim \mathcal{N}(\mu, \sigma^2)$
Gaussian prior
Inverse-Gamma prior
$\mu \sim \mathcal{N}(\gamma, \sigma^2/\nu)$
$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$
γ ν α β

Figure 3: Schematic illustration of the probabilistic UQ approaches: (a) MC dropout; (b) MVE; and (c) evidential regression.

3.1.2 Mean-variance estimation

The mean-variance estimation (MVE) method, first introduced by Nix and Weigend (1994), enables the use of a single deterministic NN to obtain the predictive uncertainty. This method largely follows the MLE framework (Section 2.3) but with slight adjustments. In MLE, the observed data are assumed to be independent and identically distributed (i.i.d.) samples from a Gaussian distribution, where the mean μ is given by a parameterized model $\hat{y} = f(x; \theta)$ of the input x, while the same constant variance σ^2 is assumed for all observed data (see Eq. (4)). In contrast, MVE uses different variances for different data points to model the uncertainty. In other words, the observed data are assumed to be drawn from the Gaussian: $\mathcal{N}(\mu(x), \sigma^2(x))$, in which both the mean μ and the variance σ^2 are parameterized models of the input x. In practice, an NN with two output nodes can be employed as the parameterized model, one node predicting the mean μ and the other for the variance σ^2 (Figure 3b).

The training process involves using MLE to optimize the NN's parameters. In this case, since the variance is not a constant, MLE is not equivalent to least-squares minimization with the loss in Eq. (7) anymore. Instead, we will need to directly minimize the NLL in Eq. (6). Once trained, the predicted mean μ by the NN gives the final prediction, and the predicted variance σ^2 serves as the uncertainty.

MVE has been adopted by Tan et al. (2023) to predict the energies of small molecules, among others. They found that it has the highest average test error when compared to other methods and suggested that this might be attributed to the harder-to-optimize NLL loss function, which has been reported in Seitzer et al. (2022).

3.1.3 Evidential regression

Amini et al. (2020) introduced deep evidential regression as a method for UQ in NNs, building on the evidence approximation approach discussed in Section 2.4. Here, we summarize and explain the key formulation of this method within the probabilistic framework, and the detailed derivation provides essential insights into its theoretical foundations. The practical formula for uncertainty estimation is given in Eq. (15).

The model evidence for an entire dataset is given in Eq. (8). To simplify the discussion, here we focus on the model evidence for a single observation (x, y) (Amini et al. 2020),

$$p(y|\xi) = \int p(y|\theta, \xi) p(\theta|\xi)\, \mathrm{d}\theta, \tag{9}$$

where we omit the conditional dependence on x for simplicity. Similar to MLE, the data is assumed to be sampled from a Gaussian likelihood:

$$p(y|\theta, \xi) = \mathcal{N}(\mu, \sigma^2), \tag{10}$$

where $\theta = (\mu, \sigma^2)$, denoting the mean and variance of the Gaussian. However, unlike in MLE, where the mean and variance are fixed constants, here the mean and variance are parameterized over ξ. Amini et al. (2020) proposed to parameterize μ and σ^2 using the Gaussian and Inverse-Gamma distributions, respectively:

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \nu^{-1}) \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \tag{11}$$

where $\xi = (\gamma, \nu, \alpha, \beta)$ denotes the four hyperparameters. Further, it is assumed that the prior $p(\theta|\xi)$ can be factorized into the product of the distributions of μ and σ^2, then it can be written as,

$$p(\theta|\xi) = \mathcal{N}(\gamma, \sigma^2 \nu^{-1}) \cdot \Gamma^{-1}(\alpha, \beta), \tag{12}$$

which is called the *Normal Inverse-Gamma* (NIG) distribution, a high-order evidential distribution. Sampling from NIG yields instances of lower-order likelihood functions from which the data is drawn.

The NIG prior in Eq. (12) is the conjugate distribution to the Gaussian likelihood in Eq. (10); therefore, the model evidence in Eq. (9) can be evaluated analytically, resulting in the Student-t distribution:

$$p(y|\xi) = \mathrm{St}\left(y\,;\gamma, \frac{\beta(1+\nu)}{\nu\alpha}, 2\alpha\right). \tag{13}$$

Recall from Section 2.4 that the optimal hyperparameters $\xi = (\gamma, \nu, \alpha, \beta)$ are obtained by maximizing the model evidence, namely Eq. (13). In the deep evidential regression method by Amini et al. (2020), the hyperparameters are further parameterized by an NN and obtained as the output of the NN with four output nodes, one for each hyperparameter (Figure 3c). So, instead of ξ, the parameters in the NN are optimized. In practice, we do not maximize Eq. (13) but minimize the equivalent NLL of Eq. (13) for numerical stability. In addition, extra regularization terms can be added to remove misleading evidence. We refer to Amini et al. (2020) for the technical details of model training.

Once trained, the final prediction can be computed from the NIG as (Amini et al. 2020)

$$\mathbb{E}[\mu] = \gamma, \tag{14}$$

and the uncertainty as

$$\mathbb{E}[\sigma^2] + \mathrm{Var}[\mu] = \frac{\beta}{\alpha - 1} + \frac{\beta}{\nu(\alpha - 1)}. \tag{15}$$

Deep evidential regression has been used by Soleimany et al. (2021) to predict molecular properties, and the obtained uncertainty has been successfully used to achieve sample-efficient training of property estimator and to guide the virtual screening for antibiotic discovery. Gruich et al. (2023) have also demonstrated its effectiveness in heterogeneous catalysis applications.

3.2 Ensemble approach

The ensemble approach, characterized by its simplicity and diverse construction methods, combines multiple models to create a more robust predictive model, surpassing individual model limitations (Zhou 2012). The frequentist ensemble approach is easy to implement and can be applied to a large number of regression algorithms. It might be computationally expensive when compared to other UQ methods, but they can be naively paralleled. A couple of methods exist to construct an ensemble, such as parameter initialization, bootstrapping, and subsampling (Figure 4).

The first type of method involves fitting models with different parameter initializations to the same dataset. Lakshminarayanan et al. (2017) proposed the use of ensembles for estimating the uncertainty of NNs. NNs of the same structure are created, but their parameters are initialized to be different. Each member of the ensemble is trained on the entire training set, meaning that all members have access to the same data.

The second type of method focuses on fitting the same model to different datasets. Bootstrapping is a widely used method to generate multiple derived datasets from a given dataset. Each subset, called a bootstrap sample, is created by randomly sampling the same number of data points from the original dataset with replacement (Hastie et al. 2009). This means that each data point has an equal probability of being selected, and the same data point can be included multiple times in the bootstrap sample. This process is repeated M times, resulting in M bootstrap samples. Then, M models are trained separately, each using one of the bootstrap samples.

Subsampling (Politis and Romano 1994; Politis et al. 1999) is an alternative to bootstrapping to create multiple derived datasets from a given dataset. It is similar to bootstrapping but with a key difference: subsampling is performed without replacement and thus each data point can only appear at most once in each subset. As a result, each sample consists of fewer data points than the original dataset.

The final ensemble prediction and uncertainty are obtained by combining the outputs $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_M$ of all members. The mean,

$$\bar{y} = \frac{1}{M}\sum_{i=1}^{M}\hat{y}_i, \tag{16}$$

gives the final prediction, and the variance,

$$\sigma^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\hat{y}_i - \bar{y})^2, \tag{17}$$

gives the predictive uncertainty.

3.3 Feature space distance approach

Another category of UQ approach is based on some distance measure in the model's feature/latent space. It assumes that data points resembling each other are positioned closer to one another in the feature space. Therefore, for a given test data point, if it is close to the training data in the feature space, the predictive uncertainty is low; otherwise, the predictive uncertainty is high.

In this approach, the uncertainty is typically obtained in a two-step process. First, given a primary model like an NN, a point estimate of the optimal model parameters is obtained using a training technique such as MLE. Second, construct another model to measure the distance between a test data

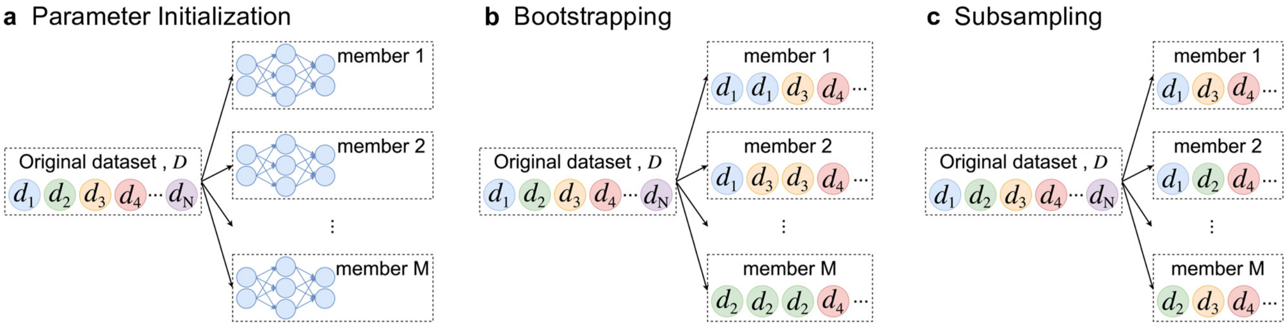

Figure 4: Schematic illustration of the ensemble UQ approaches: (a) parameter initialization, (b) bootstrapping, and (c) subsampling.

point and the training data in the primary model's feature space. When dealing with NNs, the feature space can be chosen as the last but one layer or other internal layers that are considered suitable.

3.3.1 LTAU

Loss trajectory analysis for uncertainty (LTAU) measures the distance in the Euclidean space. It begins by training an NN model to predict an atomic property y, chosen to be the forces on atoms in Vita et al. (2024). During training, besides optimizing the model parameters, the error $\epsilon_i = \|y_i - \hat{y}_i\|^2$ between the model prediction $\hat{y}_i$ and its corresponding reference y_i for each atom i is recorded as

$$T_i = \{\epsilon_i^1, \epsilon_i^2, \ldots, \epsilon_i^E\}, \tag{18}$$

where the super index denotes the training epoch at which the error is logged, and E is the total number of epochs to train the model.

After training, we get a set of errors T_i along the loss trajectory for each data point i and then convert the errors to the model's confidence score for that data point via

$$p_i = P(\epsilon_i^e \in T_i \leq atol), \tag{19}$$

which means the ratio of data points in T_i whose value is smaller than or equal to a tolerance *atol*. A reasonable choice for *atol* would be the mean absolute error (MAE) between the predictions from the trained primary model and their references. The value of p_i is in the range of [0, 1]. The uncertainty of each data point is calculated as

$$\delta_i = 1 - p_i, \tag{20}$$

which can be interpreted as the probability that the model's prediction on data point i will have an error larger than the MAE.

The uncertainty for a new data point, j, is obtained by averaging the uncertainties of its nearest K neighbors in the training data (Figure 5a):

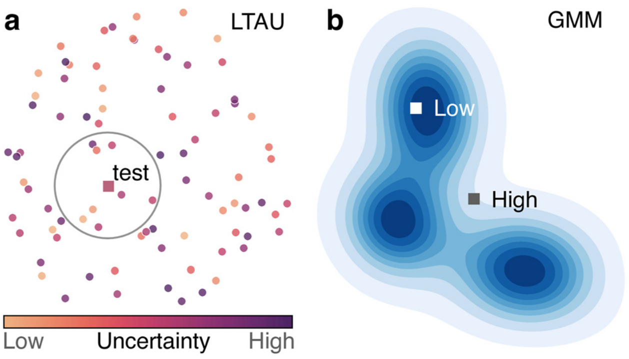

Figure 5: Schematic illustration of the UQ approaches based on feature space distance. (a) LTAU assigns the average uncertainty of neighboring training data (circles) to a test point (square). (b) GMM models the density of the training data in the feature space, and a test point has low uncertainty if it is located in a dense region.

$$\delta_j = \frac{1}{K} \sum_{i \in N_j} \delta_i, \tag{21}$$

where N_j denotes the set of neighbors. The nearest neighbors can be determined by performing a similarity search based on Euclidean distance in the feature space.

LTAU has been successfully applied to tune the training–validation gap in NN potentials for carbon materials and predict the errors in relaxation trajectories of catalysts (Vita et al. 2024).

3.3.2 Gaussian mixture model

An alternative to Euclidean distance is measuring the density. If a test point is in the dense region where the training data are located in the feature space, then this test point has low uncertainty (Figure 5b). The density can be estimated by a Gaussian mixture model (GMM).

After training the primary model such as an NN, we get a set of feature vectors $H = \{h_1, h_2, \ldots, h_N\}$, each representing a training data point in the feature space. The set of feature vectors is then used to train the second GMM model. We aim to capture the underlying structure of the feature vectors using a GMM (Bishop 2006):

$$p\left(h_i|w,\mu,\Sigma\right) = \sum_{m=1}^{M} w_m \mathcal{N}\left(h_i|\mu_m,\Sigma_m\right), \tag{22}$$

which is linear combination of M Gaussian functions of respective mean μ_m and covariance Σ_m, with weights w_m. Each Gaussian $\mathcal{N}$ is a multidimensional distribution in the feature space, and thus μ_m is a vector and Σ_m is a matrix. We assume a dataset consists of i.i.d. samples from this GMM likelihood. Then, the NLL for the dataset can be written as

$$\mathrm{NLL}\left(H|w,\mu,\Sigma\right) = -\sum_{i=1}^{N}\log\left(\sum_{m=1}^{M} w_m \mathcal{N}\left(h_i|\mu_m,\Sigma_m\right)\right), \tag{23}$$

which can be derived in the same way as from Eq. (4) to Eq. (6) for MLE. To train the GMM, we minimize the NLL w.r.t. μ, Σ, and w, meaning that we adjust the location and shape of the Gaussian distributions, as well as the weight of each member such that the GMM model best describes the density of the training data in the features space. The optimization can be performed via a gradient-based technique, or, more typically, using the expectation-maximization algorithm (Hastie et al. 2009).

Once the GMM model is trained, the uncertainty for a new data point x^* can be obtained in two steps. First, obtain its feature vector h^* using the primary model. Second, compute its NLL,

$$\mathrm{NLL}\left(h^*|w,\mu,\Sigma\right) = -\log\left(\sum_{m=1}^{M} w_m \mathcal{N}\left(h^*|\mu_m,\Sigma_m\right)\right), \tag{24}$$

and the NLL can be regarded as the predictive uncertainty (Zhu et al. 2023).

The GMM method has been employed by Zhu et al. (2023) to build NN interatomic potentials, leveraging the uncertainty estimates for active learning and efficient training data selection.

Overall, feature space distance approaches, such as LTAU and GMM, measure how far a data point is from the distribution of the training data, essentially indicating whether the point is out of distribution. The predicted uncertainty, however, does not necessarily scale with the prediction error; therefore, recalibration of the uncertainty is typically needed if one intends to use the uncertainty as a proxy of the prediction error. Demonstration of the quality of the approach is provided in Section 4.3, and further discussion of the recalibration is given in Section 4.4.

4 Performance evaluation

What makes a UQ method effective? Uncertainty is solely a property of model predictions, providing information about their *precision* – the degree to which the predicted values are concentrated around each other (Figure 6). However, uncertainty does not directly measure the *accuracy* of the predictions, i.e., how close they are to the true observations. Despite this, a key application of uncertainty is to use it as an indicator of the likely accuracy of the predictions. Ideally, a prediction with high uncertainty should indicate a large error and thus be less reliable. The degree to which the uncertainty aligns with the accuracy is called *calibration*.

An effective UQ method should be accurate, precise, and well-calibrated; in addition, it should be computationally efficient for practical usage. These four aspects evaluate UQ methods from different perspectives. In this section, we discuss performance evaluation for UQ methods, focusing on uncertainty calibration, a concept that, we believe, may be less familiar to researchers working on atomistic ML. We also examine scoring metrics for UQ evaluation, comment on the pros and cons of existing UQ methods, and provide concrete examples by drawing insights from existing benchmark studies. Furthermore, we introduce recalibration techniques to improve UQ performance.

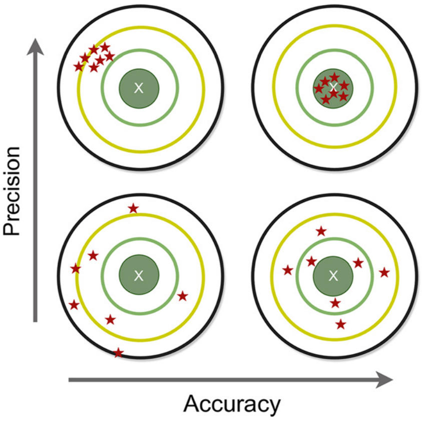

Figure 6: Uncertainty and prediction error. Loosely speaking, uncertainty gives the precision of the predictions, meaning how tightly the predictions are distributed against each other, while prediction error measures the accuracy of the predictions, meaning the distance between the prediction and the true value.

4.1 Calibration

Calibration measures the statistical consistency between the predictions and observations, a property that depends on both the predictions and observations (Gneiting et al. 2007). Uncertainty calibration has been extensively studied in the context of classification problems. Perfect calibration in this setting means that the confidence assigned to a class equals the probability of the prediction belonging to that class (Guo et al. 2017; Scalia et al. 2020). For instance, if we have 10 predictions, each with a confidence of 0.8, we expect 8 of them to be correctly classified.

Uncertainty calibration for regression is less intuitive because the model predicts continuous values, rather than discrete labels as in classification. Nevertheless, following the groundbreaking work by Gneiting et al. (2007), methods that extend the uncertainty calibration approach for classification have been proposed for regression problems (Kuleshov et al. 2018; Levi et al. 2022) and adopted in atomistic ML. Here, we discuss two such methods: *interval based* and *error based* regression uncertainty calibration. As a technical note, we will use δ to denote uncertainty in general. However, for some models, uncertainty is represented by the variance σ^2 (see Table 2). Therefore, these two notations will be used interchangeably when appropriate.

4.1.1 Interval based approach

Loosely speaking, in a regression setting, calibration means that a model prediction should fall in a given confidence interval γ% approximately γ% of the time (Kuleshov et al. 2018). For example, the model in the top panel of Figure 7 is not calibrated because only 20 % (2 out of 10) of the time the predictions are within the 90 % confidence interval, while the one in the bottom is calibrated. Formally, according to Kuleshov et al. (2018), for a given calibration dataset $D_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^N$, a regression model is calibrated if

$$\frac{\sum_{i=1}^N \mathbb{I}\left[y_i \le F_i^{-1}(p)\right]}{N} \to p \qquad \text{for all } p \in [0,1], \tag{25}$$

as $N \to \infty$. Here, $F_i = P(Y < y_i)$ denotes the CDF of the random variable Y, and F_i^{-1} is the corresponding quantile function (see Section 2.1). $\mathbb{I}[c]$ is the indicator function, which evaluates to 1 if the condition c is true and 0 otherwise. In other words, Eq. (25) means that, for a calibrated model, the empirically observed CDF from the data and the expected CDF by the model should match as the dataset size goes to infinity.

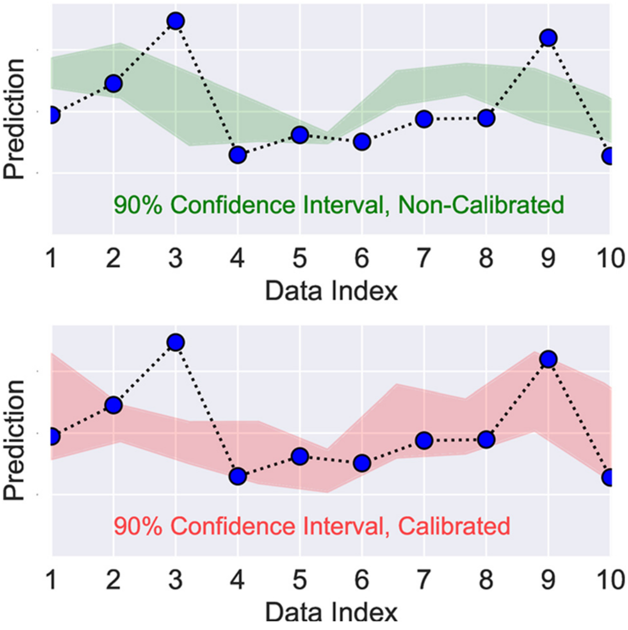

Figure 7: Illustration of a non-calibrated regression model (top) and a calibrated one (bottom). The plot is inspired by Kuleshov et al. (2018) but generated using arbitrary data.

In practice, Eq. (25) is evaluated for a selected number of p values, and the *calibration curve* is used to check the calibration level, which plots the observed CDF from the data versus the expected CDF by the model (Figure 8a). With this, the calibration curve can be obtained as follows (Kuleshov et al. 2018):

(1) Discretize the expected CDF to a set of M values $0 < p_1 < p_2 \ldots < p_M < 1$. The data is assumed to be generated from a Gaussian, $y \sim \mathcal{N}(\hat{y}, \sigma^2)$, where $\hat{y}$ and σ^2 are the predictions and the associated uncertainty, respectively (Kuleshov et al. 2018; Tran et al. 2020). For example, if an ensemble method is used, then $\hat{y}$ and σ^2 are the ensemble mean and variance, respectively. For a Gaussian with mean $\hat{y}$ and variance σ^2, the expected CDF can be readily obtained (see Section 2.1). We note that assuming a Gaussian distribution may not accurately reflect the true nature of the data in all cases.

(2) For each expected CDF p_j, compute the corresponding expected model output y_j. As mentioned above, the model output is assumed to follow a Gaussian; therefore, y_j can be readily computed using the quantile function, $y_j = Q(p_j) = F^{-1}(p_j)$, discussed in Section 2.1.

(3) For each expected CDF p_j, compute the corresponding observed CDF $\tilde{p}_j$. With y_j obtained in the previous step, $\tilde{p}_j$ is obtained as the empirical frequency $\tilde{p}_j = |\{y_i | y_i \le y_j, i = 1, 2, \ldots N\}| / N$ (left of Eq. (25)), where $|\cdot|$ denotes the size of a set, and N is the total number of data points in the calibration dataset D_{cal}. In other words, $\tilde{p}_j$ is computed as

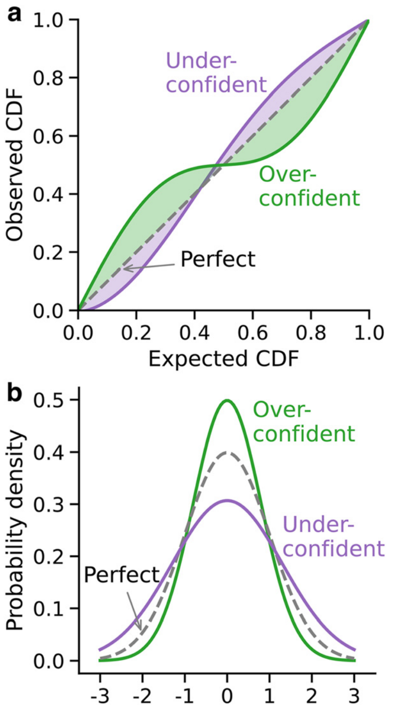

Figure 8: Calibration curves and probability density for the interval approach. (a) Calibration curves for perfectly calibrated (diagonal grey), over-confident (green) and under-confident (purple) models. (b) The Gaussian probability densities correspond to the calibration curves in (a).

the fraction of data points whose prediction y_i is smaller than or equal to y_j.

(4) Create the calibration curve by plotting $(p_j, \tilde{p}_j)$ pairs for $j = 1, 2, \ldots M$.

The calibration curve provides rich information. First, for a perfectly calibrated model as defined in Eq. (25), the calibration curve should be a diagonal line, meaning that the observed CDF from the data and the expected CDF by the model match with each other. Therefore, a model's calibration could be qualified by the closeness of its calibration curve to the diagonal line. In addition, the shape of the calibration curve could yield other insights into the predictive uncertainty of a model. A calibration curve that is above the diagonal line at low expected CDF but below at high expected CDF suggests that the model is over-confident. To understand this, let's focus on the low expected CDF region, e.g., at 0.2 in Figure 8a. Here, the expected CDF is smaller than the observed CDF, meaning that the variance in the Gaussian distribution used to construct the model is smaller than the variance in the observed data (Figure 8b). With a smaller expected variance (uncertainty), the model is over-confident. On the other hand, an under-confident model has a calibration curve that is below the diagonal line at a small expected CDF but above a large expected CDF.

In Kuleshov et al. (2018) and Levi et al. (2022), the observed CDFs are interpreted as observed confidence intervals. Because of the use of CDF, the interval here means $(-\infty, q_j]$. This is different from the commonly used notion of confidence interval, which is typically specified as an interval around the mean. For example, the 68 % confidence interval of a Gaussian distribution is $\mu \pm \sigma$. Thus, to avoid confusion, we do not use "confidence interval" but directly use CDF as is also done in Tran et al. (2020). Nevertheless, it is possible to interpret the calibration curve as the commonly used confidence interval. Instead of CDF, one can employ the probability density function (PDF), considering symmetric intervals $\mu \pm \gamma_j$ of varying confidence level $0 < \gamma_1 < \gamma_2 \ldots < \gamma_M < 1$ around the mean and examining the empirical frequency of the observed data belonging to each interval (Scalia et al. 2020).

4.1.2 Error based approach

The error based approach directly compares the predicted uncertainty σ^2 and the expected square error between the model prediction $\hat{y}$ and the observed data y, stating that, for a calibrated model, the predicted uncertainty and the expected error should match (Levi et al. 2022). Formally, a regression model is calibrated if

$$\mathbb{E}_{x,y}\left[(\hat{y}-y)^2|\sigma^2 = u\right] \to u \tag{26}$$

for any chosen positive u, where the expectation $\mathbb{E}$ is taken over the joint distribution of x and y. From the definition, no average over points with different values of u is needed; thus, in principle, for each data point, one can correctly predict the expected error. In practice, however, binning is performed to empirically evaluate Eq. (26).

Similar to the calibration curve in the interval based approach, here, a *reliability diagram* can be created to diagnose the calibration level of a model, as follows (Levi et al. 2022):

(1) Sort the data points according to their predicted uncertainty σ^2, and then divide them into M bins, $B_1, B_2, \ldots B_M$. For simplicity, the bin boundaries can be equally located from the minimum to the maximum of the uncertainty (Figure 9a).

(2) For each bin B_j, calculate the root-mean variance (RMV), $\text{RMV}(j) = \sqrt{\frac{1}{|B_j|}\sum_{i\in B_j}\sigma_i^2}$, and the root-mean-square error (RMSE), $\text{RMSE}(j) = \sqrt{\frac{1}{|B_j|}\sum_{i\in B_j}(\hat{y}_i - y_i)^2}$, where $|B_j|$ is the number of data points in bin j.

(3) Plot the RMSE(j) against the RMV(j) for each bin j. This plot is the reliability diagram (Figure 9b).

According to Eq. (26), if a model is perfectly calibrated, the RMV and RMSE should be equal for each bin; therefore, the

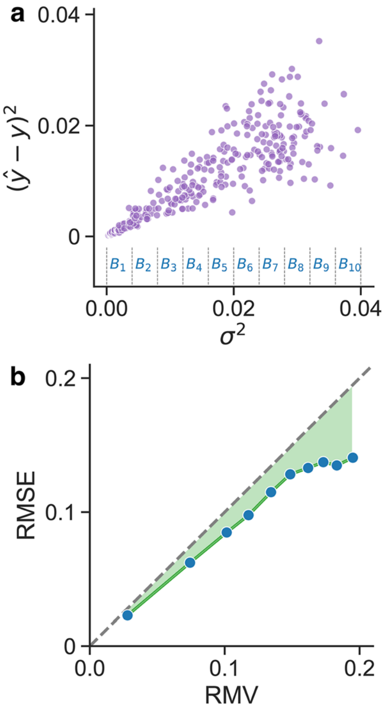

Figure 9: Reliability diagram for error based calibration approach. (a) Binning an example dataset of 400 data points into 10 equally separated intervals. (b) Calibration curve, where each blue dot represents the RMSE and RMV for each bin. A perfectly calibrated model should follow the diagonal line.

calibration curve should be a straight diagonal line in Figure 9b. A larger deviation from the diagonal line suggests a more poorly calibrated model. It is important to note that, unlike the interval based approach, here, the reliability diagram is not constrained to be within [0, 1], but instead ranges between 0 and the maximum RMV or the maximum RMSE value. Consequently, directly comparing the reliability diagrams of different models is not appropriate unless some form of normalization is performed. Furthermore, the choice of the number of bins can have a significant impact on the results. For instance, in Figure 9a, we chose 10 bins, and the last bin B_{10} consists of only three data points, which is insufficient to obtain reliable statistics for RMV and RMSE. One could consider using a smaller number of bins; however, if the number of bins is too small, the details might be averaged out.

4.2 Metrics

Calibration plots offer a qualitative way to assess UQ methods. For quantitative comparison, scoring metrics that can assign numerical scores to each UQ method become necessary. Various metrics have been proposed, and we focus on the important ones that evaluate a UQ method from four different perspectives: calibration, precision, accuracy, and efficiency.

4.2.1 Calibration error

Miscalibration area. For the interval based calibration approach, the closeness of a model's calibration curve to the perfect calibration curve (i.e., the diagonal line) can be quantified by calculating the area between them (e.g., the green area in Figure 8a), called the *miscalibration area* (Tran et al. 2020). A smaller miscalibration area indicates better calibration and a miscalibration area of 0 suggests an ideal calibration.

Expected normalized calibration error (ENCE). For the error based calibration approach, it does not make sense to calculate the area between a model's calibration curve and the perfect calibration curve and then compare across models, because the RMV and RMSE values are not bounded to be within [0, 1] and different UQ methods can have varying ranges for RMV and RMSE. To alleviate this, the ENCE can be used (Levi et al. 2022),

$$\mathrm{ENCE} = \frac{1}{M}\sum_{j=1}^{M}\frac{|\mathrm{RMV}(j) - \mathrm{RMSE}(j)|}{\mathrm{RMV}(j)}, \tag{27}$$

where M is the total number of bins used to generate the calibration curve. Similar to the miscalibration area, a smaller ENCE indicates better calibration.

Miscalibration error and ENCE summarize the reliability by aggregating/averaging the errors between the predicted and perfect calibration curves, producing an overall assessment of a model's calibration. One can also consider the maximum calibration difference between the curves to obtain the worst-case error. This becomes important in high-risk applications, e.g., in drug discovery and materials design for safety-critical systems.

4.2.2 Precision

Calibration is necessary but not sufficient for useful UQ analysis (Gneiting et al. 2007). Recall that calibration only measures the statistical consistency between the predicted uncertainty and the observed data, but does not provide information about the distribution of the predictions themselves. This aspect is related to the *precision* of the predictions, and metrics such as *sharpness* and *dispersion* have been proposed to quantify it (Kuleshov et al. 2018; Levi et al. 2022). However, according to Gneiting et al. (2007), these metrics should be considered secondary after calibration. A major reason for this is that they are properties of the predicted uncertainty alone and do not capture the relationships between uncertainty and accuracy.

Sharpness. A well-calibrated model with more precise predictions (small uncertainty estimates) would be more informative and useful than a less precise model (large uncertainty estimates) (Gneiting et al. 2007). This idea can be quantified by the *sharpness*, defined as (Kuleshov et al. 2018; Tran et al. 2020):

$$\mathrm{SHA} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}\,(F_i)}, \tag{28}$$

where Var(F_i) is the variance of the random variable whose CDF is F for data point i. In practice, this can be evaluated as $\mathrm{SHA} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\sigma_i^2}$, where σ_i^2 is the predicted variance (uncertainty) for data point i. The more precise the predictions, the sharper the model (smaller SHA value), and the sharper the better.

Dispersion. Another dimension involves the dispersion of the uncertainty. One can obtain perfectly calibrated uncertainty if a model always outputs the same constant uncertainty which matches the empirical frequency across the entire distribution (Scalia et al. 2020; Tran et al. 2020). Such an uncertainty estimate is not informative or useful because it remains unchanged regardless of the input data provided to the model. Levi et al. (2022) propose the *coefficient of variation* to measure the dispersion of the uncertainty estimates,

$$C_v = \frac{1}{\mu_\sigma}\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\sigma_i - \mu_\sigma)^2}, \tag{29}$$

where σ_i is the predicted standard deviation, μ_σ is the mean of the standard deviations and N is the total number of data points. A C_v value of 0 means the same constant uncertainty for all data points, not a useful uncertainty estimate. Higher C_v is preferred so that the uncertainty for different data points can be distinguished.

4.2.3 Accuracy

The introduction of a UQ method to a model can affect the model's prediction accuracy. For example, it has been observed that graph NNs for chemical property prediction trained with ensemble and MC dropout methods yield higher accuracy when compared with the same model trained using maximum likelihood (Scalia et al. 2020). But this may not always be the case. So, it is crucial to examine prediction accuracy as well. The two most widely used accuracy metrics are mean absolute error (MAE):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|, \tag{30}$$

and root-mean-square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{31}$$

where $\hat{y}_i$ is model prediction, y_i is the corresponding reference value, and N is the number of data points. MAE measures the average absolute difference between the predicted and reference values, treating all errors equally. On the other hand, by squaring the errors before averaging, RMSE gives higher weight to larger errors, making it more sensitive to outliers. Lower values of MAE and RMSE indicate better agreement between predictions and reference data.

4.2.4 Efficiency

Computational tractability and efficiency are essential for a UQ method to be practically usable. Even if a method is highly calibrated, precise, and accurate, it may not be suitable for real-world applications if it is computationally too demanding in terms of both time and memory. Unfortunately, efficiency is often ignored in existing studies of UQ methods for atomistic ML.

Training and inference time. A straightforward way to compare time efficiency is by tracking the total runtime of a model to obtain the prediction and uncertainty, which is usually the main concern for practical atomistic ML applications. However, a UQ method's total runtime is highly dependent on the underlying model's speed. To focus on the UQ method itself, we analyze training and inference efficiency separately: the number of models required to be trained and the number of model executions needed to obtain the uncertainty at inference. The ensemble approach is not efficient because multiple models need to be trained, and multiple model executions must be carried out to get the uncertainty at inference. Approaches such as MVE and evidential regression are on the opposite end of the spectrum: a single model at training and a single execution at inference. Methods like MC dropout lie in between these two extremes. The training/inference efficiency for all UQ methods discussed in Section 3 is listed in Table 2.

Memory. In addition to time efficiency, memory efficiency should be another consideration when evaluating the UQ methods. This is, again, largely dependent on the underlying model to which a UQ method is applied.

4.2.5 Other metrics

Besides the above-mentioned ones, other metrics have also been used to evaluate UQ performance, particularly, in assessing a model's calibration. These include ranking correlation and NLL, among others, which can be used together with the calibration metrics discussed in Section 4.2.1.

Ranking correlation. For a UQ method, we expect that a high uncertainty suggests a large prediction error. So, there should be a monotonic relationship between the uncertainty δ and the prediction error ϵ for a well-calibrated model (Tan et al. 2023; Varivoda et al. 2023). This can be quantified with Spearman's rank correlation coefficient. For a set of uncertainties and errors $\{(\delta_i, \epsilon_i)\}_{i=1}^{N}$, we first obtain ranked sequences of the uncertainties R_δ and the errors R_ϵ, separately, and then compute the Spearman's rank correlation coefficient as

$$r_s = \frac{\text{Cov}(R_\delta, R_\epsilon)}{\sigma_{R_\delta}\sigma_{R_\epsilon}}, \tag{32}$$

where Cov denotes the covariance between two variables and σ denote the standard deviation of a variable. The values of r_s are within the range of [−1, 1], with −1 or 1 suggesting a perfect monotonic relationship between the uncertainty and the error, and 0 being the worst case, indicating that there is no correlation.

NLL. NLL is a standard measure of a model's fit to the data which combines both the accuracy and the uncertainty in one measure. With a set of predictions and the associated uncertainties (i.e., variance σ^2), Eq. (6) can be directly used to obtain the NLL. Despite its popularity, NLL has been criticized for the lack of robustness (Gneiting et al. 2007). It is hypersensitive to small changes and is unbounded, with acceptable values ranging from $-\infty$ to $+\infty$ (Gneiting and Raftery 2007, Selten 1998).

4.3 Benchmark studies

Using metrics as those discussed in Section 4.2, several benchmark studies have attempted to evaluate the performance of various UQ methods on diverse atomistic ML datasets (Hirschfeld et al. 2020; Hu et al. 2022; Scalia et al. 2020; Tan et al. 2023; Tran et al. 2020; Varivoda et al. 2023). A major goal is to identify UQ methods that can perform well across metrics and datasets, thus providing practical guidance for selecting appropriate ones for chemical and materials applications. Several general observations can be made from these benchmark studies.

First, the UQ methods perform differently on different metrics; a method that works well on one metric can fall short on another. For example, Tran et al. (2020) have trained various ML models to predict the adsorption energies of small molecules on metal surfaces calculated from DFT. Although all trained ML models reported in Tran et al. (2020) have similar accuracy (MAE of ~0.20 eV) and miscalibration area (~0.13, Figure 10), their performance on precision varies a lot. For example, MC dropout is much sharper than NN ensemble (SHA: 0.09 versus 0.14), but has a lower dispersion (C_v: 0.82 versus 1.06) (see Section 3, Tran et al. (2020), for more details). Even for UQ methods that have, for example, a similar miscalibration area as illustrated in Figure 10, the shape of the calibration curves can be drastically different, indicating different modes of miscalibration. Tan et al. (2023) observed similar behavior using the rMD17 dataset (Christensen and Von Lilienfeld 2020) of energies of small molecules.

Second, performance varies by dataset; for a given metric, different UQ methods can have varying error levels across datasets. For example, Tan et al. (2023) observed that, for the rMD17 dataset of energies of small molecules, NN ensemble exhibits smaller miscalibration error than the evidential regression method (Figure 11b). Varivoda et al. (2023) have found similar behaviors for the dataset of formation energies of crystals (Jain et al. 2013) and the dataset of surface adsorption energies of metal alloys (Mamun et al. 2019). However, they found that, for the dataset of band gaps of MOFs (Rosen et al. 2021), the miscalibration errors are the same for the ensemble and evidential regression methods.

Third, ensemble methods appear to be a reliable UQ approach in general. For example, Tan et al. (2023) compared the ensemble method versus the MVE, evidential regression, and GMM deterministic methods for both in-domain (rMD17 and ammonia) and out-of-domain (silica glass) tasks. Using metrics such as MAE, Spearman correlation, and miscalibration area (Figure 11), they have found that single-deterministic methods struggle to consistently perform better across each in-domain and out-of-domain task and that the ensemble method still remains the most reliable choice. For ensemble methods, different ways to generate the ensemble can result in different performances. For example, Scalia et al. (2020) have compared three ensemble methods – NN ensemble with different parameter initialization, bootstrapping, and MC dropout – using the miscalibration area, ENCE, sharpness, and dispersion metrics (Recall from 3.1.1 that MC dropout can be regarded as an ensemble method in practice). Evaluated on various MoleculeNet benchmarking datasets (Wu et al. 2018), their finds indicate that NN ensemble with different parameter initialization and bootstrapping consistently outperform MC dropout (Figure 12).

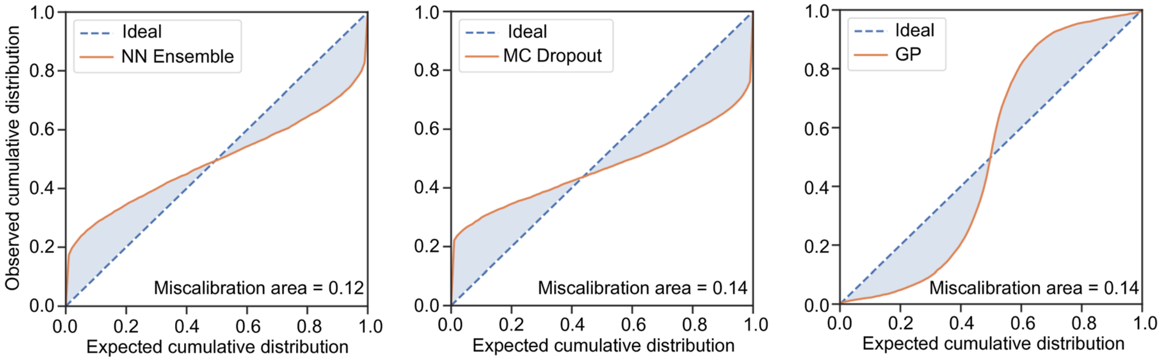

Figure 10: Comparison of calibration curves displaying miscalibration area of various UQ methods. NN ensemble, here, refers to NNs trained with random parameter initializations. GP: Gaussian process, a probabilistic modeling approach that inherently models predictive uncertainty (Rasmussen 2003). Images adapted from Tran et al. (2020) under a CC BY 4.0 DEED license.

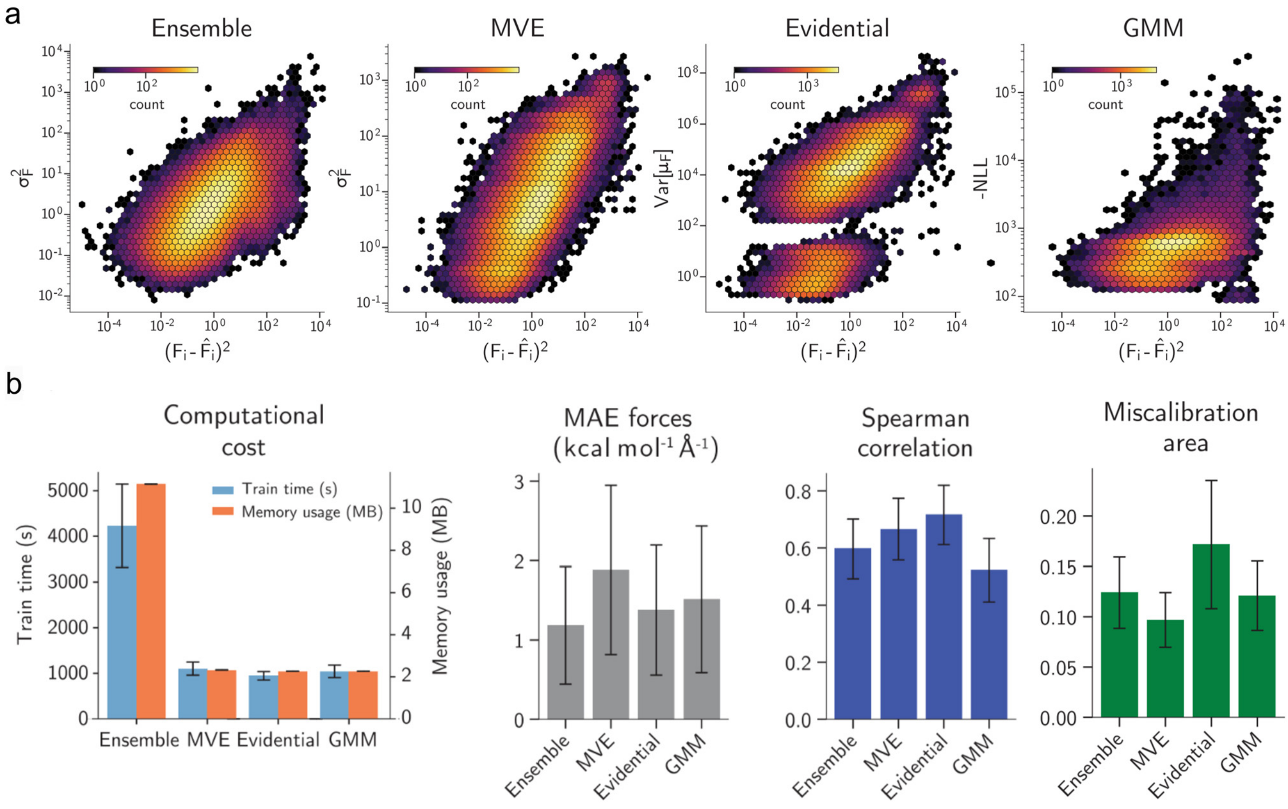

Figure 11: Comparison of various UQ methods on the rMD17 dataset. (a) Predicted uncertainties as a function of squared errors of atomic forces and (b) computational cost, MAE of atomic forces, Spearman correlation, and miscalibration area of various UQ methods. Ensemble, here, refers to NNs trained with random parameter initialization. Images adapted from Tan et al. (2023) under a CC BY 4.0 license.

Despite the robustness of the ensemble approach, there are contradictory studies showing that ensemble methods may not always be the most reliable choice. For example, Hirschfeld et al. (2020) have demonstrated that the stacking methods that sequentially combine multiple weak models to produce the uncertainty estimate can be more consistent than the ensemble approach. Heid et al. (2024) have shown that the uncertainty of a single prediction obtained from an ensemble approach cannot be directly correlated with the absolute error per atom. This is because the absolute error is distributed along a normal distribution, with its width determined by the uncertainty arising from model variance. To address this, they developed an approach that uses locally aggregated uncertainties to identify high-error local

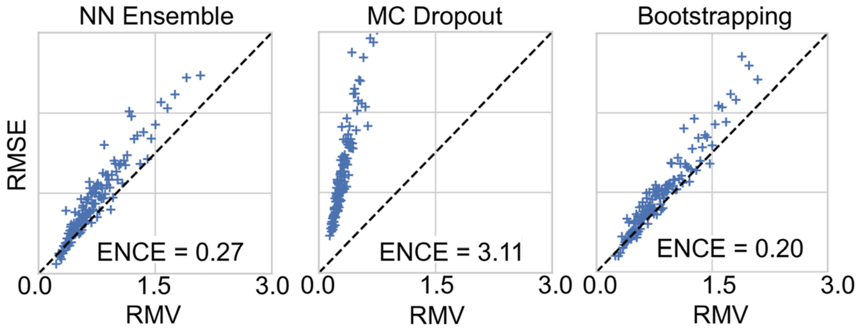

Figure 12: Comparison of error based calibration curves of three UQ methods on the QM9 dataset. NN ensemble, here, refers to NNs trained with random parameter initializations. Image adapted from Scalia et al. (2020). Copyright 2020, American Chemical Society.

substructures, enabling the resolution of absolute errors on an atomic scale.

Fourth, off-the-shelf metrics may not be directly applicable. The metrics, particularly those related to calibration discussed in Section 4.2, have been primarily developed within the ML community using non-chemical datasets. Their suitability for chemical and materials problems is not guaranteed. For instance, the error based calibration approach conventionally performs binning directly on the predicted uncertainty, represented by the variance σ^2 in Figure 9a. Chemical and materials data can have numerous small variance values. Therefore, it might be more appropriate to conduct binning after transforming the variances using a logarithmic function to avoid cluttering the bins at small variance values, as demonstrated in Figure 11a.

Benchmark studies so far have been valuable in highlighting how different UQ methods can have varied performances depending on the choice of metrics and dataset. However, these studies consistently indicate that the question of which UQ method to select in practice remains unresolved. This appears to be discouraging. Nevertheless, there are guiding principles based on ease of use, efficiency, and uncertainty propagation. Ensemble methods provide a strong baseline and are straightforward to implement, making them a go-to choice for many applications. However, single-pass models (e.g., MVE, evidential, and GMM) are generally more efficient than ensemble methods (Figure 11b), so they can be good choices for resource-bounded applications. Another consideration is UP; the chosen propagation method may make certain UQ methods more suitable. This will be further discussed in Section 5.

4.4 Recalibration

So far, we have discussed ways for evaluating UQ methods in terms of their calibration, precision, accuracy, and efficiency, and we have provided some examples. But if a model's performance is unsatisfactory, are there any approaches to improve it? The answer is yes, and, here, we will concentrate on calibration.

Informally, the calibration problem can be described as follows: given a trained model U that can generate predictive uncertainty $\delta = U(x)$, we train another recalibration model R such that the output of the composed functions $\hat{\delta} = R(U(x))$ is calibrated. The model R should be trained on a separate recalibration dataset D_{cal} distinct from the datasets used for model parameter optimization or hyperparameter tuning. Below, we discuss two recalibration methods, both of which are applicable to the interval based calibration approach introduced in Section 4.1.1.

Variance scaling. For the interval based calibration, the model prediction, y, is set to take the form of a Gaussian $y \sim \mathcal{N}(\mu, \sigma^2)$. The model can be under-confident or over-confident depending on the scale of the variance σ^2. To recalibrate it, that is, making the calibration curve move toward the diagonal line in Figure 8a, we can train a linear model $R(\sigma^2) : \hat{\sigma}^2 = a\sigma^2 + b$ to scale the variance. The parameters a and b can be determined by, e.g., minimizing a calibration NLL loss (Tan et al. 2023), $\frac{1}{2}\sum_{i=1}^{N}\left(\log[2\pi(a\sigma^2 + b)] + (y_i - \hat{y}_i)^2/(a\sigma^2 + b)\right)$, using a recalibration dataset D_{cal} consisting of N data points. Once the optimal a and b are obtained, the new variance $\hat{\sigma}^2$ for the Gaussian is known for every data point, which can then be used to regenerate the calibration plots.

Isotonic regression. The variance scaling approach still assumes that the model prediction follows a Gaussian distribution. The true observed data distribution, however, may not be Gaussian. Kuleshov et al. (2018) proposed a recalibration approach based on isotonic regression, which is effective even for non-Gaussian cases. Given a recalibration dataset D_{cal}, this approach begins by transforming it into a processed recalibration dataset $\hat{D}_{\mathrm{cal}} = \{(q_j, \tilde{q}_j)\}_{j=1}^{M}$, where q_j and $\tilde{q}_j$ are obtained using the same procedures described in Section 4.1.1. Using $\hat{D}_{\mathrm{cal}}$, an isotonic regression model $\widetilde{q}_j = R(q_j)$ is then trained. Isotonic regression is a technique for fitting a free-form line to map a sequence of inputs (q_j in this case) to a sequence of observations ($\widetilde{q}_j$ in this case) such that the fitted line is non-decreasing everywhere and lies as close to the observations as possible (Fielding et al. 1974). Isotonic regression is chosen because it accounts for the fact that the true calibration curve is monotonically increasing. Once the isotonic regression model is trained, new uncertainty $\widehat{\delta}$ can be generated, and new calibration plot can be produced.

5 Uncertainty propagation

For most chemical and materials problems, quantifying a model's predictive uncertainty is not sufficient; often, we are also interested in understanding how the uncertainty propagates to a physical QoI that can be obtained from physics-based modeling using the model. For example, if the uncertainties in energy and forces are known for an interatomic potential, and MD simulations are employed to compute a material property like thermal conductivity, we would naturally hope to know the uncertainty in the calculated thermal conductivity. Similarly, if microkinetic modeling is used to investigate chemical reaction dynamics, we need to determine how the uncertainty in reaction rates propagates to and affects the concentrations of the chemical species.

We define the uncertainty propagation (UP) problem as follows: Given a model $y = f(x; \theta)$ that can provide predictive uncertainty, we aim to determine the uncertainty in a QoI z that is a function of the model output, i.e., $z = g(y) = (g \circ f)(x; \theta) = h(x; \theta)$, where $h = g \circ f$ is defined as the composition of g and f. In other words, we investigate how the uncertainty in the model parameters θ and the training data propagate to the QoI z. Typically, g is not an ML model but rather a physics-based simulation technique, such as MD or microkinetic modeling, as mentioned above.

While UQ has been reasonably well investigated in atomistic ML for chemical and materials applications, UP remains a relatively unexplored area. Nevertheless, UP is essential for building confidence in the results. It provides a comprehensive assessment of the entire modeling pipeline, enabling the evaluation of the robustness of the final results. Furthermore, UP helps identify the most influential uncertainty sources, guiding targeted efforts to refine and improve the modeling pipeline. In this section, we introduce some of the efforts in UP for MD and microkinetic simulations.

5.1 Bayesian propagation

Given the posterior over model parameters $p(\theta|D)$ (see Section 2.2), the distribution of a QoI z can be written as

$$p(z|x, D) = \int p(z|x, \theta) p(\theta|D)\, \mathrm{d}\theta, \tag{33}$$

where $p(z|x, \theta)$ is the likelihood of θ to observe z given the composed model $z = h(x; \theta)$. The integration, generally, cannot be analytically evaluated for most QoI in MD and microkinetic simulations. To address this, again, sampling techniques can be employed. Eq. (33) can be approximated by (Angelikopoulos et al. 2012)

$$p(z|x, D) = \frac{1}{M} \sum_{i=1}^{M} p(z|x, \theta_i). \tag{34}$$

Assuming the likelihood $p(z|x, \theta)$ is a Gaussian with $h(x; \theta)$ as its mean, then the predictive mean and variance of Eq. (34) can be respectively expressed as

$$\bar{z} = \frac{1}{M} \sum_{i=1}^{M} z_i = \frac{1}{M} \sum_{i=1}^{M} h(x\,; \theta_i), \tag{35}$$

and

$$\sigma_z^2 = \frac{1}{M-1} \sum_{i=1}^{M} (z_i - \bar{z})^2. \tag{36}$$

The sampling-based Bayesian UP means selecting multiple sets of model parameters θ, computing the QoI z_i using each parameter set, and then calculating their mean as the final prediction and their variance as the uncertainty. The sampling of the parameters can be done via MC methods, such as MCMC (Berg 2004) and transitional MCMC (Ching and Chen 2007).

Practically, it can also be viewed as an ensemble approach, where each realization of the parameters is a member of the ensemble. Therefore, for the UQ methods discussed in Section 3, this UP method can be directly applied to models trained with MC dropout and the ensemble approach, but not for the others.

Using an NN interatomic potential trained with MC dropout in MD simulations, Wen and Tadmor (2020) adopted this sampling-based approach to propagate the uncertainty in atomic forces to the mechanical stress on monolayer graphene (Figure 13). As expected, the uncertainty in stress increases as the graphene layer is compressed or stretched from its equilibrium lattice parameter of 2.466 Å.

Recently, various studies have investigated the effects of uncertainty from ML models on microkinetic modeling. Li et al. (2023) was able to propagate the uncertainty in stoichiometric coefficients and parameters in the Arrhenius law (Arrhenius 1889a,b) to the concentration of chemical

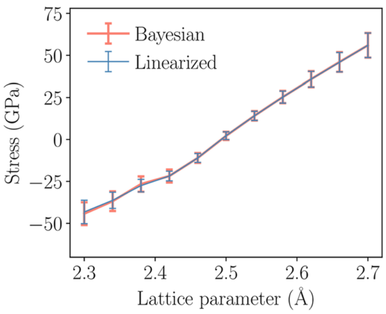

Figure 13: Uncertainty propagation in molecular dynamics computation of mechanical stresses in a monolayer graphene. Error bars indicate uncertainty level. Adapted from Wen and Tadmor (2020) with a CC BY 4.0 license.

species in biodiesel production reaction systems, using the so-called Bayesian chemical reaction NNs. In a microkinetic study of ethanol steam reforming reactions, Xu and Yang (2023) investigated the propagation of errors in binding energy predicted by ML models to kinetic properties such as reaction rates. Their results demonstrated that the preferred reaction pathway varies depending on the used ML model.

5.2 Linearized propagation

Linearized UP is a general approach that can be used together with all UQ methods discussed in Section 3. Let δ_y be the uncertainty associated with the output y of an ML model $y = f(x; \theta)$. Then for a QoI z that is a function of the model output, $z = g(y)$, we do a first-order Taylor expansion at y_0 to obtain $z = g(y_0) + \frac{\partial g}{\partial y}(y - y_0)$. With this linearization, the uncertainty in z can be expressed as (Arras 1998):

$$\delta_z = \frac{\partial g}{\partial y}\delta_y, \tag{37}$$

meaning that the uncertainty in y can be propagated to z by multiplying the gradient. In general, if g is a function of multiple independent inputs, i.e., $z = g(y_1, y_2, \ldots, y_M)$, the uncertainty in z can be written as (Arras 1998):

$$\delta_z^2 = \sum_i^M \left(\frac{\partial g}{\partial y_i}\right)^2 \delta_{y_i}^2. \tag{38}$$

Once the uncertainty is obtained from a UQ method, it can be readily propagated to the QoI z, using Eq. (37) or Eq. (38). However, there are two challenges in applying this approach in practice. First, it may not be immediately obvious how to compute the gradients of g, which typically represent physics-based simulation techniques such as MD and microkinetic modeling. Second, while the approach is exact for a linear function g, it can introduce significant errors when applied to a nonlinear function (Cho et al. 2015).

The linearized UP approach becomes very attractive if the challenges are overcome. For example, Wen and Tadmor (2020) reformulated the integration algorithm in an MD simulation, and managed to propagate the uncertainty in atomic forces to stresses. As seen from Figure 13, the predicted mean stress and uncertainty agree very well with those from the sampling-based Bayesian approach. The linearized approach is computationally more efficient than the Bayesian approach, since it only needs one model evaluation, while the Bayesian approach requires multiple model evaluations.

5.3 Sensitivity analysis

Sensitivity analysis examines how the uncertainty in a model's output can be apportioned to various sources of uncertainty in its inputs. Ideally, uncertainty and sensitivity analyses should be conducted in tandem; by working together, they can pinpoint the most critical input parameters, offering crucial insights into the model's reliability and providing guidance for further model refinement. Although these techniques have not been widely employed together with ML models yet, we expect this to happen soon in the near future. To illustrate their potential, we introduce some of their usage with classical non-ML models.

Given a QoI $z = g(y)$, we can perturb the input by some amount Δy and observe the change in the output Δz. This change, Δz, can be interpreted as the propagated uncertainty if we set the uncertainty in y as Δy. Typically, the input y lies in an M-dimensional parameter space, i.e., $y = [y_1, y_2, \ldots, y_M]$; then, depending on how Δy is chosen, we get *local sensitivity* and *global sensitivity*. Local sensitivity refers to perturbing each individual parameter y_i and observing its effects on z separately. Global sensitivity refers to exploring the entire parameter space simultaneously, considering interactions between the parameters.

Sensitivity analysis is an integral part of microkinetic modeling (Motagamwala and Dumesic 2020). Local sensitivity analysis like the derivative-based technique (Döpking et al. 2018) and global sensitivity analysis such as the Sobol' method (Sobol 2001) and the Morris method (Morris 1991) have been widely employed. For example, Bensberg and Reiher (2024) have recently leveraged these techniques to automatically refine structures, reaction paths, and energies in chemical reaction networks, and successfully identified a small number of elementary reactions and compounds that are essential for reliably describing the kinetics of the Eschenmoser–Claisen rearrangement reactions (Wick et al.

1964) of allyl alcohol and of furfuryl alcohol. In addition, their Morris sensitivity analysis also provides the uncertainty in the predicted concentrations. Similarly, Kreitz et al. (2021) applied global uncertainty assessment and sensitivity analysis to explore parametric uncertainties in microkinetic models for CO_2 hydrogenation on the (111) surface of Ni. By systematically generating numerous mechanisms, they demonstrated how UQ can identify feasible models and optimize predictions within the uncertainty space.

Sensitivity analysis has also been applied to quantify the uncertainty in QoI obtained from MD simulations. Information-theoretic approaches provide a powerful framework for this purpose (Kurniawan et al. 2022). For a QoI z that can be obtained from an MD simulation, an upper bound for the uncertainty in z can be obtained as (Dupuis et al. 2016; Pantazis and Katsoulakis 2013; Tsourtis et al. 2015):

$$|\mathbb{E}_{\theta+\Delta\theta}[z] - \mathbb{E}_{\theta}[z]| \leq \sqrt{\mathrm{Var}_{\theta}[z]}\sqrt{\Delta\theta\mathscr{I}(\theta)\Delta\theta} \tag{39}$$

upon parameter perturbation $\Delta\theta$, where Var denotes the variance, and $\mathscr{I}(\theta)$ is the Fisher information, which measures the amount of information that the trained model carries about its parameter θ (Cover and Thomas 2012; Wen 2019). This provides an efficient way to investigate the reliability of MD simulations. Using this approach, Wen et al. (2017) studied the thickness of an MoS_2 sheet and found that Eq. (39) provides a tight bound, demonstrating high reliability of the MD predictions. Although Fisher information can provide useful insights into the uncertainty in MD simulations, the overall analysis is restricted to the perturbations only in the vicinity of the equilibrium model parameters.

5.4 Data uncertainty and propagation

The discussions in the above sections have primarily focused on uncertainties and their propagation in ML model. However, uncertainty is also inherent in chemical and materials data itself. For example, for data generated using DFT, the commonly employed semi-local density functional approximations have well-known intrinsic errors (Cohen et al. 2008; Perdew and Zunger 1981) that affect the accuracy and reliability of the results. When data is derived from a single semi-local density functional approximation, there is no effective way to address data uncertainty, leaving its impact on downstream properties often unaddressed. Below, we discuss two approaches that apply systematic statistical analysis to assess data uncertainty and its propagation to downstream properties, particularly in the context of microkinetic modeling.

The first method, developed by Wellendorff et al. (2012), is an ensemble-based approach that uses the same density functional but introduces tunable parameters. This method builds on the BEEF (Bayesian error estimation functional) family (Mortensen et al. 2005), incorporating an exchange-correlation functional (BEEF-vdW) with non-local correlation terms and tunable parameters to accurately capture interactions such as van der Waals forces, which are critical for surface science and catalysis (Wellendorff et al. 2012). Uncertainty estimates are then derived from an ensemble of functionals generated by parameter perturbation, with the ensemble's standard deviation defining the uncertainty.

The second method, proposed by Walker et al. (2016), is also an ensemble-based approach. However, rather than perturbing a single functional, it employs multiple density functionals from different rungs of Perdew's "Jacob's Ladder" (Perdew and Schmidt 2001). In this approach, the energies of intermediate and transition states are calculated using various density functionals chosen for their applicability to the target system. A latent model using factor analysis (Rencher and Christensen 2012), is then developed to capture shared predictions across functionals and their unique variations. Finally, the predicted energies from the latent model are refined using a secondary probabilistic model to ensure consistency with reference data.

For both methods, once the corrected energies are obtained, they can be used in calculating downstream QoIs, and the associated uncertainty in the energies can be propagated using MC sampling as in Eq. (36). Using this approach, the first method has been widely applied in computational catalysis, from assessing reaction rate reliability (Lu et al. 2022) to identifying reaction mechanisms (Kreitz et al. 2023) and analyzing surface coverage (Wang et al. 2019). The second method has been successfully applied to determine reaction pathways in catalysis, such as identifying dominant mechanisms in the water-gas shift reaction through uncertainty propagation to turnover frequency calculations (Fricke et al. 2022; Walker et al. 2016, 2018).

While both methods can provide uncertainty in the data, the limitations should be noted. The methodology overall relies on fitting to experimental reference data or computed data with higher accuracy, such as the G3/99 dataset of experimental molecular formation energies (Curtiss et al. 2000) and the S22 dataset of intermolecular interaction energies (Jurečka et al. 2006). The reference data themselves, however, carry uncertainty. Such uncertainty is often assumed but not guaranteed to be negligible compared to DFT calculations (Wang et al. 2021). Additionally, the density functionals used or developed here are fitted to specific material properties, and thus their accuracy may not be transferable to other materials or properties (Mardirossian and Head-Gordon 2017; Medvedev et al. 2017). Thus, if the functionals in the ensemble perform poorly for a particular material or property, the resulting uncertainty, and any conclusions based on them, can be unreliable. This suggests that the pursuit of widely applicable density functionals, supported by thorough benchmark studies (Kim et al. 2024; Sheldon et al. 2021, 2024; Szaro et al. 2023), is equally vital for advancing methods to improve data uncertainty management.

6 Summary and outlook

In this work, we have provided a comprehensive overview of the UQ approaches for atomistic ML. The UQ methods are classified into three main categories: probabilistic, ensemble, and feature space distance. The similarities, differences, and connections between them were discussed to provide an overall overview of the methods. We have discussed metrics to evaluate the performance of these UQ methods from different angles, focusing on calibration, precision, accuracy, and efficiency. In addition, we have emphasized the importance of UP in downstream chemical and materials applications of the ML models.

We deliberately exclude some important but advanced topics to make the presentation more accessible and avoid further complications. For example, we chose to focus on UQ for NNs and ignore other methods such as Gaussian processes (Rasmussen 2003), which inherently provide predictive uncertainty. For the use of Gaussian processes in materials and molecular problems, we refer readers to the thorough review by Deringer et al. (2021). Additionally, we do not explicitly discuss whether the aleatoric or epistemic uncertainty is modeled by a UQ method; instead, we provide the total uncertainty, as it is the combined effect of both sources that is relevant for most practical purposes. Nevertheless, works such as Gustafsson et al. (2020), Gawlikowski et al. (2023), and Heid et al. (2023) provide further discussion on this topic.

We have identified several challenges in UQ and UP for atomistic ML, along with potential opportunities to address these challenges. First, existing benchmark studies suggest that the performance of UQ methods is highly dependent on the datasets and metrics being used. There is no universal UQ method that consistently outperforms others in all scenarios. Thus, there is a high demand for a set of best practices and guidelines for UQ in atomistic ML. These guidelines should provide recommendations on choosing appropriate UQ methods based on factors such as the nature of the dataset, the complexity of the ML model, the available computational resources, and the downstream applications of the ML model.

A second challenge is the large miscalibration of existing UQ methods. The calibration curves of many UQ methods can deviate significantly from the diagonal line, suggesting that the predicted uncertainties do not match well with the observed errors. A straightforward solution is to perform uncertainty recalibration. Uncertainty recalibration for ML models is a relatively new field, and thus it is rarely conducted in atomistic ML. We believe there is great potential to explore uncertainty recalibration techniques tailored for atomistic ML models, improving their calibration and predictive reliability.

A third pressing challenge is related to the scarcity of UP techniques. Although UQ is reasonably investigated, UP receives far less attention despite its importance in chemical and materials modeling. We suspect that this is partly due to the complexity of integrating UQ methods with physics-based simulations, which often involve solving differential equations and dealing with complex boundary conditions. A promising direction to tackle the challenge is to develop fully automatic differentiable simulation approaches. These approaches combine automatic differentiation with physics-based simulations, enabling end-to-end differentiation of entire simulation pipelines. As a result, they will allow for the seamless propagation of uncertainties from the ML models to the quantities of interest.

If the existing challenges in UQ and UP can be overcome, we foresee substantial opportunities to accelerate the adoption and development of reliable and robust atomistic ML models. This will enable the exploration of complex chemical and materials systems with quantified uncertainties, ultimately leading to more informed decision-making and accelerated discovery.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared
Conflict of interest: The authors state no conflict of interest.
Research funding: This work is supported by the National Science Foundation under Grant No. 2316667 and the startup funds from the Presidential Frontier Faculty Program at the University of Houston.
Data availability: Not applicable.

Abbreviations

CDF	Cumulative distribution function
DFT	Density functional theory
ENCE	Expected normalized calibration error
i.i.d.	Independent and identically distributed
MAE	Mean absolute error
MC	Monte Carlo
MCMC	Markov chain Monte Carlo
MD	Molecular dynamics
ML	Machine learning
MLE	Maximum likelihood estimation
NLL	Negative log-likelihood
NN	Neural network
QoI	Quantity of interest
RMSE	Root-mean-square error
UP	Uncertainty propagation
UQ	Uncertainty quantification

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al. (2021). A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf. Fusion* 76: 243–297.

Abdi, K., Celse, B., and McAuley, K. (2024). Propagating input uncertainties into parameter uncertainties and model prediction uncertainties – a review. *Can. J. Chem. Eng.* 102: 254–273.

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. (2020) Deep evidential regression. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (Eds.). *Advances in neural information processing systems*, Vol. 33, pp. 14927–14937.

Angelikopoulos, P., Papadimitriou, C., and Koumoutsakos, P. (2012). Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework. *J. Chem. Phys.* 137: 144103.

Arras, K.O. (1998). *An introduction to error propagation: derivation, meaning and examples of equation* $c_y = f_x c_x f_x^t$*, Technical Report EPFL-ASL-TR-98-01 R3*. Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne.

Arrhenius, S. (1889a). Über die dissociationswärme und den einfluss der temperatur auf den dissociationsgrad der elektrolyte. *Zeitschrift für physikalische Chemie* 4: 96–116.

Arrhenius, S. (1889b). Über die reaktionsgeschwindigkeit bei der inversion von rohrzucker durch säuren. *Zeitschrift für physikalische Chemie* 4: 226–248.

Axelrod, S., Schwalbe-Koda, D., Mohapatra, S., Damewood, J., Greenman, K.P., and Gómez-Bombarelli, R. (2022). Learning matter: materials design with machine learning and atomistic simulations. *Acc. Mater. Res.* 3: 343–357.

Back, S., Tran, K., and Ulissi, Z.W. (2019). Toward a design of active oxygen evolution catalysts: insights from automated density functional theory calculations and machine learning. *ACS Catal.* 9: 7651–7659.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373: 871–876.

Bartók, A.P., Payne, M.C., Kondor, R., and Csányi, G. (2010). Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* 104: 136403.

Batatia, I., Benner, P., Chiang, Y., Elena, A.M., Kovács, D.P., Riebesell, J., Advincula, X.R., Asta, M., Avaylon, M., Baldwin, W.J., et al (2024). A foundation model for atomistic materials chemistry. *arXiv preprint*, arXiv:2401.00096.

Behler, J. (2021). Four generations of high-dimensional neural network potentials. *Chem. Rev.* 121: 10037–10072.

Behler, J. and Parrinello, M. (2007). Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* 98: 146401.

Bensberg, M. and Reiher, M. (2024). Uncertainty-aware first-principles exploration of chemical reaction network. *J. Phys. Chem. A*: 128, 4532–4547.

Berg, B.A. (2004). *Markov chain Monte Carlo simulations and their statistical analysis: with web-based Fortran code*. World Scientific Publishing Company, Singapore.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis, Springer series in statistics*. Springer, New York.

Bernardo, J. and Smith, A. (2009). *Bayesian theory, Wiley series in Probability and statistics*. Wiley, Chichester.

Bishop, C.M. (2006). *Pattern Recognition and machine learning*. Springer-Verlag, Berlin, Heidelberg.

Ceriotti, M. (2022). Beyond potentials: integrated machine learning models for materials. *MRS Bull.* 47: 1045–1053.

Chen, C. and Ong, S.P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* 2: 718–728.

Chen, Z., Andrejevic, N., Drucker, N.C., Nguyen, T., Xian, R.P., Smidt, T., Wang, Y., Ernstorfer, R., Tennant, D.A., Chan, M., et al. (2021). Machine learning on neutron and x-ray scattering and spectroscopies. *Chem. Phys. Rev.* 2, https://doi.org/10.1063/5.0049111.

Ching, J. and Chen, Y.-C. (2007). Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *J. Eng. Mech.* 133: 816–832.

Cho, H., Luck, R., and Stevens, J.W. (2015). An improvement on the standard linear uncertainty quantification using a least-squares method. *J. Uncertain. Anal. Appl.* 3: 1–13.

Christensen, A.S. and Von Lilienfeld, O.A. (2020). On the role of gradients for machine learning of molecular energies and forces. *Mach. Learn.: Sci. Technol.* 1: 045018.

Cohen, A.J., Mori-Sánchez, P., and Yang, W. (2008). Insights into current limitations of density functional theory. *Science* 321: 792–794.

Cover, T.M. and Thomas, J.A. (2012). *Elements of information theory*. John Wiley & Sons, New Jersey.

Curtiss, L.A., Raghavachari, K., Redfern, P.C., and Pople, J.A. (2000). Assessment of Gaussian-3 and density functional theories for a larger experimental test set. *J. Chem. Phys.* 112: 7374–7383.

DeGroot, M. and Schervish, M. (2012). *Probability and statistics*. Addison-Wesley, New York.

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C.J., and Ceder, G. (2023). Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* 5: 1031–1041.

Deringer, V.L., Bartók, A.P., Bernstein, N., Wilkins, D.M., Ceriotti, M., and Csányi, G. (2021). Gaussian process regression for materials and molecules. *Chem. Rev.* 121: 10073–10141.

Döpking, S., Plaisance, C.P., Strobusch, D., Reuter, K., Scheurer, C., and Matera, S. (2018). Addressing global uncertainty and sensitivity in first-principles based microkinetic models by an adaptive sparse grid approach. *J. Chem. Phys.* 148, https://doi.org/10.1063/1.5004770.

Dupuis, P., Katsoulakis, M.A., Pantazis, Y., and Plechác, P. (2016). Path-space information bounds for uncertainty quantification and sensitivity analysis of stochastic dynamics. *SIAM/ASA J. Uncertain. Quantification* 4: 80–111.

Fedik, N., Zubatyuk, R., Kulichenko, M., Lubbers, N., Smith, J.S., Nebgen, B., Messerly, R., Li, Y.W., Boldyrev, A.I., Barros, K., et al. (2022). Extending machine learning beyond interatomic potentials for predicting molecular properties. *Nat. Rev. Chem* 6: 653–672.

Fielding, A., Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1974). Statistical inference under order restrictions. the theory and application of isotonic regression. *J. Roy. Stat. Soc. Series A (General)* 137: 92.

Fricke, C., Rajbanshi, B., Walker, E.A., Terejanu, G., and Heyden, A. (2022). Propane dehydrogenation on platinum catalysts: identifying the active sites through Bayesian analysis. *ACS Catal.* 12: 2487–2498.

Gal, Y. (2016). *Uncertainty in deep learning, PhD thesis*. University of Cambridge, Cambridge.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: *International conference on machine learning*. PMLR, Cambridge, pp. 1050–1059.

Gastegger, M., Schütt, K.T., and Müller, K.-R. (2021). Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* 12: 11473–11483.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* 56: 1513–1589.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC, New York.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* 102: 359–378.

Gneiting, T., Balabdaoui, F., and Raftery, A.E. (2007). Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.: B (Stat. Methodol.)* 69: 243–268.

Gruich, C.J., Madhavan, V., Wang, Y., and Goldsmith, B.R. (2023). Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. *Mach. Learn.: Sci. Technol.* 4: 025019.

Gubaev, K., Podryabinkin, E.V., and Shapeev, A.V. (2018). Machine learning of molecular properties: locality and active learning. *J. Chem. Phys.* 148, https://doi.org/10.1063/1.5005095.

Gull, S.F. (1989). *Developments in maximum entropy data analysis*. Springer Netherlands, pp. 53–71.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K.Q. (2017). On calibration of modern neural networks. In: *International conference on machine learning*. PMLR, Dordrecht, pp. 1321–1330.

Gustafsson, F.K., Danelljan, M., and Schon, T.B. (2020) Evaluating scalable Bayesian deep learning methods for robust computer vision. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 318–319.

Hastie, T., Tibshirani, R., Friedman, J.H., and Friedman, J.H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, 2. Springer, New York.

Heid, E. and Green, W.H. (2022). Machine learning of reaction properties via learned representations of the condensed graph of reaction. *J. Chem. Inf. Model.* 62: 2101–2110.

Heid, E., McGill, C.J., Vermeire, F.H., and Green, W.H. (2023). Characterizing uncertainty in machine learning for chemistry. *J. Chem. Inf. Model.* 63: 4012–4029.

Heid, E., Schörghuber, J., Wanzenböck, R., and Madsen, G.K. (2024). Spatially resolved uncertainties for machine learning potentials. *J. Chem. Inf. Model*, https://doi.org/10.1021/acs.jcim.4c00904.

Henkel, P. and Mollenhauer, D. (2021). Uncertainty of exchange-correlation functionals in density functional theory calculations for lithium-based solid electrolytes on the case study of lithium phosphorus oxynitride. *J. Comput. Chem.* 42: 1283–1295.

Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R., and Coley, C.W. (2020). Uncertainty quantification using neural networks for molecular property prediction. *J. Chem. Inf. Model.* 60: 3770–3780.

Honarmandi, P. and Arróyave, R. (2020). Uncertainty quantification and propagation in computational materials science and simulation-assisted materials design. *Integrat. Mater. Manuf. Innovat.* 9: 103–143.

Honarmandi, P., Paulson, N.H., Arróyave, R., and Stan, M. (2019). Uncertainty quantification and propagation in calphad modeling. *Model. Simulat. Mater. Sci. Eng.* 27: 034003.

Hu, Y., Musielewicz, J., Ulissi, Z.W., and Medford, A.J. (2022). Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Mach. Learn.: Sci. Technol.* 3: 045028.

Hüllermeier, E. and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.* 110: 457–506.

Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., et al. (2013). Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* 1, https://doi.org/10.1063/1.4812323.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature* 596: 583–589.

Jurečka, P., Šponer, J., Černỳ, J., and Hobza, P. (2006). Benchmark database of accurate (MP2 and CCSD (t) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* 8: 1985–1993.

Kim, H., Yu, N.-K., Tian, N., and Medford, A.J. (2024). Assessing exchange-correlation functionals for heterogeneous catalysis of nitrogen species. *J. Phys. Chem. C* 128: 11159–11175.

Kreitz, B., Sargsyan, K., Blöndal, K., Mazeau, E.J., West, R.H., Wehinger, G.D., Turek, T., and Goldsmith, C.F. (2021). Quantifying the impact of parametric uncertainty on automatic mechanism generation for Co_2 hydrogenation on Ni (111). *JACS Au* 1: 1656–1673.

Kreitz, B., Lott, P., Studt, F., Medford, A.J., Deutschmann, O., and Goldsmith, C.F. (2023). Automated generation of microkinetics for heterogeneously catalyzed reactions considering correlated uncertainties. *Angew. Chem., Int. Ed.* 62: e202306514.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., and Weinberger, K. (Eds.). *Advances in neural information processing systems*, 25. Curran Associates, Inc, Nevada.

Kuleshov, V., Fenner, N., and Ermon, S. (2018) Accurate uncertainties for deep learning using calibrated regression. In: *International conference on machine learning*. PMLR, pp. 2796–2804.

Kullback, S. and Leibler, R.A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22: 79–86.

Kurniawan, Y., Petrie, C.L., Williams Jr, K.J., Transtrum, M.K., Tadmor, E.B., Elliott, R.S., Karls, D.S., and Wen, M. (2022). Bayesian, frequentist, and information geometric approaches to parametric uncertainty quantification of classical empirical interatomic potentials. *J. Chem. Phys.* 156: 214103.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Adv. Neural Inf. Process. Syst.* 30.

Lejaeghere, K., Bihlmayer, G., Björkman, T., Blaha, P., Blügel, S., Blum, V., Caliste, D., Castelli, I.E., Clark, S.J., Dal Corso, A., et al. (2016). Reproducibility in density functional theory calculations of solids. *Science* 351, https://doi.org/10.1126/science.aad3000.

Levi, D., Gispan, L., Giladi, N., and Fetaya, E. (2022). Evaluating and calibrating uncertainty prediction in regression tasks. *Sensors* 22: 5540.

Li, Q., Chen, H., Koenig, B.C., and Deng, S. (2023). Bayesian chemical reaction neural network for autonomous kinetic uncertainty quantification. *Phys. Chem. Chem. Phys.* 25: 3707–3717.

Liu, Y., Kelley, K.P., Vasudevan, R.K., Funakubo, H., Ziatdinov, M.A., and Kalinin, S.V. (2022). Experimental discovery of structure–property relationships in ferroelectric materials via active learning. *Nat. Mach. Intell.* 4: 341–350.

Lu, Y., Wang, B., Chen, S., and Yang, B. (2022). Quantifying the error propagation in microkinetic modeling of catalytic reactions with model-predicted binding energies. *Mol. Catal.* 530: 112575.

MacKay, D.J. (1992a). Bayesian interpolation. *Neural Comput.* 4: 415–447.

MacKay, D.J. (1992b). A practical Bayesian framework for backpropagation networks. *Neural Comput.* 4: 448–472.

Mamun, O., Winther, K.T., Boes, J.R., and Bligaard, T. (2019). High-throughput calculations of catalytic properties of bimetallic alloy surfaces. *Sci. Data* 6: 1–9.

Mardirossian, N. and Head-Gordon, M. (2017). Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Mol. Phys.* 115: 2315–2372.

Medvedev, M.G., Bushmarinov, I.S., Sun, J., Perdew, J.P., and Lyssenko, K.A. (2017). Density functional theory is straying from the path toward the exact functional. *Science* 355: 49–52.

Morris, M.D. (1991). Factorial sampling plans for preliminary computational experiments. *Technometrics* 33: 161–174.

Mortensen, J.J., Kaasbjerg, K., Frederiksen, S.L., Nørskov, J.K., Sethna, J.P., and Jacobsen, K.W. (2005). Bayesian error estimation in density-functional theory. *Phys. Rev. Lett.* 95: 216401.

Motagamwala, A.H. and Dumesic, J.A. (2020). Microkinetic modeling: a tool for rational catalyst design. *Chem. Rev.* 121: 1049–1076.

Neal, R.M. (1993). *Probabilistic inference using Markov chain Monte Carlo methods, Technical report*. University of Toronto, Ontario.

Neal, R.M. (2003). Slice sampling. *Ann. Stat.* 31: 705–767.

Nix, D. and Weigend, A. (1994) Estimating the mean and variance of the target probability distribution. In: *Proceedings of 1994 IEEE international conference on neural networks (ICNN'94)*. IEEE.

Pakornchote, T., Ektarawong, A., and Chotibut, T. (2023). Straintensornet: predicting crystal structure elastic properties using se (3)-equivariant graph neural networks. *Phys. Rev. Res.* 5: 043198.

Pantazis, Y. and Katsoulakis, M.A. (2013). A relative entropy rate method for path space sensitivity analysis of stationary complex stochastic dynamics. *J. Chem. Phys.* 138, https://doi.org/10.1063/1.4789612.

Perdew, J.P. and Schmidt, K. (2001) Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP conference proceedings*, Vol. 577. American Institute of Physics, pp. 1–20.

Perdew, J.P. and Zunger, A. (1981). Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* 23: 5048.

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

Phenix, B.D., Dinaro, J.L., Tatang, M.A., Tester, J.W., Howard, J.B., and McRae, G.J. (1998). Incorporation of parametric uncertainty into complex kinetic mechanisms: application to hydrogen oxidation in supercritical water. *Combust. Flame* 112: 132–146.

Politis, D.N. and Romano, J.P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Stat.*: 2031–2050, https://doi.org/10.1214/aos/1176325770.

Politis, D., Romano, J.P., and Wolf, M. (1999). Weak convergence of dependent empirical measures with application to subsampling in function spaces. *J. Stat. Plann. Inference* 79: 179–190.

Psaros, A.F., Meng, X., Zou, Z., Guo, L., and Karniadakis, G.E. (2023). Uncertainty quantification in scientific machine learning: methods, metrics, and comparisons. *J. Comput. Phys.* 477: 111902.

Rao, Z., Tung, P.-Y., Xie, R., Wei, Y., Zhang, H., Ferrari, A., Klaver, T., Körmann, F., Sukumar, P.T., Kwiatkowski da Silva, A., et al. (2022). Machine learning–enabled high-entropy alloy discovery. *Science* 378: 78–85.

Rasmussen, C.E. (2003). Gaussian processes in machine learning. In: *Summer school on machine learning*. Springer, New York, pp. 63–71.

Reagan, M.T., Najm, H.N., Pébay, P.P., Knio, O.M., and Ghanem, R.G. (2005). Quantifying uncertainty in chemical systems modeling. *Int. J. Chem. Kinet.* 37: 368–382.

Rencher, A.C. and Christensen, W.F. (2012). *Methods of multivariate analysis*, 2nd ed. John Wiley & Sons, Chichester.

Rosen, A.S., Iyer, S.M., Ray, D., Yao, Z., Aspuru-Guzik, A., Gagliardi, L., Notestein, J.M., and Snurr, R.Q. (2021). Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter* 4: 1578–1597.

Ruiz, E., Rodríguez-Fortea, A., Tercero, J., Cauchy, T., and Massobrio, C. (2005). Exchange coupling in transition-metal complexes via density-functional theory: comparison and reliability of different basis set approaches. *J. Chem. Phys.* 123, https://doi.org/10.1063/1.1999631.

Scalia, G., Grambow, C.A., Pernici, B., Li, Y.-P., and Green, W.H. (2020). Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* 60: 2697–2717.

Schienbein, P. (2023). Spectroscopy from machine learning by accurately representing the atomic polar tensor. *J. Chem. Theory Comput.* 19: 705–712.

Schwalbe-Koda, D., Tan, A.R., and Gómez-Bombarelli, R. (2021). Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun.* 12: 5104.

Seitzer, M., Tavakoli, A., Antic, D., and Martius, G. (2022) On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In: *International conference on learning representations*.

Selten, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Exp. Econ.* 1: 43–61.

Sheldon, C., Paier, J., and Sauer, J. (2021). Adsorption of CH4 on the Pt (111) surface: random phase approximation compared to density functional theory. *J. Chem. Phys.* 155, https://doi.org/10.1063/5.0071995.

Sheldon, C., Paier, J., Usvyat, D., and Sauer, J. (2024). Hybrid RPA: DFT approach for adsorption on transition metal surfaces: methane and ethane on platinum (111). *J. Chem. Theory Comput.* 20: 2219–2227.

Sobol, I.M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simulat.* 55: 271–280.

Soleimany, A.P., Amini, A., Goldman, S., Rus, D., Bhatia, S.N., and Coley, C.W. (2021). Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* 7: 1356–1367.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15: 1929–1958.

St. John, P.C., Guan, Y., Kim, Y., Kim, S., and Paton, R.S. (2020). Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* 11: 1–12.

Szaro, N.A., Bello, M., Fricke, C.H., Bamidele, O.H., and Heyden, A. (2023). Benchmarking the accuracy of density functional theory against the random phase approximation for the ethane dehydrogenation network on Pt (111). *J. Phys. Chem. Lett.* 14: 10769–10778.

Tan, A.R., Urata, S., Goldman, S., Dietschreit, J.C.B., and Gómez-Bombarelli, R. (2023). Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Comput. Mater.* 9, https://doi.org/10.1038/s41524-023-01180-8.

Tavazza, F., DeCost, B., and Choudhary, K. (2021). Uncertainty prediction for machine learning models of material properties. *ACS Omega* 6: 32431–32440.

Tian, Y., Xue, D., Yuan, R., Zhou, Y., Ding, X., Sun, J., and Lookman, T. (2021). Efficient estimation of material property curves and surfaces via active learning. *Phys. Rev. Mater.* 5: 013802.

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., and Ulissi, Z.W. (2020). Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn.: Sci. Technol.* 1: 025006.

Tsourtis, A., Pantazis, Y., Katsoulakis, M.A., and Harmandaris, V. (2015). Parametric sensitivity analysis for stochastic molecular systems using information theoretic metrics. *J. Chem. Phys.* 143, https://doi.org/10.1063/1.4922924.

Unke, O.T. and Meuwly, M. (2019). Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theor. Comput.* 15: 3678–3693.

van der Oord, C., Sachs, M., Kovács, D.P., Ortner, C., and Csányi, G. (2023). Hyperactive learning for data-driven interatomic potentials. *npj Comput. Mater.* 9: 168.

Varivoda, D., Dong, R., Omee, S.S., and Hu, J. (2023). Materials property prediction with uncertainty quantification: a benchmark study. *Appl. Phys. Rev.* 10, https://doi.org/10.1063/5.0133528.

Villegas, M., Augustin, F., Gilg, A., Hmaidi, A., and Wever, U. (2012). Application of the polynomial chaos expansion to the simulation of chemical reactors with uncertainties. *Math. Comput. Simulat.* 82: 805–817.

Vita, J.A., Samanta, A., Zhou, F. and Lordi, V. (2024). Ltau-ff: loss trajectory analysis for uncertainty in atomistic force fields, *arXiv preprint arXiv: 2402.00853*.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *Ann. Stat.*: 1378–1402, https://doi.org/10.1214/aos/1176349743.

Walker, E., Ammal, S.C., Terejanu, G.A., and Heyden, A. (2016). Uncertainty quantification framework applied to the water–gas shift reaction over pt-based catalysts. *J. Phys. Chem. C* 120: 10328–10339.

Walker, E.A., Mitchell, D., Terejanu, G.A., and Heyden, A. (2018). Identifying active sites of the water-gas shift reaction over Titania supported platinum catalysts under uncertainty. *ACS Catal.* 8: 3990–3998.

Wang, H. and Sheen, D.A. (2015). Combustion kinetic model uncertainty quantification, propagation and minimization. *Prog. Energy Combust. Sci.* 47: 1–31.

Wang, B., Chen, S., Zhang, J., Li, S., and Yang, B. (2019). Propagating DFT uncertainty to mechanism determination, degree of rate control, and coverage analysis: the kinetics of dry reforming of methane. *J. Phys. Chem. C* 123: 30389–30397.

Wang, A., Kingsbury, R., McDermott, M., Horton, M., Jain, A., Ong, S.P., Dwaraknath, S., and Persson, K.A. (2021). A framework for quantifying uncertainty in DFT energy corrections. *Sci. Rep.* 11: 15496.

Wellendorff, J., Lundgaard, K.T., Møgelhøj, A., Petzold, V., Landis, D.D., Nørskov, J.K., Bligaard, T., and Jacobsen, K.W. (2012). Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B – Condens. Matter Mater. Phys.* 85: 235149.

Wen, M. (2019). *Development of interatomic potentials with uncertainty quantification: applications to two-dimensional materials, PhD thesis*. University of Minnesota, Minnesota.

Wen, M. and Tadmor, E.B. (2020). Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj Comput. Mater.* 6: 124.

Wen, M., Shirodkar, S.N., Plecháč, P., Kaxiras, E., Elliott, R.S., and Tadmor, E.B. (2017). A force-matching stillinger-weber potential for MoS2: parameterization and fisher information theory based sensitivity analysis. *J. Appl. Phys.* 122, https://doi.org/10.1063/1.5007842.

Wen, M., Blau, S.M., Spotte-Smith, E.W.C., Dwaraknath, S., and Persson, K.A. (2021). Bondnet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* 12: 1858–1868.

Wen, M., Blau, S.M., Xie, X., Dwaraknath, S., and Persson, K.A. (2022). Improving machine learning performance on small chemical reaction data with unsupervised contrastive pretraining. *Chem. Sci.* 13: 1446–1458.

Wen, M., Spotte-Smith, E.W.C., Blau, S.M., McDermott, M.J., Krishnapriyan, A.S., and Persson, K.A. (2023). Chemical reaction networks and opportunities for machine learning. *Nat. Comput. Sci.* 3: 12–24.

Wen, M., Horton, M.K., Munro, J.M., Huck, P., and Persson, K.A. (2024). An equivariant graph neural network for the elasticity tensors of all seven crystal systems. *Digit. Discov.* 3: 869–882.

Wick, A., Felix, D., Steen, K., and Eschenmoser, A. (1964). Claisen'sche umlagerungen bei allyl-und benzylalkoholen mit hilfe von acetalen des n, n-dimethylacetamids. vorläufige mitteilung. *Helv. Chim. Acta* 47: 2425–2429.

Wiener, N. (1938). The homogeneous chaos. *Am. J. Math.* 60: 897–936.

Wu, Z., Ramsundar, B., Feinberg, E.N., Gomes, J., Geniesse, C., Pappu, A.S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.* 9: 513–530.

Xu, W. and Yang, B. (2023). Microkinetic modeling with machine learning predicted binding energies of reaction intermediates of ethanol steam reforming: the limitations. *Mol. Catal.* 537: 112940.

Zahrt, A.F., Henle, J.J., Rose, B.T., Wang, Y., Darrow, W.T., and Denmark, S.E. (2019). Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* 363, https://doi.org/10.1126/science.aau5631.

Zaverkin, V., Holzmüller, D., Christiansen, H., Errica, F., Alesiani, F., Takamoto, M., Niepert, M., and Kästner, J. (2024). Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials. *npj Comput. Mater.* 10: 83.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*, 1st ed. Chapman & Hall/CRC, Florida.

Zhu, A., Batzner, S., Musaelian, A., and Kozinsky, B. (2023). Fast uncertainty estimates in deep learning interatomic potentials. *J. Chem. Phys.* 158, https://doi.org/10.1063/5.0136574.