nature communications

Article

https://doi.org/10.1038/s41467-025-67663-y

Evidential deep learning for interatomic potentials

Received: 15 May 2025

Accepted: 5 December 2025

Published online: 20 December 2025

Check for updates

Han Xu[1,2,7], Taoyong Cui[1,3,7], Chenyu Tang[1,7], Jinzhe Ma[1,4], Dongzhan Zhou[1], Yuqiang Li[1], Xiang Gao[2], Xingao Gong[5], Wanli Ouyang[1], Shufei Zhang[1] & Mao Su[1,6]

Machine learning interatomic potentials have been widely used to facilitate large-scale molecular simulations with accuracy comparable to ab initio methods. To ensure the reliability of the simulation, the training dataset is iteratively expanded through active learning, where uncertainty serves as a critical indicator for identifying and collecting out-of-distribution data. However, existing uncertainty quantification methods tend to involve either expensive computations or compromise prediction accuracy. Here we show an evidential deep learning framework for interatomic potentials with a physics-inspired design. Our method provides uncertainty quantification without significant computational overhead or decreased prediction accuracy, consistently outperforming other methods across a variety of datasets. Furthermore, we demonstrate applications in exploring diverse atomic configurations, using examples including water and universal potentials. These results highlight the potential of our method as a robust and efficient alternative for uncertainty quantification in molecular simulations.

Molecular dynamics (MD) simulation provides atomic insights into physical and chemical processes and has become an indispensable research tool in computational physical science[1–3]. Classical MD simulation uses an empirical potential function to determine interatomic forces[4,5], which is computationally efficient but not accurate enough, especially when polarization or many-body interactions are important[6]. In contrast, the ab initio approach for modeling atomic interactions is based solely on fundamental physical principles, leading to generally higher accuracy and transferability[7,8], but the high computational cost limits the size of systems that can be simulated. To achieve both efficiency and accuracy, machine learning interatomic potentials (MLIPs) have been proposed[9–12], which allow to learn ab initio interatomic potentials and performing MD simulations with much lower computational cost. MLIPs have been successfully applied in the study of amorphous solid[13], catalysis[14], chemical reaction[15], and more.

One of the primary challenges to MLIP-based MD simulations lies in the construction of the training dataset, which should include various configurations that may appear during the simulation. Inadequate training data will lead to decreased accuracy or even failure of the simulations[16,17]. This challenge limits the application of MLIP-based MD simulations. Active learning based on uncertainty quantification (UQ) plays a crucial role in constructing training sets for MLIPs[18–21]. During active learning, configurations with higher uncertainties are sampled to enrich the training set. This process usually needs to be repeated dozens or more times[19], and the computational cost required for UQ could be considerable. Therefore, a robust yet efficient method for UQ is desired.

[1]Shanghai Artificial Intelligence Laboratory, Shanghai, China. [2]The State Key Laboratory of Chemical Engineering, College of Chemical and Biological Engineering, Zhejiang University, Hangzhou, China. [3]The Chinese University of Hong Kong, Hong Kong, China. [4]School of Physical Science and Technology, ShanghaiTech University, Shanghai, China. [5]Key Laboratory of Computational Physical Sciences (Ministry of Education), Institute of Computational Physical Sciences, State Key Laboratory of Surface Physics, Department of Physics, Fudan University, Shanghai, China. [6]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. [7]These authors contributed equally: Han Xu, Taoyong Cui, Chenyu Tang.
e-mail: zhangshufei@pjlab.org.cn; sumao@pjlab.org.cn

A variety of UQ methods have been developed for MLIPs. Moment tensor potential[22] uses an extrapolation parameter to estimate uncertainty, but this method does not apply to deep neural network models. Gaussian approximation potential[23] utilizes Gaussian process regression to provide UQ along with its predictions. However, the primary limitation of the Gaussian approximation potential lies in its computational cost, which scales cubically with the dataset size. Ensemble methods[24] are quite reliable for UQ, but also suffer from computational burdens due to the training of multiple models. The computational issue of the ensemble methods can be alleviated by weight-sharing. For example, Kellner et al. proposed direct propagation of shallow ensembles (DPOSE)[25]. However, this approximation may worsen the known issue of ensemble overconfidence, especially in larger models[26]. Single-model methods, such as Monte Carlo dropout[27–29], Gaussian mixture models (GMM)[30], and mean-variance estimation (MVE)[31], mitigate the computational issue, but their performances are still not satisfactory[32]. The development of efficient UQ has been advanced by methods such as loss trajectory analysis for uncertainty (LTAU)[33] and last-layer prediction rigidity (LLPR)[34]. LTAU extracts uncertainty from training trajectories. LLPR proposes a prediction rigidity formalism to obtain uncertainties and employs several approximations to reduce computational complexity.

Evidential deep learning[35,36] is a promising alternative, which estimates uncertainty through a single forward pass and requires minimal extra computational resources. Another advantage of evidential deep learning is that it can estimate aleatoric and epistemic uncertainties separately. Aleatoric uncertainty arises from intrinsic noise in the data and cannot be evaded or reduced. In contrast, epistemic uncertainty reflects the fidelity of the model in its representation of the data (excluding aleatoric effects) and decreases as the number of training samples increases[37]. The ability of evidential deep learning to distinguish between these two types of uncertainty is particularly beneficial for active learning, where we want to sample data with high epistemic uncertainty rather than aleatoric uncertainty. However, recent attempts[32,38] trying to integrate evidential deep learning with MLIPs result in unsatisfactory performance. Failures may be attributed to inappropriate design in model architecture.

In this work, we reexamine the uncertainty associated with MLIPs from a physical perspective and propose a framework for UQ based on evidential deep learning. We call this framework the evidential interatomic potential (eIP). The performance of eIP is evaluated across various datasets and benchmarked with other UQ methods, demonstrating outstanding performance with minimal additional computational cost. We then extend the application of eIP to uncertainty-driven dynamics (UDD) simulations[39–42], enabling the efficient exploration of the diverse atomic configurations. This approach (UDD) modifies the potential energy surface by assigning lower energies to high-uncertainty configurations, thereby making them more accessible. Lastly, we use eIP to train a universal potential and demonstrate its ability for concurrent UQ during simulations. While other UQ methods can achieve this, eIP offers advantages in efficiency and reliability.

Results

Preliminary

Machine learning interatomic potential (MLIP). MLIPs are used to predict energy and forces within a given atomic configuration. For a system comprising N atoms, MLIPs typically take the atomic species $Z \in \mathbb{Z}^N$ and coordinates $R \in \mathbb{R}^{N\times 3}$ as input and output the total potential energy E. The forces $F \in \mathbb{R}^{N\times 3}$ exerted on the atoms are derived by calculating the negative gradient of E with respect to the coordinates. The primary distinction among various MLIPs lies in the algorithm used to convert the input information into vectorized features that represent the local atomic environments. These features are designed to be invariant or equivariant under translation, rotation, and permutation.

Aleatoric and epistemic uncertainty. Two categories of uncertainty can be modeled in deep learning[43]. Aleatoric uncertainty arises from inherent noise in data labels. In the context of MLIPs, data labels are obtained from ab initio calculations. Although highly rigorous ab initio calculations could limit the aleatoric uncertainty, they are often computationally prohibitive in practice. As a result, MLIPs are often trained on multiple datasets of varying quality, which introduces label noise. For example, in the MPtrj dataset, aleatoric uncertainty may stem from inconsistent Hubbard U correction or varying convergence criteria[26]. In contrast, epistemic uncertainty is caused by a lack of knowledge, typically due to insufficient data. This type of uncertainty can be reduced by adding more training data, often through active learning. For the sake of simplicity, the term uncertainty in the following results refers to epistemic uncertainty, unless otherwise specified. We further discuss aleatoric uncertainty in Supplementary Note 2.

Evidential deep learning. Evidential deep learning is an efficient method to estimate the uncertainty of the results predicted by neural networks. Starting from a maximum likelihood perspective, the targets are assumed to be drawn from a Gaussian distribution but with unknown mean and variance (μ, σ^2). A Gaussian prior is placed on the unknown mean μ and an Inverse-Gamma prior on the unknown variance σ^2, leading to the Normal Inverse-Gamma distribution with a set of parameters $\mathbf{m} = (\gamma, \nu, \alpha, \beta)$. Neural networks are then trained to infer $\mathbf{m}$, and the prediction, aleatoric, and epistemic uncertainty are calculated as[35]:

$$\underbrace{\mathbb{E}[\mu] = \gamma}_{\text{prediction}}, \quad \underbrace{\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}}_{\text{aleatoric}}, \quad \underbrace{\mathrm{Var}[\mu] = \frac{\beta}{\nu(\alpha - 1)}}_{\text{epistemic}}. \qquad (1)$$

Framework of eIP

As illustrated in Fig. 1, eIP consists of an MLIP block for predicting energy and force, and an evidential quantile regression block for UQ. The eIP framework extends a regular MLIP by feeding its equivariant features into a lightweight network that outputs prior parameters **m**. The loss function is central to eIP. We used a mean absolute error (MAE) loss for energy fitting, whereas force predictions are optimized by a composite loss that combines negative log-likelihood (NLL) with a regularization term. Further details are provided in the "Methods" section. In designing eIP, we considered the following points, which are indispensable to achieving robust performances.

Locality. In most MLIPs, the potential energy is calculated as the sum of atomic contributions, $E = \sum_{i=1}^{N} E_i$, with the model learning the mapping from the local environment of the atom i to E_i. Therefore, we estimate the uncertainty associated with E_i rather than the total potential energy E. However, we do not have the ground truth for E_i. Fortunately, we can adapt the atomic forces instead of E_i to estimate the uncertainty per atom.

Directionality. We attribute uncertainty in MLIP predictions to the inadequate learning of the local atomic environment. Consequently, this uncertainty is directionally dependent. Our model produces a separate uncertainty value for each Cartesian component of the atomic force, rather than a single value per atom or system. This is illustrated using a three-atom toy system in Supplementary Note 1, where the uncertainty varies with direction. Beyond MLIPs, the directional dependence is also crucial for predicting other non-scalar properties, such as dipole moments, dielectric tensors, and Hamiltonians. In the following experiments, we employ the equivariant backbone PaiNN[44] to extract equivariant features and output the parameters of the Normal Inverse-Gamma prior. We also apply eIP to other equivariant backbones to demonstrate the generality, and the results are provided in Supplementary Note 5.

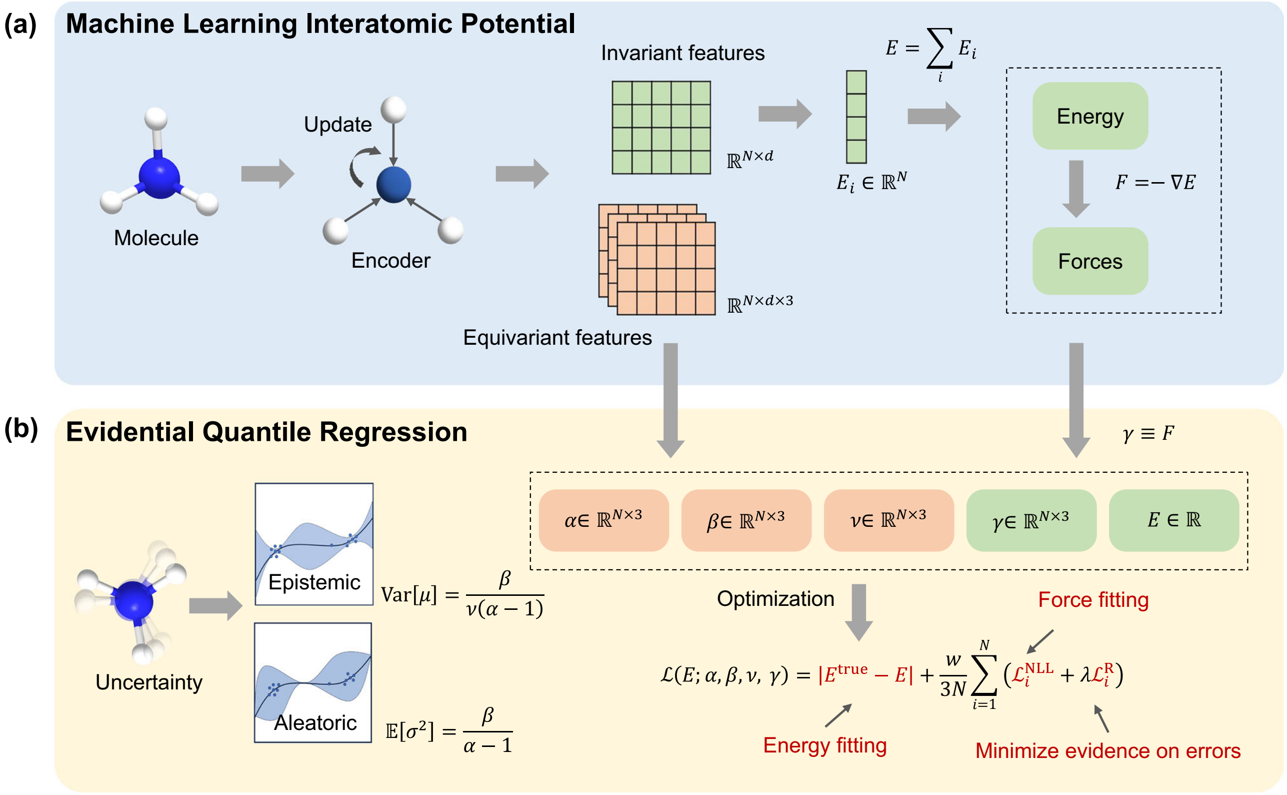
(a)
Machine Learning Interatomic Potential
Molecule
Update
Encoder
Invariant features
$\mathbb{R}^{N\times d}$
Equivariant features
$\mathbb{R}^{N\times d\times 3}$
$E = \sum_i E_i$
$E_i \in \mathbb{R}^N$
Energy
$F = -\nabla E$
Forces
(b)
Evidential Quantile Regression
$\gamma \equiv F$
$\alpha \in \mathbb{R}^{N\times 3}$
$\beta \in \mathbb{R}^{N\times 3}$
$\nu \in \mathbb{R}^{N\times 3}$
$\gamma \in \mathbb{R}^{N\times 3}$
$E \in \mathbb{R}$
Uncertainty
Epistemic
$\mathrm{Var}[\mu] = \frac{\beta}{\nu(\alpha-1)}$
Aleatoric
$\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha-1}$
Optimization
Force fitting
$\mathcal{L}(E;\alpha,\beta,\nu,\gamma) = |E^{\mathrm{true}} - E| + \frac{w}{3N}\sum_{i=1}^{N}(\mathcal{L}_i^{\mathrm{NLL}} + \lambda\mathcal{L}_i^{\mathrm{R}})$
Energy fitting
Minimize evidence on errors

Fig. 1 | Framework of eIP. a A typical equivariant interatomic potential model extracting both invariant and equivariant features. The invariant features are used to output the potential energy. **b** Evidential quantile regression. The equivariant features are used to output the parameters for uncertainty quantification. The model is optimized using a mean absolute error (MAE) loss for energy, and a composite loss that combines negative log-likelihood (NLL) with a regularization term for force.

Quantile regression. The original evidential deep learning framework for regression[35] assumes that the targets are drawn from a Gaussian distribution, which may not adequately describe the target distribution of MLIPs. To alleviate this limitation, we adopt the Bayesian quantile regression framework[45] that combines evidential deep learning with quantile regression, enabling UQ without relying on Gaussian assumptions. Unlike traditional regression that minimizes mean squared error, our approach learns a specified quantile q of the force distribution. This is achieved by employing an asymmetric Laplace distribution, which is subsequently reformulated as a scale mixture of Gaussians to facilitate Bayesian inference. The computational procedure of Bayesian quantile regression is analogous to the original evidential framework. The key difference lies in the loss functions, which are specifically designed for quantile estimation. Our implementation details are provided in the "Methods" section.

Experiments

ISO17 dataset. We started by assessing the performance of eIP using the ISO17 dataset, which comprises MD trajectories of $C_7O_2H_{10}$ isomers. This dataset is divided into in-distribution (ID) and out-of-distribution (OOD) subsets, making it particularly suitable for uncertainty quantification (UQ). In the ID scenario (known molecules/unknown conformations), the test molecules are also present in the training set. In contrast, the OOD scenario (unknown molecules/unknown conformations) involves test molecules that are not in the training set. The training set contains 400,000 conformations, which is a substantial amount for such small molecules. Therefore, we also explore the impact of training data volume. Specifically, we train the model using 1%, 5%, 30%, and 100% of the training data, respectively. Figure 2a–d shows the scatter plots that compare uncertainties with force errors for different amounts of training data, demonstrating positive correlations in both ID and OOD scenarios. The mean uncertainty and mean absolute error (MAE) for force predictions are shown in Fig. 2e, f, respectively. As expected, both metrics decrease with an increase in the amount of training data. Furthermore, we evaluated the reliability of UQ using additional metrics, including Spearman's rank correlation coefficient and the area under the receiver operating characteristic curve (ROC-AUC), whose definitions are provided in the "Methods" section. As shown in Fig. 2g, h, both Spearman's rank correlation coefficient and ROC-AUC improve as the amount of training data grows. In the ID scenario, Spearman's rank correlation coefficients ranging from 0.74 to 0.86 and ROC-AUC values ranging from 0.86 to 0.93 indicate the strong performance of eIP. In the OOD scenario, although the test set molecules are absent from the training set, the evaluation metrics still provide reasonable uncertainty estimates.

Silica glass dataset. We then evaluate eIP's performance for more complex systems using a silica glass dataset, which comprises large bulk structures. Given the challenges in partitioning large structures into ID and OOD datasets, we adopted the dataset partition scheme consistent with the previous study[32]. We also compare eIP with other UQ methods, including ensemble, Monte Carlo dropout, Gaussian mixture model (GMM), and Mean-variance estimation (MVE), whose implementations are provided in Supplementary Note 6. Figure 3a shows the scatter plots of uncertainties versus force errors and indicates that all methods achieve positive correlations. Figure 3b presents the computational efficiency analysis of the five methods. Despite the good performance of the ensemble method, it incurs higher computational costs during both training and inference due to the need for

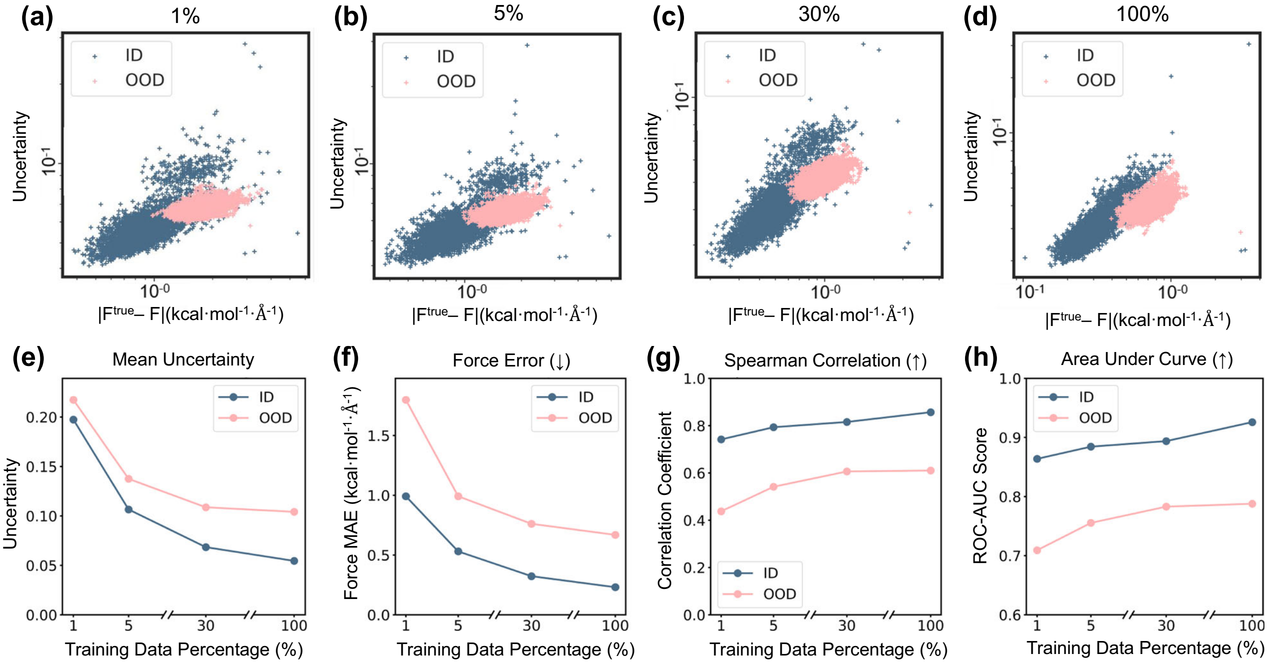

Fig. 2 | Results on ISO17 dataset with increasing data volume. a–**d** Scatter plots of uncertainties versus force errors using 1%, 5%, 30%, and 100% of the training data, respectively. Each point corresponds to the averaged uncertainty/error in a molecule. **e** Mean uncertainty on the test set. **f** Force mean absolute errors (MAEs) on the test set. **g** Spearman's rank correlation coefficient between uncertainty and force error. **h** ROC-AUC scores.

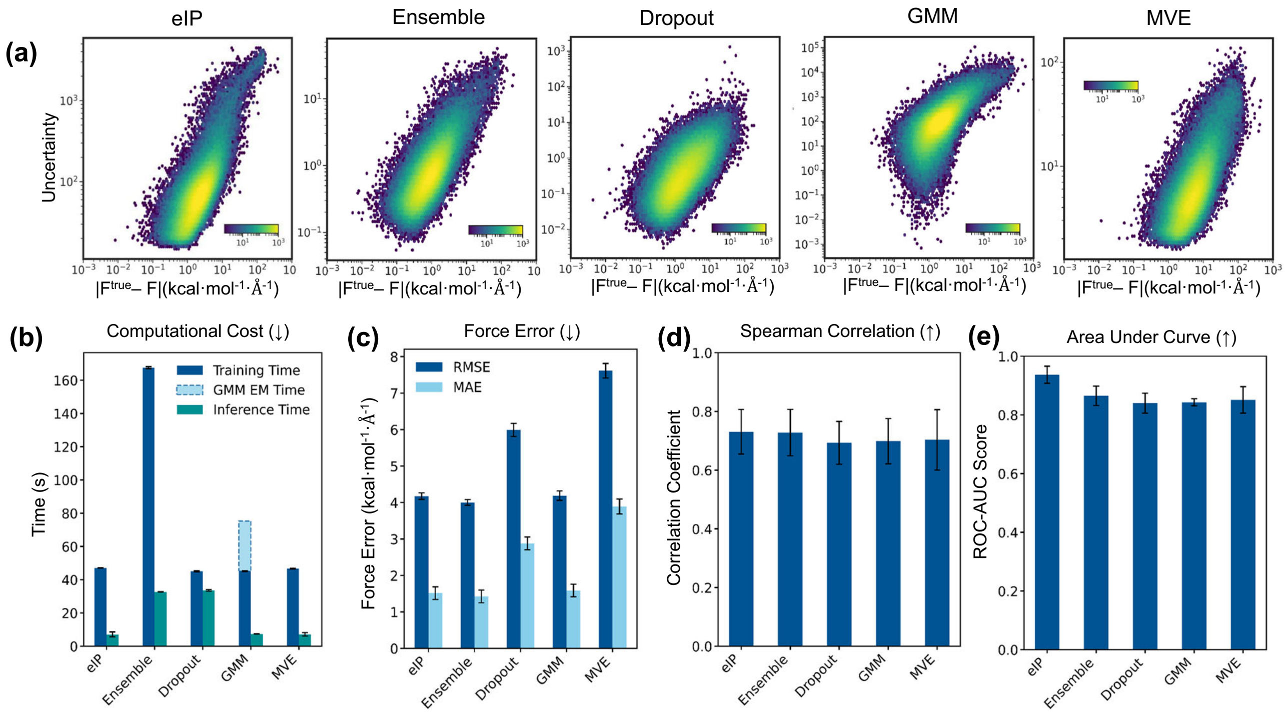

Fig. 3 | Comparing eIP with other uncertainty quantification methods on silica glass dataset. a Hexbin plots of uncertainties versus atomic force errors. **b** Computational costs. The training time here refers to the time required for each epoch. The inference time includes the time cost of computing uncertainty. The light blue dashed bar denotes the time cost of the expectation-maximization (EM) algorithm. **c** Force mean absolute errors (MAEs) and root mean square errors (RMSEs) on the test set. **d** Spearman's rank correlation coefficient between uncertainty and force error. **e** ROC-AUC scores. While all five methods achieve strong Spearman's rank correlations and ROC-AUC scores, ensemble, dropout, and GMM require longer computation times; dropout and MVE exhibit much lower accuracy in force prediction. Error bar denotes the standard deviation from 5 independent experiments. Source data are provided as a Source Data file.

multiple independent MLIPs. Since our implementation uses four independent models, training and inference times are approximately four times longer than for a single model. It is worth noting that the actual time cost is highly dependent on implementation. In practice, the actual time required to train an ensemble can be less than the number of models implies. The Monte Carlo dropout method requires multiple inferences to obtain uncertainty, and the corresponding computational cost is similar to the ensemble method. GMM quantifies uncertainty through an iterative expectation-maximization algorithm and therefore requires additional computation time. Both MVE and eIP have minimal training and inference times, comparable to those of a normal MLIP. Regarding the force prediction accuracy shown in Fig. 3c, ensemble, GMM, and eIP achieve the lowest errors, while dropout and MVE exhibit larger errors. Figure 3d, e further illustrate the comparison of Spearman's correlation and ROC-AUC, respectively. Notably, Fig. 3e shows that eIP performs even better than the ensemble method on the ROC-AUC metric.

Applications

Active learning with eIP. UQ plays a key role in active learning for training set construction. The quality of the training set is particularly crucial for MLIP, as the accuracy of MLIPs can significantly decrease when encountering unseen atomic configurations, leading to the collapse of simulations[16]. Figure 4a illustrates a typical active learning workflow for MLIPs, where the data points with high uncertainty are iteratively explored to enrich the training set. In addition, uncertainty-driven dynamics (UDD) simulation[39] can be employed to enhance sampling efficiency. In UDD simulations, the potential energy surface is modified so that the atomic configurations with higher uncertainties are assigned lower potential energies, and consequently, these structures become more accessible, as indicated in Fig. 4b. The implementation of UDD simulation with eIP is provided in the "Methods" section.

We demonstrate the active learning process with eIP, using a water dataset as an example. In each iteration, we performed standard

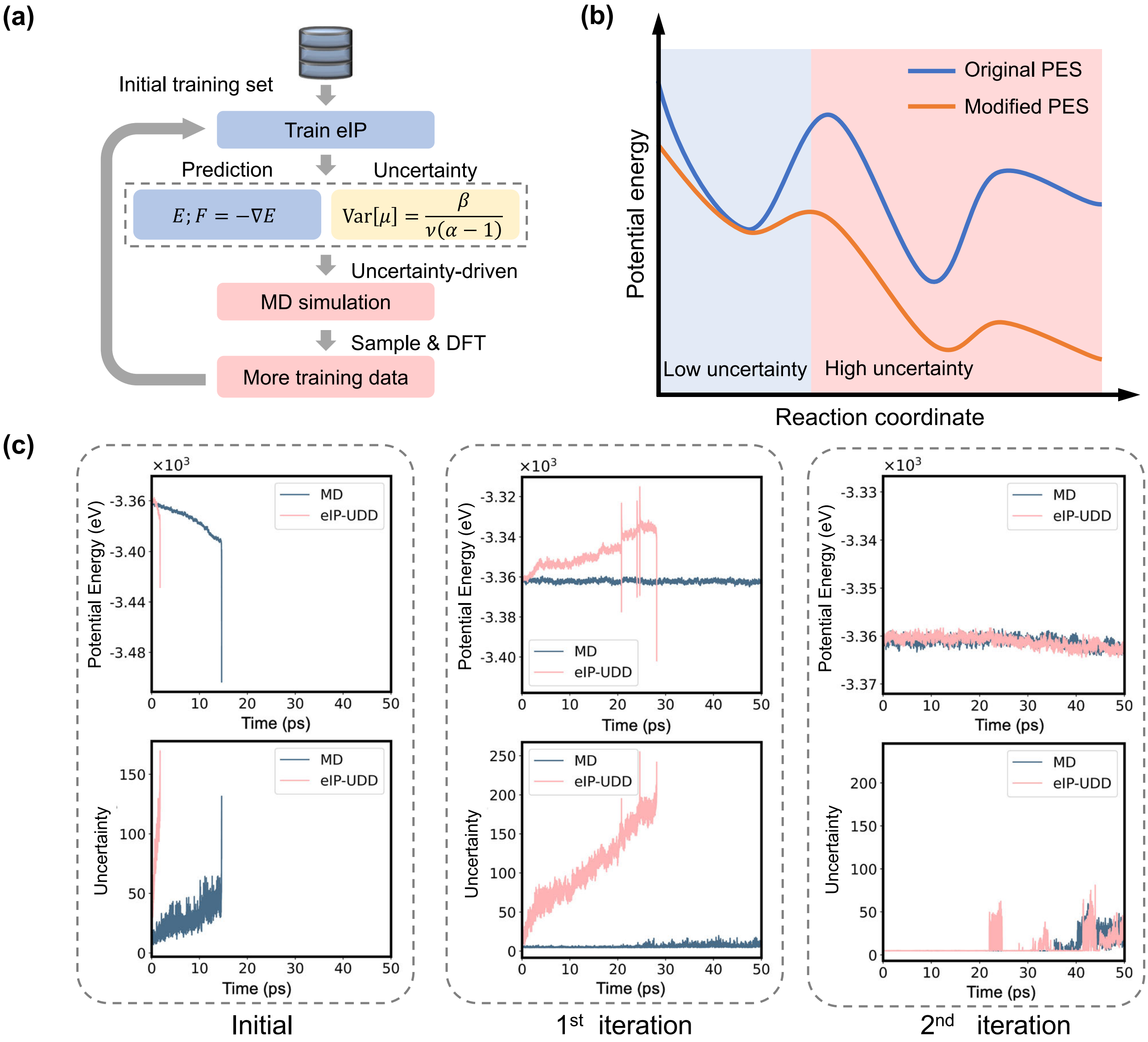

Fig. 4 | Active learning with eIP. a Workflow. Potential energy and uncertainty are calculated simultaneously by eIP. **b** Illustration of uncertainty-driven dynamics (UDD). The potential energy surface (PES) is adaptively modified according to uncertainty, with the potential energy in high-uncertainty regions being reduced to facilitate enhanced sampling. **c** Simulation results in each generation. The evolution of potential energy and uncertainty over time is shown for both conventional MD and eIP-UDD simulations. The uncertainty for each configuration is the mean atomic uncertainty determined by Eq. (7). In MD simulations, the PES remains unmodified, whereas in eIP-UDD simulations, the PES is modified based on the uncertainty from eIP.

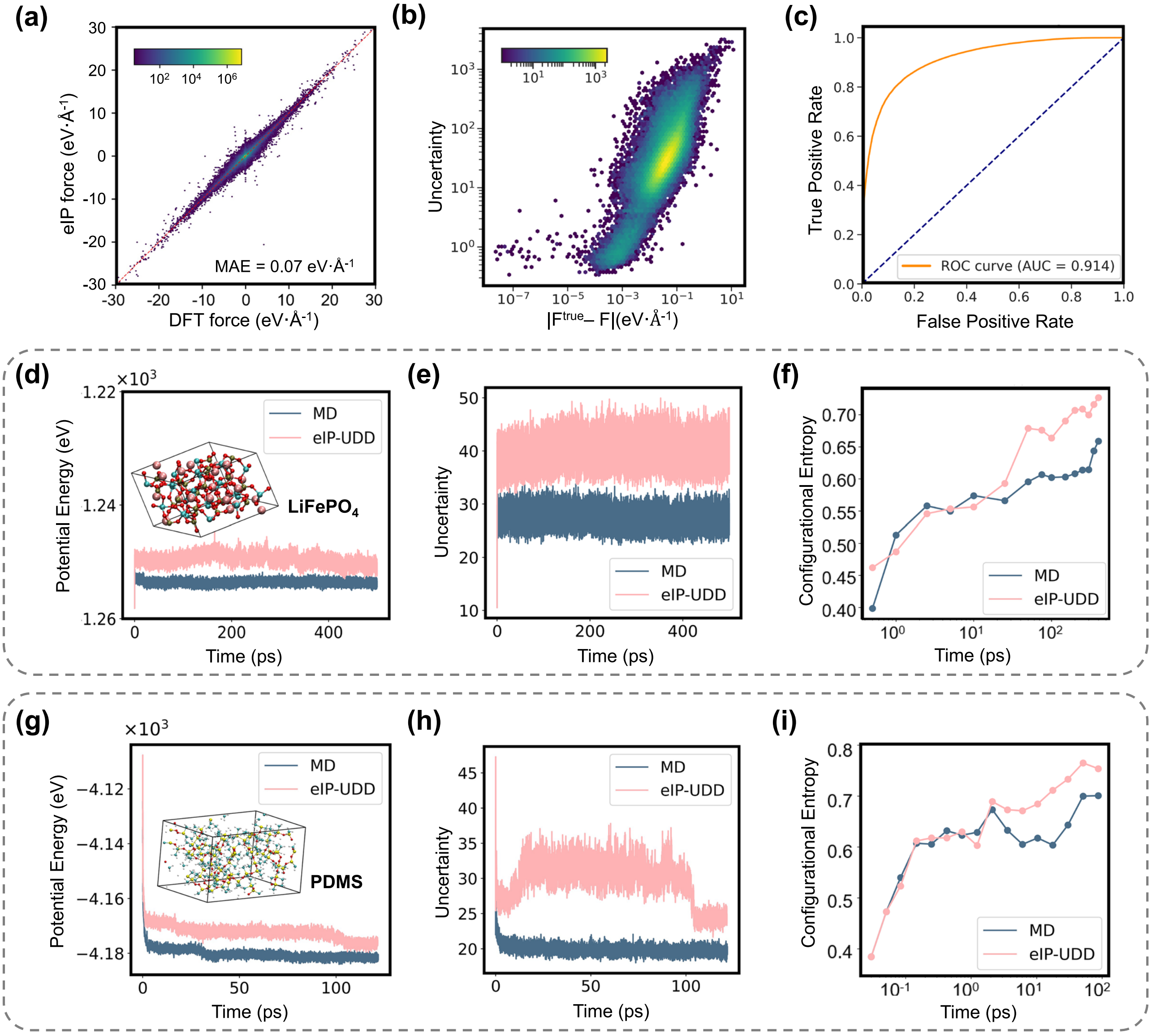

Fig. 5 | Universal potential with eIP. a Comparison of atomic forces between eIP prediction and ground truth. The model is trained on the Materials Project Trajectory (MPtrj) dataset[46]. **b** Hexbin plots of uncertainties versus atomic force errors. The Spearman's rank correlation coefficient is 0.76. **c** ROC curve. The ROC-AUC score is 0.914. **d**–**f** Simulation results of $LiFePO_4$. **g**–**i** Simulation results of polydimethylsiloxane (PDMS). The potential energy curves (**d**) and (**g**) indicate that both MD and eIP-UDD simulations are stable, demonstrating the effectiveness of the universal potential. The uncertainty curves (**e**) and (**h**) reveal that eIP-UDD configurations exhibit higher uncertainty levels for both materials. The evolutions of configurational entropy (**f**) and (**i**) further confirm that eIP-UDD simulations generate more diverse configurations than conventional MD simulations.

MD and UDD simulations using our eIP model. We refer to the UDD simulation driven by the uncertainty quantified by eIP as the eIP-UDD simulation. Changes in uncertainty and energy over simulation time are illustrated in Fig. 4c, where the uncertainty for each configuration is the mean atomic uncertainty determined by Eq. (7). The initial training set comprises 1000 configurations sampled from a classical MD simulation trajectory generated using an empirical force field. The abnormal energy fluctuations suggest that both the MD and eIP-UDD simulations collapse very early. After the first iteration, the MD simulation remains stable for the full 50 ps. Although the eIP-UDD simulation collapses after 20 ps, the uncertainty increases over time, indicating that more previously unseen configurations are explored during the eIP-UDD simulation. After the second iteration, both the MD and the eIP-UDD simulations remain stable for the full 50 ps. We also observe that the uncertainty does not increase significantly, and this may suggest that configurations are explored sufficiently around certain local minima. The UDD parameters can be adjusted to further explore a broader configuration space. Details of the settings of UDD parameters are provided in Supplementary Note 7 and Supplementary Figs. 9–12.

Application of eIP in universal MLIP. Finally, we explored the performance of eIP in universal MLIPs. To this end, we trained the model on the Materials Project Trajectory (MPtrj) dataset[46]. The hexibin plots and the ROC curve in Fig. 5a–c demonstrate the performance of eIP on such a large dataset. We then conducted eIP-UDD simulations to test the performance of eIP in enhanced sampling. We selected two distinct materials as examples, namely lithium iron phosphate ($LiFePO_4$) and polydimethylsiloxane (PDMS). $LiFePO_4$ is a mature commercial cathode material for lithium-ion batteries, while PDMS is a widely applied organosilicon polymer material. These materials serve as benchmarks for evaluating the configurational sampling performance of eIP-UDD simulations for both inorganic crystalline and organic polymeric systems. To evaluate the diversity of the generated configurations, we calculated the configurational entropy for each trajectory, as detailed in the "Methods" section. For each material, changes in potential energy, uncertainty, and configurational entropy over simulation time are shown in Fig. 5d–i. In Fig. 5d, the initial $LiFePO_4$ configuration was a pre-optimized structure sourced from the Materials Project. No significant decrease in potential energy was observed at the beginning of the simulation. The brief rise in potential energy during the early stage of the eIP-UDD simulation occurs due to the modified potential energy surface that forces the system to escape the original local minimum. The initial PDMS configuration in Fig. 5g was built in-house (see "Methods") and not pre-relaxed. The decrease in potential energy corresponds to the structural relaxation process. In Fig. 5e, h, the trajectory of the eIP-UDD simulation has a higher uncertainty than that of the conventional MD simulation, as expected. The results of the configurational entropy in Fig. 5f, i further prove that the eIP-UDD simulations have obtained more diverse configurations.

Discussions

UQ is a critical topic in various fields of machine learning, particularly in scientific applications such as molecular simulations based on MLIP. Conventional UQ methods suffer from either high computational costs or decreased prediction accuracy. In this work, we propose a single-model UQ method, called eIP, which achieves both efficiency and accuracy, as demonstrated by extensive experiments in various applications. The eIP framework incorporates locality, directionality, and quantile regression, all of which are essential for achieving optimal results. This is evident from the ablation study presented in Supplementary Note 3, where the absence of any single component leads to a noticeable decline in performance.

Although ensemble methods have been widely used in active learning, they typically require training four or more models simultaneously. In practice, this process usually involves dozens or more iterations and takes a significant amount of time and computational resources to obtain a satisfactory training set. As a result, single-model UQ methods, such as eIP, have the potential to save several months in applications, making eIP a more efficient alternative when time constraints and computational resources are a significant concern. In addition, for large-scale simulations, ensemble methods require a significant amount of computation to evaluate the reliability of MLIP-based MD simulations, while eIP facilitates real-time assessment without incurring noticeable additional costs.

Methods

Formulism of eIP

We employ quantile regression with maximum likelihood estimation to better model the uncertainty of MLIPs. Quantile regression is solved by minimizing the tiled loss for a given quantile q:

$$\mathcal{L}_i = \rho_q(\epsilon_i) = \max(q\epsilon_i, (q-1)\epsilon_i), \tag{2}$$

where ϵ_i denotes the residual for observation i.

The quantile q follows an asymmetric Laplace distribution with mean μ, variance σ, and an asymmetrical parameter equal to the quantile q[47]. The likelihood function can be expressed as a scalar mixture of Gaussians[48,49] $\mathcal{N}(\mu+\tau z, \omega\sigma z)$, where $\tau = \frac{1-2q}{q(1-q)}$, $\omega = \frac{2}{q(1-q)}$, $z \sim \exp\left(\frac{1}{\sigma}\right)$.

We assume that the atomic forces $F \in \mathbb{R}^{N\times 3}$ come from a Gaussian distribution, but the mean and variance are unknown. For instance, the x-component of the force on the atom i follows:

$$f_{ix} \sim \mathcal{N}(\mu_{ix} + \tau z_{ix}, \omega\sigma_{ix} z_{ix}). \tag{3}$$

By placing a Gaussian prior on the unknown mean μ_{ix} and an Inverse-Gamma prior on the unknown variance σ_{ix}^2, we obtain the Normal-Inverse-Gamma (NIG) evidential prior $p(\mu_{ix}, \sigma_{ix}|\mathbf{m}_{ix})$ with a set of parameters $\mathbf{m}_{ix} = (\gamma_{ix}, \nu_{ix}, \alpha_{ix}, \beta_{ix})$[35,45]. As a result, γ_{ix} is equal to the predicted force

$$\mathbb{E}[\mu_{ix}] = \int_{\mu_{ix}=-\infty}^{\infty} \mu_{ix} p(\mu_{ix})\, \mathrm{d}\,\mu_{ix} = \gamma_{ix}, \tag{4}$$

and the x-component of epistemic uncertainty for the atom i is

$$\mathrm{Var}[\mu_{ix}] = \int_{\mu_{ix}=-\infty}^{\infty} \mu_{ix}^2 p(\mu_{ix})\, \mathrm{d}\,\mu_{ix} - (\mathbb{E}[\mu_{ix}])^2 = \frac{\beta_{ix}}{\nu_{ix}(\alpha_{ix}-1)}. \tag{5}$$

The y- and z-components are computed similarly. We define the uncertainty σ_i associated with the atom i as

$$\sigma_i = \sqrt{\left(\frac{\beta_{ix}}{\nu_{ix}(\alpha_{ix}-1)}\right) + \left(\frac{\beta_{iy}}{\nu_{iy}(\alpha_{iy}-1)}\right) + \left(\frac{\beta_{iz}}{\nu_{iz}(\alpha_{iz}-1)}\right)}. \tag{6}$$

The uncertainty for a configuration composed of N atoms is determined by computing the average:

$$\sigma = \frac{1}{N}\sum_{i=1}^{N} \sigma_i. \tag{7}$$

It should be noted that averaging can lead to a loss of local information. While the maximum uncertainty value is an alternative, it is susceptible to intrinsic errors in uncertainty quantification. The weighted quantile is a more robust choice, but it requires careful selection of the weights.

The model is trained by maximizing the probability $p(f_{ix}|\mathbf{m}_{ix})$. Marginalizing out the likelihood parameters μ_{ix} and σ_{ix}^2:

$$p(f_{ix}|\mathbf{m}_{ix}) = \int_{\sigma_{ix}^2=0}^{\infty}\int_{\mu_{ix}=-\infty}^{\infty} p(f_{ix}|\mu_{ix}, \sigma_{ix}^2) p(\mu_{ix}, \sigma_{ix}^2|\mathbf{m}_{ix})\, \mathrm{d}\mu_{ix}\mathrm{d}\,\sigma_{ix}^2. \tag{8}$$

By placing the NIG prior on μ_{ix} and σ_{ix}^2, this integral has an analytical solution:

$$p(f_{ix}|\mathbf{m}_{ix}) = \mathrm{St}\left(f_{ix}; \gamma_{ix} + \tau z_{ix}, \frac{2\beta_{ix}(1+\omega\nu_{ix}z_{ix})}{\nu_{ix}\alpha_{ix}}, 2\alpha_{ix}\right), \tag{9}$$

where $\mathrm{St}(f; \mu_{\mathrm{St}}, \sigma_{\mathrm{St}}^2, \upsilon_{\mathrm{St}})$ is the Student t-distribution evaluated at f with location parameter μ_{St}, scale parameter σ_{St}^2, and υ_{St} degree of freedom. We then obtain the negative log-likelihood (NLL) loss function[45]:

$$\begin{aligned}\mathcal{L}_{ix}^{\mathrm{NLL}} &= -\log p(f_{ix}^{\mathrm{true}}|\mathbf{m}_{ix}) \\ &= \frac{1}{2}\log\left(\frac{\pi}{\nu_{ix}}\right) - \alpha_{ix}\log(\Omega) \\ &\quad + \left(\alpha_{ix} + \frac{1}{2}\right)\log\left((f_{ix}^{\mathrm{true}} - (\gamma_{ix} + \tau z_{ix}))^2 \nu_{ix} + \Omega\right) \\ &\quad + \log\left(\frac{\Gamma(\alpha_{ix})}{\Gamma(\alpha_{ix} + \frac{1}{2})}\right).\end{aligned} \tag{10}$$

where $\Omega = 4\beta_{ix}(1 + \omega z_{ix} v_{ix})$, $z_{ix} = \frac{\beta_{ix}}{\alpha_{ix}-1}$, and $\Gamma(\cdot)$ is the gamma function.

We use an evidence regularizer[35] so that the model tends to output low confidence when the predictions are incorrect:

$$\mathcal{L}_{ix}^{\mathrm{R}} = \rho_q(f_{ix}^{\mathrm{true}} - \gamma_{ix}) \cdot \Phi_{ix}, \tag{11}$$

where $\Phi_{ix} = \left(2\nu_{ix} + \alpha_{ix} + \frac{1}{\beta_{ix}}\right)$ is the model confidence[45]. When predictions are inaccurate, the model learns to reduce its confidence by outputting lower values for ν and α, or a higher value for β. Consequently, as demonstrated by the quantitative results in Supplementary Table 1, this regularization term effectively mitigates overconfidence.

The y- and z-components are computed similarly. Finally, the overall loss function, including the L1 loss for energy prediction, is:

$$\mathcal{L} = |E^{\mathrm{true}} - E| + \frac{w}{3N}\sum_{i=1}^{N}\sum_{a\in\{x,y,z\}}\left(\mathcal{L}_{ia}^{\mathrm{NLL}} + \lambda\mathcal{L}_{ia}^{\mathrm{R}}\right), \tag{12}$$

where w and λ are hyperparameters to adjust the weighting of each term. The effects of these parameters on the results are discussed in Supplementary Note 4.

The eIP model in this work is implemented on the PaiNN[44] backbone as an example, but it is also applicable to other equivariant backbones (see Supplementary Note 5). In contrast to the standard PaiNN model, the eIP model incorporates an additional evidential block, which takes the equivariant features as input and produces the output α, β, and ν, as illustrated in Supplementary Fig. 7. Since the evidential block is lightweight compared to the message-passing layers in the backbone, the additional computational overhead is minimal. The model parameters are trained by minimizing the overall loss function Eq. (12).

Datasets

ISO17 dataset. The ISO17 dataset[50] was obtained from http://quantum-machine.org/datasets/. We adopted the original splitting strategy for the training, validation, and test sets. For training sets of different sizes, the smaller training sets were randomly sampled from the largest training set, containing 400,000 conformations.

Silica glass dataset. The silica glass dataset is obtained from a previously published study[32]. It comprises 1691 configurations, each containing 699 atoms (233 Si and 466 O atoms). These configurations were generated through MD simulations with a force-matching potential[51] under various conditions, followed by density functional theory (DFT) calculations to obtain energies and forces. We adopted the same dataset splitting scheme as described in the ref. 32. Partitioning these structures into ID and OOD datasets is challenging, as it is difficult to find configurations with atomic environments entirely distinct from one another. To reflect a more generalized evaluation under more extreme conditions, the training set includes only structures generated under low-temperature, low-deformation-rate conditions, while the test set contains structures extracted from trajectories at higher temperatures and higher deformation rates.

Water dataset. The initial water training set is taken from our previous work[17]. It comprises 1000 configurations sampled from classical MD trajectories with the SPC/E force field[52]. Each configuration contains 288 atoms with periodic boundary conditions. During active learning, we ran UDD simulations at 300 K and sampled 1000 configurations for each iteration. The energies and forces are determined using density functional theory (DFT) calculations employing the cp2k software package[53] with the PBE-PAW-DFT-D3 method[54–56].

MPtrj dataset. The MPtrj dataset[46] is a collection of MD trajectories designed for training a universal potential. It comprises millions of configurations covering 89 elements, and the energies and forces are determined using DFT calculations. We adopted the original splitting strategy with an 8:1:1 training, validation, and test ratio.

Evaluation metrics

Spearman's rank correlation coefficient. Spearman's rank correlation is a non-parametric measure of the strength and direction of association between two ranked variables. Unlike Pearson's correlation, which accesses linear relationships, Spearman's rank correlation evaluates how well the relationship between two variables can be described using a monotonic function. We expect a larger error to be associated with a higher uncertainty, and their correlation does not necessarily need to be linear. Therefore, Spearman's rank correlation coefficient was used to assess the reliability of the uncertainty. A coefficient of 1 means perfect correlation, and a coefficient of 0 indicates that there is no correlation between the ranks of the two variables.

Area under the receiver operating characteristic curve. The receiver operating characteristic (ROC) curve is a graphical representation of a classifier's performance. The area under the ROC curve (ROC-AUC) provides a complementary evaluation metric for UQ that avoids the possible limitations of using Spearman's rank correlation coefficient alone. Following the approach of a previous study[32], we designed a classification task in which predictions with high errors are expected to exhibit high levels of uncertainty. An error threshold (ε_c) and an uncertainty threshold (U_c) are defined to classify data points. A data point is classified as a true positive (TP) if both its true error and estimated uncertainty exceed their respective thresholds ($\varepsilon > \varepsilon_c$ and $U > U_c$); a false positive (FP) if the error is below its threshold but the uncertainty is above ($\varepsilon \le \varepsilon_c$ and $U > U_c$); a true negative (TN) if both are below their thresholds ($\varepsilon \le \varepsilon_c$ and $U \le U_c$); and a false negative (FN) if the error is above its threshold but the uncertainty is below ($\varepsilon > \varepsilon_c$ and $U \le U_c$). We set the threshold to be at the 20th percentile as in ref. 32. The ROC-AUC score ranges from 0 to 1, with a score of 1 denoting a perfect classifier and 0.5 indicating performance no better than random choice.

Configurational entropy. Configurational entropy[57,58] quantifies the number of ways that atoms in a system can be arranged. High entropy indicates that the system is likely to take on many different arrangements, whereas low entropy implies a more ordered, less random state. We used configurational entropy as a metric to measure the diversity of configurations obtained during MD and UDD simulations. The configurational entropy is defined as the Shannon entropy[59,60]:

$$S_{\mathrm{conf}} = -\sum_i p_i \log p_i, \tag{13}$$

where p_i is the probability of the system being in state i. We then projected states onto a discretized order parameter grid and calculated the frequency of these order parameters within a simulation trajectory. For $\mathrm{LiFePO_4}$, the selected order parameters were the P-O-Fe angle and the $\mathrm{PO_4}$ tetrahedral distortion. For PDMS, we selected the end-to-end distance and the radius of gyration as the order parameters. To determine the probability distribution, the order parameter space was discretized into an $N_e \times N_e$ grid, and the frequency of configurations within each grid cell was calculated. The configurational entropy was normalized by dividing it by the maximum possible entropy value, $2\log(N_e)$, resulting in values between 0 and 1. A larger grid size N_e offers a finer resolution but may suffer from statistical noise, while a smaller N_e provides more robust statistics at a lower resolution. We used $N_e = 40$ for all reported results. Varying the value of N_e does not significantly affect the results, as the configurational space was sampled sufficiently in our simulations.

Molecular dynamics (MD) simulations

MD simulations were performed using the Atomic Simulations Environment (ASE) Python library[61]. The simulations are set with a timestep of 0.1 fs in the canonical (NVT) ensemble. The Berendsen thermostat[62] was used with a coupling temperature of 300 K and a decaying time constant τ of 100 fs. The atomic velocities were initialized according to the Boltzmann distribution at 300 K. The initial water configuration was selected from the water test set. The $\mathrm{LiFePO_4}$ configuration was obtained from the Materials Project, comprising 168 atoms in the unit cell. The PDMS configuration was constructed using three polymer chains with a polymerization degree of 25 and a density of 0.97 $\mathrm{g \cdot cm^{-3}}$, containing 759 atoms in total. All systems were modeled with periodic boundary conditions.

Uncertainty-driven dynamics (UDD) simulations

The UDD simulation technique utilizes a bias energy that favors configurations with higher uncertainties. Kulichenko et al. introduce a bias energy[39] defined as:

$$E_{\text{bias}}(\sigma^2) = A\left[\exp\left(-\frac{\sigma^2}{NB^2}\right) - 1\right], \tag{14}$$

where the parameters A and B are chosen empirically. The bias force F_{bias} is then determined by calculating the negative gradient of the bias energy:

$$F_{\text{bias}} = -\nabla(E_{\text{bias}}(\sigma^2)) = -E_{\text{bias}}(\sigma^2)'\nabla\sigma^2. \tag{15}$$

By leveraging eIP for UQ, the gradient of σ can be obtained through automatic differentiation.

Notably, the bias force could become exceptionally large, leading to the collapse of molecular simulations. We found that limiting the magnitude of the bias forces using a clipping strategy proved not effective. To prevent this issue, we incorporate a Gaussian term to limit the magnitude of the bias force with two additional empirically chosen parameters C and D:

$$F_{\text{bias}}^{\text{limited}} = F_{\text{bias}}\frac{D}{\sqrt{2\pi}C}\exp\left(\frac{-F_{\text{bias}}^2}{2C^2}\right). \tag{16}$$

This adjustment of bias force implies a new bias energy formulation and ensures more stable UDD simulations. Detailed discussions about the empirical parameters A, B, C, and D are provided in the Supplementary Note 6. Finally, the combined force $F + F_{\text{bias}}^{\text{limited}}$ is used to guide the simulations toward configurations with higher uncertainties, enhancing the sampling for more diverse atomic configurations.

Data availability

The ISO17 datasets are publicly available (see "Methods"). The Silica Glass datasets are available at[32]. The raw data of error-uncertainty plots and MD simulation trajectories generated in this study have been deposited in figshare[63]. Source data are provided with this paper.

Code availability

The source code for reproducing the key findings in this work is available at Zenodo (https://doi.org/10.5281/zenodo.17730621) and GitHub (https://github.com/xuhan323/eIP). It is licensed under `Apache License 2.0`, which allows users to use, modify, and distribute the code freely, provided that proper attribution is given to the original authors. This open source approach improves the reproducibility of our results and facilitates further research in this area.

References

1. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
2. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
3. Warshel, A. Molecular dynamics simulations of biological reactions. *Acc. Chem. Res.* **35**, 385–395 (2002).
4. Cornell, W. D. et al. A second-generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).
5. MacKerell Jr, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
6. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
7. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471 (1985).
8. Huang, B., von Rudorff, G. F. & von Lilienfeld, O. A. The central role of density functional theory in the AI age. *Science* **381**, 170–175 (2023).
9. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
10. Noé, F., Tkatchenko, A., Müller, K.-R. & Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
11. Manzhos, S. & Carrington Jr, T. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **121**, 10187–10217 (2020).
12. Keith, J. A. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
13. Deringer, V. L. et al. Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
14. Galib, M. & Limmer, D. T. Reactive uptake of N2O5 by atmospheric aerosol is dominated by interfacial processes. *Science* **371**, 921–925 (2021).
15. Zeng, J., Cao, L., Xu, M., Zhu, T. & Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **11**, 5713 (2020).
16. Fu, X. et al. Forces are not enough: benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transact. Mach. Learn. Res.* https://openreview.net/forum?id=A8pqQipwkt (2023).
17. Cui, T. et al. Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. *Nat. Commun.* **16**, 1891 (2025).
18. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
19. Zhang, Y. et al. Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **253**, 107206 (2020).
20. Yuan, X. et al. Active learning to overcome the exponential-wall problem for effective structure prediction of chemical-disordered materials. *npj Comput. Mater.* **9**, 12 (2023).
21. Moon, J. et al. Active learning guides the discovery of a champion four-metal perovskite oxide for oxygen evolution electrocatalysis. *Nat. Mater.* **23**, 108–115 (2024).
22. Novikov, I. S., Gubaev, K., Podryabinkin, E. V. & Shapeev, A. V. The MLIP package: moment tensor potentials with MPI and active learning. *Mach. Learn.: Sci. Technol.* **2**, 025002 (2020).
23. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: a brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057 (2015).
24. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems.* **30** (2017).
25. Kellner, M. & Ceriotti, M. Uncertainty quantification by direct propagation of shallow ensembles. *Mach. Learn.: Sci. Technol.* **5**, 035006 (2024).
26. Bilbrey, J. A., Firoz, J. S., Lee, M.-S. & Choudhury, S. Uncertainty quantification for neural network potential foundation models. *npj Comput. Mater.* **11**, 109 (2025).
27. Gal, Y. & Ghahramani, Z.*Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*, 1050–1059 (PMLR, 2016).
28. Wen, M. & Tadmor, E. B. Uncertainty quantification in molecular simulations with dropout neural network potentials. *npj comput. Mater.* **6**, 124 (2020).

29. Thaler, S., Mayr, F., Thomas, S., Gagliardi, A. & Zavadlav, J. Active learning graph neural networks for partial charge prediction of metal-organic frameworks via dropout Monte Carlo. *npj Comput. Mater.* **10**, 86 (2024).
30. Zhu, A., Batzner, S., Musaelian, A. & Kozinsky, B. Fast uncertainty estimates in deep learning interatomic potentials. J. Chem. Phys. **158**, 164111 (2023).
31. Nix, D. A. & Weigend, A. S.*Estimating the mean and variance of the target probability distribution*, **1**, 55–60 (IEEE, 1994).
32. Tan, A. R., Urata, S., Goldman, S., Dietschreit, J. C. & Gómez-Bombarelli, R. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Comput. Mater.* **9**, 225 (2023).
33. Vita, J. A., Samanta, A., Zhou, F. & Lordi, V. LTAU-FF: loss trajectory analysis for uncertainty in atomistic force fields. *Mach. Learn.: Sci. Technol.* **6**, 015048 (2025).
34. Bigi, F., Chong, S., Ceriotti, M. & Grasselli, F. A prediction rigidity formalism for low-cost uncertainties in trained neural networks. *Mach. Learn.: Sci. Technol.* **5**, 045018 (2024).
35. Amini, A., Schwarting, W., Soleimany, A. & Rus, D. Deep evidential regression. *Adv. Neural Inf. Process. Syst.* **33**, 14927–14937 (2020).
36. Soleimany, A. P. et al. Evidential deep learning for guided molecular property prediction and discovery. *ACS Cent. Sci.* **7**, 1356–1367 (2021).
37. Hüllermeier, E. & Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach. Learn.* **110**, 457–506 (2021).
38. Wollschläger, T., Gao, N., Charpentier, B., Ketata, M. A. & Günnemann, S. *Uncertainty estimation for molecules: desiderata and methods*, 37133–37156 (PMLR, 2023).
39. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
40. van der Oord, C., Sachs, M., Kovács, D. P., Ortner, C. & Csányi, G. Hyperactive learning for data-driven interatomic potentials. *npj Comput. Mater.* **9**, 168 (2023).
41. Zaverkin, V. et al. Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials. *npj Comput. Mater.* **10**, 83 (2024).
42. Tan, A. R., Dietschreit, J. C. & G¢mez-Bombarelli, R. Enhanced sampling of robust molecular datasets with uncertainty-based collective variables. *J. Chem. Phys.* **162**, 034114 (2025).
43. Kendall, A. & Gal, Y. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in neural information processing systems* **30** (2017).
44. Schütt, K., Unke, O. & Gastegger, M.*Equivariant Message Passing for the Prediction of Tensorial Properties and Molecular Spectra*, 9377–9388 (PMLR, 2021).
45. Hüttel, F. B., Rodrigues, F. & Pereira, F. C. Deep evidential learning for Bayesian quantile regression. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2308.10650 (2023).
46. Deng, B. et al. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
47. Yu, K. & Zhang, J. A three-parameter asymmetric Laplace distribution and its extension. *Commun. Stat.-Theory Methods* **34**, 1867–1879 (2005).
48. Kotz, S., Kozubowski, T. & Podgorski, K.*The Laplace Distribution and Generalizations: A Revisit with Applications wo Communications, Economics, Engineering, and Finance*. (Springer Science & Business Media, 2012).
49. Kozumi, H. & Kobayashi, G. Gibbs sampling methods for Bayesian quantile regression. *J. Stat. Comput. Simul.* **81**, 1565–1578 (2011).
50. Schütt, K. et al. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems* **30** (2017).
51. Urata, S., Nakamura, N., Aiba, K., Tada, T. & Hosono, H. How fluorine minimizes density fluctuations of silica glass: molecular dynamics study with machine-learning assisted force-matching potential. *Mater. Des.* **197**, 109210 (2021).
52. Berendsen, H. J., Grigera, J. R. & Straatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
53. Khne, T. D. et al. CP2K: An electronic structure and molecular dynamics software package-quickstep: efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 194103 (2020).
54. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
55. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
56. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements h-pu. *J. Chem. Phys.* **132**, 154104 (2010).
57. Karplus, M. & Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules* **14**, 325–332 (1981).
58. Peter, C., Oostenbrink, C., Van Dorp, A. & Van Gunsteren, W. F. Estimating entropies from molecular dynamics simulations. *J. Chem. Phys.* **120**, 2652–2661 (2004).
59. Baxa, M. C., Haddadian, E. J., Jha, A. K., Freed, K. F. & Sosnick, T. R. Context and force field dependence of the loss of protein backbone entropy upon folding using realistic denatured and native state ensembles. *J. Am. Chem. Soc.* **134**, 15929–15936 (2012).
60. Klyshko, E. et al. Functional protein dynamics in a crystal. *Nat. Commun.* **15**, 3244 (2024).
61. Larsen, A. H. et al. The atomic simulation environment-a Python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002 (2017).
62. Berendsen, H. J., Postma, J. V., Van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).
63. Xu, H. Evidential Deep Learning for Interatomic Potential https://figshare.com/articles/dataset/Evidential_Deep_Learning_for_Interatomic_Potential/28805819 (2025).

Acknowledgments

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project(2025ZD0121802), Shanghai Committee of Science and Technology, China (Grant No. 23QD1400900), and the National Natural Science Foundation of China (Grant No. 12404291). H.X., T.C., and C.T. did this work during their internship at Shanghai Artificial Intelligence Laboratory.

Author contributions

M.S. and S.Z. conceived the idea and led the research. H.X. and T.C. developed the eIP code and trained the models. H.X. and J.M. performed the experiments and analyses. C.T. developed the active learning workflow and performed the molecular dynamics simulations. Y.L., X.Gao, and X.Gong contributed technical ideas for datasets and experiments. D.Z. and W.O. contributed technical ideas for designing and training the models. H.X., C.T., and M.S. wrote the first draft. All authors discussed the results and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-67663-y.

Correspondence and requests for materials should be addressed to Shufei Zhang or Mao Su.

Peer review information *Nature Communications* thanks Jenna Bilbrey, Johannes Dietschreit and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.