Contents

List of Tables

List of Figures

S1 Hyperparameters

The hyperparameter selections for the small molecule dataset and the water dataset used in this work are summarized in Table S1 and Table S2, respectively. The implementation of eIP simply involves a few additional fully connected layers, which are referred to as the evidential layer. Our experiments show that using two evidential layers with a hidden state size of 128 yields satisfactory results. This only leads to a modest increase in the total parameter count and inference time, as shown in Figure S1.

The choice of the hyperparameters appeared in the loss function (Equation 9 in the main text) may influence eIP's preference for energy and force prediction and uncertainty estimation.

The hyperparameter λ trades off uncertainty inflation with model fit. Setting λ with a small value yields an over-confident estimation while setting λ too large results in over-inflation [1]. The influence of varying λ on mean absolute error (MAE) and uncertainty is shown in Figure S2.

The hyperparamneter q is the quantile in quantile regression. When the quantile q equals 0.5, the titled loss function reduces to the absolute error function [2]. For quantiles above 0.5, underestimation is penalized less than overestimation, and vice versa for quantiles below 0.5. The influence of varying q on MAE and uncertainty is shown in Figure S3.

The hyperparameter θ balances the prediction of energy and force. A larger θ will prefer the energy prediction, while a smaller θ will prefer the force prediction as well as uncertainty estimation. When θ is very large, it may lead to training failure, as the energy loss is more sensitive to parameter variations. This design allows for a trade-off between the accuracy of energy and force prediction and the reliability of uncertainty estimation. The influence of varying θ on MAE and uncertainty is shown in Figure S4.

Table S1 Hyperparameters for small molecules dataset.

Hyperparameters	Search space	Best value
Number of interaction blocks in backbone	3, 4, 5	3
Hidden state size in backbone	128, 256	128
Number of evidential layers	2, 3, 4	2
Hidden state size in evidential layer	128, 256	128
Cutoff distance	5.0, 6.0	5.0
Batch size	2, 4	2
Initial learning rate	1e-5, 5e-4, 1e-4, 5e-3	5e-4
Learning rate strategy	-	StepLR
Learning rate decay ratio	-	0.5
Learning rate step size	150, 100, 250	200
Epochs	500, 1000, 1500, 2000	1000
θ	0.00001, 0.0001, 0.001, 0.01, 0.1, 1	0.1
λ	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100	1
q	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	0.4

Table S2 Hyperparameters for water dataset.

Hyperparameters	Search space	Best value
Number of interaction blocks in backbone	3, 4, 5	3
Hidden state size in backbone	128, 256	128
Number of evidential layer	2, 3, 4	2
Hidden state size in evidential layer	128, 256	128
Cutoff distance	5.0, 6.0	5.0
Batch size	2, 4	2
Initial learning rate	1e-5, 5e-4, 1e-4, 5e-3	5e-4
Learning rate strategy	-	StepLR
Learning rate decay ratio	-	0.5
Learning rate step size	150, 100, 250	150
Epochs	500, 1000, 1500, 2000	1000
θ	0.00001, 0.0001, 0.001, 0.01, 0.1, 1	0.1
λ	0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100	0.1
q	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9	0.6

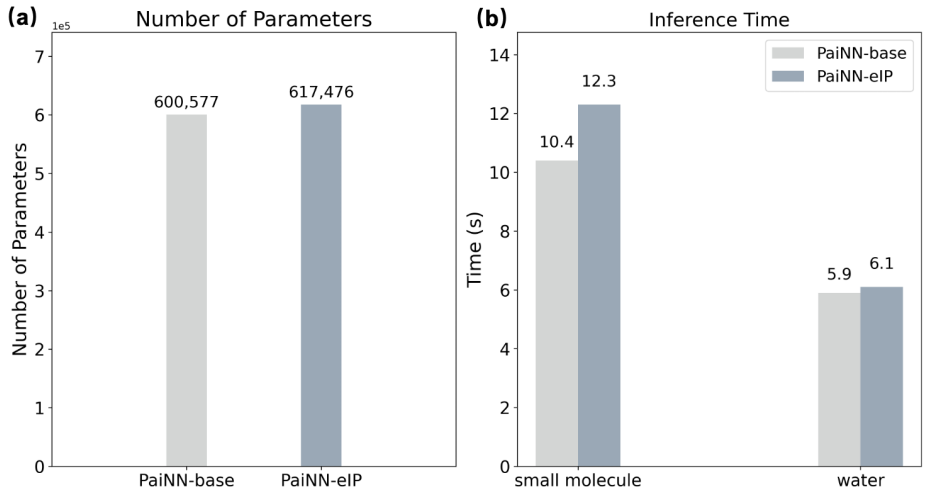

Fig. S1 Parameter count and inference time of eIP. (a) Parameter count of the eIP model (617,476) and its backbone model, PaiNN (600,577). The additional layers introduced by eIP result in an parameter increase of only 2.81%. (b) Inference time comparisons for small molecule dataset and water dataset.

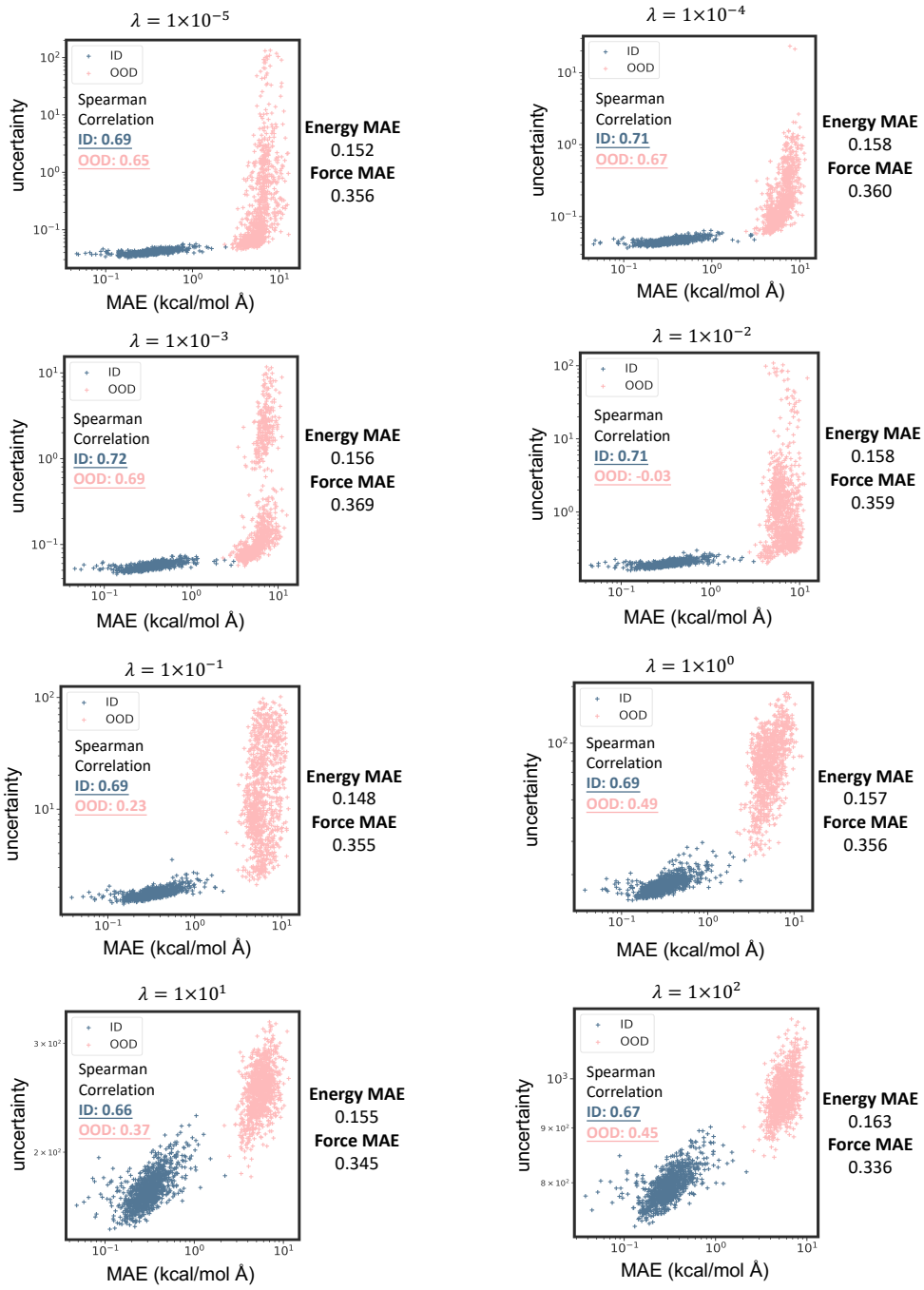

Fig. S2 Results of eIP with different λ. As the parameter λ varies from small to large, the uncertainty estimated by eIP increases, while the predicted results of eIP (energy and force MAE, measured in kcal/mol and kcal/mol/Å, respectively) do not exhibit significant variations. The experiments were performed on small molecule dataset with PaiNN backbone.

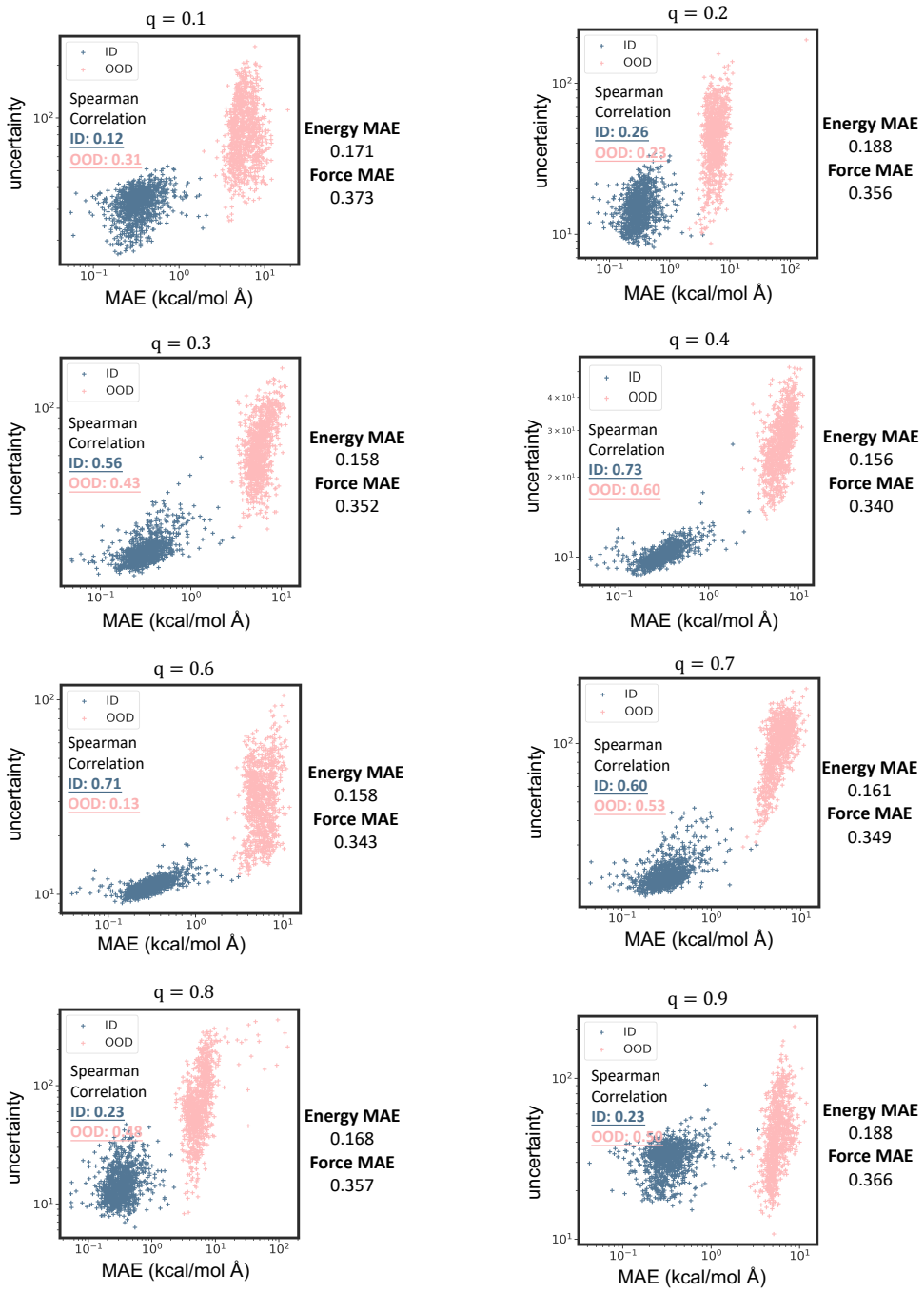

Fig. S3 Results of eIP with different q. As the parameter q varies from small to large, the uncertainty estimated by eIP initially improves before deteriorating, achieving optimal performance around 0.4-0.6, while the predicted results of eIP (energy and force MAE, measured in kcal/mol and kcal/mol/Å, respectively) do not exhibit significant variations.The experiments were performed on small molecule dataset with PaiNN backbone.

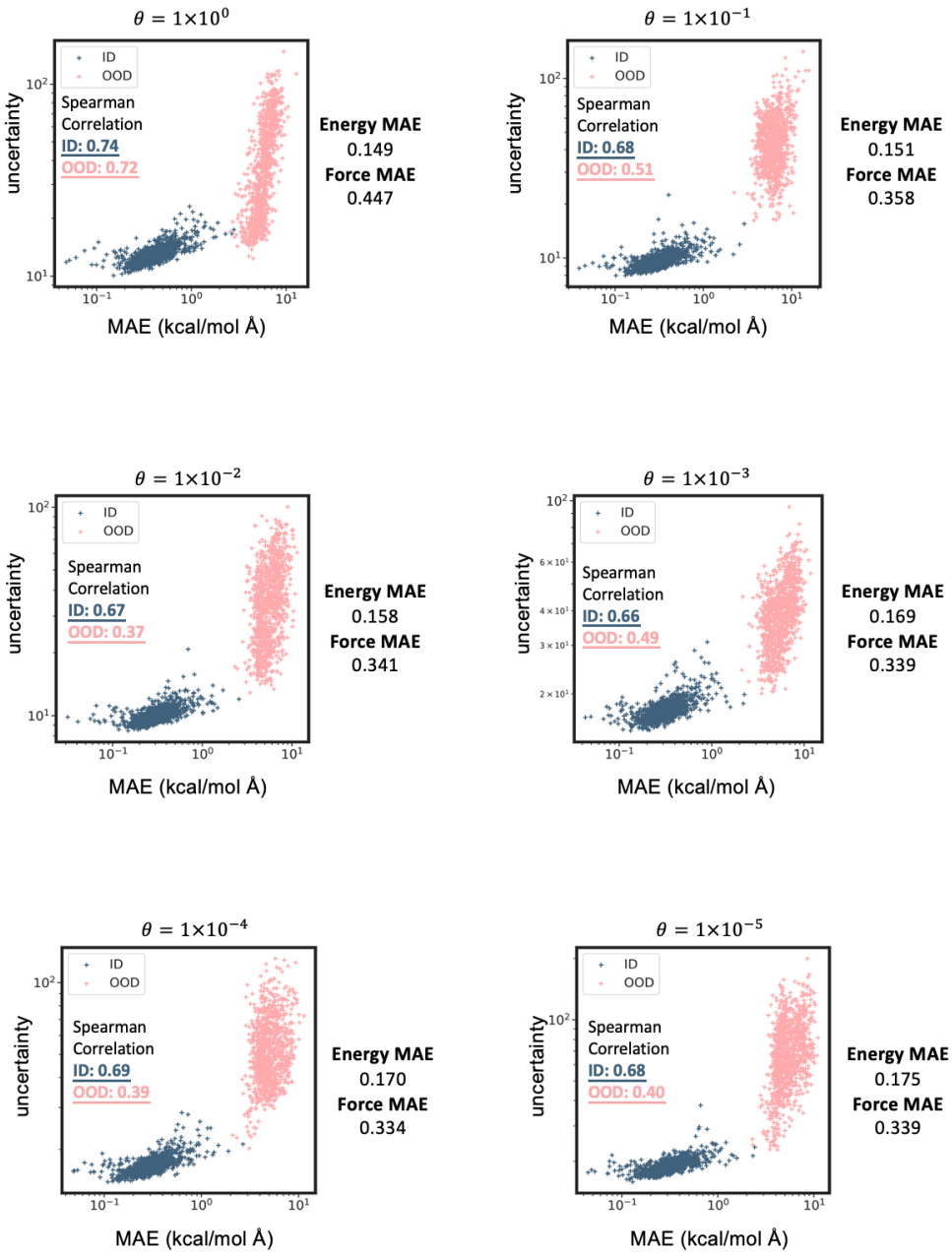

Fig. S4 Results of eIP with different θ. As the parameter θ varies from large to small, the energy MAE (measured in kcal/mol) increases while the force MAE (measured in kcal/mol/Å) initially decreases but shows a slight increase later, and the uncertainty estimated by eIP do not exhibit significant variations. The optimal values are observed around 0.1. The experiments were performed on small molecule dataset with PaiNN backbone.

S2 Ablation Study

We conduct ablation experiments to investigate the significance of locality, equivariance, and quantile regression in the design of eIP. Specifically, when ablating locality, we chose global embedding to estimate the uncertainty associated with atomic forces [3]. When ablating equivariance, we replaced the PaiNN backbone with SchNet, which is an invariant model. When ablating quantile regression, we utilized standard deep evidential regression to estimate uncertainty. The results are shown in Figure S5 and Table S3.

Since changing the backbone to SchNet during the ablation of equivariance resulted in reduced accuracy of energy and force prediction, we conduct additional tests using SchNet as the backbone of eIP. The tests include sensitivity analysis to hyperparameters λ (Figure S6) and q (Figure S7), and comparison of MAE and computational time with ensemble and dropout methods (Table S4).

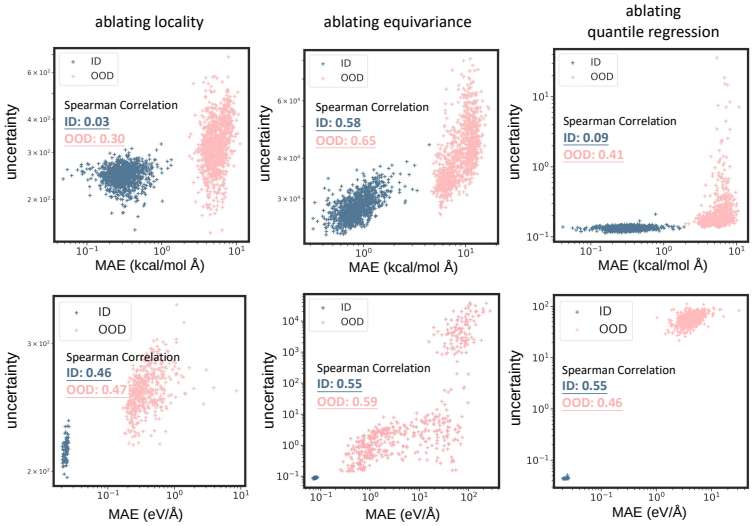

Fig. S5 Results of force MAE versus uncertainty in ablation study on small molecule dataset (top row) and water dataset (bottom row). Ablating locality of eIP results in indistinguishable ID and OOD uncertainties (left column). Ablating equivariance slightly improves the distinction between ID and OOD uncertainties, but not significantly, and introduces unreasonable clustering in the OOD distribution of the water dataset [3] (middle column). Ablating quantile regression reduces the clarity of uncertainty distinction between ID and OOD data in small datasets, but a more distinct separation can be found in the water dataset (right column).

Table S3 Results of energy and force MAEs in ablation study. Energy and force MAEs on small molecule dataset are reported in unit of [kcal/mol] and [kcal/mol/Å], respectively. Energy and force MAEs on water dataset is reported in unit of [eV] and [eV/Å], respectively. Ablating locality yields a comparable accuracy to eIP. Ablating equivariance leads to a notable decrease in predictive accuracy due to the substitution of SchNet for PaiNN as the backbone. Ablating quantile regression results in a decline in predictive accuracy. Standard deviations are calculated from three independent experiments. The results of eIP are shown in bold.

		ablating locality	ablating equivariance	ablating quantile regression	eIP
Small Molecule	Energy	0.155±0.017	0.402±0.142	0.178±0.164	**0.155±0.008**
	Force	0.337±0.326	1.104±0.326	0.358±1.063	**0.338±0.003**
Water	Energy	0.079±0.163	1.095±0.120	0.528±0.062	**0.082±0.039**
	Force	0.023±0.028	0.095±0.021	0.223±0.133	**0.020±0.001**

Table S4 MAE and computing time of three uncertainty estimation methods on small molecule dataset and water dataset with SchNet backbone. The energy and force MAEs for the validation sets are reported. "Training time" refers to the time consumption of each epoch, while "Inference time" denotes the duration required to process the entire test set. For the ensemble method, four differently initialized models are trained. For the dropout method, inferences are performed four times. Standard deviations are calculated from three independent experiments. The force MAEs of SchNet-eIP have been reduced compare those of ensembles.

		Ensemble	Dropout	SchNet-eIP
Small Molecule	Energy (kcal/mol)	0.403±0.033	3.691±0.073	0.402±0.142
	Force (kcal/mol/Å)	1.007±0.073	4.143±1.516	1.104±0.326
	Training time (s)	24.402±0.447	7.139±0.306	7.649±0.011
	Inference time (s)	32.398±0.098	32.849±0.086	7.604±0.748
Water	Energy (eV)	1.096±0.033	2.177±0.033	1.095±0.120
	Force (eV/Å)	0.082±0.033	0.241±0.033	0.095±0.021
	Training time (s)	69.433±0.104	18.458±0.033	18.675±0.073
	Inference time (s)	23.484±1.062	23.098±0.120	4.367±0.130

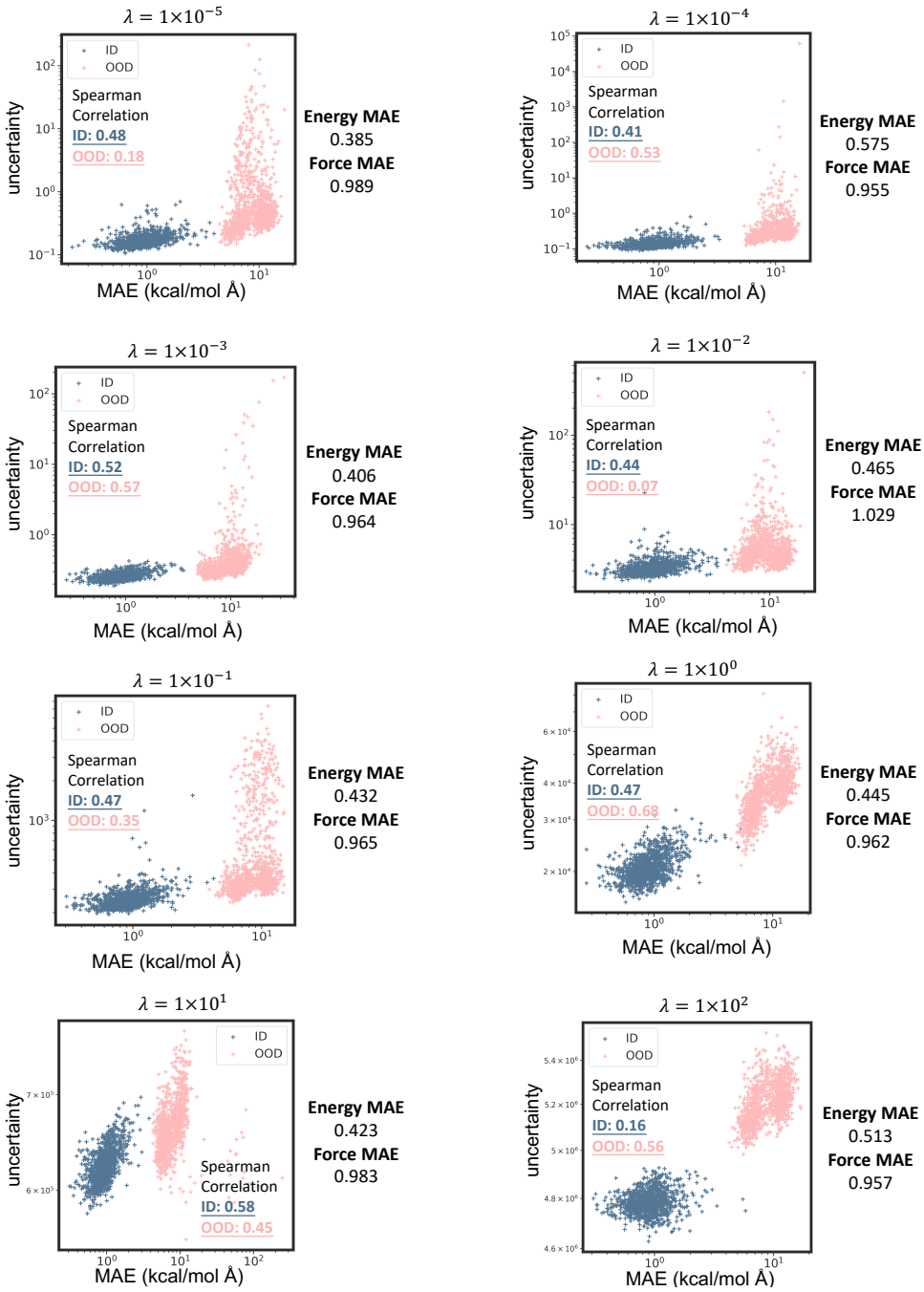

Fig. S6 Results of SchNet-eIP with different λ. As the parameter λ varies from small to large, the uncertainty of SchNet-eIP output increases, while the predicted results of eIP (Energy MAE and Force MAE, measured in kcal/mol and kcal/mol/Å, respectively) do not exhibit significant variations. However, comparing to the result of PaiNN-eIP, the uncertainty between ID and OOD dataset is indistinguishable and lack stability. The experiments were performed on small molecule dataset.

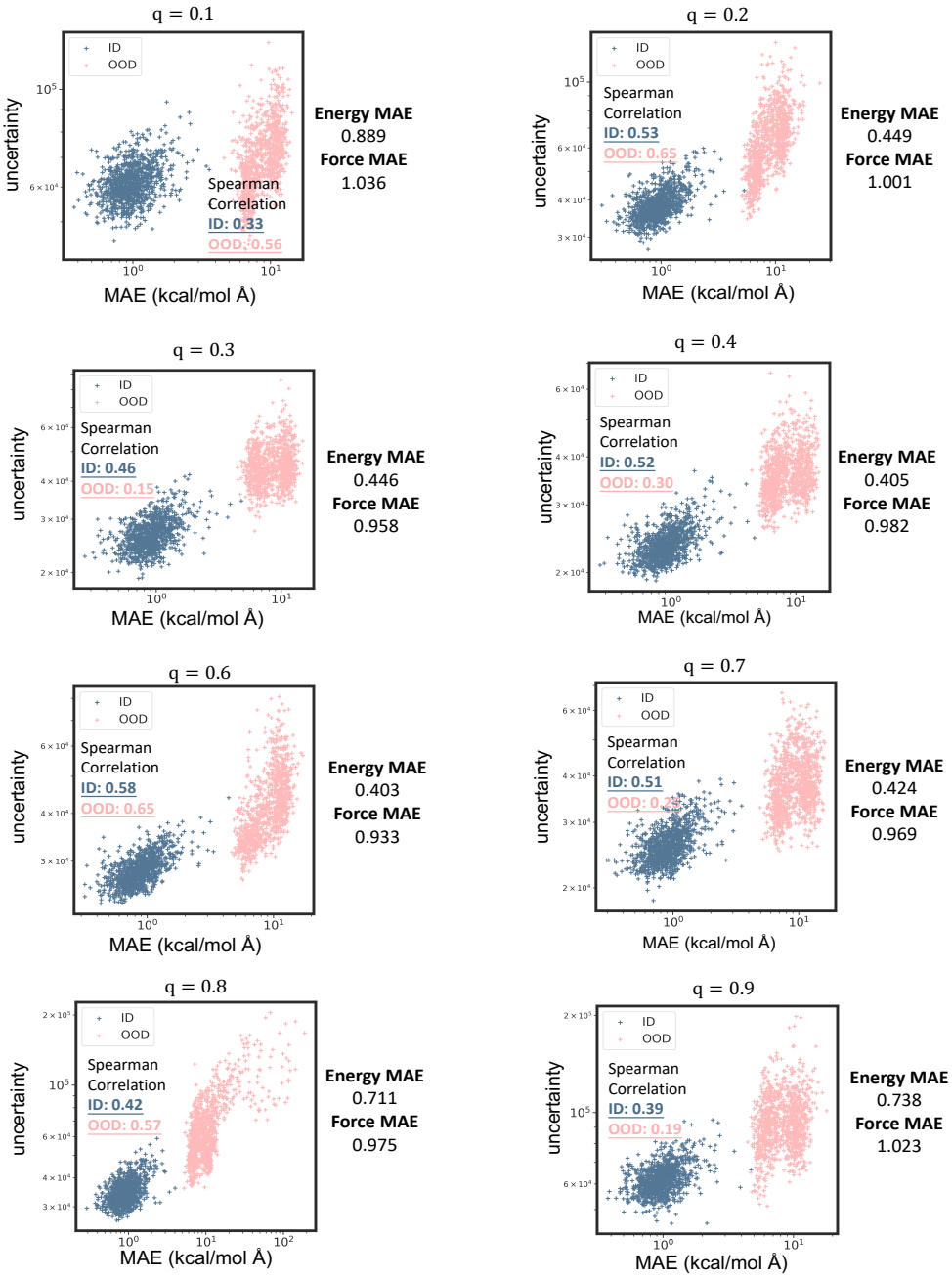

Fig. S7 Results of SchNet-eIP with different q. As the parameter q varies from small to large, the uncertainty in the model output initially improves before deteriorating, achieving optimal performance around 0.4-0.6, while the predicted results of SchNet-eIP (Energy MAE and Force MAE, measured in kcal/mol and kcal/mol/Å, respectively) do not exhibit significant variations. However, comparing to the result of PaiNN-eIP, the uncertainty between ID and OOD dataset is indistinguishable. The experiments were performed on small molecule dataset.

S3 Uncertainty-driven dynamics (UDD)

The use of UDD simulations enables more efficient sampling in the configuration space. We use the uncertainty estimated by eIP to modify the potential energy surface and run UDD simulations. We also run standard MD simulations for comparison. The configurations generated during at different simulation times are visualized in Figure S8, indicating that eIP-based UDD simulations are able to generate more diverse atomic structures than standard MD simulations.

We have conducted four iterations during the active learning process on the liquid water dataset. The result after the first iteration are shown in the main text. Figure S9 shows the uncertainty and potential energy variations after more iterations. Figure S10 visualized the sample structures in each iterations.

Selecting the appropriate parameters A, B, C, and D in the bias force (Equations 14 and 16 in the main text) is crucial yet manageable for UDD simulations. For a practical guidance, the parameter A is chosen to be of the same magnitude as the potential energy E. This ensures that the bias energy $E_{\text{bias}}(\sigma^2)$ remains proportional and does not overwhelm the system dynamics. The parameter B controls the exponential decay rate in the bias energy equation. A larger value of B slows the decay, which leads to a broader response to the change of σ. The parameter B should be balanced to capture rapid fluctuations without compromising the stability of the bias energy. The parameter C and D affects the exponential decay rate of the limited bias force $\mathbf{F}_{\text{bias}}^{\text{limited}}$. A larger value of C results in a faster decay, limiting the values of bias force for larger $\mathbf{F}_{\text{bias}}$. The parameter D scales the overall magnitude of $\mathbf{F}_{\text{bias}}^{\text{limited}}$. Empirically, D should be set so that the maximum bias force should be at least smaller than 0.3 times the typical maximum interatomic forces, providing a safeguard against unrealistic force magnitudes that could cause collapses of UDD simulations.

During the active learning process shown in Figure S9 and Figure S10, we set $A = 2,000$, $B = 20$, $C = 1.0$, and $D = 7.0$ for the first and second iterations and $A = 10,000$, $B = 20$, $C = 1.0$, and $D = 7.0$ for the third and fourth iterations in order to balance sampling stability and efficiency. More results of UDD simulations with different values of A, B, C, and D are shown in Figure S11 to Figure S14.

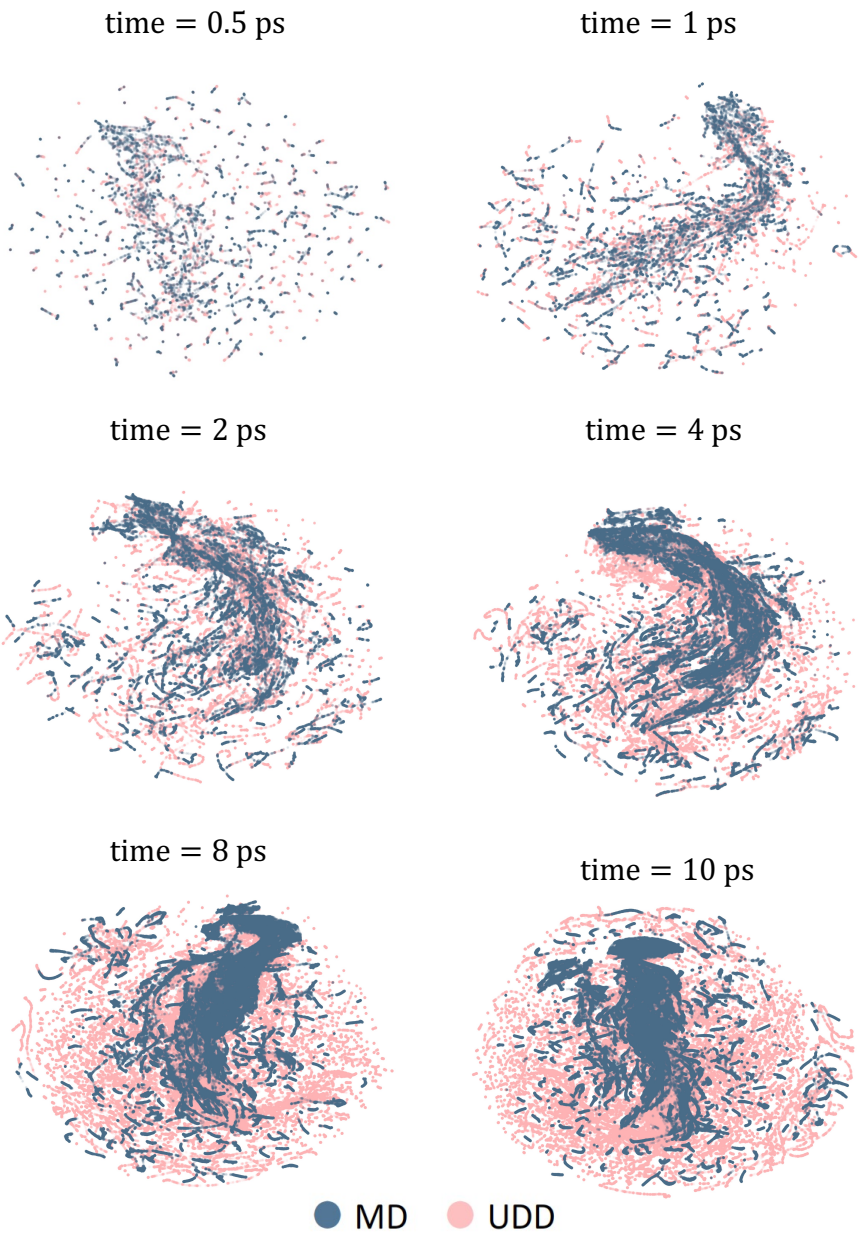

Fig. S8 Comparisons of UDD and MD sampling. The atomic structure distributions at different simulation times show that the eIP-based UDD simulations can significantly enhance the exploration of a system's configuration space compared to traditional unbiased MD simulations. Starting from the same initial configuration, UDD can gradually samples a much larger configuration space that covers the sampling of MD simulation entirely as the simulation time increases. This ensures the applicability of eIP-based UDD for active learning. The visualizations of the UDD and MD sampling trajectories are achieved by using SOAP descriptor and UMAP dimensionality reduction method.

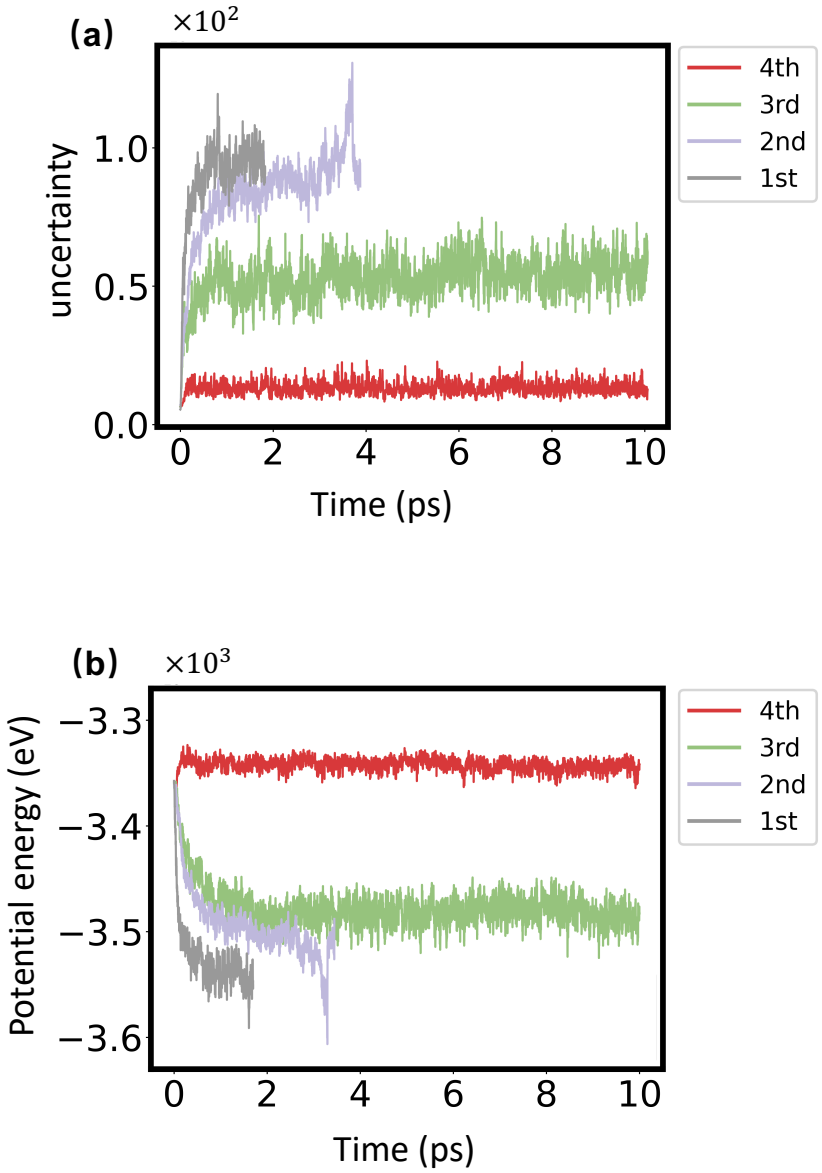

Fig. S9 UDD simulation results after more iterations. (a) The variation of uncertainty over simulation time. (b) The variation of potential energy over simulation time.

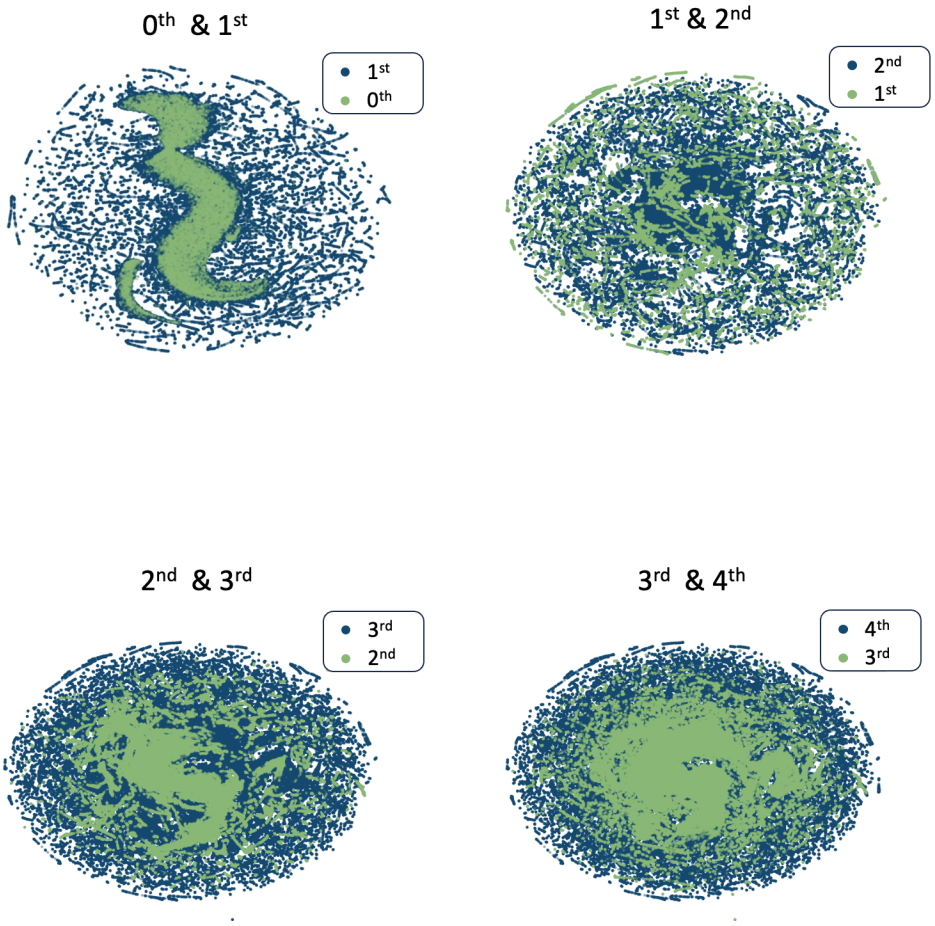

Fig. S10 Visualization of sampled structures in each iteration. The structures in the initial training set, randomly sampled from simulation trajectories, are labeled by 0^{th}. The structures sampled at the $n^{\text{th}}(n > 0)$ iteration are labeled by n^{th}.

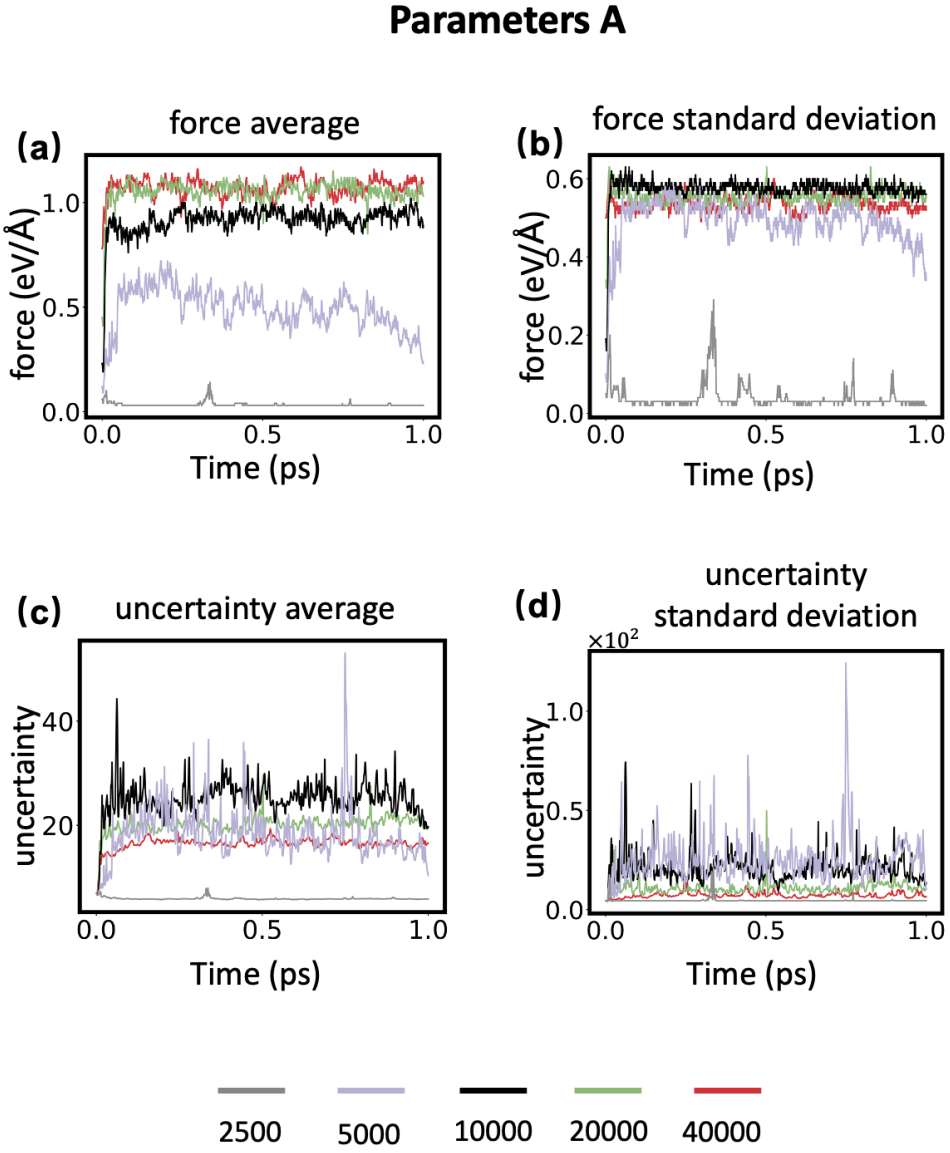

Fig. S11 Results of UDD with different parameter A. Different colored curves represent different A values when $B = 20$, $C = 1.0$, and $D = 7.0$. We use the model trained after four active learning iterations. (a) The variation of force average over simulation time. (b) The variation of force standard deviation over simulation time. (c) The variation of uncertainty average over simulation time. (d) The variation of uncertainty standard deviation over simulation time.

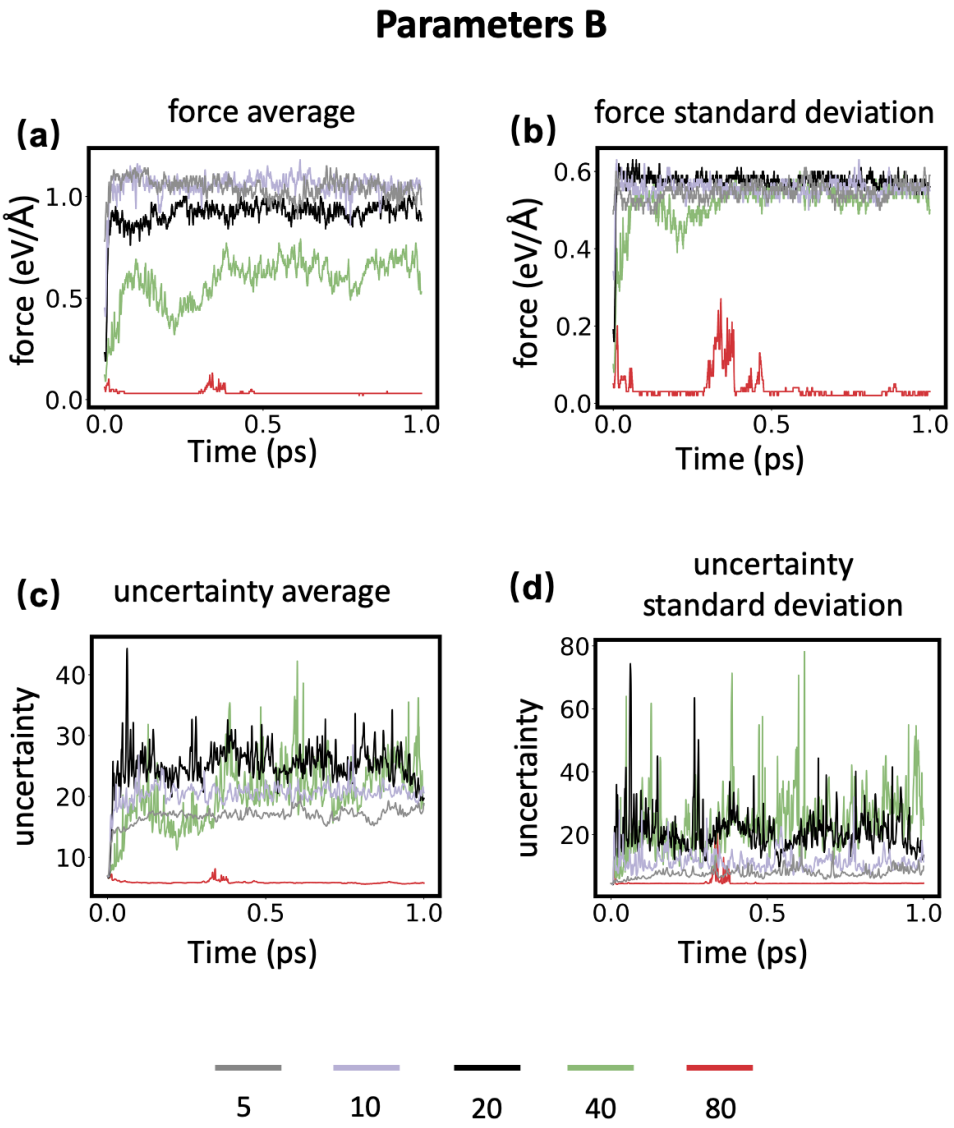

Fig. S12 Results of UDD with different parameter B. Different colored curves represent different B values when $A = 10,000$, $C = 1.0$, and $D = 7.0$. We use the model trained after four active learning iterations. (a) The variation of force average over simulation time. (b) The variation of force standard deviation over simulation time. (c) The variation of uncertainty average over simulation time. (d) The variation of uncertainty standard deviation over simulation time.

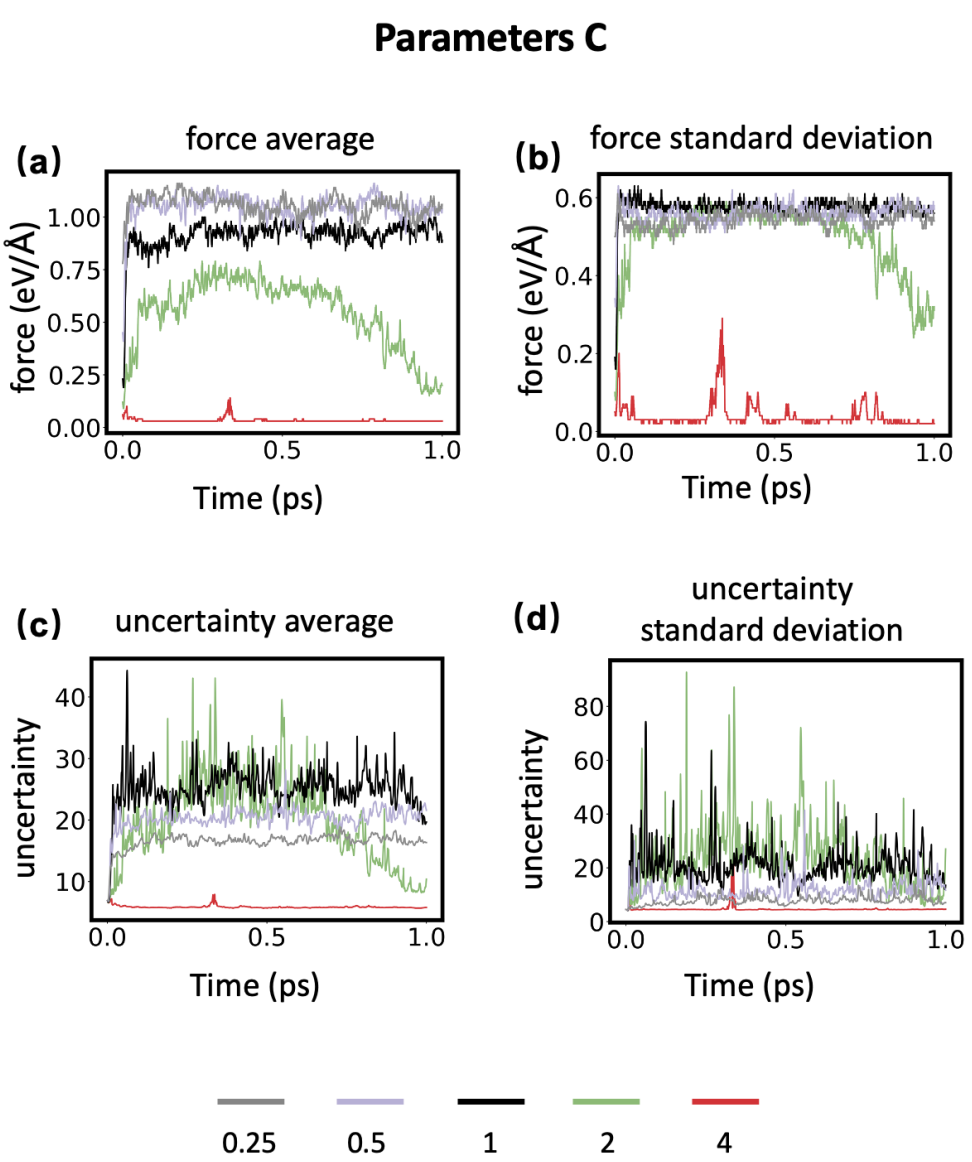

Fig. S13 Results of UDD with different parameter C. Different colored curves represent different C values when $A = 10,000$, $B = 20.0$, and $D = 7.0$. We use the model trained after four active learning iterations. (a) The variation of force average over simulation time. (b) The variation of force standard deviation over simulation time. (c) The variation of uncertainty average over simulation time. (d) The variation of uncertainty standard deviation over simulation time.

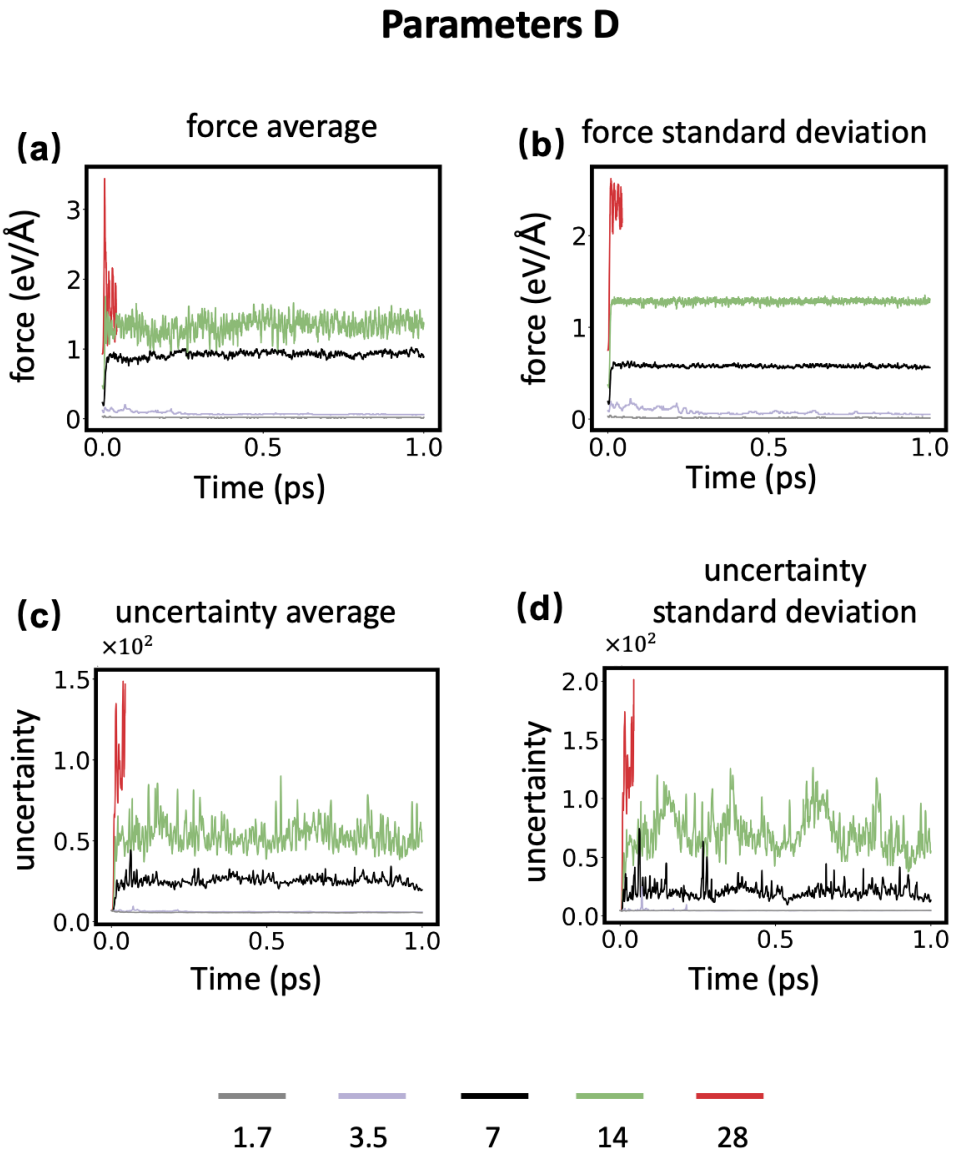

Fig. S14 Results of UDD with different parameter D Different colored curves represent different D values when $A = 10,000$, $B = 20$, and $C = 1.0$. We use the model trained after four active learning iterations. (a) The variation of force average over simulation time. (b) The variation of force standard deviation over simulation time. (c) The variation of uncertainty average over simulation time. (d) The variation of uncertainty standard deviation over simulation time.

References

[1] Amini, A., Schwarting, W., Soleimany, A. & Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems* **33**, 14927–14937 (2020).

[2] Hüttel, F. B., Peled, I., Rodrigues, F. & Pereira, F. C. Modeling censored mobility demand through censored quantile regression neural networks. *IEEE Transactions on Intelligent Transportation Systems* **23**, 21753–21765 (2022).

[3] 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 Computational Materials* **9**, 225 (2023).