

Supplementary information for
Single-model uncertainty quantification in neural network potentials does not
consistently outperform model ensembles

Aik Rui Tan,¹ Shingo Urata,² Samuel Goldman,³ Johannes C. B. Dietschreit,¹ and
Rafael Gómez-Bombarelli^{*1, a)}

¹⁾*Department of Materials Science and Engineering, Massachusetts Institute of Technology*

²⁾*Innovative Technology Laboratories, AGC Inc.*

³⁾*Computational and Systems Biology, Massachusetts Institute of Technology*

(Dated: 27 November 2023)

^{a)}Electronic mail: rafagb@mit.edu

I. NEURAL NETWORK INTERATOMIC POTENTIALS (NNIPS)

We used Polarizable atom interaction Neural Network (PaiNN), an equivariant message-passing NN to learn interatomic potentials of molecules and materials. The NNIP architecture was the same for all systems, we used 3 convolutional layers with a hidden dimension size of 128 each, 20 learnable Gaussians, a cutoff of 5.0 Å and shifted softplus activation function. The 128-dimensional feature vectors were then passed through 2 dense layers of size 64 before being summed to yield a final output value. For a specific system, all models were trained on different splits of the data set using the Adam optimizer with an initial learning rate of 10^{-4} . A scheduler was used to reduce the learning rate by a factor of 2 if the error on the validation set stagnates for 30 epochs.

For the rMD17 data set, all models were trained and validated on 800 and 200 data points (molecular conformations) in each fold/split for 500 epochs with a batch size of 64. The models were then tested on 1000 data points. 5-fold cross validation was performed for each molecule and uncertainty quantification (UQ) method, where the indices for the train and test set for each fold was provided by Christensen et al. For the ammonia data set, first generation models were trained on 62 geometries and validated on 16 geometries to assess the models’ ability to bootstrap on few data points. The test set contains 129 geometries with energies ranging from 0 to 70 kcal/mol. The models were trained for 500 epochs with a batch size of 64. For the silica data set, the models were trained on a data set containing 1318 geometries (train: validation at a ratio 80: 20) and tested on 373 structures from a test set with energies ranging from 0 - 10,000 kcal/mol (more details in Supplementary Section II). They were trained for 1000 epochs with a batch size of 3.

A. Ensemble-based method

5 independent NNs are used for rMD17 and ammonia data sets, whereas 4 independent NNs are used for the silica glass data set. During training of the NNs in the ensembles, the loss $\mathcal{L}$ is then computed as the mean absolute error of the predicted and target values of a batch size N

$$\mathcal{L} = \frac{1}{M} \frac{1}{N} \sum_m^M \left[\sum_s^N \left(w_E |E_{s,m} - \hat{E}_{s,m}| + w_F |\mathbf{F}_{s,m} - \hat{\mathbf{F}}_{s,m}| \right) \right] \quad (1)$$

where w_E and w_F are the weights of the energy and force trade-offs during training. In all models, the weights were set to 0.1 and 1.0, respectively.

B. Mean-variance estimation (MVE)

NNs in the MVE method are trained using a negative log-likelihood (NLL) loss function of the normal distribution instead of a regular mean absolute error (MAE). In our work, the NLL loss function was applied only on the force prediction while a regular mean absolute error (MAE) was applied on the energy outputs. This increases the prediction accuracy of the NN. Similar to the ensemble method, the uncertainty estimates are described only by the variances of the atomic forces.

$$\text{MAE}_E = \frac{1}{N} \sum_s^N |E_s - \mu_{E_s}| \quad (2)$$

$$\text{NLL}_F = \frac{1}{2} \frac{1}{N} \frac{1}{n} \sum_s^N \sum_i^n \log(2\pi\sigma_{F_i}^2) + \frac{(\mathbf{F}_i - \mu_{F_i})^2}{2\sigma_{F_i}^2} \quad (3)$$

$$\mathcal{L} = w_E \text{MAE}_E + w_F \text{NLL}_F \quad (4)$$

where w_E and w_F are set to 0.1 and 1.0, respectively. Note that N is the number of systems (i.e. molecules or structures) in a single batch, while n is the number of atoms within each system.

C. Deep evidential regression

For training of the deep evidential NNs, the loss function used is computed as

$$\text{MAE}_E = \frac{1}{N} \sum_s^N |E_s - \gamma_{E_s}| \quad (5)$$

$$\text{NLL}_F = \frac{1}{N} \frac{1}{n} \sum_s^N \sum_i^n \frac{1}{2} \log \left(\frac{\pi}{\nu_i} \right) - \alpha_i \log [2\beta_i(1 + \nu_i)] \quad (6)$$

$$\begin{aligned} &+ \left(\alpha_i + \frac{1}{2} \right) \log \left\{ (\mathbf{F}_i - \gamma_{F_i})^2 \nu_i + [2\beta_i(1 + \nu_i)] \right\} + \log \left(\frac{\Gamma(\alpha_i)}{\Gamma(\alpha_i + \frac{1}{2})} \right) \\ &+ \lambda |\mathbf{F}_i - \gamma_{F_i}| \cdot (2\nu_i + \alpha_i) \\ \mathcal{L} &= w_E \text{MAE}_E + w_F \text{NLL}_F \end{aligned} \quad (7)$$

where λ , which we set to 0.2, controls the trade-off between minimizing evidence on errors (i.e. lowering ν and α for incorrect predictions) and maximizing model fit. Again, w_E and w_F are set to 0.1 and 1.0, respectively. N is the number of systems (i.e. molecules or structures) in a single batch, while n is the number of atoms within each system.

D. Gaussian mixture model (GMM)

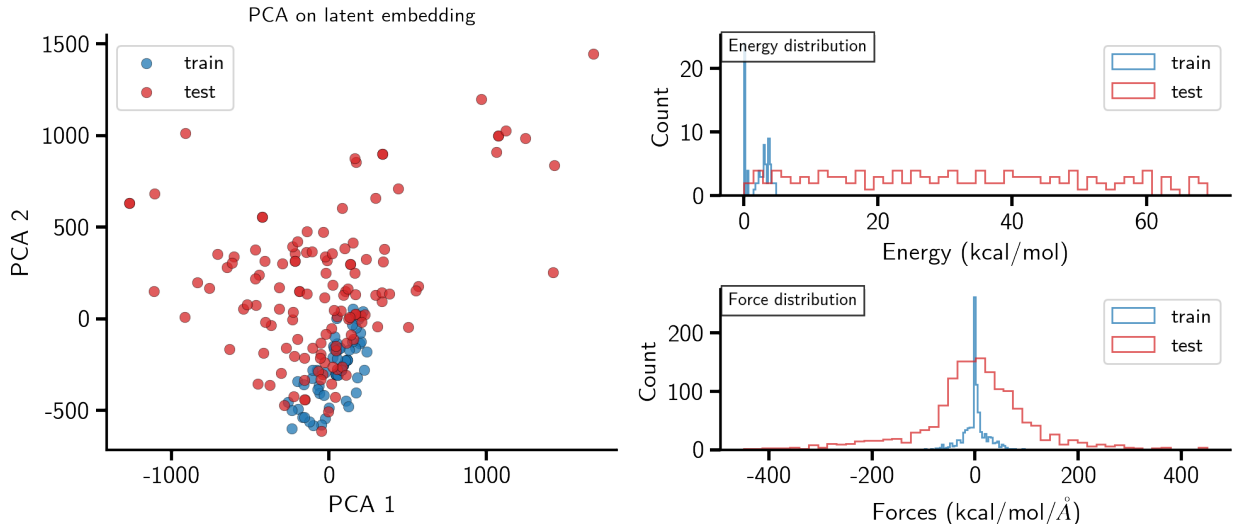
In GMM method, a NN is trained with a regular mean absolute error (MAE) function as shown in Supplementary Equation 1 where $M = 1$, $w_E = 0.1$, and $w_F = 1.0$

II. DATASET

A. Ammonia

In the ammonia dataset, the train (+ validation) and test sets are comprised of 78 and 129 geometries, respectively. The train set geometries are generated from hessian displacement in the direction of normal mode vectors on initial molecular conformers generated using RDKit, whereas the test set is independently picked using adversarial sampling during trial runs of this work. We have built a training set that comprises of low energy geometries near the ground state and a test set that comprises higher energy geometries. This is reasonable, since the training set excludes the type of flat geometries one encounters near the inversion

barrier and further the split by energy in this small molecule also ensures that the training set contains no information about strongly stretched bonds or distorted angles. To ensure that the test set is indeed a good out-of-domain data set, we further analyzed the atomic environments in the train and test sets, as shown in Supplementary Figure 1, by training a simple NN on both the train and test sets and extracting the latent features/embeddings from the last message passing layer before the per-atom prediction of the output. As it turns out, since ammonia is such a small molecule consisting of 4 atoms, partitioning by energy does separate distinct atomic environments into in-domain (ID) and out-of-domain (OOD).



Supplementary Figure 1. Difference between ammonia geometries comprised within the train and test sets are shown. In the left figure, the latent embedding/feature for each geometry is obtained from a NN trained on both the train and test sets. Latent embedding of each geometry is extracted from the immediate layer before the per-atom prediction from the trained NN. Principal component analysis (PCA) is then performed on the latent embedding to visualize the components in 2 dimensions. Note that each point represents the latent embedding of one molecule, not atom. On the right, the subfigures show distributions of the per-molecule energy and per-atom forces for train and test sets of ammonia.

B. Silica

The silica glass (SiO_2) data set is sampled from multiple long molecular dynamics trajectories subjected to different conditions including temperature equilibration, cooling, uniaxial tension, uniaxial compression, biaxial tension, biaxial compression, and biaxial shear at different rates using the force-matching potential described in Urata et al.¹. Each amorphous

silica structure consists of 699 atoms (233 Si and 466 O atoms). In cooling simulations, 5 silica systems were cooled from 2300 K to 300 K at a rate of 1 K/ps and frames were saved every 25 ps. As for the temperature equilibration simulations, the amorphous silica structures were equilibrated at 300 K, 500 K, 1000 K, 1500 K, 2000 K, 2500 K, 3000 K, and 3500 K for 11 ns, and the frames were saved every 100 ps from the last 10 ns of the simulations. Note that the cooling simulations are separate from the temperature equilibration simulations. In the uniaxial deformation condition, we did two different simulations, one each for uniaxial tension and uniaxial compression. Specifically, the silica systems were subjected to a constant strain rate of $\pm 10^{-4} \text{ ps}^{-1}$ (+ indicating tension and $-$ indicating compression), reaching a final strain magnitude of ± 0.30 in one direction and the frames were sampled every 100 ps corresponding to a change in strain of ± 0.1 . In the biaxial tension and compression cases (separate simulations), systems are subjected to a constant strain rate of $\pm 0.0001 \text{ ps}^{-1}$ in the x-direction up to ± 0.30 final strain magnitude. In the y-direction, strain rates of $\pm 2.5 \times 10^{-5} \text{ ps}^{-1}$, $\pm 5.0 \times 10^{-5} \text{ ps}^{-1}$, $\pm 7.5 \times 10^{-5} \text{ ps}^{-1}$, and $\pm 10^{-4} \text{ ps}^{-1}$ were applied, reaching final strain magnitudes of ± 0.075 , ± 0.015 , ± 0.0225 , and ± 0.3 , respectively. In the biaxial shear simulations, tensile stresses are applied in the x-direction to the systems at a strain rate of $\pm 10^{-4} \text{ ps}^{-1}$ up to a maximum strain of -0.3 . In the y-direction, compressive stresses are applied at strain rates of $2.5 \times 10^{-5} \text{ ps}^{-1}$, $5.0 \times 10^{-5} \text{ ps}^{-1}$, $7.5 \times 10^{-5} \text{ ps}^{-1}$, and 10^{-4} ps^{-1} up to maximum strains of 0.075, 0.015, 0.0225, and 0.3, respectively (see Supplementary Table I for tabulated details of the simulations and Supplementary Figure 2 for the data distributions). Frames were sampled every 100 ps. All molecular dynamics (MD) simulations were done using the LAMMPS package using temperature and pressure controls of the Nosé-Hoover thermostat and barostat. Additionally, 101 structures with energies between 0 and 10,000 kcal/mol (14 kcal/mol/atom) were picked from adversarial sampling during trial runs in this work.

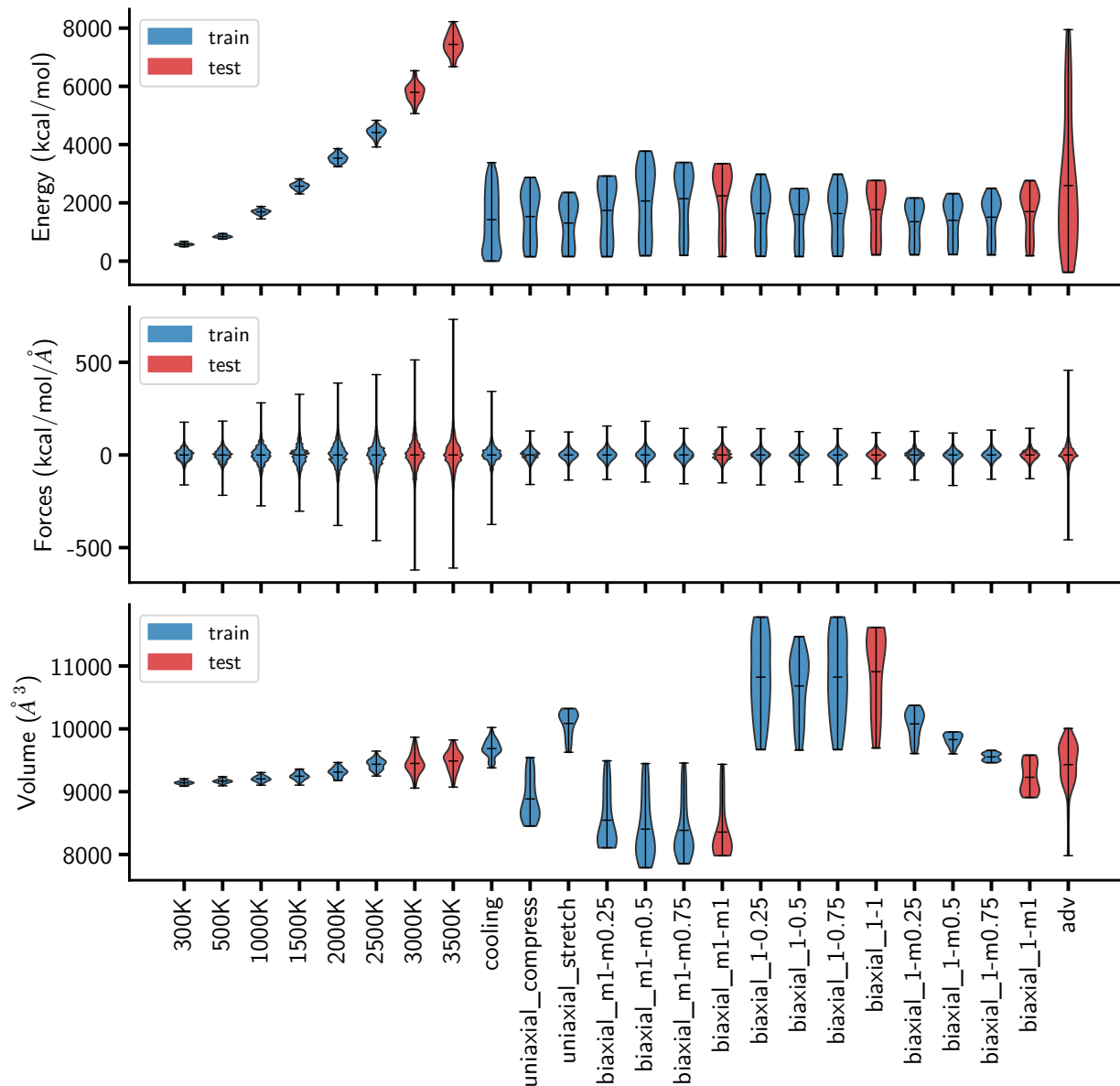
From these structures, we partitioned them into train and test sets, depending on the type of simulations, visualized in Supplementary Figure 2. Train set contains 1318 structures whereas the test set contains 373 structures. No structures appear in both the train and test sets. In the silica glass system with 699 atoms, it is difficult to find all-atom configurations with atomic environments that are completely different from another all-atom configuration. For example, a structure with an energy at 1000 kcal/mol and another at 8000 kcal/mol will still have a considerable overlap in local atomic environments. Hence, partitioning

large structures into ID and OOD data sets is harder. We thus partition the structures in a way that reflects a more generalized challenge in more extreme conditions, where the test set now contains structures extracted from higher temperature and higher deformation rate trajectories, and the training set only those generated under low-temperature, low-deformation conditions. This is also done in Ref². Distributions of the atomic environments, energy and forces are shown in Supplementary Figure 3.

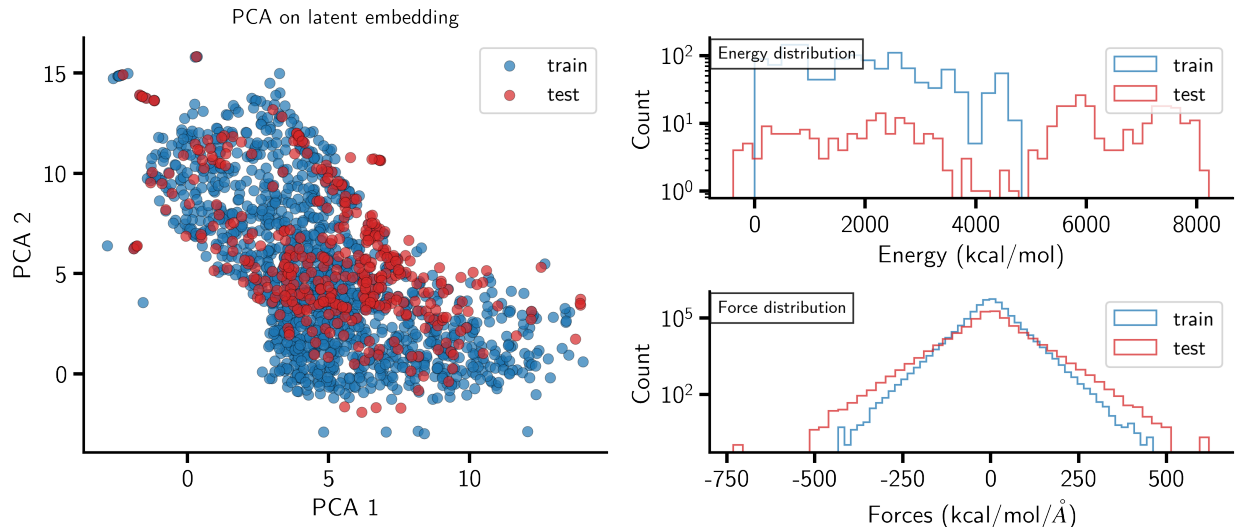
In total, 1691 silica structures were sampled from the molecular dynamic trajectories and adversarial sampling. DFT calculations were performed on the structures using the Vienna Ab-initio Simulation Package (VASP) within the projector-augmented wave (PAW) method. The Perdew-Burke-Ernzerhof (PBE) functional within the generalized gradient approximation (GGA) was adopted as the exchange-correlation functional. The kinetic energy cutoff for plane waves was restricted to 600 eV. The threshold for the energy convergence within a self-consistent field (SCF) cycle was set to 10^{-5} eV.

Simulation Type	Name	Temperature (K)	Heating / cooling rate (K/ps)	Strain rate in x direction (10^{-4} /ps)	Strain rate in y direction (10^{-4} /ps)	Num frame
Temperature equilibration	300 K	300	-	-	-	100
	500 K	500	-	-	-	100
	1000 K	1000	-	-	-	100
	1500 K	1500	-	-	-	100
	2000 K	2000	-	-	-	100
	2500 K	2500	-	-	-	99
	3000 K	3000	-	-	-	98
	3500 K	3500	-	-	-	92
Cooling	cooling	-	1.00	-	-	410
Uniaxial deformation	uniaxial_compress	300	-	-1.00	-	30
	uniaxial_stretch	300	-	1.00	-	30
Biaxial compression	biaxial_m1-m0.25	300	-	-1.00	-0.25	30
	biaxial_m1-m0.5	300	-	-1.00	-0.50	30
	biaxial_m1-m0.75	300	-	-1.00	-0.75	30
	biaxial_m1-m1	300	-	-1.00	-1.00	30
Biaxial stretch	biaxial_1-0.25	300	-	1.00	0.25	23
	biaxial_1-0.5	300	-	1.00	0.50	26
	biaxial_1-0.75	300	-	1.00	0.75	23
	biaxial_1-1	300	-	1.00	1.00	22
Biaxial shear	biaxial_1-m0.25	300	-	1.00	-0.25	28
	biaxial_1-m0.5	300	-	1.00	-0.50	29
	biaxial_1-m0.75	300	-	1.00	-0.75	30
	biaxial_1-m1	300	-	1.00	-1.00	30
Adversarial sampling	adv	-	-	-	-	101
						1691

Supplementary Table I. Table containing breakdown of the initial silica glass dataset (does not contain newly sampled structures from the active learning loops).



Supplementary Figure 2. Violin plots showing energy, force and cell volume distributions of the initial silica glass dataset. Horizontal marks in each violin plot denote the maximum, mean, and minimum value of the distributions. Dataset contains density functional theory (DFT) calculations of silica glass frames extracted from molecular dynamics (MD) trajectories performed under various conditions (See Supplementary Table I). All silica systems contain 699 atoms. In the biaxial tension cases, the first and second numbers after the underscore (“_”) indicate the strain rates in the magnitude of $1 \times 10^{-4} \text{ ps}^{-1}$ in the x- and y-directions, respectively. “m” denotes minus, meaning that a compressive strain is applied. The term “adv” indicates that the structures are obtained from adversarial sampling.

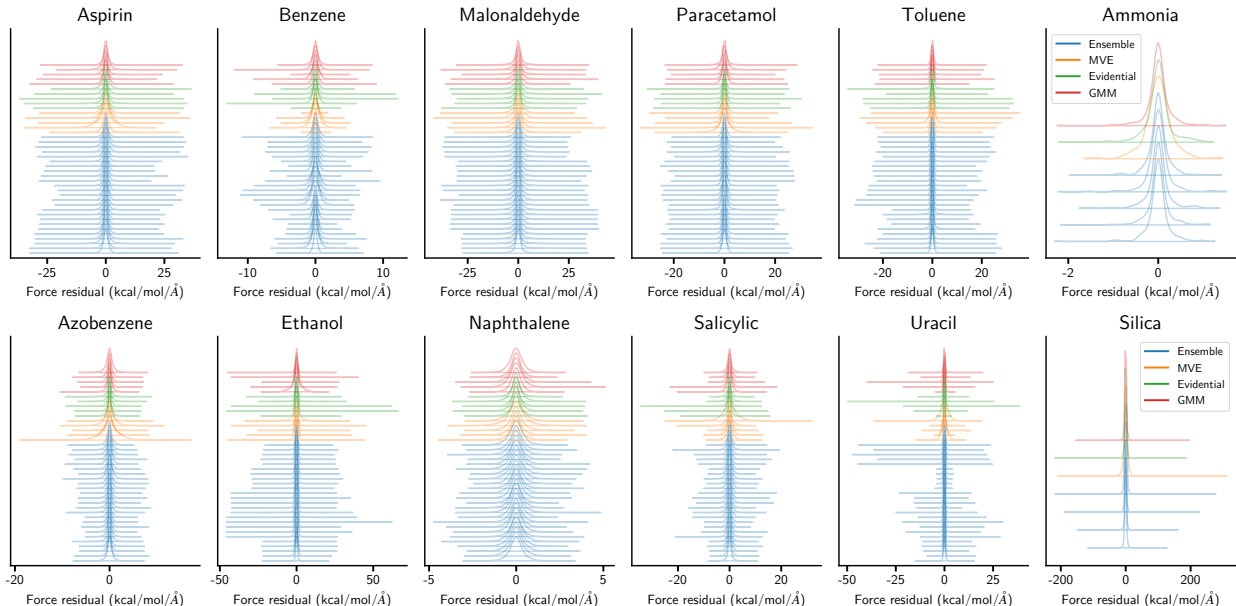


Supplementary Figure 3. Difference between silica glass structures comprised within the train and test sets are shown. In the left figure, the latent embedding/feature for each atom within a structure is obtained from a NN trained on both the train and test sets. Latent embedding of each atom is extracted from the immediate layer before the per-atom prediction from the trained NN. Principal component analysis is then performed on the latent embedding to visualize the components in 2 dimensions. Note that each point represents the latent embedding of one atom, not the entire structure. On the right, the subfigures show distributions of the per-molecule energy and per-atom forces for train and test sets of silica glass.

III. ADVERSARIAL SAMPLING

In the adversarial sampling strategy, a perturbation, δ is applied to randomly chosen structures, x from the initial training data. The perturbation is then iteratively updated using gradient ascent technique to maximize the uncertainty of the perturbed structures, $U(x_\delta)$, as defined in Equation 15 in the main paper. For our purposes, the perturbation update was optimized for 80 iterations at a normalized temperature kT of 0.7 kcal/mol/atom. The learning rates were set to 5×10^{-5} and 3×10^{-5} for ammonia and silica, respectively. 20 adversarial conformers were sampled at each generation (i.e., active learning loop) for ammonia while 100 adversarial conformers were sampled for silica. Finally, we set a cutoff of 100 kcal/mol for new adversarial samples for ammonia such that the geometries are not extremely distorted. For silica, the cutoff was set at 5000 kcal/mol (7.2 kcal/mol/atom) for generation 2 and increased to 10000 kcal/mol (14.3 kcal/mol/atom) at generation 3.

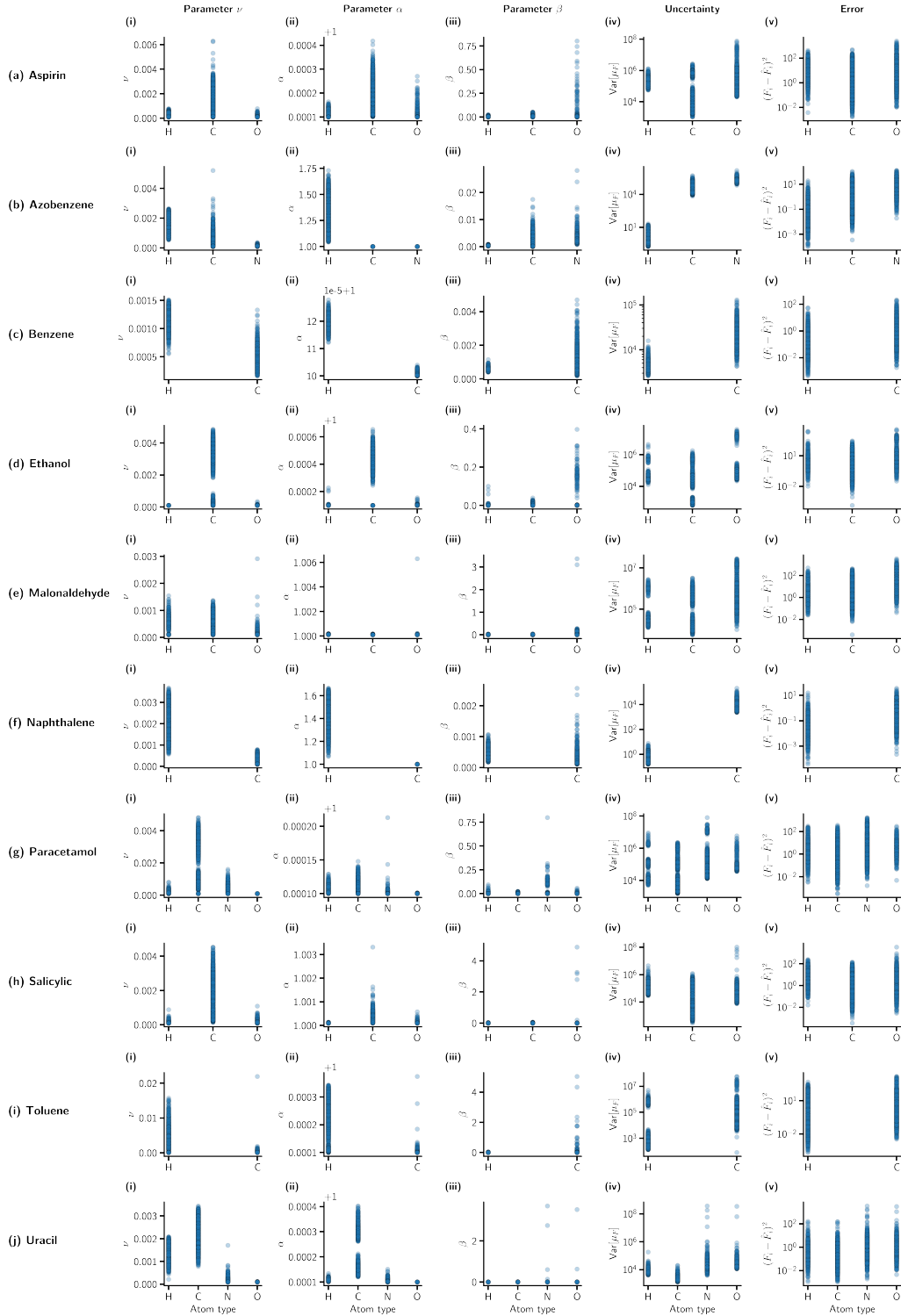
IV. POTENTIAL BIASES IN NN ENSEMBLES



Supplementary Figure 4. Distributions of residuals of predicted forces of individual neural networks in the NN ensembles. For each molecule in the rMD17 data sets, one NN ensemble or one single NN each is trained on each fold split of the data. Each ensemble contains 5 NNs, which totals 25 NNs. There are a total of 5 and 4 NNs in the ensemble for ammonia and silica, respectively. Force residual for each atom i is calculated as $\frac{1}{3}\sum_{x,y,z}(\hat{\mathbf{F}}_i - \mathbf{F}_i)$, where residuals in all directions are averaged.

In terms of performance, it is not uncommon for individual models to occasionally outperform ensembles, especially when comparing peak performances. In order to understand if there could exist biases in the NN ensembles or single NNs that could make the models seem better or worse, residuals of the predicted forces of all models are visualized in Supplementary Figure 4. Residuals of the predicted forces are calculated as $\frac{1}{3}\sum_{x,y,z}(\hat{\mathbf{F}}_i - \mathbf{F}_i)$, where x, y, z indicates the Cartesian direction, while $\hat{\mathbf{F}}_i$ and $\mathbf{F}_i$ indicate the predicted and ground truth force of atom i , respectively. From the figure, the residuals of the predicted forces, for all data sets, are symmetrically distributed around the zero point. This symmetry suggests the absence of systematic biases in the model predictions. The extent of the tails in the distribution reflects the variability in prediction errors but is also dependent on the data splits and weight initialization.

V. BIMODAL DISTRIBUTION IN EVIDENTIAL UNCERTAINTY



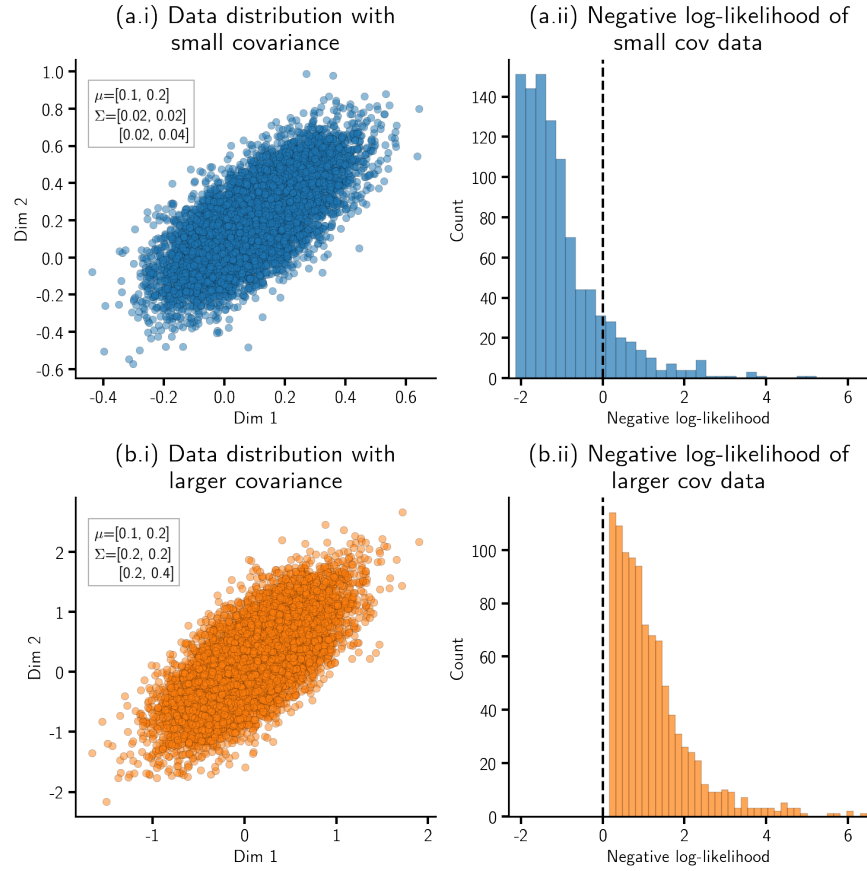
Supplementary Figure 5. Values of predicted evidential parameters (a) ν , (b) α , (c) β , (d), epistemic uncertainty, $Var[\mu_F]$, and (e), atomic force squared error, $\frac{1}{3}\Sigma_{x,y,z}(\hat{\mathbf{F}}_i - \mathbf{F}_i)^2$, as a function of atom types for molecules in the rMD17 data sets. As shown in Equation 9 of the main paper, epistemic uncertainty is calculated as $Var[\mu_F] = \frac{\beta}{\nu(\alpha-1)}$.

As seen in Figure 2a of the main paper (for uncertainty-error distributions of individual molecules, see Supplementary Figures 10 and 11), there is a very distinct bimodal distribution for evidential uncertainty against atomic force errors. When visualizing the evidential parameters α , β , and ν as a function of the atom types (Supplementary Figure 5), we found that the distributions of these parameters, especially α , depend heavily on the atom type in molecules such as azobenzene (Supplementary Figure 5b) and naphthalene (Supplementary Figure 5f). This then causes the epistemic uncertainty, $Var[\mu_F] = \frac{\beta}{\nu(\alpha-1)}$, to form distinct distributions w.r.t. the atom type due to small differences in the denominator values. We further found that this dependency to have no correlation with the predicted atomic forces and its corresponding error. While we still cannot find a definite reason for such behavior, we believe it is caused by the complicated loss function (Supplementary Equation 6) the neural network is trying to minimize.

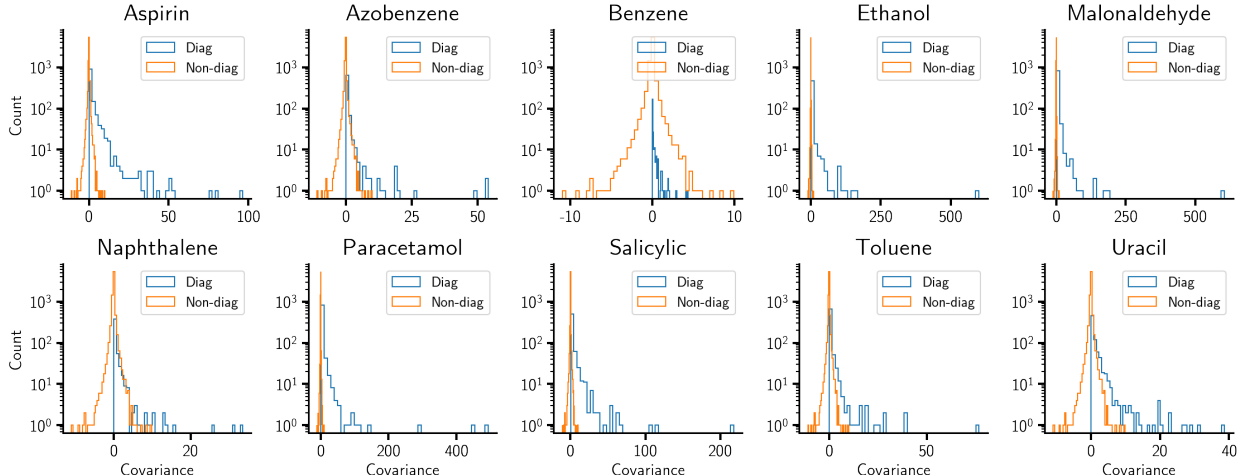
VI. NON-POSITIVE NEGATIVE LOG-LIKELIHOOD VALUES IN GAUSSIAN MIXTURE MODELS (GMM)

Another anomaly in the results is the non-positive negative log-likelihood values in GMM. In the GMM uncertainty-error distribution shown in Figure 2a of the main paper, the negative values are not shown since all values are scaled by the minimum value to yield positive quantities for plotting on a log scale. The occurrence of the non-positive negative log-likelihood values is not impossible. Intuitively, likelihood is between 0 and 1, in which case the logarithm of likelihood will yield values between negative infinity and 0. By applying a sign inversion to that, only non-negative values will appear (0 - infinity). However, likelihood is defined as the probability density data viewed as a function of the parameters of a statistical model. In cases where (co)variance of a univariate/multivariate normal distribution is very small, the probability density of observing a high probability value could become more than 1. This seems to defy our intuition that probabilities can only exist between 0 and 1, but note that we are taking here the logarithm of probability densities and not probabilities. A likelihood more than 1 thus yields a NLL value lower than zero. This is better illustrated in Supplementary Figure 6, showing the effect of covariance magnitudes on the negative log-likelihood values. We have also plotted the distribution of covariances from GMMs fitted on the rMD17 dataset in Supplementary Figure 7 to show that the covariances

contain very small values close to zero that causes the non-positive negative log-likelihood values.



Supplementary Figure 6. **(a)** For a 2D distribution with very small covariances, negative log-likelihood values for some observed data are lower than zero. **(b)** With higher covariances but same mean values, the negative log-likelihood values are higher than zero.



Supplementary Figure 7. Distributions of covariances of Gaussian mixture models (GMM) fitted on NN latent features for rMD17 dataset. The covariances are separately visualized as diagonal and non-diagonal components.

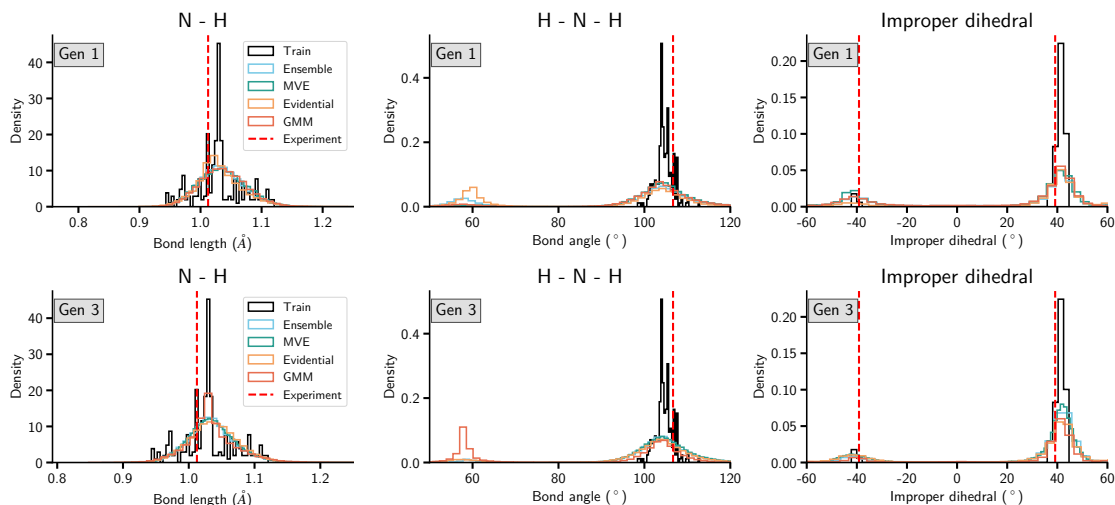
VII. BOND GEOMETRY ANALYSIS OF STRUCTURES FROM NNIP MD

To evaluate the efficacy of the Molecular Dynamics (MD) simulations conducted using the trained Neural Network Interatomic Potentials (NNIPs), we scrutinize the bond geometries of ammonia and silica glass structures. In the case of ammonia, we depict in Supplementary Figure 8 the relative frequency of bond distances between nitrogen (N) and hydrogen (H) atoms, bond angles (H - N - H), and the improper dihedral (out-of-plane angle). The improper dihedral in ammonia is determined as the angle between the planes (H, N, H) and (H, H, H). For silica glass, we display the relative frequency of all atom pair distances, namely silicon-silicon (Si-Si), oxygen-oxygen (O-O), and silicon-oxygen (Si-O), alongside bond angles Si-O-Si and O-Si-O (Supplementary Figure 9).

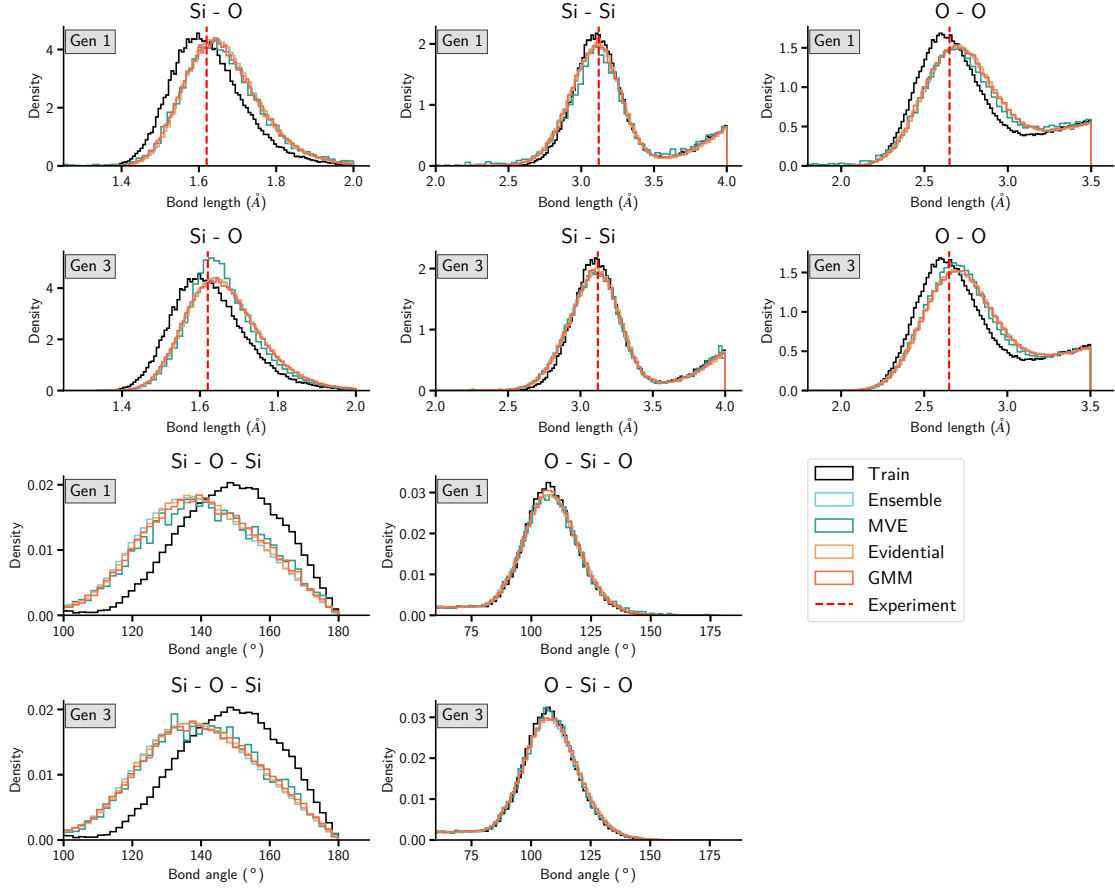
Supplementary Figure 8 shows bond geometry information from the training data and MD simulations of the NVT ensemble at 500 K using NNIPs of the different UQ methods. Bond length and bond angle distributions show very similar ranges to those of the training data. However, unphysical H-N-H bond angles seem to occur for some NN-based MD geometries. Those angles are around 60° small, which causes the involved H atoms to be around $0.8 \text{ \AA}$ apart. Intriguingly, the NNIPs associated with the Gaussian Mixture Model (GMM) method yield a higher frequency of such geometries in the 3rd generation; however, the fraction of MD simulations reaching successful completion increases when compared to

the 1st generation (as depicted in Figure 3g of the main paper). This is because the 1st generation NNIPs actually produced more geometries with H-N-H angles even smaller than 60° , which causes the simulations to fail and thus those geometries are not included in this bond angle analysis. It is also important to highlight that the PaiNN architecture does not distinguish between geometries with improper dihedrals around 40° and -40° . Nonetheless, a majority of geometries exhibits dihedrals around 40° because the initial configurations for the MD simulations were sampled from the training data, which predominantly features geometries with dihedrals around 40° , and ammonia inversions occurred only rarely during the MD simulations.

For the silica glass analysis (Supplementary Figure 9), all MD simulations were carried out with constant volume at 2500K. Note that the configurations of the initial data set were generated with the force-matching potential from Urata et al., energies and forces were computed with Density Functional Theory (DFT) (detailed description in Supplementary Section II B). The experimentally measured values of Si-O, O-O, and Si-Si atom pair distances derived from X-ray diffraction by Mozzi and Warren³ are depicted as red dashed lines. As one can see in Supplementary Figure 9, the Si-O and O-O atom pair distances from the NN-driven MD simulations have shifted peaks relative to the training data, even in the third generation of NNIPs. A similar observation applies to the Si-O-Si bond angle. However, our NNIPs yield mean bond distance values more closely aligned with experimental values compared to the initial data set and the ranges of sampled distances is the same as in the training data. Those differences arise, since the training data, initially generated using the force-matching potential, exhibited minimum energy Si-O and O-O atom pair distances that differed from those of the DFT-functional. Thus, all NNIPs properly learned the DFT PES. The slight overestimation of bond lengths by the NNIPs compared to the experimental values can be easily explained, as GGA exchange-correlation functionals are known for their tendency to overestimate atomic distances and thus generate lattice constants larger than observed experimentally⁴.



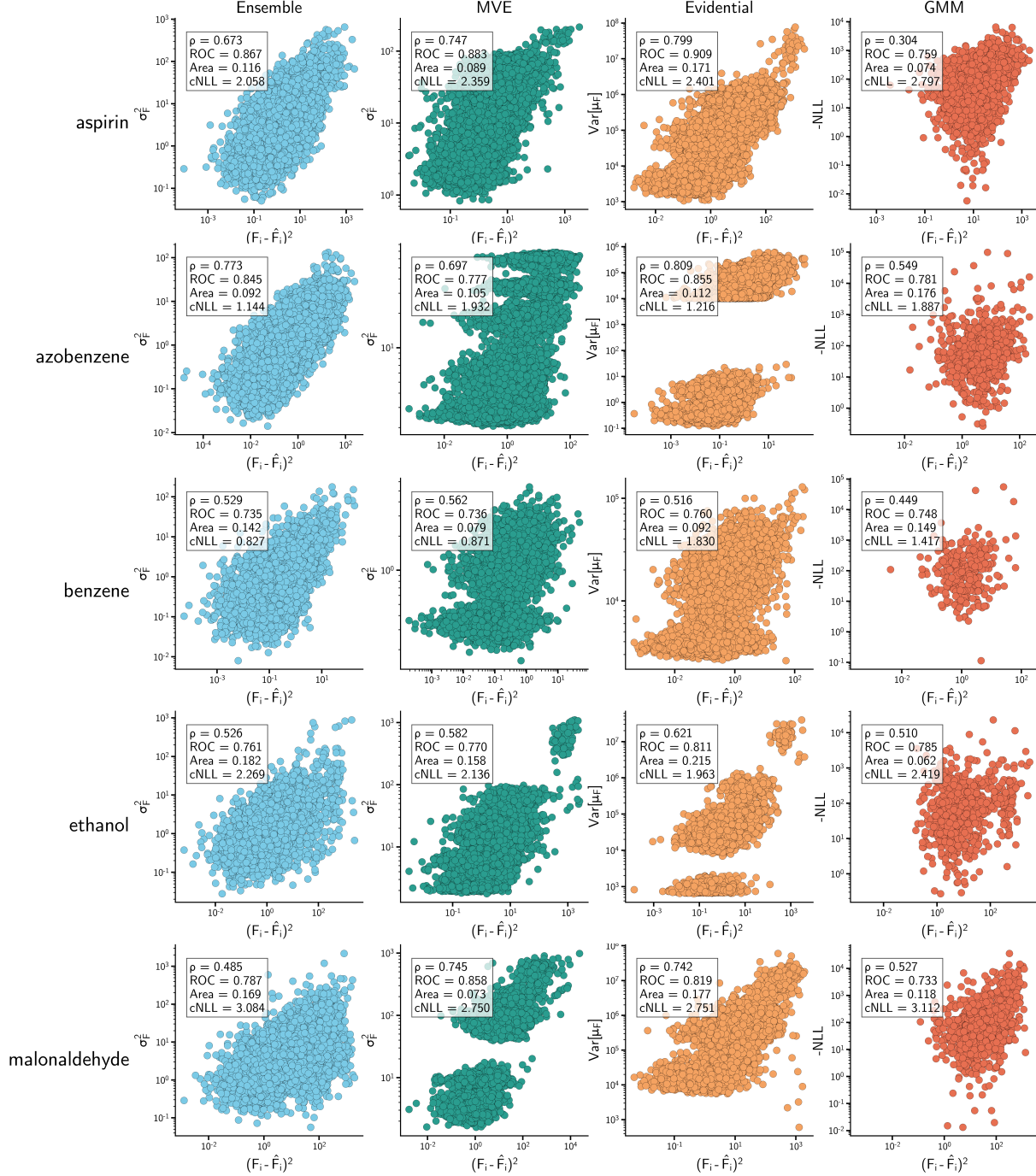
Supplementary Figure 8. N-H bond length, H-N-H bond angle, and improper dihedral angle distributions in the initial ammonia data set and NVT MD simulations at 500 K produced using NNIPs corresponding to the UQ methods in generations 1 and 3. The generations are shown in the box on the top left corner of each subfigure. Experimental values for 300 K, shown as vertical red dashed lines in the subfigures, for N-H bond distance, H-N-H bond angle, and improper dihedral are at 1.0124 Å, 106.67°, and $\pm 39.142^\circ$, respectively⁵.



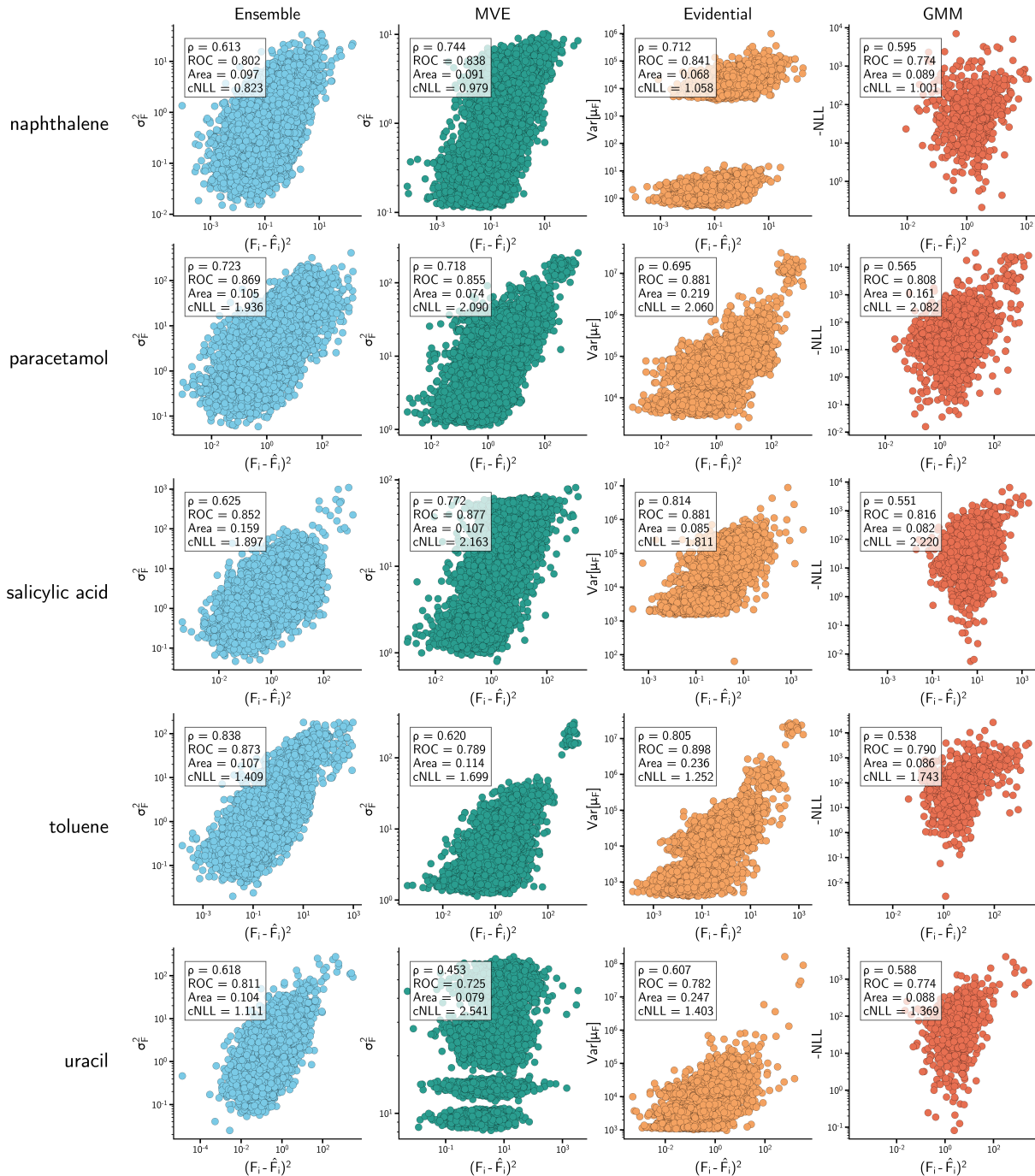
Supplementary Figure 9. Atom pair distance and bond angle comparison in silica glass structures of initial training data (see Supplementary Section IIB for details) and NVT MD simulations at 2500 K produced using NNIPs corresponding to the UQ methods in generations 1 and 3. The generations are shown in the box on the top left corner of each subfigure. Subfigures in rows 1 and 2 show the Si-Si, Si-O, and O-O distance distributions, whereas rows 3 and 4 show the bond angle distributions of Si-O-Si and O-Si-O. Experimental results by Mozzi and Warren³ for the Si-O, Si-Si, and O-O pair distances are shown as vertical red dashed lines at 1.62 Å, 3.12 Å, and 2.65 Å, respectively.

VIII. ADDITIONAL RESULTS

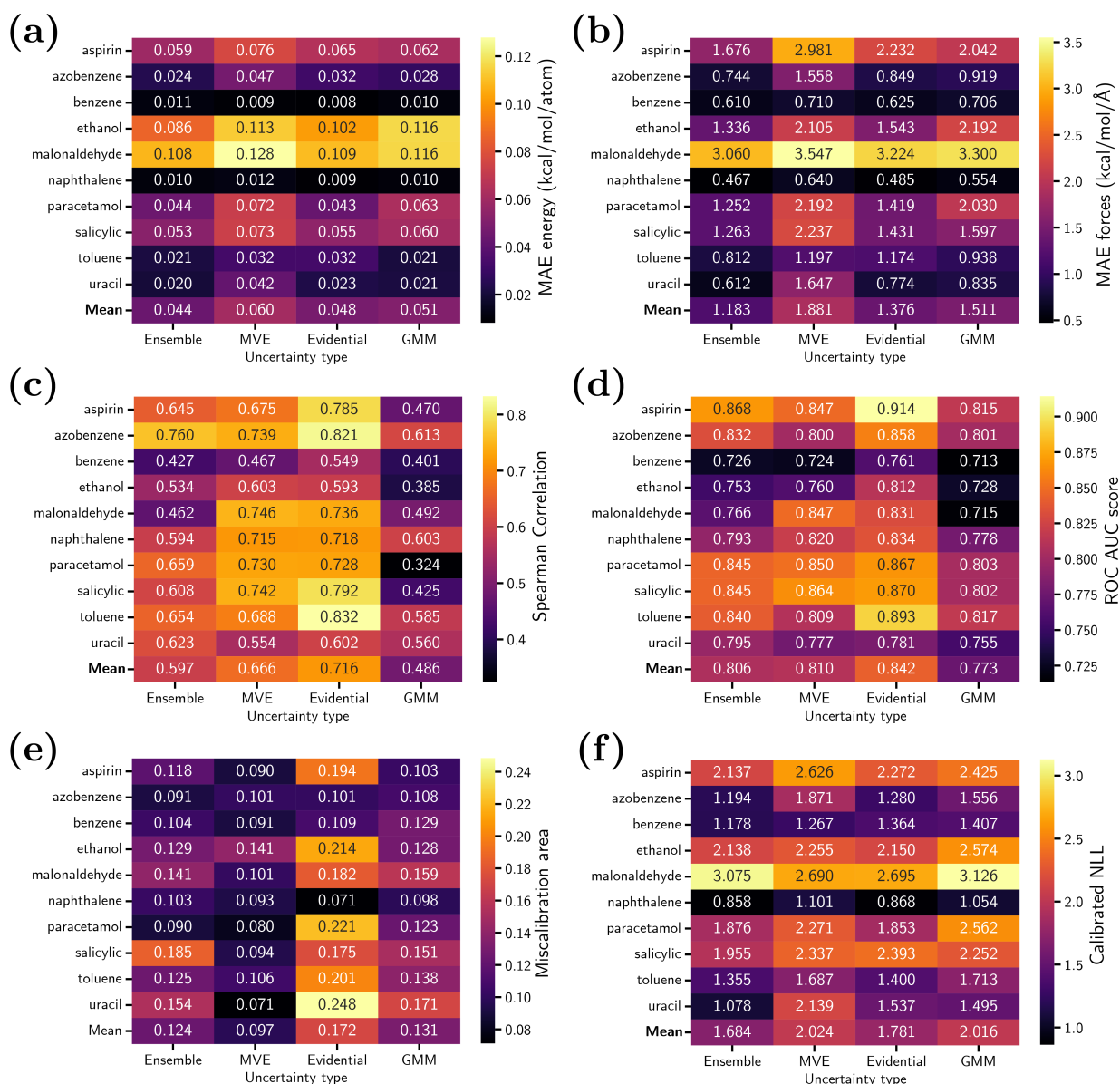
A. rMD17 dataset



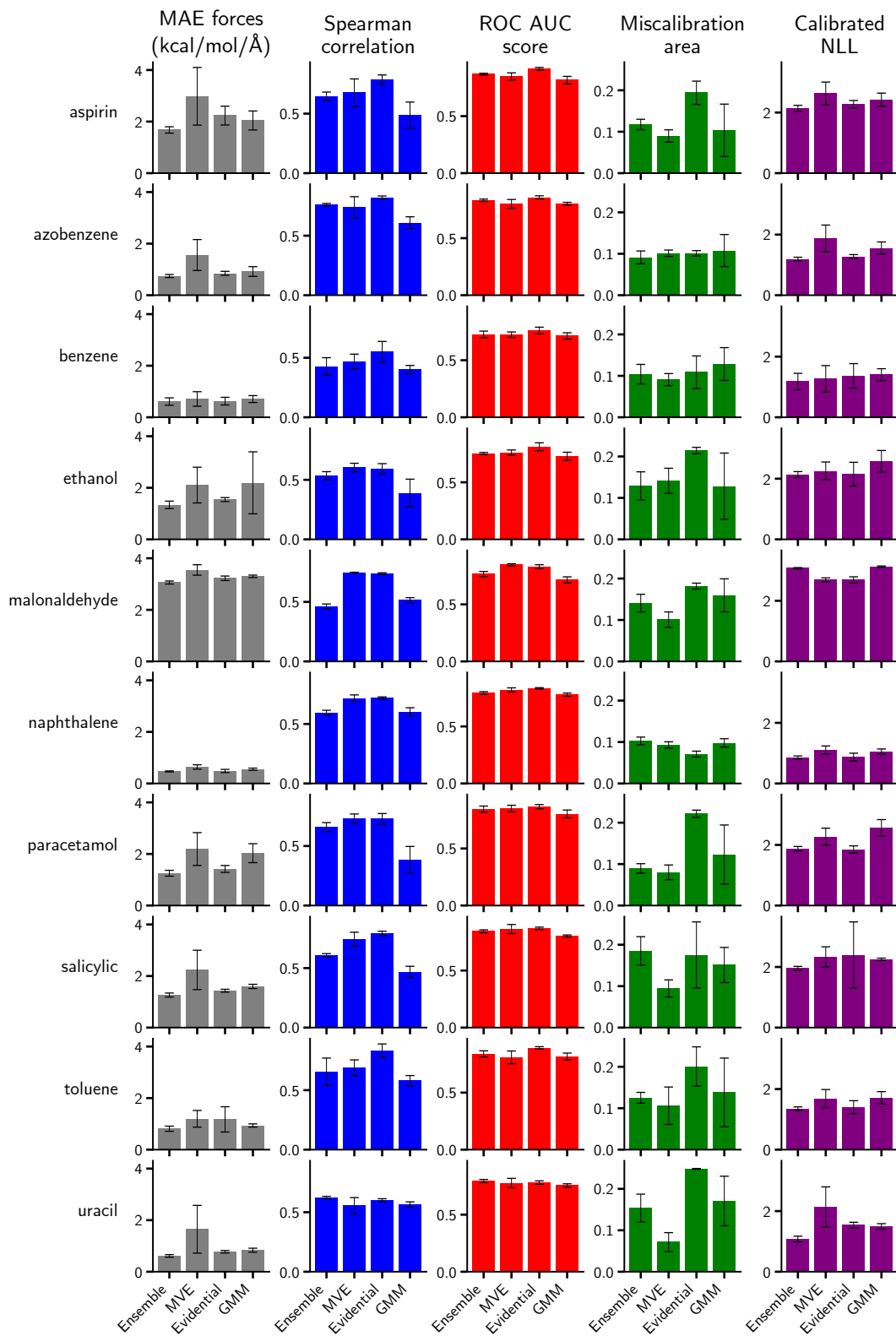
Supplementary Figure 10. Scatter plots showing (predicted) uncertainties versus squared errors of atomic forces of the aspirin, azobenzene, benzene, ethanol, and malonaldehyde molecules in the rMD17 dataset for each considered uncertainty estimation method. Note that since the uncertainties (-NLL) for the GMM method contain negative values, all uncertainties are scaled by the minimum value to yield positive quantities for plotting of the log scale figure.



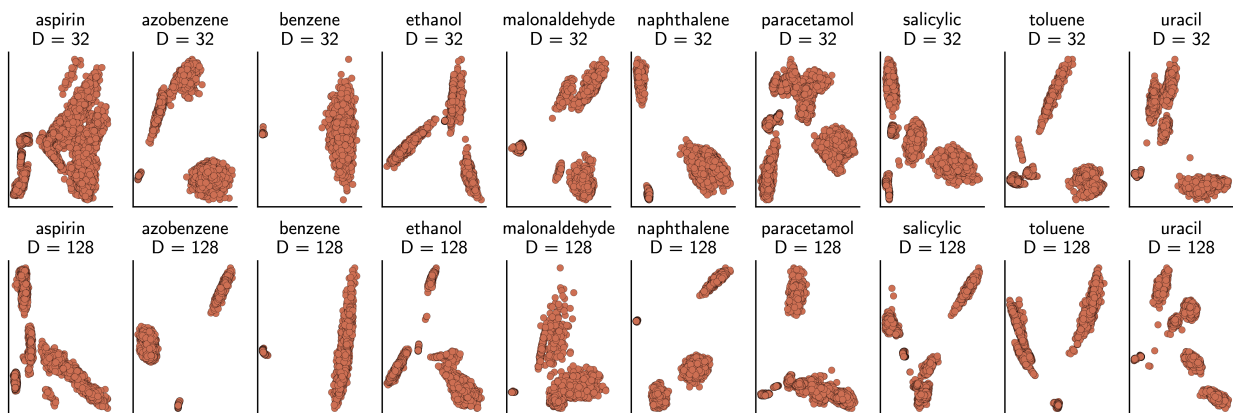
Supplementary Figure 11. Scatter plots showing (predicted) uncertainties versus squared errors of atomic forces of the naphthalene, paracetamol, salicylic acid, toluene, and uracil molecules in the rMD17 dataset for each considered uncertainty estimation method. Note that since the uncertainties (-NLL) for the GMM method contain negative values, all uncertainties are scaled by the minimum value to yield positive quantities for plotting of the log scale figure.



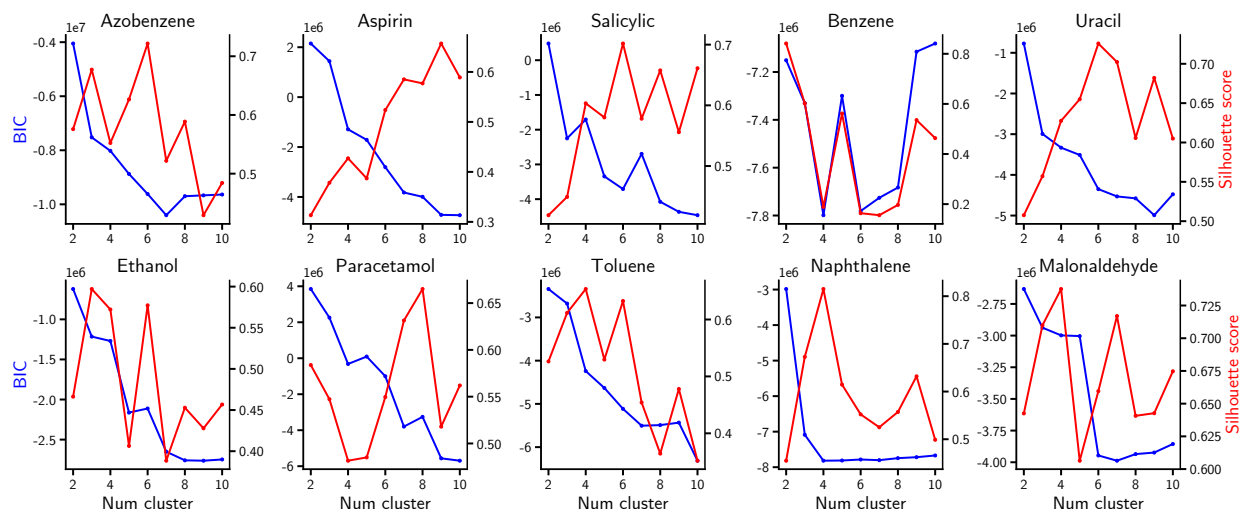
Supplementary Figure 12. Heatmaps of prediction accuracies and evaluation metrics for uncertainty estimation on the test sets of rMD17 dataset for each considered uncertainty estimation methods. Each cell represents the average for all splits on each molecule. (a) Mean absolute error of energy in unit kcal/mol/atom. (b) Mean absolute error of atomic forces in unit kcal/mol/Å. (c) Spearman's rank correlation coefficient. (d) Area under the curve of receiver operating characteristics curve (ROC AUC score). (e) Miscalibration area. (f) Calibrated negative log-likelihood score.



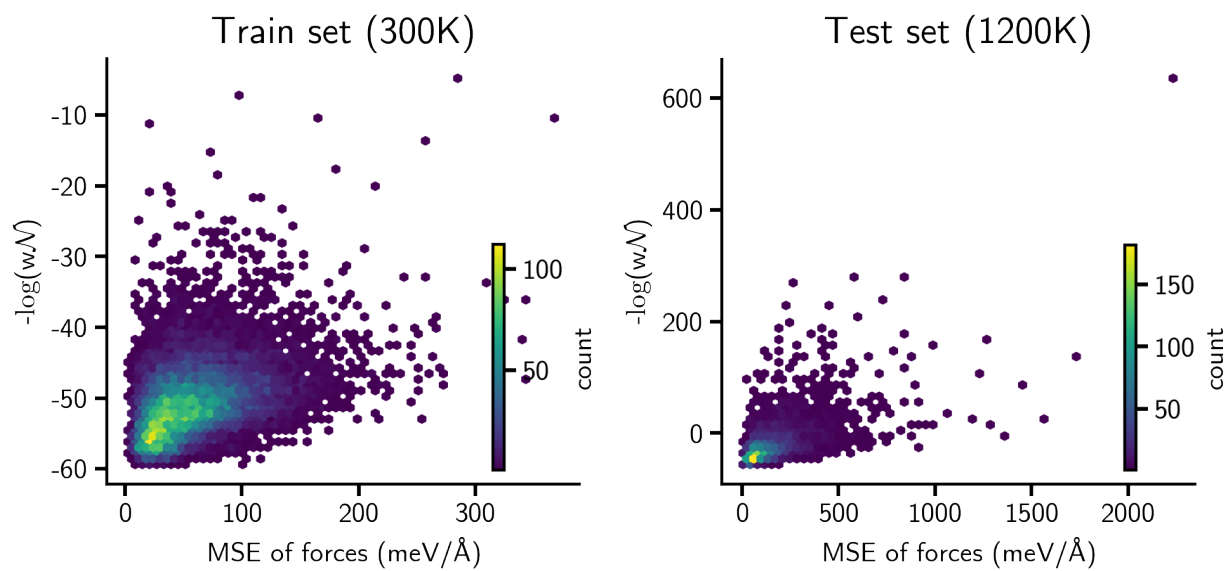
Supplementary Figure 13. Bar charts showing easier comparison of the prediction accuracy (mean absolute error of atomic forces) and uncertainty estimation evaluation metrics on all molecules in the rMD17 dataset for each considered uncertainty estimation methods. Error bars denote the standard deviation across the 5-fold test sets.



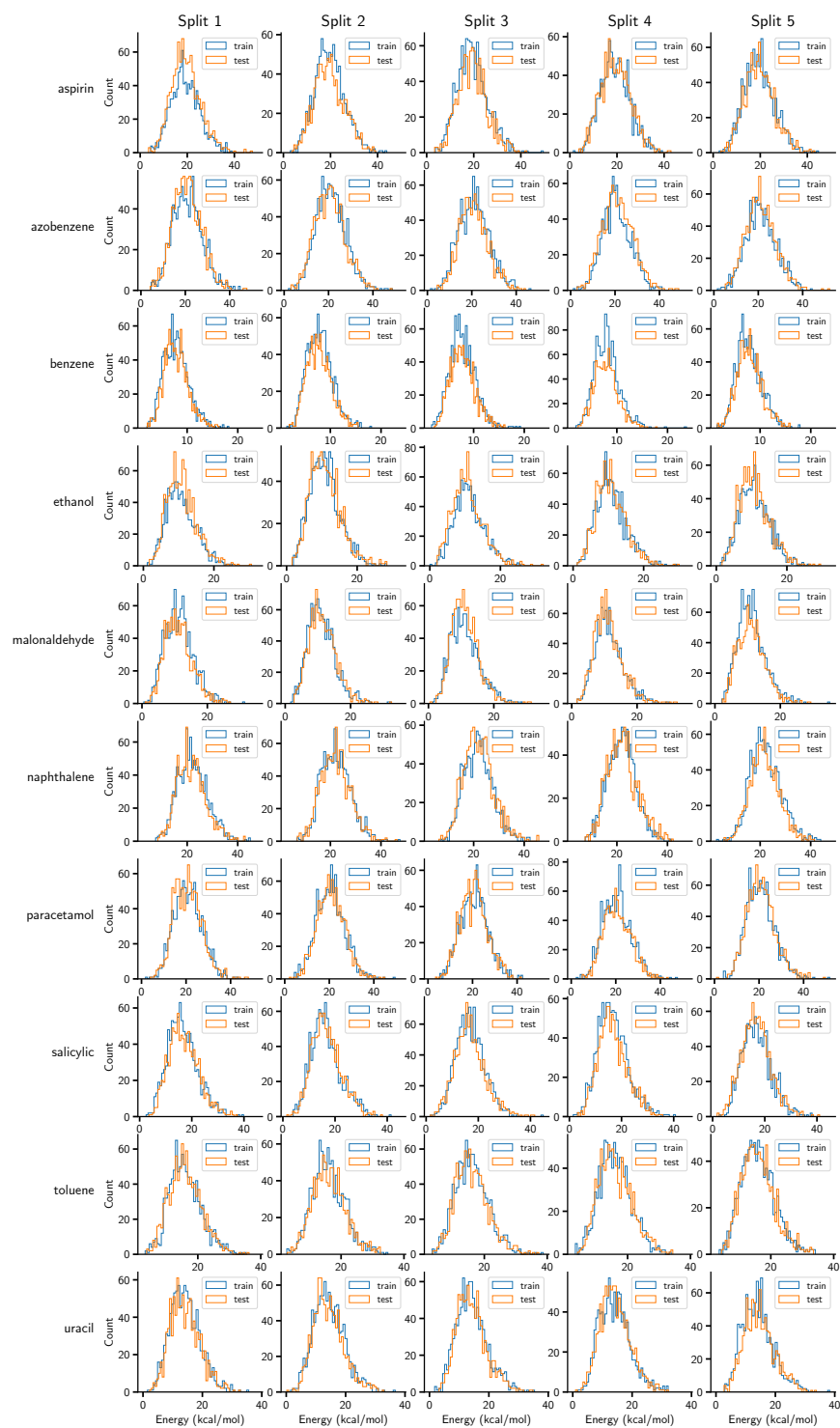
Supplementary Figure 14. Principal component analysis (PCA) of the latent embeddings in neural networks (NNs) used for Gaussian mixture models (GMM) on each molecule in the rMD17 dataset (for the train set of one split), illustrating that the data set can be fitted well with a mixture of multivariate Gaussians.



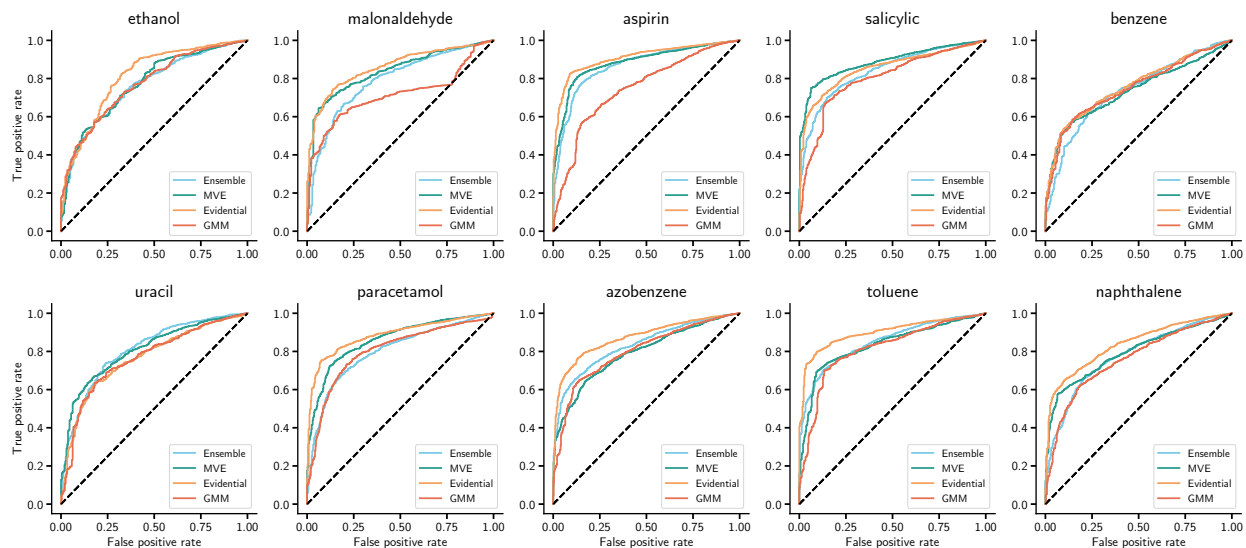
Supplementary Figure 15. Plots showing Bayesian information criteria (BIC) and silhouette scores with respect to the number of clusters/components used for fitting latent embeddings in neural networks (NNs) used for GMM. The number of clusters can then be determined for optimal fitting of the GMMs by examining the tradeoff between the BIC and silhouette scores. A model that fit well would ideally have lower BIC and higher (closer to 1) silhouette scores.



Supplementary Figure 16. Replication of the 3BPA molecule results in the original paper proposing GMM⁶ for uncertainty estimation scheme.



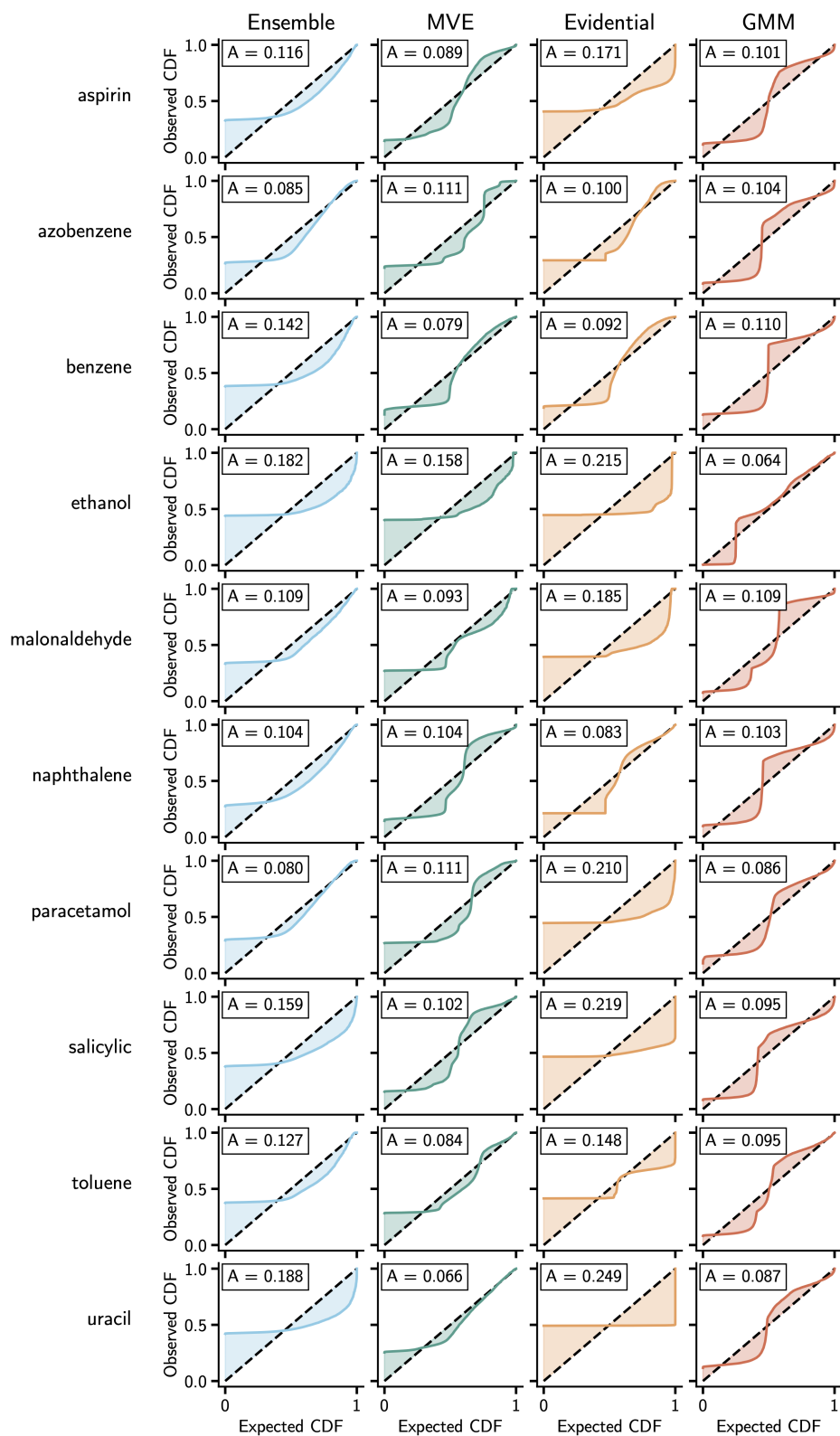
Supplementary Figure 17. Energy distribution of train and test sets of the rMD17 dataset for all splits (across) of all molecules (down).



Supplementary Figure 18. Receiver operating characteristics (ROC) curve of each considered uncertainty estimation method for all molecules in the rMD17 dataset. The threshold specified for the “target” (true error) is at 20th percentile of the distribution.

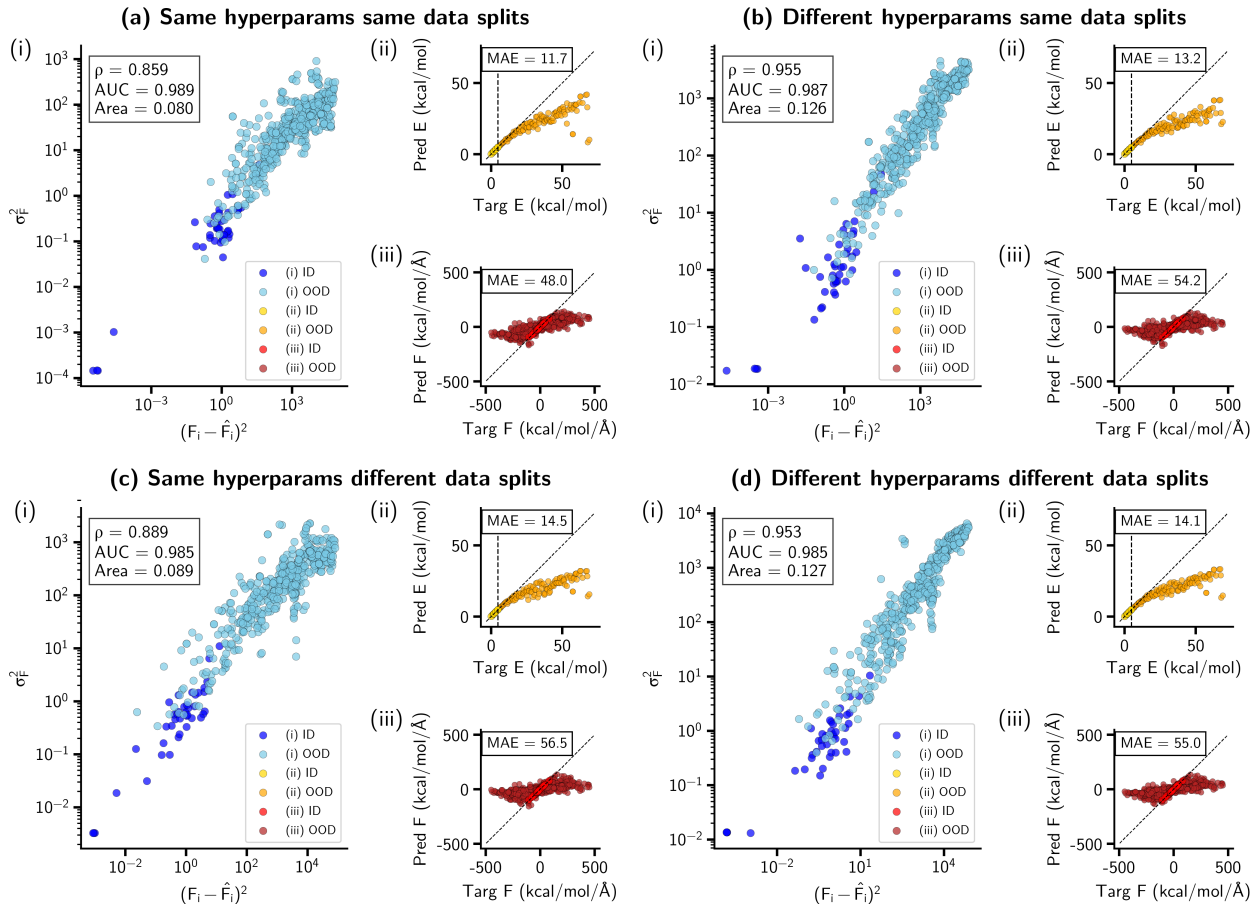
	20th percentile				50th percentile			
	Ensemble	MVE	Evidential	GMM	Ensemble	MVE	Evidential	GMM
aspirin	0.853	0.840	0.921	0.783	0.815	0.834	0.887	0.733
azobenzene	0.853	0.826	0.890	0.807	0.883	0.873	0.915	0.821
benzene	0.700	0.716	0.764	0.702	0.702	0.733	0.777	0.698
ethanol	0.745	0.765	0.780	0.693	0.755	0.800	0.779	0.679
malonaldehyde	0.763	0.866	0.849	0.745	0.741	0.887	0.876	0.786
naphthalene	0.786	0.840	0.843	0.781	0.788	0.868	0.867	0.800
paracetamol	0.837	0.862	0.874	0.773	0.822	0.870	0.854	0.690
salicylic	0.835	0.884	0.904	0.794	0.811	0.876	0.901	0.738
toluene	0.842	0.821	0.924	0.821	0.808	0.839	0.912	0.784
uracil	0.793	0.771	0.787	0.758	0.805	0.779	0.802	0.779
Mean	0.801	0.819	0.854	0.766	0.793	0.836	0.857	0.751

Supplementary Table II. Area under the curve of receiver operating characteristics (ROC AUC) score of the considered uncertainty estimation methods in the rMD17 dataset using the 20th and 50th percentile to determine the binary cutoff in the true error (target) distribution.

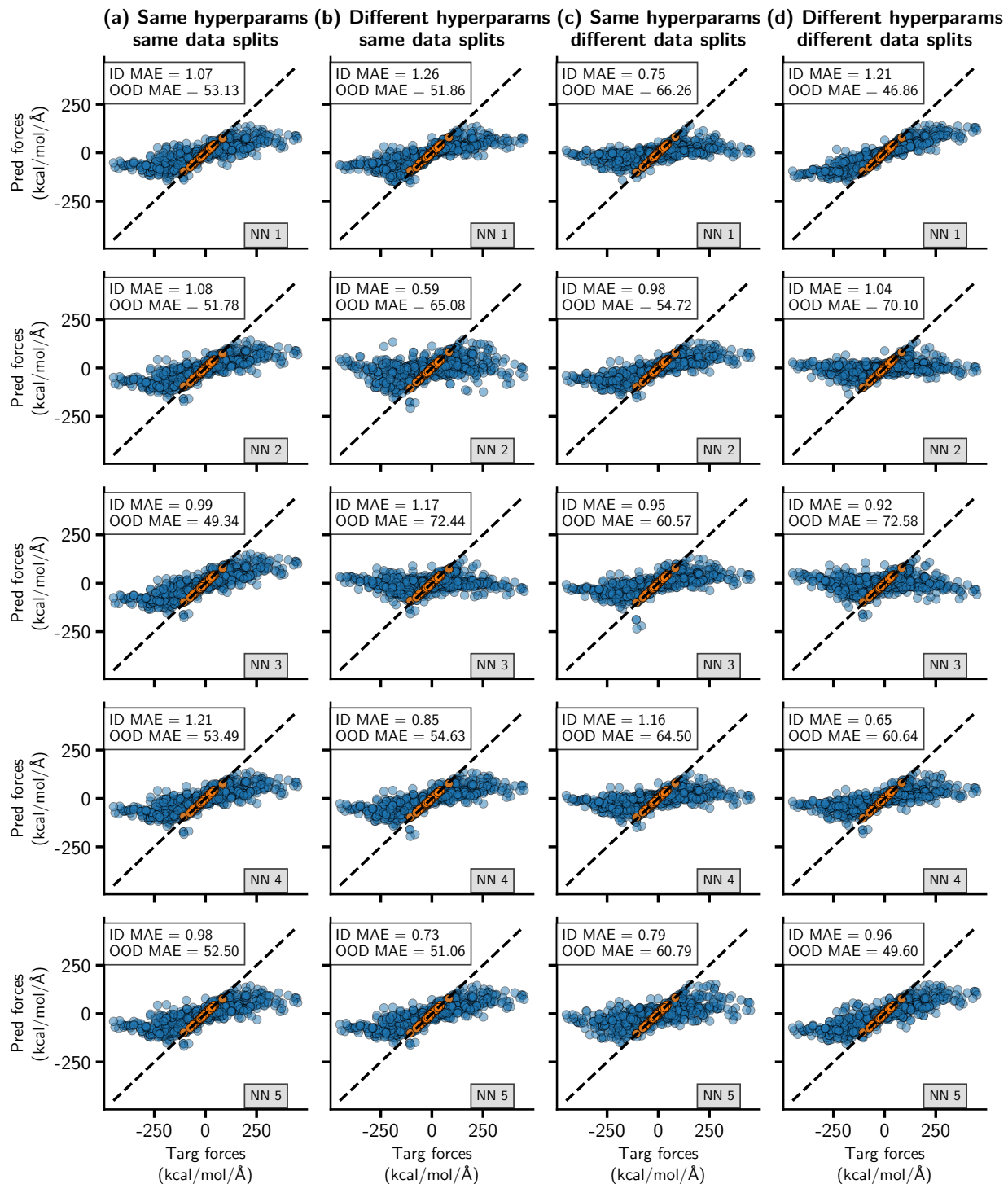


Supplementary Figure 19. Calibration curves of each considered uncertainty estimation method (across) for all molecules in the rMD17 dataset. Boxes at the top left of each subfigure indicates the miscalibration area, which is the shaded area between the calibration curve and the ideal curve (dashed).

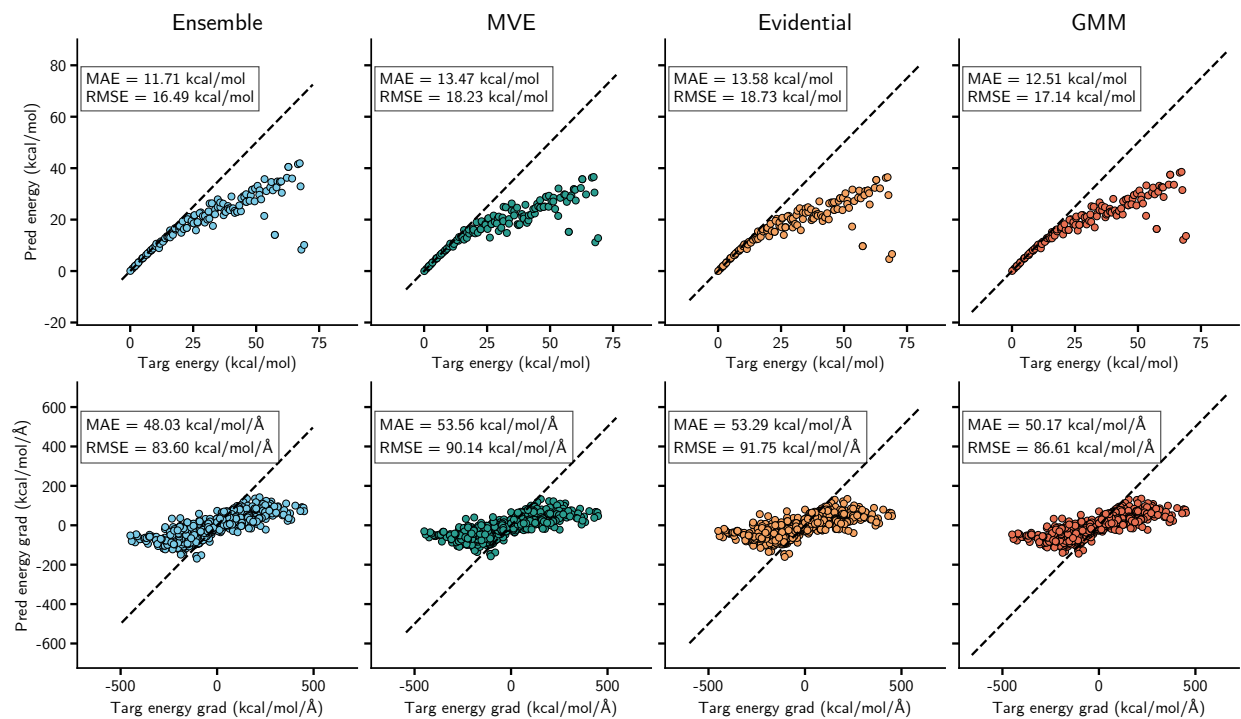
B. Ammonia dataset



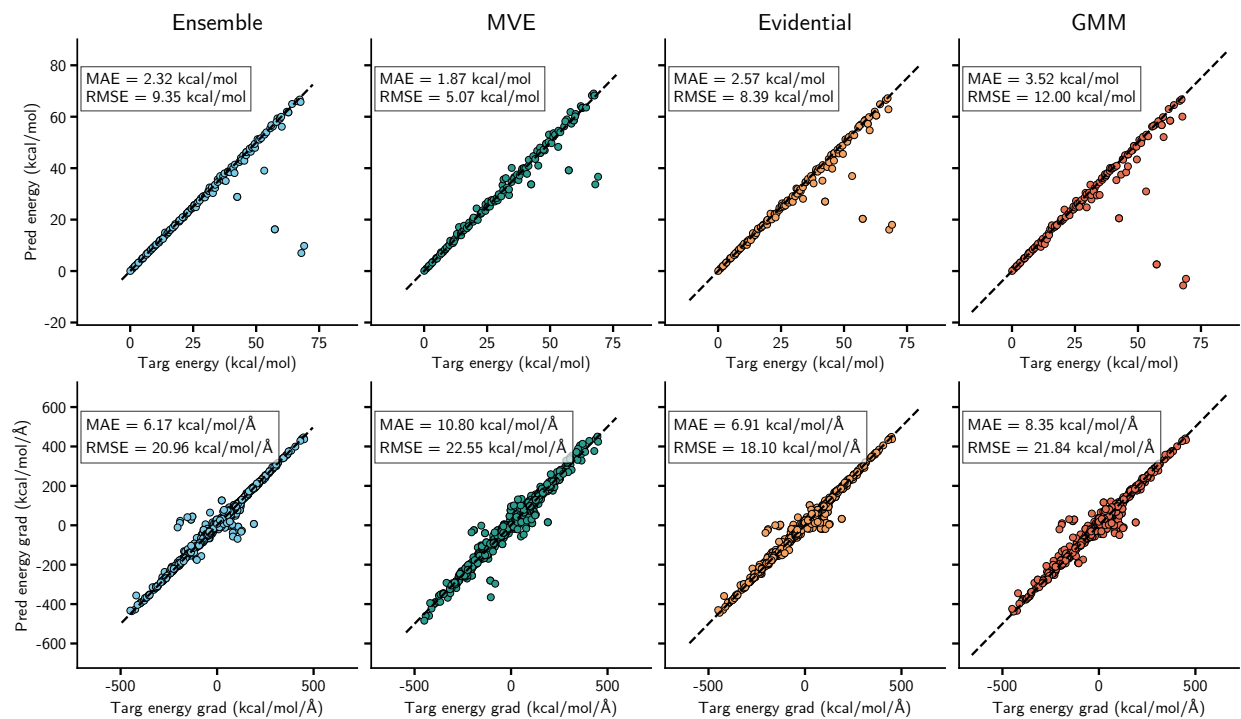
Supplementary Figure 20. 4 different types of ensemble methods, each containing 5 neural network models are shown. All ensembles are trained only on the initial training data of ammonia, and results shown above are tested on a testing set. The different colors in each subplot differentiates between in-domain (ID) and out-of-domain (OOD) data, partitioned by an energy cutoff of 4.85 kcal/mol, the maximum energy value available in the training set. The 4 ensembling methods are: **(a)** only random initialization, **(b)** hyperparameter diversification, **(c)** bagging, and **(d)** all three methods combined. Recurring subfigures (i) show the uncertainty vs squared errors of atomic forces in the test dataset, and predicted energy and forces are shown in subfigures (ii) and (iii). The calibrated negative log-likelihood uncertainty metric is not shown since the bagging method splits validation set differently. As can be seen from **(d)** and uncertainty metrics, UQ performance is the best for a combination of all three methods, but the gain over the other methods is not significant. In the paper, we have opted to use the random initialization scheme rather than one of the more elaborate schemes such that we could do a baseline comparison with the single NNs.



Supplementary Figure 21. Parity plots showing prediction of atomic forces for all individual neural networks from the 4 different types of ensembling methods: (a) random initialization, (b) hyperparameter diversification, (c) bagging, and (d) all three methods combined. In-domain (ID) points are colored in orange, whereas out-of-domain (OOD) points are colored in blue. ID and OOD points are partitioned by an energy cutoff of 4.85 kcal/mol, the maximum energy value available in the training set.

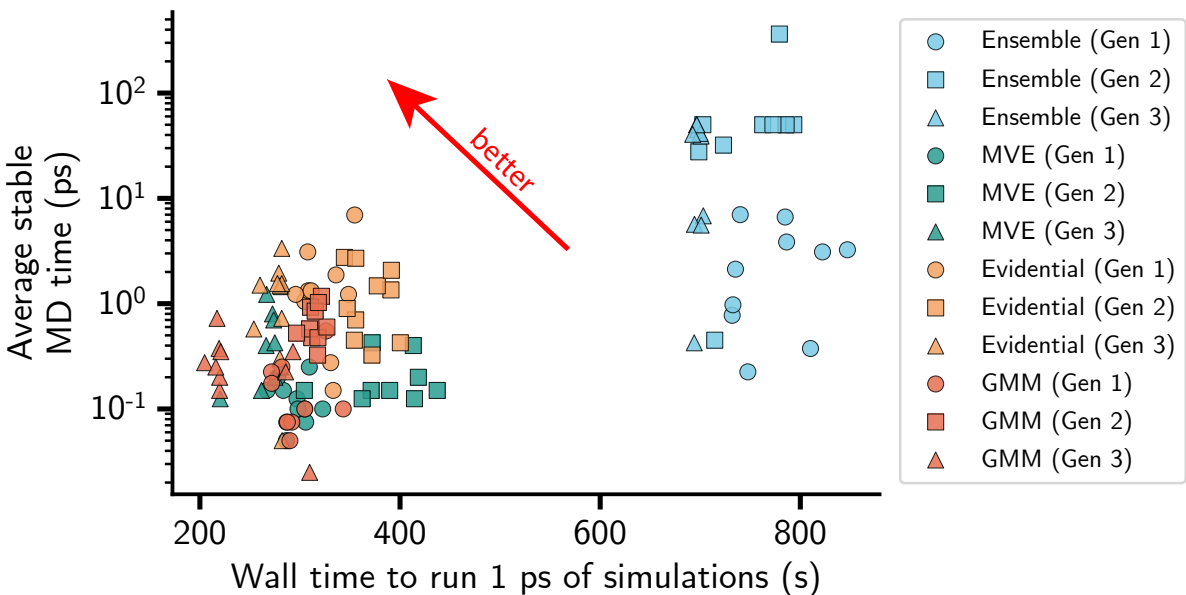


Supplementary Figure 22. Parity plot showing predicted energy and forces on the test set of ammonia (containing a large amount of OOD data) using the NN models trained in generation 1 (trained and validated only on 78 geometries).

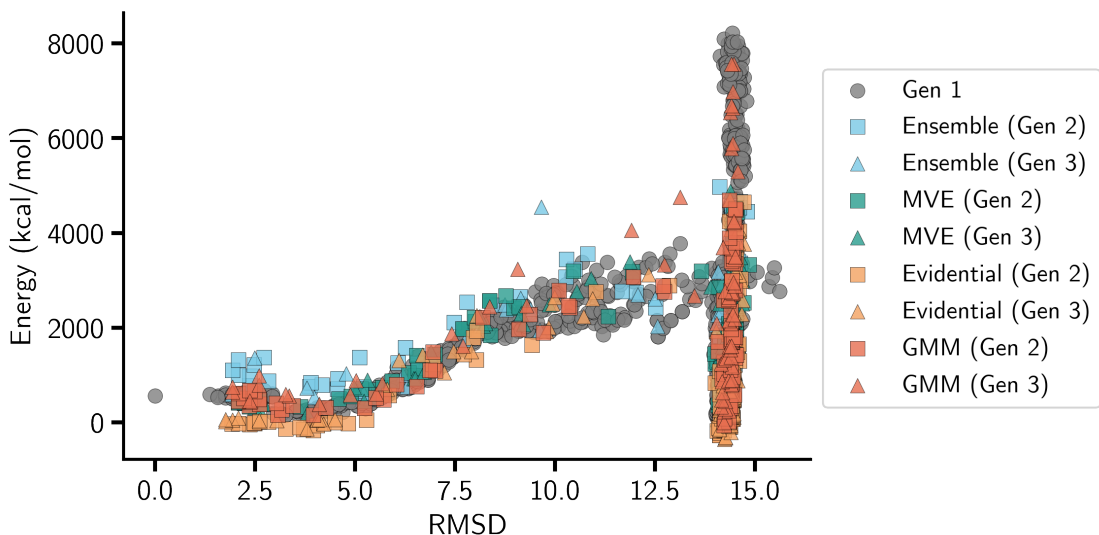


Supplementary Figure 23. Parity plot showing predicted energy and forces on the same test set of ammonia as in Supplementary Figure 22 using the models trained in generation 3 (trained on 78 initial geometries plus 40 adversarial samples).

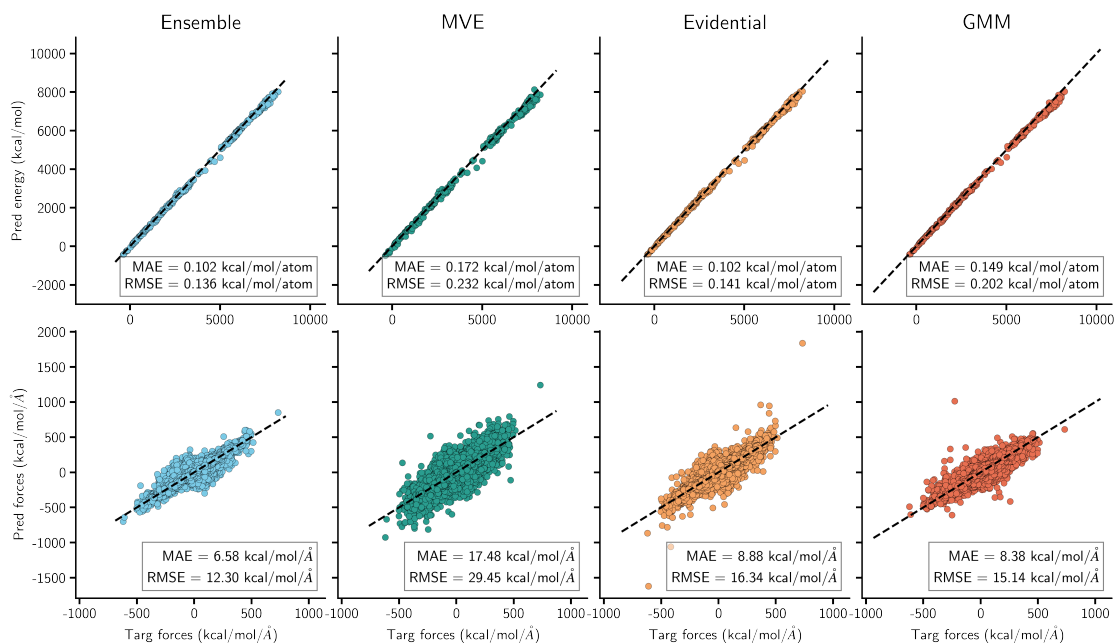
C. Silica dataset



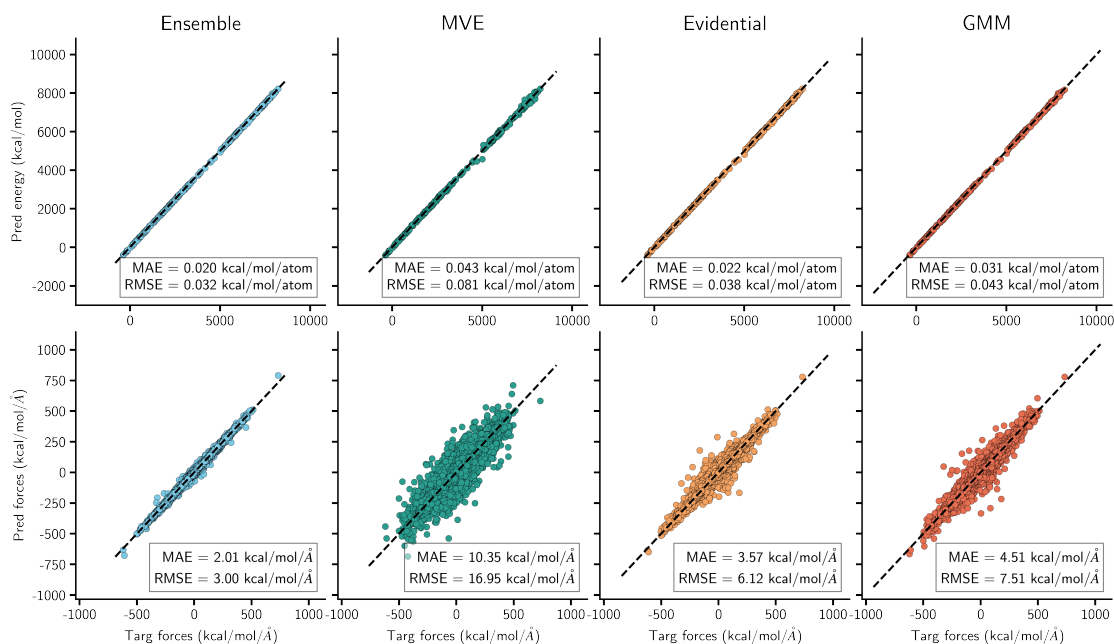
Supplementary Figure 24. Scatter plots showing the length of stable MD trajectories against the wall time taken to run 1 ps of simulations for NN models in each considered uncertainty estimation methods in different generations. Every marker indicates a NVT trajectory simulated at 2500K.



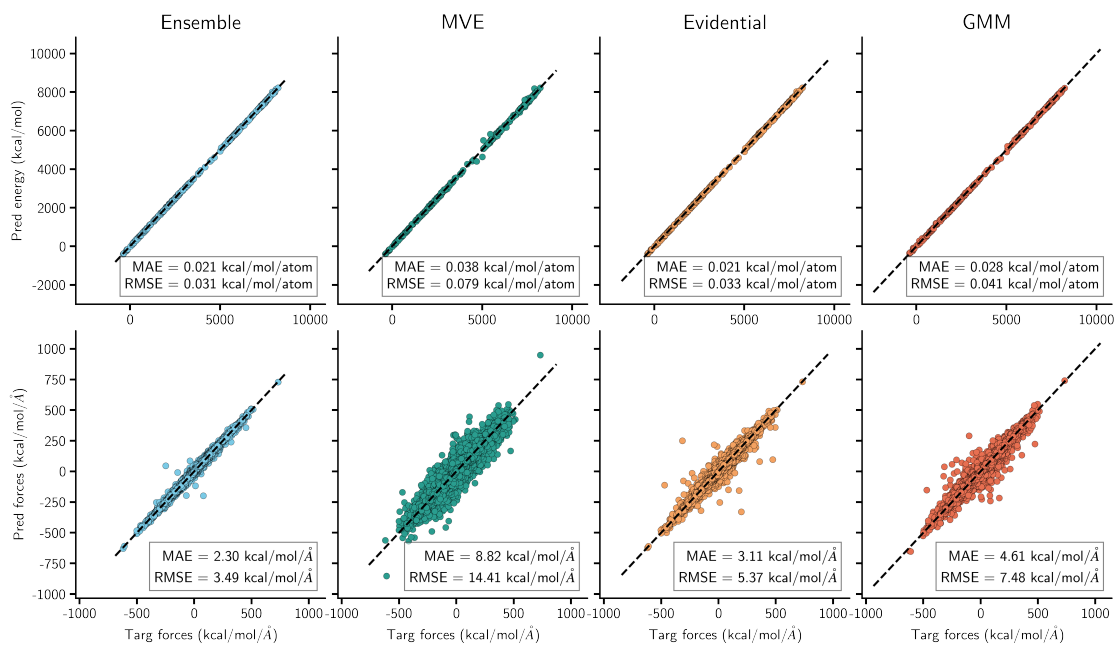
Supplementary Figure 25. Energy against root mean squared deviation (RMSD) of silica structures in training data and new adversarial samples generated by the respective uncertainty estimation methods. RMSD of all silica structures are computed against a silica structure sampled at a 300 K temperature equilibration simulation in the training data.



Supplementary Figure 26. Parity plot showing predicted energy and forces on the test set of silica using the models trained in generation 1 (trained on 1318 initial training data).



Supplementary Figure 27. Parity plot showing predicted energy and forces on the test set of silica using the models trained in generation 2 (trained on 1318 initial training data plus 100 adversarial samples).



Supplementary Figure 28. Parity plot showing predicted energy and forces on the test set of silica using the models trained in generation 3 (trained on 1318 initial training plus and 200 adversarial samples).

SUPPLEMENTARY REFERENCES

- ¹S. Urata, N. Nakamura, K. Aiba, T. Tada, and H. Hosono, “How fluorine minimizes density fluctuations of silica glass: Molecular dynamics study with machine-learning assisted force-matching potential,” *Materials & Design* **197**, 109210 (2021).
- ²D. P. Kovács, C. V. D. Oord, J. Kucera, A. E. Allen, D. J. Cole, C. Ortner, and G. Csányi, “Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE,” *J. Chem. Theory Comput.* **17**, 7696–7711 (2021).
- ³R. L. Mozzi and B. E. Warren, “The structure of vitreous silica,” *J. Appl. Crystallogr.* **2**, 164–172 (1969).
- ⁴G. I. Csonka, J. P. Perdew, A. Ruzsinszky, P. H. T. Philipsen, S. Lebègue, J. Paier, O. A. Vydrov, and J. G. Ángyán, “Assessing the performance of recent density functionals for bulk solids,” *Phys. Rev. B* **79**, 155107 (2009).
- ⁵H. Gerhard, “Electronic spectra and electronic structure of polyatomic molecules,” Van Nostrand, Princeton (1966).
- ⁶A. Zhu, S. Batzner, A. Musaelian, and B. Kozinsky, “Fast uncertainty estimates in deep learning interatomic potentials,” *J. Chem. Phys.* **158**, 164111 (2023), https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0136574/17060517/164111_1_5.0136574.pdf.