Electronic Supplementary Information:

Graph Neural Network Interatomic Potential Ensembles with Calibrated Aleatoric and Epistemic Uncertainty on Energy and Forces

Jonas Busk (jbusk@dtu.dk), Mikkel N. Schmidt, Ole Winther, Tejs Vegge and Peter Bjørn Jørgensen

A Additional results

Here we include additional calibration results from the experiments on the ANI-1x and Transition1x datasets presented in Section 3 of the main paper.

Confidence curves can be used to evaluate the ranking ability of the model and to estimate the drop in error as a percentage of the high uncertainty instances are removed[17,25], which is especially important in applications such as active learning where high uncertainty instances are iteratively added to the training set to improve the model. A confidence curve is generated by sorting predictions by uncertainty in decreasing order and computing the error as a function of removing a percentage of the most uncertain predictions. For a well calibrated model the confidence curve is expected to decrease monotonically. An oracle curve representing perfect ranking can be generated by sorting the prediction by error instead and the confidence curve can be summarised by computing the area between the confidence and oracle curves (AUCO). However, we do not expect the uncertainty predictions to produce a perfect ranking with respect to the errors since instances with high predicted uncertainty can still have small empirical errors.

Quantile-calibration plots compare the quantiles of the predictive distribution with the quantiles of the empirical distribution and is a way to evaluate distribution calibration averaged over the data[20]. If the predictive distribution matches the empirical distribution, the quantile-calibration curve corresponds to the identity function and forms a line along the diagonal of the plot. Assuming a symmetric distribution, the confidence interval can be evaluated instead of the quantile. The quantile-calibration curve can be summarised by the sum of squared errors (SSE) between the predicted and empirical quantiles.

To further check the distribution assumptions averaged over the data, a histogram the errors normalised by the predicted uncertainties (z-scores) along with the assumed standard distribution can also be plotted.

A.1 Additional ANI-1x results

Additional calibration plots for the ensemble model trained on ANI-1x with NLL loss on energy and forces are presented in Figure A.1. The confidence curves for energy and forces are both monotonically decreasing indicating good ranking ability. The confidence curves also show a significant drop in error on both energy and forces when removing the top $\sim 5\%$ highest uncertainty predictions. This confirms the observation from the reliability diagrams in Figure 2, that there are a few instances with very high error but they are correctly identified and assigned high uncertainty by the model.

The energy quantile-calibration plot shows that assuming a normal distribution percentiles of the predicted distributions corresponds well to the empirical distribution and the symmetry at the 0.5 percentile indicates that the model is unbiased overall. In the case of the forces, we assume the distribution of the component-wise errors is normal and unbiased. Furthermore, if the component-wise force errors are normally distributed, the squared L2 norm of the 3-dimensional force errors should follow a chi-square distribution with 3 degrees of freedom. Consequently, we plot the symmetric version of the quantile-calibration plot for the component-wise force errors using a normal distribution and a regular quantile plot for the squared L2 norm of the force errors using a chi-square distribution as they allow for easier comparison. In both cases, the uncertainty estimates look fairly well calibrated with regards to the assumed distributions. This is also apparent from the corresponding histograms of normalised errors plotted along with the reference distributions.

A.2 Additional Transition1x results

Additional calibration plots for the ensemble model trained on Transition1x with NLL loss on energy and forces are presented in Figure A.2. The energy confidence curve is generally decreasing, but like the corresponding reliability diagram shown in Figure 3 it is not perfectly consistent whereas the forces confidence curve look more consistent and monotonically decreasing. In both cases, removing the highest uncertainty instances results in a large drop en error.

The energy quantile-calibration plot shows that the error is fairly well distribution calibrated with a slight overestimation of the uncertainty on average. The same applies to the forces where the error is also fairly well distribution calibrated but with some overestimation of the uncertainty. This is consistent with the results of the LZV analysis described in Section 3.4.

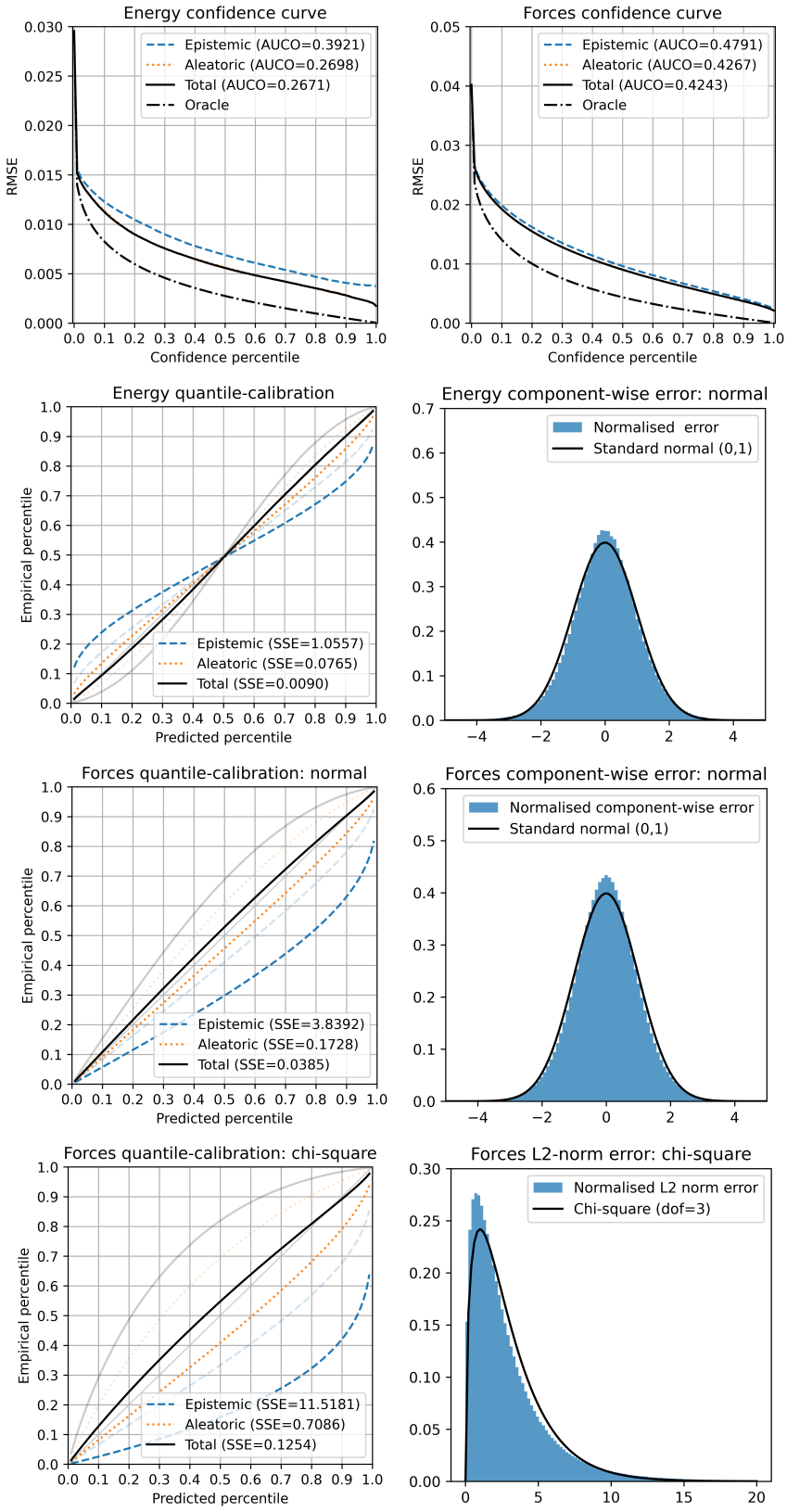

Fig. A.1 Additional calibration results on the ANI-1x dataset of energy and forces for an ensemble of $M = 5$ models trained with NLL loss on both energy and forces. To illustrate the effect of recalibration, the transparent curves show results before applying recalibration whereas the solid curves show results after recalibration.

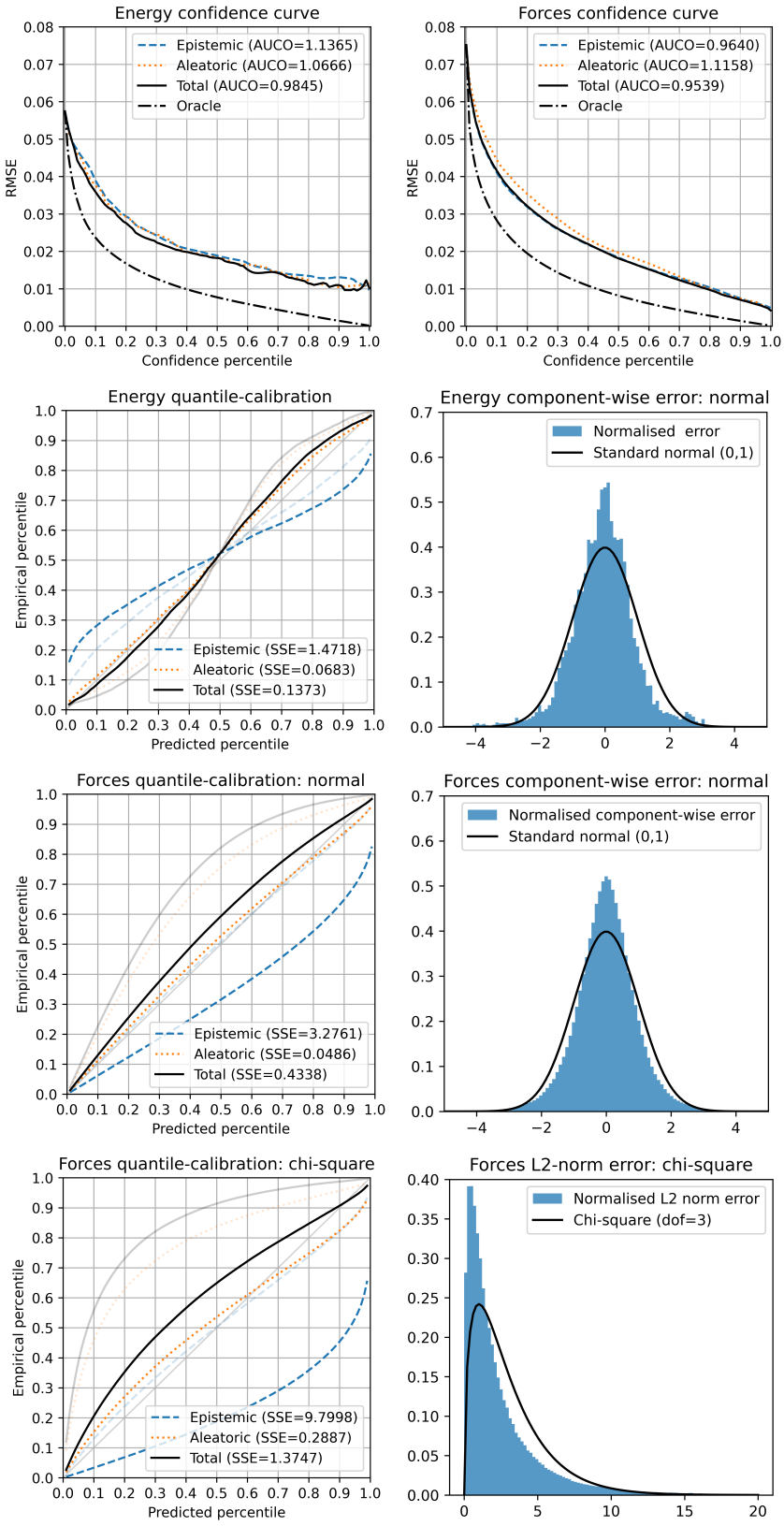

Fig. A.2 Additional calibration results on the Transition1x dataset of energy and forces for an ensemble of $M = 5$ models trained with NLL loss on both energy and forces. To illustrate the effect of recalibration, the transparent curves show results before applying recalibration whereas the solid curves show results after recalibration.