Contribution ID: **17** Type: **Talk**

Evaluating uncertainty estimations of Gaussian process regression-based machine learning interatomic potentials

Thursday 29 August 2024 16:30 (30 minutes)

Machine learning interatomic potentials (MLIPs) are machine learning (ML) models that map molecular configurations to corresponding energies and potentially forces, replacing highly accurate but expensive quantum chemical calculations. Quantum chemical calculations are carried out to calculate reference energies for only a few molecular configurations. When the ML model predicts an unseen sample, it introduces a new error, which highly depends on how different the sample is from the training data. Therefore, it is desirable to provide an uncertainty estimation for every model prediction. On the one hand, it is important to know the range of magnitude of a potential error of a prediction; on the other hand, it can be utilized in active learning (AL) schemes to identify samples for which the model is uncertain and add them as training samples. For MLIPs based on Gaussian process regression (GPR), the most commonly used uncertainty estimation is the variance of the predictive distribution provided by the GPR model. Besides that, ensemble-based uncertainties are possible. Although these GPR uncertainty measures have been applied to AL, it has rarely been studied how well or whether they correlate with the actual error, and it is not always clear whether AL strategies provide an actual improvement over simply adding samples randomly. In our study, we consider GPR models with Coulomb representations as well as Smooth Overlap of Atomic Positions (SOAP) representations as inputs. Our aim is to evaluate how the GPR variance and ensemble-based uncertainties correlate with the actual error. Furthermore, we want to find out how much additional information can be extracted from a pool of candidate samples, compared to selecting samples randomly, when we utilize the different uncertainty measures to iteratively add the most uncertain samples as training samples. This AL scheme is called uncertainty sampling. For different datasets and uncertainty estimations, we observed substantial differences in how error and uncertainties correlate. We find that, no matter if the correlation of an uncertainty estimation closely aligns with the theoretical expectations or not, the amount of additional information that can be extracted from the candidate pool via uncertainty sampling is highly limited and often close to zero.

Primary author: HOLZENKAMP, Matthias (Uni Wuppertal)

Co-author: Prof. ZASPEL, Peter (Uni Wuppertal)

Presenter: HOLZENKAMP, Matthias (Uni Wuppertal)