**Efficient Data Selection Methods for the Development of Machine Learned Potentials**

Jan Finkbeiner, Samuel Tovey & Christian Holm

Institute for Computational Physics - University of Stuttgart

---

**Background: Atomistic ML-Potentials**

Method can be summarized in 3 steps:

1. Run multiple MD simulations at different temperatures and pressures to optimally sample configuration space.
2. Select configurations by sampling uniformly in:
   - global energy $E = \sum E_i$
   - forces $f_i$
   - atomic energies $E_i$.
3. Future work: Perform Ab Initio single point calculations on selected configurations.

---

**Selection Methods**

Method can be summarized in 3 steps:

1. Run multiple MD simulations at different temperatures and pressures to optimally sample configuration space.
2. Select configurations by sampling uniformly in:
   - global energy $E = \sum E_i$
   - forces $f_i$
   - atomic energies $E_i$.
3. Future work: Perform Ab Initio single point calculations on selected configurations.

---

**References**


---

**Test-Dataset Performance**

The figure shows a comparison of errors in force predictions for NN-models (of different sizes with different hyperparameter choices) trained on data-sets that were selected by the different selection methods. From left to right we show plots for the maximum error (Max), root-mean-square error (RMSE), and the mean absolute error (MAE) in the ML models.

---

**Simulation Results**

The figure shows the comparison of the Na-Na radial distribution function computed from the reference Born-Meyer-Huggins-Tosi-Fumi (BMHTF) simulation and NN-driven (left) and GPR-driven (right) MDs trained on 512 (NN) or 128 (GPR) configurations selected via the global energy and atomic energies selection method.

We see that, in the case of the global data-selection methods, the radial distribution functions contain non-physical short range peaks implying the training data did not sufficiently represent the potential energy surface. In the case of the atomic energies however, no such short range peaks arise and the function fits that of the reference BMHTF potential.