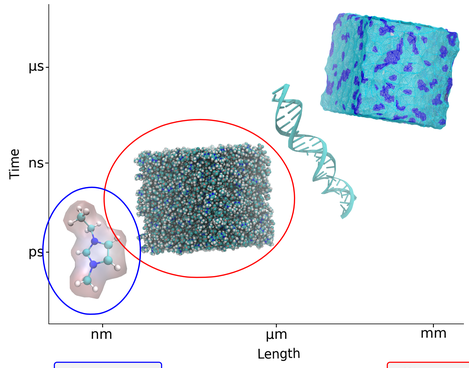
Efficient data selection methods for the development of machine learned potentials

Jan Finkbeiner, Samuel Tovey & Christian Holm

Institute for Computational Physics - Universiy of Stuttgart

University of Stuttgart
Germany

Background: Atomistic ML-Potentials



Ab Initio

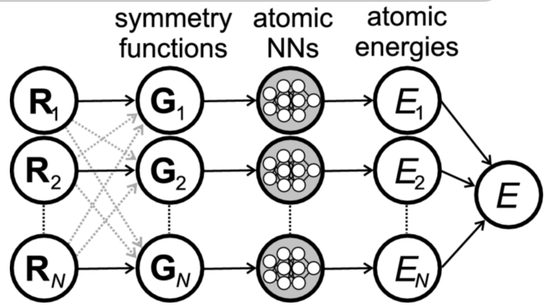
$$M_n \ddot{R} = -\nabla \langle \Psi | H_e | \Psi \rangle$$
$$H_e | \Psi \rangle = E_e | \Psi \rangle$$

Classical

$$M\ddot{R} = -\nabla V(R)$$

$\Longrightarrow$ Find $V(R)$ as function only of the atomic coordinates, atomic types and cell size, without taking electronic structure explicitly into account in data-driven way:

$$\langle \Psi | H_e | \Psi \rangle \stackrel{!}{=} V(R)$$



[1]

Descriptors/Symmetry functions:

- SOAP [2, 3]

$$\rho^{(i)}_{\mathrm{SOAP}}(\mathbf{R}) = \sum_{j=1}^{N^{(i)}_{\mathrm{env}}} \exp\left(-\alpha \left|\mathbf{R} - \mathbf{R}_j\right|^2\right)$$
$$\approx \sum_{j=1}^{N^{(i)}_{\mathrm{env}}} \sum_{n=0}^{n_{\max}} \sum_{l=0}^{l_{\max}} \sum_{m=-l}^{l} c^{(i)}_{nlm} g_n(R) Y_{lm}(\hat{\mathbf{R}})$$
$$p^{(i)}_{nn'l} \equiv \frac{1}{\sqrt{2l+1}} \sum_m c^{(i)}_{nlm} \left(c^{(i)}_{n'lm}\right)^*$$

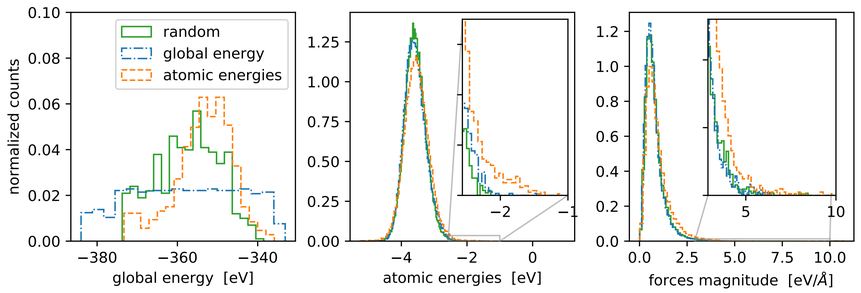
- Graph NNs, e.g. SchNet [4]

Fitting methods:

- GPR
- NNs

Selection Methods

Distribution comparisons for random selection and uniform selection in global and atomic energies



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 $E = \sum_i^N E_i$
   - forces $\|\mathbf{f}_i\|$
   - atomic energies E_i
3. Future work: Perform Ab Initio single point calculations on selected configurations

$\Longrightarrow$ Figure shows that atomic energies sampling enhances sampling of low probability configurations/configurations of the long tail

References

[1] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". In: *The Journal of Chemical Physics* 134.7 (2011), p. 074106.

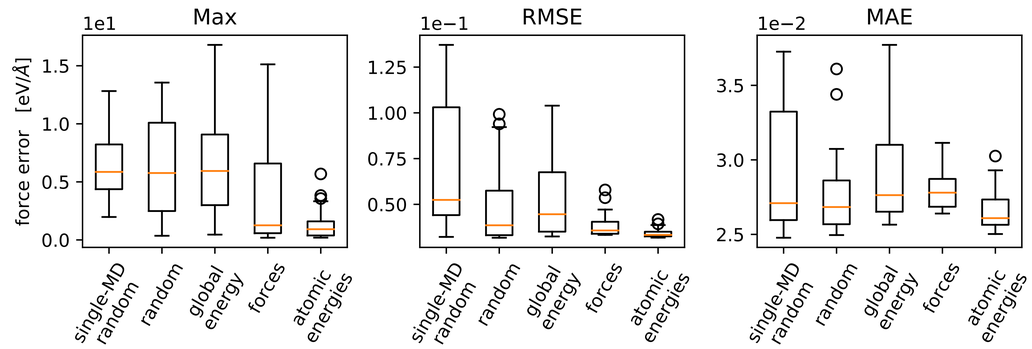
[2] Albert P. Bartók et al. "Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons". In: *Physical Review Letters* 104 (13 2010), p. 136403.

[3] Albert P. Bartók, Risi Kondor, and Gábor Csányi. "On representing chemical environments". In: *Phys. Rev. B* 87 (18 2013), p. 184115.

[4] K. T. Schütt et al. "SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, 992–1002. ISBN: 9781510860964.

Test-Dataset Performance

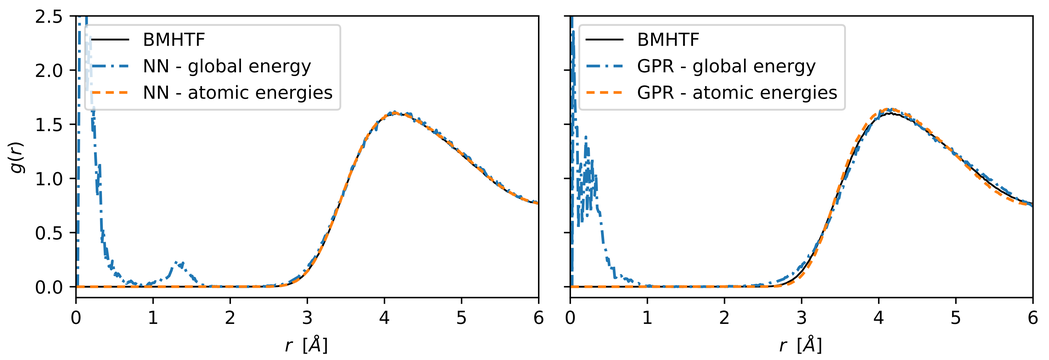
The figure shows a comparison of errors in force predictions for NN-models (of different sizes with different hyperparameter choices) trained on data-sets that were selected by the different selection methods. From left to right we show plots for the maximum error (Max), root-mean-square error (RMSE),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



One can see a clear trend in the prediction errors: The models trained on data-sets that were generated by the force or atomic energy selection methods achieve lower Max and RMSE errors than the other methods (random or based on global energy).

Simulation Results

To further analyse the influence of the selection methods, molecular dynamics (MD) simulations of the sodium chloride (NaCl) system with 1000 atoms for up to 1000 ps in the NVT ensemble at 1400K driven were performed. This should expose the models to a wide variety of new configurations.



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.