EFFICIENT DATA SELECTION METHODS FOR THE DEVELOPMENT OF MACHINE LEARNED POTENTIALS

Jan Finkbeiner, Samuel Tovey & Christian Holm
Institute for Computational Physics
University of Stuttgart
Allmandring 3, 70569, Stuttgart, DE
{jfinkbeiner, stovey, holm}@icp.uni-stuttgart.de

ABSTRACT

We present an investigation into data selection methods for the efficient sampling of configuration space as applied to the development of inter-atomic potentials for scale bridging in molecular dynamics (MD) simulations. This investigation suggests that the most efficient sampling techniques are those that incorporate information on an atomic level such as forces or atomic energies. Finally, we generate an inter-atomic potential for the a sodium chloride system using each data selection technique and find that the global selection methods result in non-physical simulations.

1 INTRODUCTION

In the development of any supervised regression model it is important to consider how to generate training data to optimally represent your target function. In the case where generation of this data is expensive, it is further necessary to considering how to minimize the number of points required for effective training. One application where this is especially prevalent is the generation of inter-atomic potentials for use in molecular dynamics (MD) simulations. In these applications, expensive ab-initio data is generated on small atomic system over short time scales before a machine learning algorithm such as Gaussian process regression (GPR) (Bartók et al. (2010)) or a neural network (NN) (Behler & Parrinello (2007); Schütt et al. (2018)) is used to fit a function that maps atomic environments to a system energy and atomic forces. Once constructed, these potentials can be used for classical MD simulations and employed on larger atomistic systems for longer time scales. In these simulations, the machine learned models can achieve near ab-initio accuracy with a computational complexity of $\mathcal{O}(N)$ where N is the number of atoms in the system (Tovey et al. (2020); Sivaraman et al. (2020)). The fitting procedure involves the deconstruction of the global energy into contributions from atomic environments. In order to do this, the atomic environments are transformed into so-called descriptors that are then passed into a machine learning algorithm. These descriptors encode symmetries present in the potential such as rotation, translation and particle exchange. In order to develop an effective model, it is necessary that the training data contain as many unique atomic environments as possible such that one maximally samples configuration space. Whilst this may be stated simply, it is not clear in practice what constitutes a unique atomic environment, or how best to assess an environment for uniqueness. There are a number of factors in the development of these inter-atomic potentials which are currently being investigated including the choice of descriptor (Bartók et al. (2013); Behler & Parrinello (2007); Behler (2011); Gastegger et al. (2018); Lindsey et al. (2017); Rupp et al. (2012); Samanta (2018); Seko et al. (2014); Shapeev (2016); Takahashi et al. (2018); Thompson et al. (2015); Zhu et al. (2016)), or the machine learning algorithm used in the fitting procedure (Rupp et al. (2012); Schütt et al. (2017); Balabin & Lomakina (2011); Bartók et al. (2010); Behler & Parrinello (2007)). In the case of data selection for these models, it is often the case that training data is sampled uniformly in time from long, expensive, ab-initio MD simulations (Cole et al. (2020); Shao et al. (2020)), with some notable steps being taken in the direction of active learning (Sivaraman et al. (2020)), descriptor space metrics (De et al. (2016)), and in the direct manipulation of atomic structure to induce rare events as in the so-called RAG sampling procedure (Choi & Jhi (2020)).

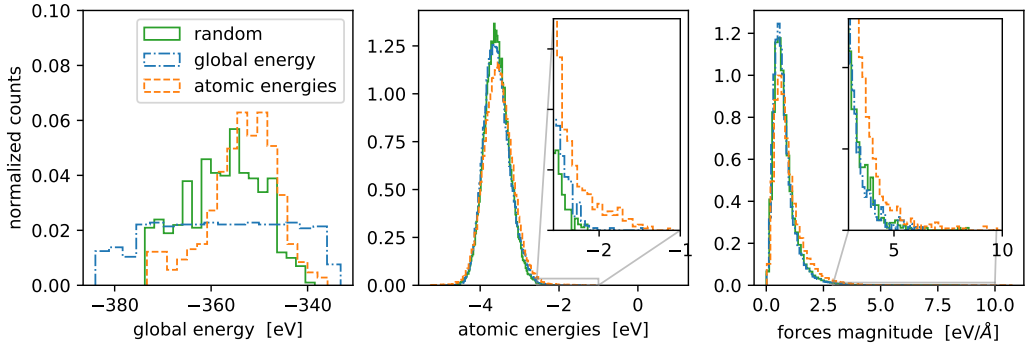

Figure 1: Resulting distributions of global energy (left), atomic energies (middle) and force magnitude (right) in data-sets for different selection methods presented in Section 2. Each data-set consists of 512 configurations selected via the random-selection method (solid, green), sampling uniformly in global energy (dash-dotted, blue), or sampling uniformly in atomic energies (dashed, orange). Insets show zoom-ins in the long-tail of the distributions.

In this work, several approaches to selecting data for use in the development of an inter-atomic potential are presented and tested in order to understand what factors might impact model performance. The methods studied incorporate both global properties of a configuration, as well as the local, atomic information, in order to understand which approach leads to a more accurate model. Data selected by each method is used to train an inter-atomic potential on simulated molten sodium chloride which is then assessed based on the root mean square error (RMSE), maximum error, and mean absolute error (MAE) of the model predictions.

2 Data Selection Methods

In training an inter-atomic potential on global energy data we make an underlying assumption that this energy may be decomposed into atomic contributions as

$$E = \sum_{i}^{N} \epsilon_i, \tag{1}$$

where ϵ_i is the contribution of the i^{th} atomic environment and N is the number of atoms in the system. As was briefly mentioned in the introduction, the key to data selection in the development of machine learned inter-atomic potentials is sufficient sampling of configuration space such that all unique configurations resulting in some ϵ_i above are realised in the training data. In order to ensure that large parts of configuration space are sampled, we perform several MD simulations over different sets of constant parameters, in our case, fixed atom number, pressure, and temperature. On this pool of samples we apply one of the selection methods below to identify configurations that contain relevant atomic environments.

2.1 Random Selection

In the random sampling approach, configurations over the full MD trajectory are sampled at random and used as training, test, and validation data. As a benchmark for the performance of the sampling methods over different simulations, we also use this method on data from a single MD simulation which we refer to as single-MD random sampling.

2.2 Global Energy Selection

In ab-initio simulations, the Schrödinger equation is solved numerically using density functional theory (DFT) (Burke (2012)) in order to determine the total energy of the system of atoms. It is on this energy, along with the forces on each of the atoms, that the machine learning algorithm

employed will train. Therefore, it is rational to select training data by sampling uniformly across the energy values as illustrated in Figure 1. Whilst this approach ensures the existence of unique global energies in the training data, it is not necessarily true that these configurations contain within them unique atomic environments.

2.3 Atomic Energy Selection

Whilst in an ab-initio simulation the concept of a atomic energy in Equation 1 is controversial, in a classical simulation it can be written simply as the summation of terms in the inter-atomic potential calculation up to a defined cutoff as

$$\epsilon_i = \sum_{j}^{N_{\text{pairs}}} U(r_{ij}, r_c), \tag{2}$$

where U is a function determining the potential energy between two atoms i and j, r_{ij} is the pairwise distance between the atoms i and j, r_c is the short range cutoff of the potential, and the summation runs over all atoms forming an interacting pair with atom i. Due to the similarity between Equation 2 and the fundamental assumption made in Equation 1, this is a candidate for data selection. In our study, we generate data-sets by selecting configurations uniformly based on atomic energies.

2.4 Force Selection

Along a similar line of thinking to the previous method, force sampling also looks to an atomic property to choose interesting configurations. In this case, there is the added benefit of forces being quantum mechanical observables, and therefore available in an ab-initio simulation through the Hellmann-Feynman theorem (Güttinger (1932); Feynman (1939)). In this method, rather than the atomic energy being studied, the net force on the i^{th} atom is used to indicate a unique environment.

3 Results

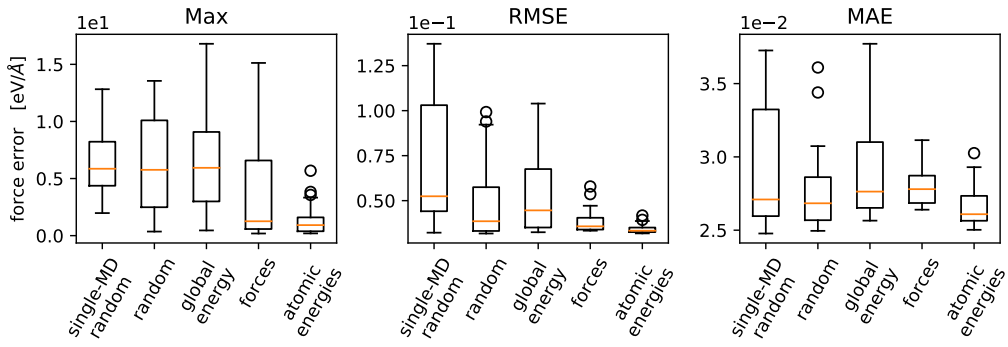

Figure 2: Comparison of errors in force predictions for NN-models trained on data-sets that were selected by the different methods presented in Section 2. From left to right we show plots for the maximum error (Max), root-mean-square error (RMSE), and the mean absolute error (MAE) in the ML models.

In order to test the effectiveness of each selection method data was generated from 100 atom classical MD simulations of molten sodium chloride using a Born-Meyer-Huggins-Tosi-Fumi (BMHTF) potential (Tosi & Fumi (1964); Fumi & Tosi (1964); Mayer (1933); Born & Mayer (1932); Huggins & Mayer (1933)) for temperatures and pressures ranging from 900 K to 2200 K and $1 \cdot 10^{-3}$ bar to $5 \cdot 10^{4}$ bar respectively.

We initially evaluated the performance of the selection methods by training models on data chosen by each of them and then calculating the maximum error (Max), root mean square error (RMSE), and the mean absolute error (MAE) of the model force predictions with respect to test sets. For each

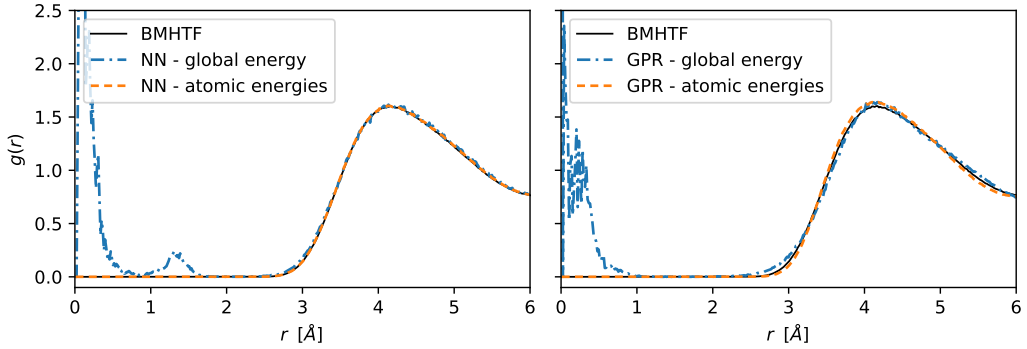

Figure 3: Comparison of the Na-Na radial distribution function computed from a reference BMHTF simulation and NN-driven (left) and GPR-driven (right) MDs with 500 ion-pairs trained on 512 (NN) or 128 (GPR) configurations selected via the global energy and atomic energies selection method.

selection method, 6 models of differing size and input parameters (see Appendix) were trained. Each model was tested on 20'480 configurations consisting of data-sets spanning different temperatures chosen in equal parts by the selection methods described in Section 2.

Figure 2 shows 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 root mean square errors. One explanation for the improved performance of the atomic selection approaches might be in their distribution of data. While the global energy selection method appears to produce more diverse global energies, the atomic energy selection method contains these so-called 'fat tails' in the atomic energy distribution. The appearance of these fat tails implies the inclusion of less probable configurations. In the case of global energy selection, the contribution of these atomic energies may be diluted in the summation over all environments, whereas with the atomic approach, they are identified and added to the training data.

As a further validation of the selection techniques, the trained models were used in 1000 atom MD simulations at 1400 K in an NVT ensemble run for up to 1000 ps. This investigation exposes the models to a number of varying configurations from large regions of configuration space thereby assessing robustness. In order to eliminate possible algorithm dependence, neural network and Gaussian process regression models were used (see Appendix). The radial distribution functions (RDF) of these simulations were then compared with BMHTF model under the same conditions. 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. The success of the atomic property selection methods highlights the robustness that accompanies the low RMSE and Max errors seen in Figure 2.

4 Conclusion

We have shown that data selection based on atomic energies or forces yields more accurate inter-atomic potentials than those trained on global energies or randomly chosen configurations. We found that the calculated RMSE values of models trained using the atomic energies data selection method was consistently lower than any other approach. Furthermore, potentials trained on global properties produced non-physical results when used in an MD simulation, whereas those trained on atomic properties reproduced the reference data. These results suggest that selecting data based on atomic properties results in more accurate and robust machine learned potentials as opposed to global property selection methods.

ACKNOWLEDGMENTS

The authors acknowledge financial support from the German Funding Agency (Deutsche Forschungsgemeinschaft DFG) under Germany's Excellence Strategy EXC 2075-390740016, and S. Tovey was supported by a LGF stipend of the state of Baden-Württemberg

REFERENCES

Roman M. Balabin and Ekaterina I. Lomakina. Support vector machine regression (ls-svm) an alternative to artificial neural networks (anns) for the analysis of quantum chemistry data. *Physical Chemistry Chemical Physics*, 13:11710–11718, 2011. doi: 10.1039/C1CP00051A.

Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters*, 104: 136403, 2010. doi: 10.1103/PhysRevLett.104.136403.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Phys. Rev. B*, 87:184115, May 2013. doi: 10.1103/PhysRevB.87.184115. URL https://link.aps.org/doi/10.1103/PhysRevB.87.184115.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011. doi: 10.1063/1.3553717. URL https://doi.org/10.1063/1.3553717.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.*, 98:146401, Apr 2007. doi: 10.1103/PhysRevLett.98.146401. URL https://link.aps.org/doi/10.1103/PhysRevLett.98.146401.

Max Born and Joseph E. Mayer. Zur gittertheorie der ionenkristalle. *Zeitschrift für Physik*, 75(1): 1–18, Jan 1932. ISSN 0044-3328. doi: 10.1007/BF01340511. URL https://doi.org/10.1007/BF01340511.

Kieron Burke. Perspective on density functional theory. *The Journal of Chemical Physics*, 136(15): 150901, 2012. doi: 10.1063/1.4704546. URL https://doi.org/10.1063/1.4704546.

Young-Jae Choi and Seung-Hoon Jhi. Efficient training of machine learning potentials by a randomized atomic-system generator. *The Journal of Physical Chemistry B*, 124(39):8704–8710, 2020. doi: 10.1021/acs.jpcb.0c05075. URL https://doi.org/10.1021/acs.jpcb.0c05075. PMID: 32910653.

Daniel J. Cole, Letif Mones, and Gábor Csányi. A machine learning based intramolecular potential for a flexible organic molecule. *Faraday Discuss.*, pp. –, 2020. doi: 10.1039/D0FD00028K. URL http://dx.doi.org/10.1039/D0FD00028K.

Sandip De, Albert P. Bartók, Gábor Csányi, and Michele Ceriotti. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.*, 18:13754–13769, 2016. doi: 10.1039/C6CP00415F. URL http://dx.doi.org/10.1039/C6CP00415F.

R. P. Feynman. Forces in molecules. *Phys. Rev.*, 56:340–343, Aug 1939. doi: 10.1103/PhysRev.56.340. URL https://link.aps.org/doi/10.1103/PhysRev.56.340.

F.G. Fumi and M.P. Tosi. Ionic sizes and born repulsive parameters in the nacl-type alkali halides—i: The huggins-mayer and pauling forms. *Journal of Physics and Chemistry of Solids*, 25(1):31–43, 1964. ISSN 0022-3697. doi: https://doi.org/10.1016/0022-3697(64)90159-3. URL https://www.sciencedirect.com/science/article/pii/0022369764901593.

M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand. wacsf—weighted atom-centered symmetry functions as descriptors in machine learning potentials. *The Journal of Chemical Physics*, 148(24):241709, 2018. doi: 10.1063/1.5019667. URL https://doi.org/10.1063/1.5019667.

P. Güttinger. Das verhalten von atomen im magnetischen drehfeld. *Zeitschrift für Physik*, 73(3): 169–184, Mar 1932. ISSN 1434-601X. doi: 10.1007/BF01351211. URL https://doi.org/10.1007/BF01351211.

Maurice L. Huggins and Joseph E. Mayer. Interatomic distances in crystals of the alkali halides. *The Journal of Chemical Physics*, 1(9):643–646, 1933. doi: 10.1063/1.1749344. URL https://doi.org/10.1063/1.1749344.

Rebecca K. Lindsey, Laurence E. Fried, and Nir Goldman. Chimes: A force matched potential with explicit three-body interactions for molten carbon. *Journal of Chemical Theory and Computation*, 13(12):6222–6229, 2017. doi: 10.1021/acs.jctc.7b00867. URL https://doi.org/10.1021/acs.jctc.7b00867.

Joseph E. Mayer. Dispersion and polarizability and the van der waals potential in the alkali halides. *The Journal of Chemical Physics*, 1(4):270–279, 1933. doi: 10.1063/1.1749283. URL https://doi.org/10.1063/1.1749283.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012. doi: 10.1103/PhysRevLett.108.058301. URL https://link.aps.org/doi/10.1103/PhysRevLett.108.058301.

Amit Samanta. Representing local atomic environment using descriptors based on local correlations. *The Journal of Chemical Physics*, 149(24):244102, 2018. doi: 10.1063/1.5055772. URL https://doi.org/10.1063/1.5055772.

K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. 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, pp. 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Múller. Schnet a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24): 241722, 2018. doi: 10.1063/1.5019779.

Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8(1): 13890, 2017. ISSN 2041-1723. doi: 10.1038/ncomms13890. URL https://doi.org/10.1038/ncomms13890.

Atsuto Seko, Akira Takahashi, and Isao Tanaka. Sparse representation for a potential energy surface. *Phys. Rev. B*, 90:024101, Jul 2014. doi: 10.1103/PhysRevB.90.024101. URL https://link.aps.org/doi/10.1103/PhysRevB.90.024101.

Yunqi Shao, Matti Hellström, Are Yllö, Jonas Mindemark, Kersti Hermansson, Jörg Behler, and Chao Zhang. Temperature effects on the ionic conductivity in concentrated alkaline electrolyte solutions. *Phys. Chem. Chem. Phys.*, 22:10426–10430, 2020. doi: 10.1039/C9CP06479F. URL http://dx.doi.org/10.1039/C9CP06479F.

Alexander V. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016. doi: 10.1137/15M1054183. URL https://doi.org/10.1137/15M1054183.

Ganesh Sivaraman, Anand Narayanan Krishnamoorthy, Matthias Baur, Christian Holm, Marius Stan, Gábor Csányi, Chris Benmore, and Álvaro Vázquez-Mayagoitia. Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Computational Materials*, 6(1):104, Jul 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00367-7. URL https://doi.org/10.1038/s41524-020-00367-7.

Akira Takahashi, Atsuto Seko, and Isao Tanaka. Linearized machine-learning interatomic potentials for non-magnetic elemental metals: Limitation of pairwise descriptors and trend of predictive power. *The Journal of Chemical Physics*, 148(23):234106, 2018. doi: 10.1063/1.5027283. URL https://doi.org/10.1063/1.5027283.

A.P. Thompson, L.P. Swiler, C.R. Trott, S.M. Foiles, and G.J. Tucker. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *Journal of Computational Physics*, 285:316 – 330, 2015. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2014.12.018. URL http://www.sciencedirect.com/science/article/pii/S0021999114008353.

M.P. Tosi and F.G. Fumi. Ionic sizes and born repulsive parameters in the nacl-type alkali halides—ii: The generalized huggins-mayer form. *Journal of Physics and Chemistry of Solids*, 25(1):45–52, 1964. ISSN 0022-3697. doi: https://doi.org/10.1016/0022-3697(64)90160-X. URL https://www.sciencedirect.com/science/article/pii/002236976490160X.

Samuel Tovey, Anand Narayanan Krishnamoorthy, Ganesh Sivaraman, Jicheng Guo, Chris Benmore, Andreas Heuer, and Christian Holm. Dft accurate interatomic potential for molten nacl from machine learning. 5 2020. doi: 10.26434/chemrxiv.12355160.v1. URL https://chemrxiv.org/articles/preprint/DFT_Accurate_Interatomic_Potential_for_Molten_NaCl_from_Machine_Learning/12355160.

Li Zhu, Maximilian Amsler, Tobias Fuhrer, Bastian Schaefer, Somayeh Faraji, Samare Rostami, S. Alireza Ghasemi, Ali Sadeghi, Migle Grauzinyte, Chris Wolverton, and Stefan Goedecker. A fingerprint based metric for measuring similarities of crystalline structures. *The Journal of Chemical Physics*, 144(3):034203, 2016. doi: 10.1063/1.4940026. URL https://doi.org/10.1063/1.4940026.

A APPENDIX

A.1 SCHNET-HYPERPARAMETERS

The Neural Network models are based on the SchNet-architecture (Schütt et al. (2017), Schütt et al. (2017)). All hyperparameters used for the different model sizes are summarized in Table 1 and 2. If not specified, the default value was chosen.

A.2 GAP-HYPERPARAMETERS

To generate the GPR models the GAP suite (Bartók et al. (2010), Bartók et al. (2013)) of the QUIP software package was used. GAP suite is available for non-commercial use from www.libatoms.org. The model uses a the SOAP descriptor. Table A.2 shows the hyperparameters for the SOAP descriptor.

Table 1: Hyperparameters of the SchNet-Model types

NAME	Small	Medium	Large
n_interactions	3	4	4
n_atom_basis	32	48	64
n_filters	32	64	128
n_gaussians	3	48	64
n_in	32	48	64
elements	(11, 17)	(11, 17)	(11, 17)
n_neurons	[48, 48]	[128, 64, 32]	[128, 128, 64, 32]
n_layers	3	4	5

Table 2: Training hyperparameters and other parameters of all SchNet-Models

NAME	VALUE
cutoff radius	6.0 $\mathring{A}$
optimizer	Adam
learning rate	5e-4

Table 3: Hyperparameters of the SOAP descriptor for the GAP model

NAME	VALUE
n_max	8
l_max	6
atom_sigma	0.825
zeta	4
cutoff	6.5
cutoff_transition_width	0.5
delta	1.0