Improving the accuracy and efficiency of Machine Learning derived interatomic potentials

Nisarg Kaushikkumar Joshi

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

In

Chemical Engineering

University of Washington

2021

Committee:

Jim Pfaendtner

David Beck

Program Authorized to Offer Degree:

Chemical Engineering

University of Washington

Abstract

Improving the accuracy and efficiency of
machine learning derived interatomic potentials

Nisarg Kaushikkumar Joshi

Chair of the Supervisory Committee:
Jim Pfaendtner
Department of Chemical Engineering

Understanding molecules and molecular interactions has been an active field of research in many fields like materials discovery, catalysis design, biomedicine, and drug design. Molecular dynamics (MD) simulation is one of the tools which facilitates the study of molecules at the atomistic levels and predict their properties. However, there is always a trade-off between the accuracy and computational cost among different MD simulations. Moreover, MD simulations are also limited to the size and complexity of the system. To overcome these limitations, there have been many developments in the research communities where machine learning (ML) methods are used to develop interatomic potentials for molecules. These ML methods also face obstacles to accurately represent molecular systems. In this thesis, we propose a solution to cross the hurdles faced by ML

models. We present, the use of enhanced sampling methods to generate training data for developing interatomic potentials. We illustrate the generalizability of ML derived interatomic potentials and show how enhanced sampling methods improve the accuracy of the potentials which can be used for different thermodynamic ranges.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to thank everyone who has contributed and supported me throughout this thesis journey. To begin with, I would like to thank my academic advisor Jim Pfaendtner, for providing mentorship and welcoming me in his research group. All his support and excellent advice has provided me invaluable experience in this journey. I would like to thank David Beck, for introducing me to the world of Data Science. It was through his instruction, I got involved in my research field. I truly appreciate your constant help and support throughout my thesis program.

Thanks to Chowdhury Ashraf for being a great mentor for this thesis. Your mentorship helped me learn the basics of the field and I really appreciate your guidance and efforts in helping me troubleshoot so many errors. Your encouragement in solving the problems and working on solutions was a great learning experience for me. Thanks to Orion Dollar for always being available to discuss Data Science topics with and working on all the great projects. I look forward working with you for the upcoming projects and learning so many new topics. Thanks to Janani Sampath, Kaylyn Torkelson, Sarah Alamdari, Xin Qi, Jessica Kong, Sabiha Rustam, Luke Gibson, Nadia Intan, and Stephanie Hare through whom I have learned and grown my knowledge about different aspects of molecular simulations in the group.

I would like to acknowledge the DIRECT program at UW for providing an invaluable experience in teaching me Data Science. I would like to thank my friends who have always

been with me through my thick and thins and always having my back whenever I needed them. Thanks to all the instructors at UW, from whom I have learned different skills and thanks to the Department of Chemical Engineering for supporting me throughout the thesis.

Finally, I would like to thank my parents without whom I wouldn't have made this far. Their constant support and love were the driving force for me to accomplish my goals without any fear.

Chapter 1. Introduction

Molecular dynamics (MD) simulations are widely used for understanding molecules and determining their properties at the atomistic level. A good representation of the potential energy surface (PES) of the system, helps in obtaining thermodynamic and dynamical properties[1]. MD simulations help in understanding the conformational changes of molecules from the atomic configurations and obtaining PES for the systems. However, there is an accuracy vs computational cost trade-off for MD simulations. *Ab initio* molecular dynamics (AIMD) provides a good accuracy of the system, closer to density functional theory (DFT) accuracy, but is limited to the size of the system and is computationally expensive. While on the other end, classical MD can handle large systems and provide estimation of bulk properties but is limited with low accuracy.

Recent developments in machine learning (ML) aim at overcoming these limitations of MD simulations. Advances in ML have helped in constructing interatomic potentials from the configuration of atoms. Different potentials are developed based on the ML model and the type of input it takes for predicting the energies and forces of the system. Neural network potential (NNP)[2], gaussian approximation potential (GAP)[3] and moment tensor potential (MTP)[4] are some of the well known interatomic potentials developed though ML. Each of these methods use a different set of descriptors to be passed as an input for the ML model. Descriptors are chemical fingerprints obtained from the atoms and their neighbors within a chemical environment of a molecular system which are used as input for the model.

Machine learning potentials (MLP) are developed by training the model on the configurations labeled with their energies and forces. The trained model helps in obtaining a potential which could be used for obtaining different interactions of the molecules just from the atomic positions and removes the need to run the calculations again. MLPs are computationally less expensive compared to AIMD while simulating bulk systems. MLPs are accurate and have transferable bulk properties, thus overcoming the limitations of classical MD. MLP has facilitated understanding molecules and reactions at atomistic levels and getting better insights into the properties of molecules. The advantages of MLPs are they can be used to speed up sampling-intensive path-integral molecular dynamics simulations at the accuracy of the reference electronic structure method. Also, MLPs can be used for systems with larger molecules after training on smaller molecule system thus, saving the computational time and costs to predict energies and forces[5]. MLPs can also operate to obtain properties at different thermodynamic parameters.

1.1 Challenges in developing MLPs

MLPs have proved to have many advantages over the conventional electronic structure methods or empirical forcefields for studying atomistic simulations. However, there are also many challenges that come for reproducing PES for molecular systems. To develop a converged MLP, models require large training data, and it can be infeasible when expensive ab initio data is required[6]. There have been issues with building an accurate MLP that describes complex PES. The more complex the system, more reference data is required, which ends up adding the number of descriptors. Constructing MLP for organic molecules has proved to be more challenging compared

to solids and materials[7]. MLPs are also unable to give new insights which are not provided in the training data[8].

To overcome these challenges, there have been different solutions suggested in the literature. Watanabe et al.[9] discusses different sampling and optimization methods to construct better potentials. MLPs are often limited to the accuracy of the system when predicting energies and forces for new configurations as usually molecular structures are majorly described through local chemical information[10]. Developments have been made to develop MLPs by taking longer range interactions into account like long range electrostatic[11,12] and van der Waals[13] interactions. To improve the performance of the model, descriptors are optimized in a differential[14] or hybrid[15] fashion. Data science methods like principal component analysis (PCA)[16] have also been used to obtain the important components of the training data and thus reducing the training time. While other efforts like using gaussian density function (GDF)[17] has been used for weighting and overcoming the sampling bias in MLP have been developed. Recently, in order to avoid generating large datasets, active learning[18,19] approach has been used for developing MLP by generating 'on-the-fly' via sampling. This approach significantly reduces the training data and thus the training time, and only runs calculations for the configurations with low accuracy.

In this work, we present a solution to overcome the challenges faced while developing MLP. We use enhanced sampling methods as the training data for ML models. Enhanced sampling methods helps in obtaining far from equilibrium data and the potential developed with it, shows effective results and accuracy which is close to that obtained from the conventional methods. Furthermore,

we check the validity of the potential by obtaining different properties for the system and compare them with the properties from the training data.

Chapter 2. Methods

All the developments for MLPs hint that it is important to have a good training dataset. If the dataset is not well described or large enough, the accuracy of MLP is severely affected. Obtaining a large dataset is not always possible if the training data is from computationally expensive methods like ab initio simulations.

For these limitations, we present a solution by using enhanced sampling to generate the training data. Enhanced sampling methods help in exploring the energy profile of the system by adding a history dependent bias potential to the Hamiltonian of the system. Enhanced sampling methods are categorized into collective variable and collective variable free methods[20]. There are different methods such as metadynamics[21], umbrella sampling, replica exchange[22] and many others which are used to remove the energy barrier and explore the energy surface of the system. This bias potential is added to the system based on the collective variables (CV) or degrees of freedom of the system. Enhanced sampling pushes the system away from the local minima such that it does not resample the minima again and explores the energy profile of the system.

In this work, we obtained MLP by training the model on 1000 water molecules data generated from MD simulations by equilibrating the system with isobaric and canonical ensembles. From

the training, we observe that MLP fails to accurately represent the PES for the system. Later, we add metadynamics to the system and generate new training data on which the ML model is trained.

2.1 Generating training data

We obtain training data by initializing a 1000 molecules water system using Packmol[23]. All the simulations are carried out using the LAMMPS software package. The initial geometry for LAMMPS[24] was created using VMD[25]. After obtaining the initial geometry, the simulations were carried out in LAMMPS with TIP3P water model and Lennard Jones and coulomb potential. The system was equilibrated with NPT at 300K temperature using CSVR thermostat and Nose-Hoover barostat for 10ns and later NVT simulation was carried out at 300K temperature using CSVR thermostat for 10ns. The dimensions of the box are 31.72 Å × 31.72 Å × 31.72 Å under periodic boundary conditions.

We perform well tempered ensemble (WTE) simulations for minimizing the energy barrier of the system. We take the system discussed above and add metadynamics to the system. We use potential energy as the CV for enhanced sampling. The width of the gaussian was obtained from the standard deviation of the potential energy of the system from the NVT simulation. Lastly, we carried out the simulations for different biasfactor (16, 32, 64, 256) and obtained the training data.

To validate the training data, we obtained the radial distribution function (RDF) plots for the training data. As shown below, the RDF plots for NVT and WTE simulations agree with the RDF

plot of water at 300K. The potential energy CV plot was obtained which shows a gaussian distribution.

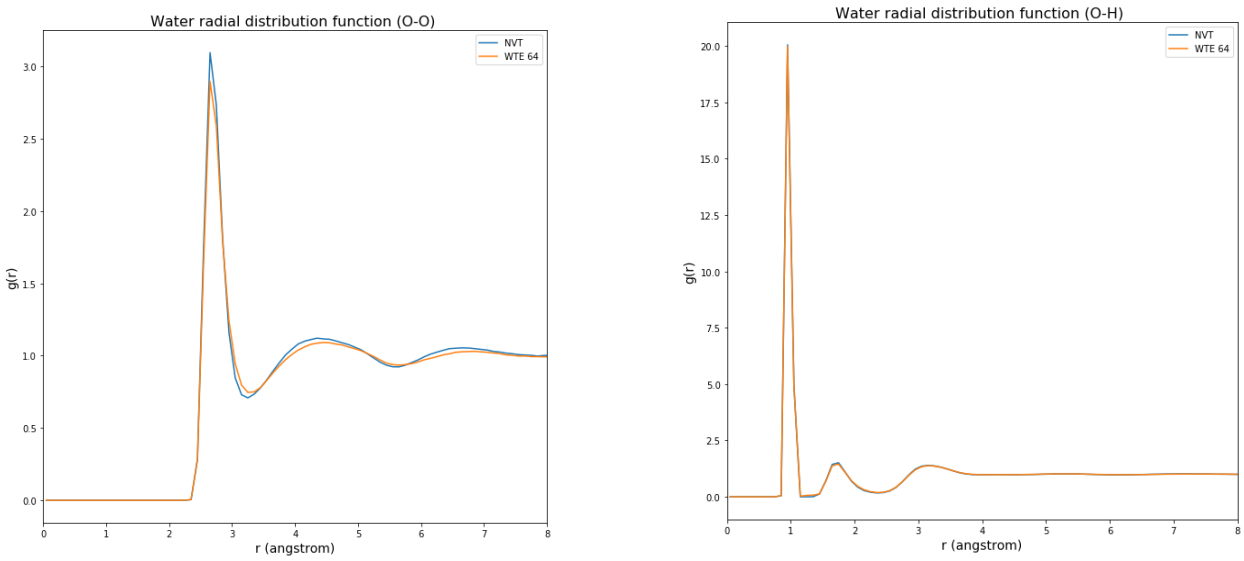


Figure 1: RDF g(r) (O-O) & (O-H) plots for training data, water system after NVT and WTE simulations

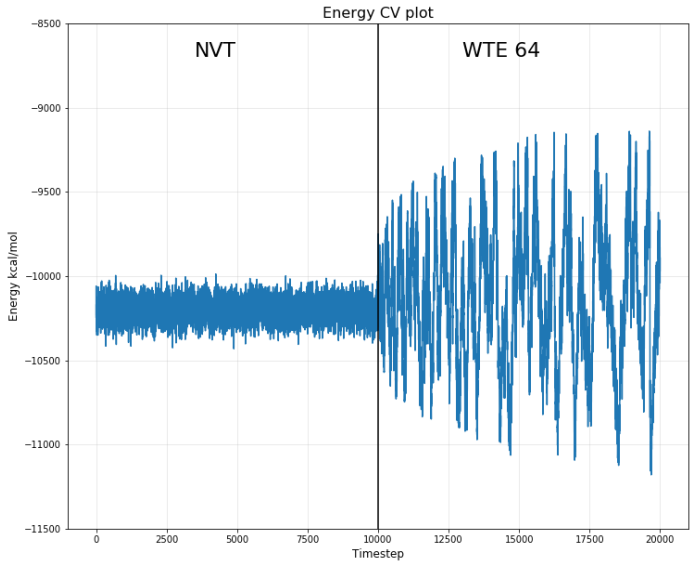


Figure 2: Energy profile for NVT and WTE simulations from the training data

2.2 Training ML model

To develop a MLP for the water system we use the DeePMD-kit[26] package which implements deep learning neural network representation for atomic interactions. DeePMD-kit uses atomic environment descriptors as input for training the neural network. The atomic environment is created by defining a cut-off radius and taking the interactions of atoms within the cut-off radius. By developing descriptors through the atomic environment, DeePMD-kit preserves the rotational, translational and permutational symmetry of the system.

DeePMD-kit constructs local frames and records local coordinates for each atom. The descriptor information is given as:

$$D_{ij}{}^{\alpha} = \left\{\frac{1}{R_{ij}}, \frac{x_{ij}}{R_{ij}}, \frac{y_{ij}}{R_{ij}}, \frac{z_{ij}}{R_{ij}}\right\} \text{(total information)}$$

$$D_{ij} = \frac{1}{R_{ij}} \quad \text{(radial information)} \tag{1}$$

Here, when $\boldsymbol{\alpha}$ is 0, only radial information of the atoms is provided. While $\boldsymbol{\alpha}$ = {1, 2, 3, full} provides radial plus angular information of the atoms. R_{ij} denotes the relative position of atom i with respect to atom j and (x_{ij}, y_{ij}, z_{ij}) are the coordinates of the relative position between the atoms. The total coordinate information provides the bonded interactions between the atoms and radial coordinate information provides the non-bonded interactions between the atoms.

The NN trained on the descriptors predicts the atomic energies of the coordinates and the total energy of the system is obtained by summing the atomic energies of all the atoms. DeePMD-kit tries to achieve maximum accuracy by minimizing the loss function:

$$L\left(p_\varepsilon, p_f, p_\xi\right) = \frac{p_\xi}{N}\Delta E^2 + \frac{p_f}{3N}\sum_i |\Delta \mathrm{F_i}|^2 + \frac{p_\varepsilon}{9N}\|\Delta \Xi\|^2 \qquad (2)$$

where ΔE, ΔF_i, and $\Delta\ \Xi$ are root mean square (RMS) errors of energy, force, and virial respectively. $p_\varepsilon, p_f, p_\xi$ are the prefactors which change during the optimization process in training and obtained from the learning rate. The model is optimized by implementing TensorFlow's Adam stochastic gradient descent method.

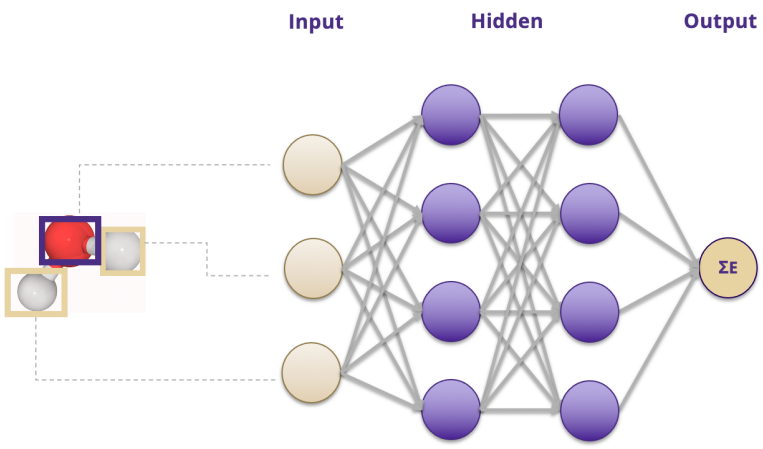


Figure 3: MLP model construction. The input for the model are atomic coordinates which are set as descriptors and feed into the network

The ML model is trained by converting the training data generated to frames. Each coordinate frame is labeled with its corresponding energy and force. Along with coordinates, energies and forces, DeePMD-kit also requires the dimensions of the box for the system and a file which contains the atom types of the system. After converting the training data to appropriate format, the NN is trained with input scripts which describe the hyperparameters for training. Here, you define the cut-off radius for the atomic environment, number of hidden layers and other training parameters. We trained four models for each of the simulations, changing the seed value in the training script to have a different initialization for the training. The ML models were trained with 40,000 frames for all the simulations unless mentioned otherwise. The frames were randomly shuffled before feeding them to the NN.

Chapter 3. Results and Discussions

In this work, we developed MLP from the training data generated through enhanced sampling methods. To verify the inefficiency of MLP with limited data, we trained the model with NVT simulation data and obtained results. As shown in the figure below, MLP are not able to represent the system accurately from NVT simulation data. We obtained high fluctuations in temperature while performing NVT simulation using MLP.

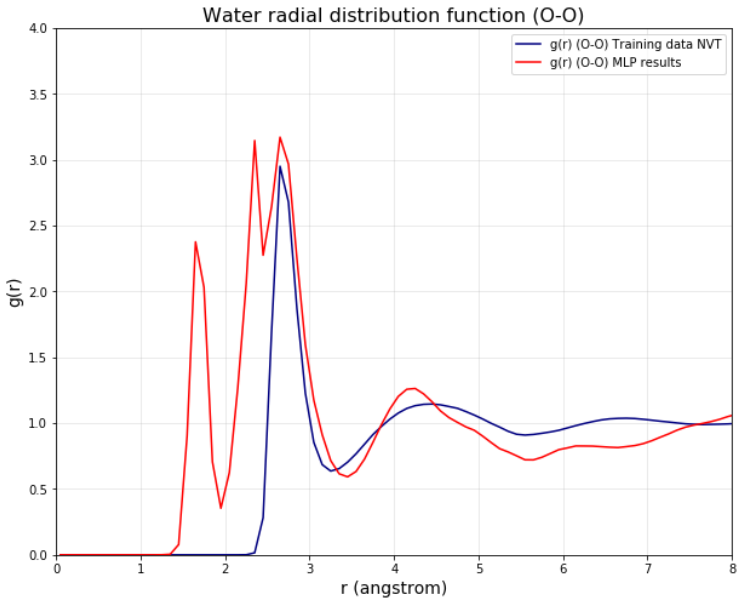


Figure 4: RDF results for MLP obtained with NVT training data

From the RDF results, we verify that MLP is unable to represent the systems trained with NVT dataset. We develop MLPs from the enhanced sampling data. The MLPs were developed using data from WTE (γ = 16, 64, 256). All the results show that the accuracy of the MLP is much improved by using enhanced sampling as the training data. Here, we also observe that good accuracy for the potential can be obtained even by training the model for less time. We trained the model with WTE data for 5 epochs instead of 25 epochs which has been used as a standard training time for our work. The results obtained show good accuracy for the model. Hence, we can say that enhanced sampling methods not only provide accurate representation of molecular systems but also provides accurate models with less training.

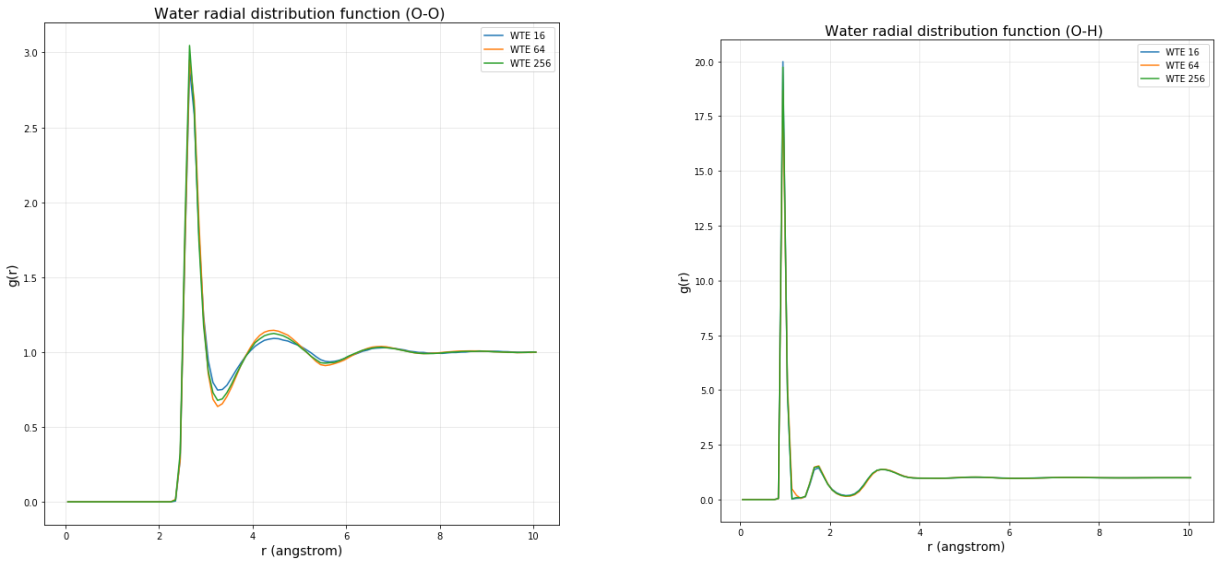


Figure 5: RDF results for MLP obtained with enhanced sampling training data

All the simulations for testing the potential were carried out using LAMMPS. We used NVT simulation at 300K with the deepmd potential in LAMMPS. We also verify the transferability of MLP by changing the thermodynamic parameters and running the simulations at higher temperatures. Results show that MLP can represent systems at higher temperature values than the temperature it was trained on.

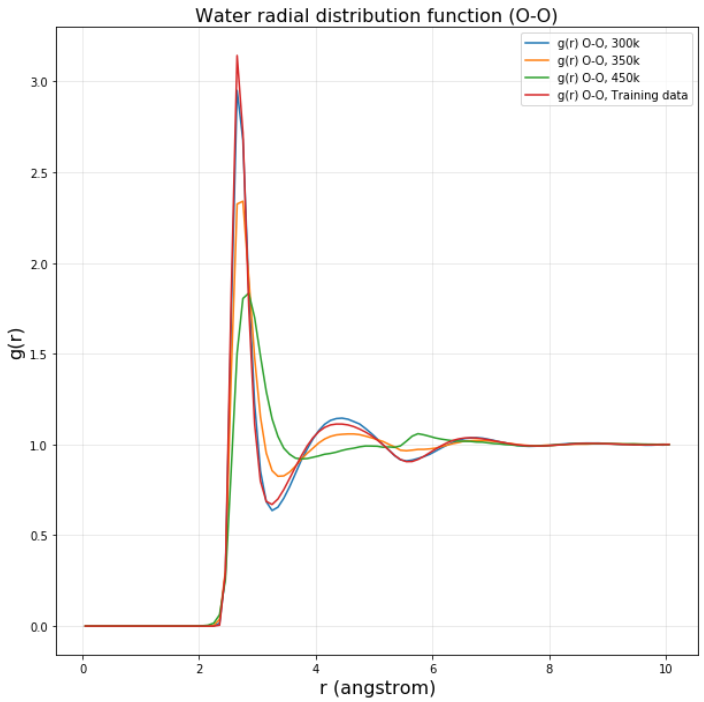


Figure 6: Different temperature RDF plots for MLP trained with enhanced sampling training data

We obtain the self-diffusion coefficient of water calculated from MLP trained with enhanced sampling data obtained at 300K temperature. Self-diffusion coefficients were calculated from the mean square displacement (MSD) obtained using LAMMPS. The results obtained are in agreement with the self-diffusion coefficients for TIP3P potential[27,28].

	Self-diffusion coefficient (training data)	Self-diffusion coefficient (obtained from MLPs)
WTE 16	5.8877 (10^{-5} cm^2/s)	5.7194 (10^{-5} cm^2/s)
WTE 64	5.9008 (10^{-5} cm^2/s)	5.6933 (10^{-5} cm^2/s)
WTE 256	5.9747 (10^{-5} cm^2/s)	5.9137 (10^{-5} cm^2/s)

Table 1: Self-diffusion coefficient of water obtained from MLP results

Chapter 4. Conclusion

In this work, we have presented a method to develop MLP. There are many challenges faced while obtaining MLPs for systems. We propose the use of enhanced sampling methods to obtain training datasets for the ML models. Using enhanced sampling methods, overcomes the limitations of developing MLPs and improves the accuracy of the model. We verify the hypothesis for the challenges faced in MLP by using NVT training data for liquid water system with TIP3P potential. Enhanced sampling not only improves the quality of the training data but also succeeds in delivering converged models trained for less epochs.

In addition, we explain the workflow for developing MLP, generating training data and obtaining results with the developed potential. We provide analytical details on building chemical environment descriptors for the deep learning architecture used by DeePMD-kit in the supplementary information. We believe there are many applications where enhanced sampling generated training data could be used for developing MLPs. Our future work plans are to expand our approach for different systems. We plan to develop MLP for organic molecules where enhanced sampling methods could play a crucial role. Moreover, we also plan to automate the process by minimizing human intervention and developing a tool to automize the entire process for developing MLPs.

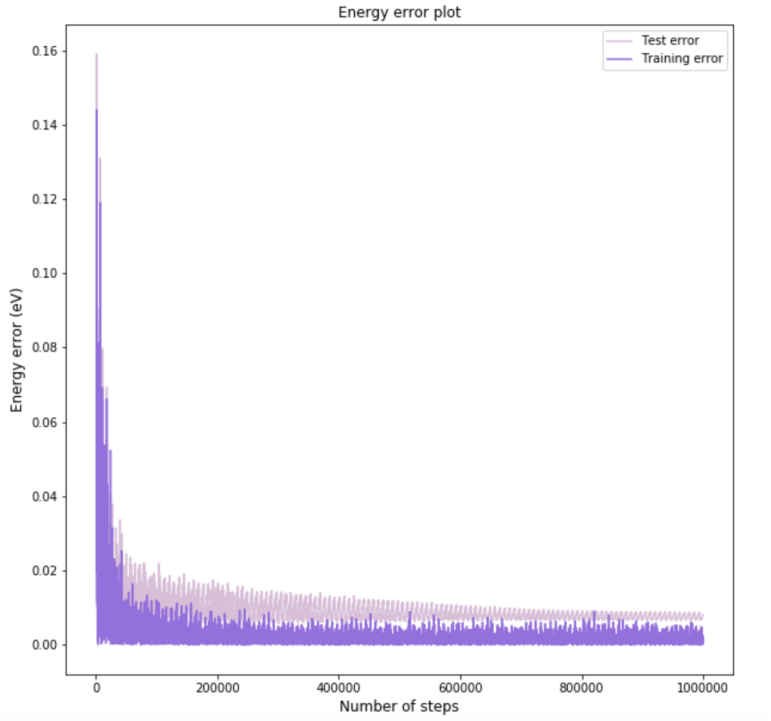
Supplementary Material

All the training data was generated using LAMMPS package. The ML model was trained using DeePMD-kit package on NVIDIA GeForce RTX 2080 Ti GPUs. We used 40,000 frames obtained from 1000 molecules water system among which 38,000 frames were used for training and 2,000 were used for testing set. The neural network architecture contains 3 hidden layers with size [60 ,60, 60]. The cut-off radius is set to 6 Å. We train four different models for developing each potential, where each model is initialized with a different seed value. The model was trained for 25 epochs. After training, the model is saved, and the state of the model is frozen using DeePMD-kit commands which is later used as a potential for running the simulations.

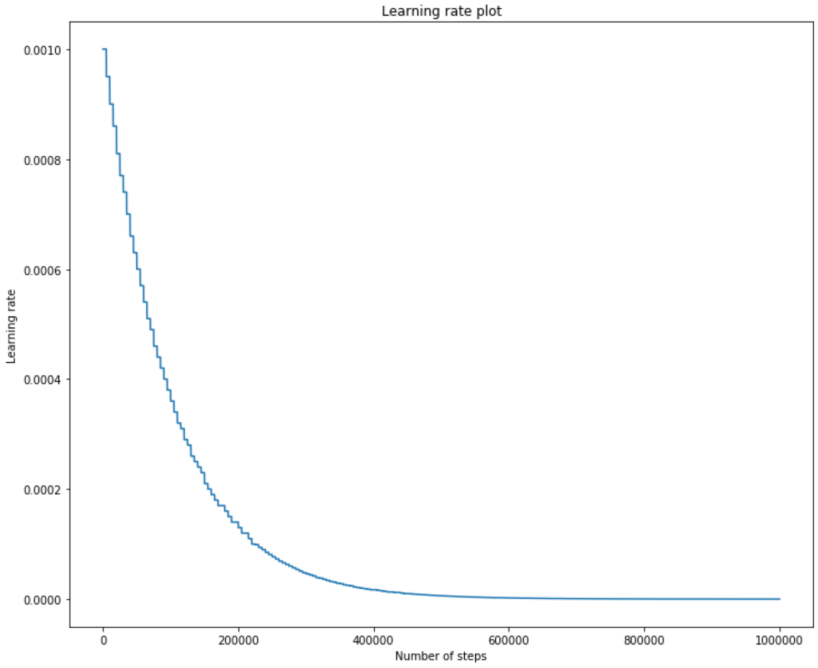
The validation of the MLP was done with LAMMPS using deepmd pair_style. All the MD simulations with the developed MLP were run on NVIDIA GeForce RTX 2080 Ti GPUs.

```
{
    "_comment": " model parameters",
    "model": {
        "type_map":     ["H", "O"],
        "descriptor" :{
            "type":             "se_a",
            "sel":              [92, 46],
            "rcut_smth":        5.80,
            "rcut":             6.00,
            "neuron":           [25, 50, 100],
            "resnet_dt":        false,
            "axis_neuron":      16,
            "seed":             1,
            "_comment":         " that's all"
        },
        "fitting_net" : {
            "neuron":           [240, 240, 240],
            "resnet_dt":        true,
            "seed":             1,
            "_comment":         " that's all"
        },
        "_comment":     " that's all"
    },
"water_se_a.json" 70L, 1373C
```

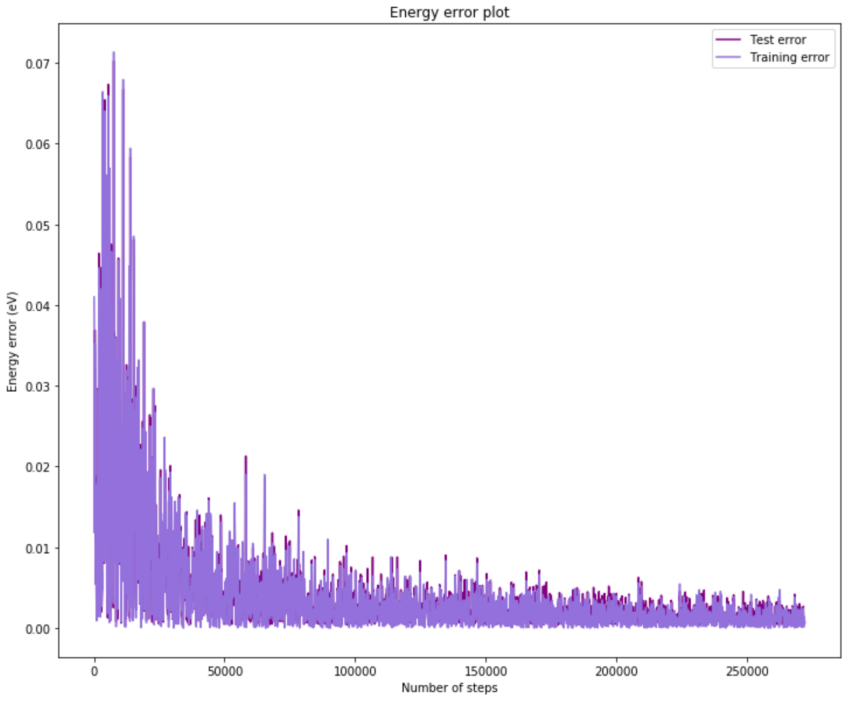
S. I. Figure 1: DeePMD-kit training parameters



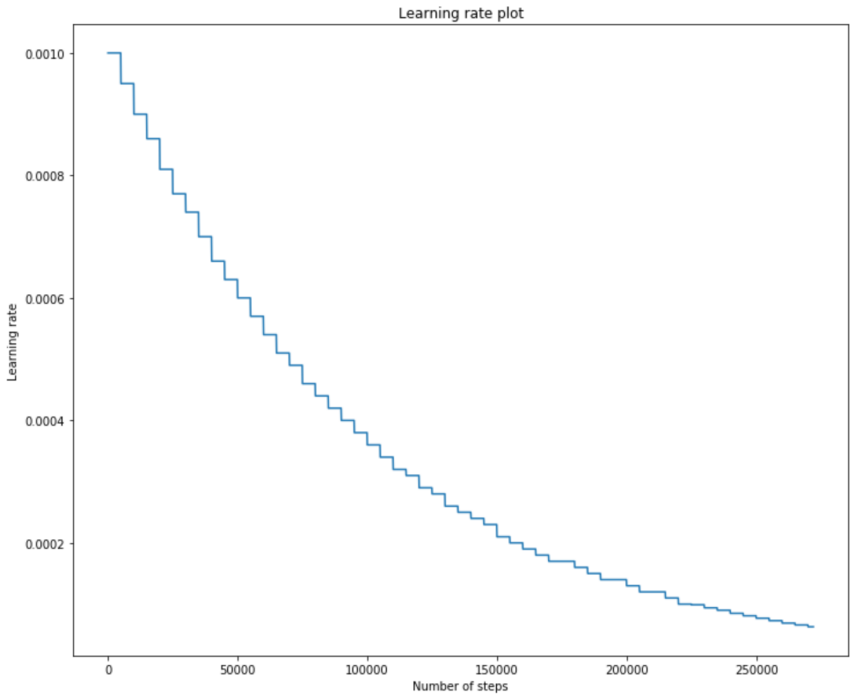
S. I. Figure 2: The root mean square testing errors are plotted with the training errors while performing training for data generated through NVT simulation



S. I. Figure 3: Learning rate plot for training data generated with NVT simulation



S.I. Figure 4: The root mean square testing errors are plotted with the training errors while performing training for data generated through WTE simulation



S. I. Figure 5: Learning rate plot for training data generated with WTE simulation

Bibliography

1. Unke OT, Chmiela S, Sauceda HE, et al. Machine Learning Force Fields. *Chem Rev*. Published online March 11, 2021. doi:10.1021/acs.chemrev.0c01111

2. Behler J, Parrinello M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys Rev Lett*. 2007;98(14):146401. doi:10.1103/PhysRevLett.98.146401

3. Bartók AP, Csányi G. Gaussian Approximation Potentials: a brief tutorial introduction. *arXiv:150201366 [cond-mat, physics:physics]*. Published online February 5, 2020. Accessed June 13, 2021. http://arxiv.org/abs/1502.01366

4. Shapeev AV. Moment Tensor Potentials: a class of systematically improvable interatomic potentials. *Multiscale Model Simul*. 2016;14(3):1153-1173. doi:10.1137/15M1054183

5. Gastegger M, Marquetand P. Molecular Dynamics with Neural-Network Potentials. *arXiv:181207676 [physics, stat]*. Published online December 18, 2018. Accessed May 18, 2021. http://arxiv.org/abs/1812.07676

6. Vassilev-Galindo V, Fonseca G, Poltavsky I, Tkatchenko A. Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules. *The Journal of Chemical Physics*. 2021;154(9):094119. doi:10.1063/5.0038516

7. Gkeka P, Stoltz G, Barati Farimani A, et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *J Chem Theory Comput*. 2020;16(8):4757-4775. doi:10.1021/acs.jctc.0c00355

8. Behler J. Perspective: Machine learning potentials for atomistic simulations. *J Chem Phys*. 2016;145(17):170901. doi:10.1063/1.4966192

9. Watanabe S, Li W, Jeong W, et al. High-dimensional neural network atomic potentials for examining energy materials: some recent simulations. *J Phys Energy*. 2020;3(1):012003. doi:10.1088/2515-7655/abc7f3

10. Noé F, Tkatchenko A, Müller K-R, Clementi C. Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry*. 2020;71(1):361-390. doi:10.1146/annurev-physchem-042018-052331

11. Artrith N, Morawietz T, Behler J. High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys Rev B*. 2011;83(15):153101. doi:10.1103/PhysRevB.83.153101

12. Unke OT, Meuwly M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments and Partial Charges. *J Chem Theory Comput*. 2019;15(6):3678-3693. doi:10.1021/acs.jctc.9b00181

13. Han J, Zhang L, Car R, E W. Deep Potential: a general representation of a many-body potential energy surface. *CiCP*. 2018;23(3). doi:10.4208/cicp.OA-2017-0213

14. Gao H, Wang J, Sun J. Improve the performance of machine-learning potentials by optimizing descriptors. *J Chem Phys*. 2019;150(24):244110. doi:10.1063/1.5097293

15. Goryaeva AM, Maillet J-B, Marinica M-C. Towards better efficiency of interatomic linear machine learning potentials. *Computational Materials Science*. 2019;166:200-209. doi:10.1016/j.commatsci.2019.04.043

16. Cubuk ED, Malone BD, Onat B, Waterland A, Kaxiras E. Representations in neural network based empirical potentials. *J Chem Phys*. 2017;147(2):024104. doi:10.1063/1.4990503

17. Toward Reliable and Transferable Machine Learning Potentials: Uniform Training by Overcoming Sampling Bias | The Journal of Physical Chemistry C. Accessed June 13, 2021. https://pubs.acs.org/doi/full/10.1021/acs.jpcc.8b08063

18. Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*. 2017;140:171-180. doi:10.1016/j.commatsci.2017.08.031

19. Zhang L, Lin D-Y, Wang H, Car R, E W. Active Learning of Uniformly Accurate Inter-atomic Potentials for Materials Simulation. *Phys Rev Materials*. 2019;3(2):023804. doi:10.1103/PhysRevMaterials.3.023804

20. Yang YI, Shao Q, Zhang J, Yang L, Gao YQ. Enhanced sampling in molecular dynamics. *J Chem Phys*. 2019;151(7):070902. doi:10.1063/1.5109531

21. Valsson O, Tiwary P, Parrinello M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu Rev Phys Chem*. 2016;67(1):159-184. doi:10.1146/annurev-physchem-040215-112229

22. Bernardi RC, Melo MCR, Schulten K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochimica et Biophysica Acta (BBA) - General Subjects*. 2015;1850(5):872-877. doi:10.1016/j.bbagen.2014.10.019

23. Martínez L, Andrade R, Birgin EG, Martínez JM. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry*. 2009;30(13):2157-2164. doi:10.1002/jcc.21224

24. Plimpton S. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *Journal of Computational Physics*. 1995;117(1):1-19. doi:10.1006/jcph.1995.1039

25. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*. 1996;14(1):33-38. doi:10.1016/0263-7855(96)00018-5

26. Wang H, Zhang L, Han J, E W. DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Computer Physics Communications*. 2018;228:178-184. doi:10.1016/j.cpc.2018.03.016

27. Mark P, Nilsson L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J Phys Chem A*. 2001;105(43):9954-9960. doi:10.1021/jp003020w

28. Mark P, Nilsson L. Structure and dynamics of liquid water with different long-range interaction truncation and temperature control methods in molecular dynamics simulations. *Journal of Computational Chemistry*. 2002;23(13):1211-1219. doi:10.1002/jcc.10117