NE452 Project Report

Training a Machine Learning Force Field to Mimic the Classic TIP3P Force Field

By

Yunheng Zou

Submitted in fulfilment of NE452

Instructor: P-N. Roy

# Acknowledgement

I would like to thank Professor Pierre-Nicholas Roy for his support and supervision of this work.

# Abstract

Deep learning and machine learning techniques are revolutionizing many scientific areas. This report presents a pipeline to train a machine learning force field based on a potential energy surface dataset. In this report, we demonstrate how to use this pipeline to train an ANI neural network as well as a simple Multi-Layer Perceptron to approximate the Tip3P potential for H2O monomer and dimer systems. We also compared both the potential and force from the Tip3P and NN force fields to confirm their accuracy.

# Table of Content

Contents

[Acknowledgement ii](#_Toc163931676)

[Abstract ii](#_Toc163931677)

[Table of Content ii](#_Toc163931678)

[List of Figures iii](#_Toc163931679)

[Introduction 1](#_Toc163931680)

[Methodology 1](#_Toc163931681)

[Data Generation 1](#_Toc163931682)

[Water Monomer Dataset Preparations 1](#_Toc163931683)

[Wate Dimer Dataset Preparations 2](#_Toc163931684)

[Neural Network Architecture 3](#_Toc163931685)

[Multi-Layer Perceptron 3](#_Toc163931686)

[ANI 4](#_Toc163931687)

[Loss Function 5](#_Toc163931688)

[Results 5](#_Toc163931689)

[Conclusions 7](#_Toc163931690)

[Impact 7](#_Toc163931691)

[Future Work 7](#_Toc163931692)

[Code availability 7](#_Toc163931693)

[References 8](#_Toc163931694)

# List of Figures

[Figure 1. Compare NN based Force Field with DFT calculations. (Gao et al., 2020) 1](#_Toc163931695)

[Figure 2. Code snippet of the potential energy and force calculation based on configurations 2](#_Toc163931696)

[Figure 3. Multi-Layer Perceptron architecture 3](#_Toc163931697)

[Figure 4. MLP PyTorch Implementation 3](#_Toc163931698)

[Figure 5. ANI architecture for H2O monomer 4](#_Toc163931699)

[Figure 6. Code Snippet of the Customized Loss Function 5](#_Toc163931700)

[Figure 7. MLP Potential and Force prediction compared to tip3p ground truth. A: MLP potential on Monomer. B: MLP force on Monomer. C: MLP Potential on Dimer. D. MLP force on Dimer 6](#_Toc163931701)

[Figure 8. ANI Potential and Force prediction compared to tip3p ground truth. A: ANI potential on Monomer. B: ANI force on Monomer. C: ANI Potential on Dimer. D. ANI force on Dimer 6](#_Toc163931702)

# Introduction

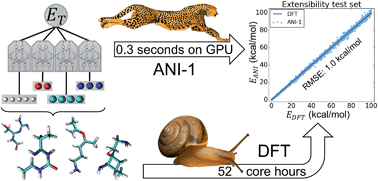
Machine learning force fields (ML-FFs) represent a novel approach to constructing interatomic potentials that combine the accuracy of ab-initio/DFT quantum mechanical calculations with the computational efficiency of classical force fields. ML-FFs utilize machine learning techniques to map atomic environments to potential energy surfaces or interatomic forces, instead of relying on predefined analytical functional forms typical of traditional force fields. They provide improved accuracy when trained with ab initio or DFT datasets. Compared to DFT and ab initio methods, ML-FFs achieve an acceleration of 3-5 orders of magnitude, as they eliminate the need for complex computations. Potentials and forces can be obtained directly by forwarding through the network.

Figure 1. Compare NN based Force Field with DFT calculations. (Gao et al., 2020)

# Methodology

## Data Generation

### Water Monomer Dataset Preparations

Water monomer configurations are generated by randomly assigning oxygen and two hydrogen atoms into a 2x2x2 Angstrom box. A total of 40,000 random configurations are generated. These configurations are used as inputs to an OpenMM simulation (Eastman et al., 2017) with TIP3P (LAMMPS, 2024) as the force field. The simulation accurately applies the potential calculations defined by the force field to calculate the potential and the forces, as shown in Figure 2. Configurations, their associated forces, and energies are saved together in a NumPy .npz file.A screen shot of a computer code

Description automatically generated

Figure 2. Code snippet of the potential energy and force calculation based on configurations

### Wate Dimer Dataset Preparations

Like how water monomer is prepared, water dimer also involves creating configuration, and use configuration to compute associated potentials and forces. However, due to 2 molecules are involved, randomly assign each atom in a box will not create valid water molecules. To generate valid water dimer structure, following methodologies are applied:

**Molecular Configuration:** Each water molecule is generated using fixed geometrical parameters with the flexibility in bond lengths. The O-H bond length, typically 0.9572 Å, is varied randomly within a range of ±0.5 Å, resulting in bond lengths between 0.4572 Å and 1.4572 Å. The H-O-H bond angle is maintained at 104.5 degrees.  
**Orientation and Positioning:** For generating a water dimer, two water molecules are created independently with their respective variable bond lengths. The second molecule is subjected to a random rotation using a 3D rotation matrix, which helps simulate random molecular orientations.

**Intermolecular Arrangement:** The relative position of the second molecule to the first is defined by translating it within a preset distance range (typically 2.5 Å to 3.5 Å due to hydrogen bonding) along the x-axis, with additional random adjustments along the y and z axes to simulate a natural fluid environment. These translations incorporate small random components to enhance the variability in the dimer configurations.

The approach ensures that the generated configurations are physically realistic and are varied enough to provide a robust training set for machine learning models simulating the TIP3P water model.

A 70%,15%,15% split is performed on the dataset generated to separate the training, validation, and testing sets.

## Neural Network Architecture

Two Neural Network (NN) architectures are utilized in this report. The first is a simple Multi-Layer Perceptron (MLP), and the second is the Accurate Neural Network Engine for Molecular Energies (ANI), a specialized NN developed for machine learning force fields (MLFF), as outlined by Smith et al. (2017) and Behler & Parrinello (2007).

Both models are implemented using PyTorch, a high-performance Python library for constructing neural networks, (Paszke et al., 2019). ANI is specifically imported from TorchANI, a PyTorch-based wrapper focused on MLFF, as described by Gao et al. (2020).

### Multi-Layer Perceptron

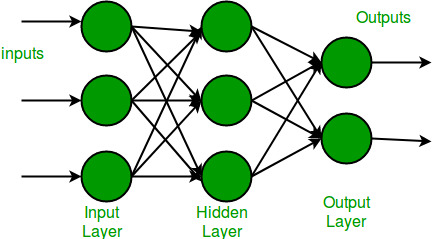


Figure 3. Multi-Layer Perceptron architecture

A Multi-Layer Perceptron (MLP) is a type of artificial neural network comprising multiple layers of neurons, typically including one input layer, one or more hidden layers, and one output layer. Each neuron in a layer connects to every neuron in the subsequent layer, forming a fully connected network. This structure allows it to be trained with backpropagation to learn complex patterns within the data, such as the relationships between forces and potential in relation to geometry.A computer screen shot of a program

Description automatically generated

Figure 4. MLP PyTorch Implementation

The input for this model consists of 3D coordinate information in the format (Batch, Atoms, 3). The output is the predicted potential energy. Forces are computed as the gradient of the predicted potential, aligning with the physical principle that force is the negative derivative of potential energy.

A black text with black letters

Description automatically generated with medium confidence (1)

### ANI

A diagram of a diagram of energy

Description automatically generated

Figure 5. ANI architecture for H2O monomer

ANI, compared to MLP, is a more physically informed network. It is spatially invariant, input order invariant, and facilitates easy transfer learning between different atomic systems. The fundamental concept of this network is that the total energy E is the sum of atomic contributions Ei.

In the initial step, atomic numbers, along with coordinate information (q1, q2, q3) are transformed into symmetry functions G. Each atom’s G is used to feed into a subnet S that is unique to the atomic number. Note that for different atom, S architecture can still (and usually is) be same, while after training, weight associated with each element is stored separately. Thus, each element will have a corresponding subnet Sz. Each subnet provides Ei which is summed to calculate the total energy Etotal. (Behler & Parrinello, 2007)

The symmetry function G, which accounts for radial and angular features, creates an approximation of the local atomic environment. This serves as a solid foundation for the subnet to learn and accurately approximate the atomic energy contribution Ei, as detailed by Smith et al. (2017) and Behler & Parrinello (2007)."

## Loss Function

As the primary goal of using Machine Learning Force Fields (MLFF) is to provide accurate predictions of both potential energy and, more importantly, force, to guide molecular dynamics simulations, the loss functions are constructed to encompass both these aspects. A computer screen shot of a program code

Description automatically generated

Figure 6. Code Snippet of the Customized Loss Function

According to equation (1), an accurate prediction of potential should lead to a precise prediction of force, and vice versa. However, to better regularize network behavior, the loss is formulated as a combination of these two aspects, as depicted in Figure 6. Weights ηpotential and η­force ​are assigned to each term, facilitating improved learning and adaptation of the neural networks. In fact, by experiments, training ANI with only potential led to inaccurate prediction of forces. By fine-tuning the network with a focus on force over a few epochs, the network becomes regularized and begins to produce the correct force values.

# Results

Results displayed in Figures 7 and 8 were obtained by training the model and applying it to the test set, which comprises 15% of the total dataset. The MLP, although relatively small, achieved a high R score for potential but exhibited a large variance in force predictions, as shown in Table 1. This model was trained with force regularizations. As the system scales up, an increase in force variations is anticipated. In contrast, the ANI performed commendably well. Both potential and forces aligned closely with expectations, indicating that the model successfully learned to predict Tip3P potentials. Furthermore, significantly fewer outliers were observed compared to MLP, further substantiating that the model has effectively learned the underlying patterns rather than merely fitting them to the numerical data.

A collage of graphs

Description automatically generated

Figure 7. MLP Potential and Force prediction compared to tip3p ground truth. A: MLP potential on Monomer. B: MLP force on Monomer. C: MLP Potential on Dimer. D. MLP force on Dimer

A group of graphs showing different colored lines

Description automatically generated

Figure 8. ANI Potential and Force prediction compared to tip3p ground truth. A: ANI potential on Monomer. B: ANI force on Monomer. C: ANI Potential on Dimer. D. ANI force on Dimer

|  |  |  |
| --- | --- | --- |
|  | Monomer R^2 | Dimer R^2 |
| MLP\_Force | 0.899 | 0.713 |
| MLP\_Potential | 0.979 | 0.985 |
| ANI\_Force | 0.998 | 0.898 |
| ANI\_Potential | 1.000 | 0.999 |

Table 1. R^2 comparison between MLP and ANI for H2O monomer and dimer system

# Conclusions

## Impact

This study has demonstrated the efficacy of using machine learning techniques to create accurate and computationally efficient force fields for molecular dynamics simulations. By training both a Multi-Layer Perceptron (MLP) and an Accurate Neural network engine for Molecular Energies (ANI) on the Tip3P potential for H2O monomer and dimer systems, we have shown that these machine learning models can learn the pattern hidden behind the dataset. Although the dataset is prepared by Tip3P, the same pipeline adapts to any datasets including DFT and ab initio ones. The successful application of these models confirms their potential to enhance the accuracy and efficiency of force field calculations in molecular simulations, paving the way for their application in more complex systems.

## Future Work

Future research will focus on expanding the application of the developed machine learning models to more complex molecular systems and different types of potential energy surfaces. There is also an opportunity to explore the integration of other machine-learning techniques that could further improve the accuracy and generalizability of the force fields. Additionally, further development can be done to integrate such machine-learned potential with common simulation toolkits and libraries such as OpenMM and LAMMPS. This will enable machine-learned potential to compute important system characteristics such as radial distribution functions that are experimentally measurable, which allows further confirmations.

## Code availability

Code to reproduce the results is available on GitHub: <https://github.com/yunhzou/MLFFOpenMM>

# References

Behler, J., & Parrinello, M. (2007).Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, *98*(14). https://doi.org/10.1103/physrevlett.98.146401

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R., & Pande, V. S. (2017). OpenMM 7: Rapid development of high performance algorithms for Molecular Dynamics. *PLOS Computational Biology*, *13*(7). https://doi.org/10.1371/journal.pcbi.1005659

Gao, X., Ramezanghorbani, F., Isayev, O., Smith, J. S., & Roitberg, A. E. (2020). Torchani: A free and open source pytorch-based deep learning implementation of the ANI neural network potentials. *Journal of Chemical Information and Modeling*, *60*(7), 3408–3415. https://doi.org/10.1021/acs.jcim.0c00451

LAMMPS. (2024). *8.4.3. tip3p water modelℑ*. 8.4.3. TIP3P water model - LAMMPS documentation. https://docs.lammps.org/Howto\_tip3p.html

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc. Retrieved from http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Smith, J. S., Isayev, O., & Roitberg, A. E. (2017). Ani-1: An extensible Neural Network Potential with DFT accuracy at force field computational cost. *Chemical Science*, *8*(4), 3192–3203. https://doi.org/10.1039/c6sc05720a