Machine Learned Conformer Energy Prediction via Approximating Pairwise Potential

Curtis Wu, Carmen Matar

May 12, 2025

1 Abstract

Accurately predicting molecular conformer energies is essential for material science and drug discovery; however, traditional methods for calculating and predicting these energies can be computationally demanding or lack flexibility for chemically diverse molecular systems. In this report, we discuss our approach and results in creating a C++ implementation of an interpretable multi-layer perceptron model that learns directly from the Accurate Neural Network Interatomic Potentials (ANI-1) dataset, to produce near quantum mechanical accuracy for predicting conformer energies by approximating the atomic pairwise potential. After training and validating on the model, the test set earned a mean absolute error of 6.67 Hartree, significantly outperforming the tested Lennard-Jones potential and demonstrating its ability to learn meaningful chemical interactions, such as bond stabilization and energy decay at larger interatomic distances. Limitations were observed, such as a decrease in model accuracy for larger molecules with more heavy atoms; however, its interpretable nature and exceptional accuracy highlight the benefits of machine learning and data-driven approaches to achieve computationally efficient and accurate energy predictions in diverse chemical systems.

2 Introduction

Traditional methods for calculating and predicting molecular energies and force fields include quantum mechanical methods, such as Density Functional Theory (DFT), and classical force fields like Lennard-Jones and Merck Molecular Force Field (MMFF94). In a paper titled "Molecular dynamics simulations and drug discovery," authors Jacob Durrant and Andrew McCammon discuss the

usage of quantum-mechanical calculations and data in molecular dynamics simulations and the importance of energy parameters in molecular behavior replication [2]. However, these traditional methods also have limitations. Quantum mechanical approaches are accurate but require ample computational resources, especially for large-scale simulations [1]. Because of their high computational cost, QM approaches are impractical for small systems. Classic force fields' simpler analytical forms offer computational speed but have limited accuracy for intricate molecular systems. A classic force field approach also lacks flexibility for some chemically diverse molecular systems due to strict parameters.

A machine learning approach is achievable because a model can learn directly from given data without being bound to analytical forms. Moreover, if designed correctly, models can achieve accuracy near QM results while lowering computational expense. Multi-layer perceptions (MLPs) in neural networks effectively model molecular interactions, but sometimes are uninterpretable, hindering trustworthiness and adoption.

2.1 Motivation

With these strengths and limitations in mind, this project aims to develop an explainable neural force field model that learns atomic interaction potentials directly from data and clearly evaluate whether this learned model can match or surpass the accuracy of a classical Lennard-Jones potential approach.

3 Objectives

Our first objective is to ensure our neural network model is implemented entirely in C++ using low-level libraries to predict molecular conformer energies from learned pairwise atomic interactions. Our second objective is to train and validate our model using the ANI-1 dataset. Lastly, we evaluate and quantify the accuracy of our learned neural force field. We aim to maintain computational efficiency and transparency in our model's predictions.

4 Data

To train and validate our model, we use ANI-1 because of its accurate molecular energy calculations for small organic molecules. According to Smith et al., the ANI-1 dataset uses Normal Mode Sampling (NMS) applied to each molecule in the GDB-11 database with no more than eight heavy atoms comprised of carbon, nitrogen, and oxygen [3]. A separate article by Smith et al. states that the electronic structures are calculated with wB97X DFT density functional with 6-31G(d) basis set in the Gaussian 09 electronic structure package [4]. The dataset comprises 17.2 million conformations from approximately 58 thousand small molecules [3].

ANI-1 has subsets based on the number of heavy atoms present in each molecule. The ANI-1 tar file contains files s01 through s08 [5]. For example, subset s01 contains one heavy atom. Moreover, each subset has a specific number of molecular conformers and unique molecules, which helps ensure our model trains in multiple molecular environments. Figure 1 shows further specifications for subsets one through five. From the data, we randomly selected one hundred conformers for each molecule in subsets one through four, resulting in 5900 conformers. Our test set includes molecular conformers from a randomly selected unique molecule with a set of heavy atoms. Therefore, our training set comprises 5600 conformers and a test set of 300 conformers. Figure 2 shows the energy distribution of subsets one through three. We did not include subset four in this plot because of the large file size.

Subset (# of Heavy Atoms)	# of Conformers	# of Unique Molecules
s01	10,800	3
s02	50,962	13
s03	151,200	14
s04	651,936	29
s05	1,813,151	69

Figure 1: Specifications for ANI-1 Subsets 1 - 5

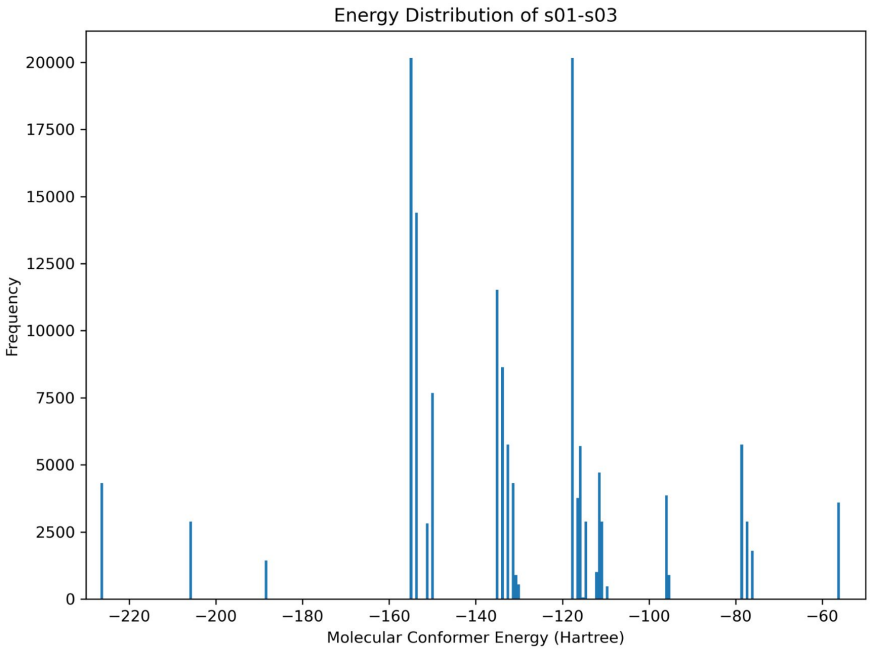


Figure 2: Energy Distribution for ANI-1 Subsets 1 - 3

5 Methodology

We began by preprocessing the ANI-1 dataset. The ANI-1 subset files originate as .h5 files. Our code repository comprises .cpp files, except for a single Python script to convert .h5 files to a CSV format. The Python script ensures seamless integration of the ANI-1 data into the rest of our C++ code. The Python script also parses through data in the .h5 files. We extracted pairwise atomic data, including atomic types, interatomic distances, and total conformer-level energies. The extracted data is an essential unit of input for our model. Further preprocessing is done in our main.cpp file, where we calculate the mean and standard deviation to calculate normalized interatomic distances for stable neural network training.

As for input representation, we encode each atomic pair using one-hot encoded vectors for atom types C, H, O, and N, along with normalized interatomic distances. The result is an input vector of dimension nine per atomic pair, with four values from atom one, four from atom two, and one value for the distance. Our neural force field model is a multi-layer perceptron (MLP) with the architecture 9, 256, 128, 64, 1. Beyond the dimension nine layer, the model contains other hidden layers. For example, the first hidden layer consists of 256 neurons, and the last layer is a single

scalar output predicting the energy contribution of the atomic pair. Therefore, our MLP takes in an atomic pair and outputs an energy for the pair, which is then summed across all pairs to estimate the total conformer energy. This approach imitates the process of classic force fields, like Lennard-Jones, which aggregate pairwise potentials; however, unlike the classic force fields, we do not need fixed equations, as our neural network learns energy contributions directly from the ANI-1 data.

To implement the neural network, we created a `MLP` class to hold all functionalities needed for the training of our model, with the header file snippet shown in Fig. 3. It includes functions such as `predict_pair_energy` and `forward_pair` for inference, `backward_pair` and `apply_gradients` for training, and utilities such as saving and loading model weights. In terms of training, The total predicted energy for a molecule M composed of N atomic pairs is the sum of the individual MLP predictions for each pair:

$$E_{\text{pred}}^{(M)} = \sum_{p=1}^{N^{(M)}} E_{\text{pair},p}^{(M)}$$

where $E_{\text{pair},p}^{(M)}$ is the MLP's output for the p-th pair in molecule M. The model is trained by minimizing the Mean Squared Error (MSE) between the predicted total molecular energy and the true (ground truth) molecular energy $E_{\text{true}}^{(M)}$. For a single molecule M, the loss is:

$$\mathcal{L}^{(M)} = \frac{1}{2}\left(E_{\text{pred}}^{(M)} - E_{\text{true}}^{(M)}\right)^2$$

The gradients with respect to the weights and biases of the MLP are computed using backpropagation. For a given layer l in the network (where l ranges from the first hidden layer to the output layer), the gradient contributions from each pair p are accumulated. The accumulated gradient for the weight and bias $w_{jk}^{(l)}$ connecting neuron k in layer $l-1$ to neuron j in layer l is given by:

$$\Delta W_{jk}^{(l)} = \sum_{p=1}^{N^{(M)}} \left(\delta_{j,p}^{(l)} \cdot a_{k,p}^{(l-1)}\right)$$

$$\Delta B_{j}^{(l)} = \sum_{p=1}^{N^{(M)}} \delta_{j,p}^{(l)}$$

After processing all $N^{(M)}$ pairs in a molecule M and accumulating these gradients $\Delta W_{jk}^{(l)}$ and $\Delta B_{j}^{(l)}$ for all weights and biases in the network, the MLP parameters are updated using gradient descent:

$$W_{jk}^{(l)} \leftarrow W_{jk}^{(l)} - \eta \cdot \Delta W_{jk}^{(l)}$$

```
class MLP {
public:
    MLP(int input_dim, const std::vector<int>& hidden_layers, int output_dim, double lr);

    Vector encode_input(int i1, int i2, double norm_d, int num_atom_types);
    double predict_pair_energy(const Vector& input);
    void forward_pair(const Vector& input);
    void backward_pair(const Vector& input, double grad, std::vector<Matrix>& dw, std::vector<Vector>& db);
    void apply_gradients(const std::vector<Matrix>& dw, const std::vector<Vector>& db);
    void save_weights(const std::string& filename) const;
    void load_weights(const std::string& filename);


    const std::vector<int>& get_layer_sizes() const;

private:
    std::vector<Matrix> weights;
    std::vector<Vector> biases;
    std::vector<Vector> activations;
    std::vector<Vector> zs;
    std::vector<int> layer_sizes;
    double learning_rate;

    double relu(double x);
    double relu_deriv(double x);
};
```

Figure 3: Code snippet of the header file for the `MLP` class.

$$B_j^{(l)} \leftarrow B_j^{(l)} - \eta \cdot \Delta B_j^{(l)}$$

where η is the pre-defined learning rate.

6 Results and Discussion

The model was trained on a dataset of 5,600 conformers for 1,000 epochs using a learning rate of 0.0001. The training loss, measured by Mean Squared Error (MSE), is shown in Fig. 4 and converged to approximately 11. The trained model was then evaluated on a test set of 300 conformers, achieving a Mean Absolute Error (MAE) of 6.67 Hartree. However, performance varied notably depending on the number of heavy atoms in the test set conformers. As illustrated in Fig. 5, conformers with 4 heavy atoms—typically exhibiting true energy values around -210 Hartree—show larger discrepancies between predicted and actual energies. Quantitatively, the MAE for the 4-heavy-atom subset was 11.22 Hartree, whereas the MAE for the subset with 2 or 3 heavy atoms was significantly lower at 4.40 Hartree, indicating the model's reduced accuracy on larger molecules. Nonetheless, the model still significantly outperforms a baseline Lennard-Jones potential, which yields an MAE of 2,039 Hartree on the same test set.

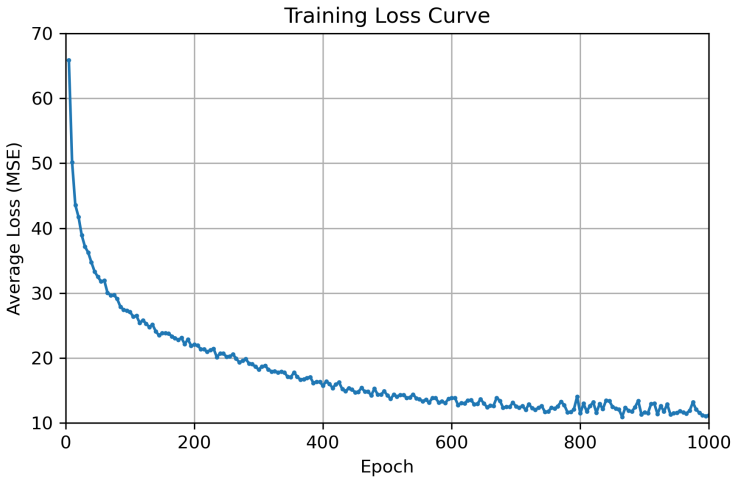


Figure 4: Training loss curve of the MLP model for the 5600 conformers training set.

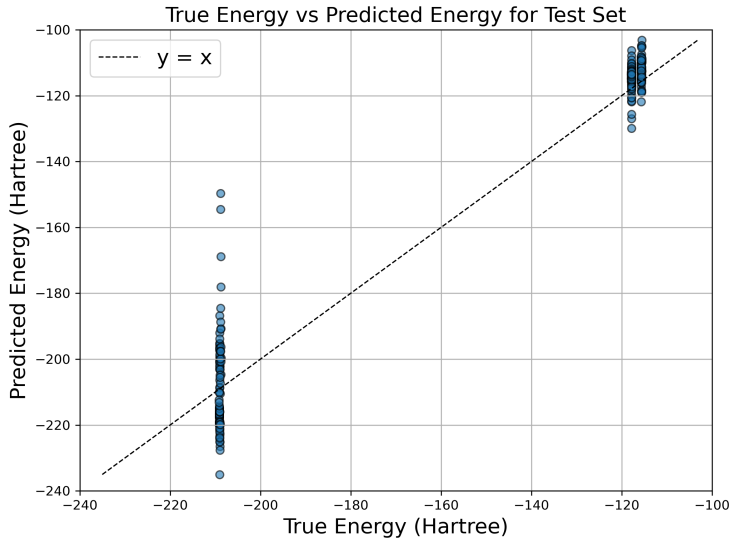
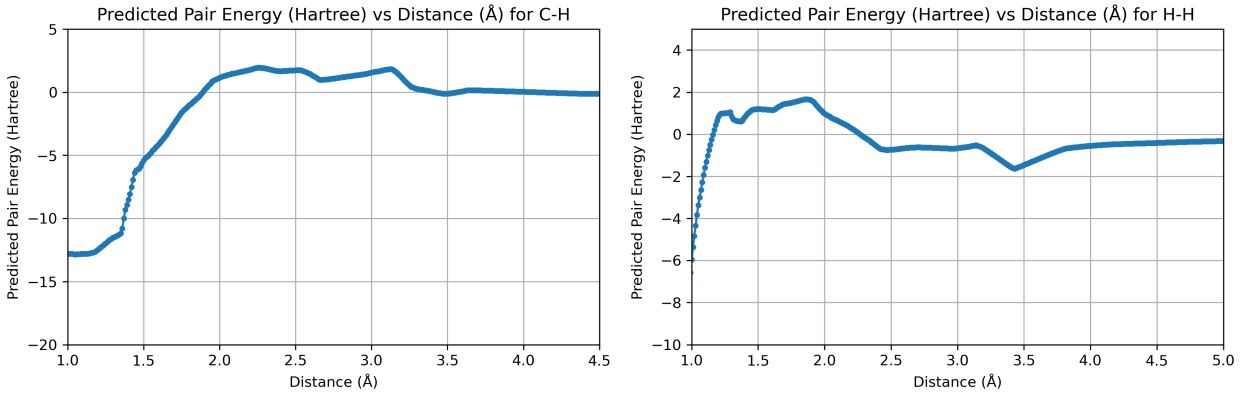


Figure 5: True Energy vs. Predicted Energy scatter plot showing inference results of the trained model on the 300 conformers test set.

To further assess the interpretability of our model, we analyzed its predictions of pairwise atomic interaction energies for C–H and H–H pairs across a range of interatomic distances, as shown in Fig. 6. The results reveal that the model predicts negative energy values near typical bond lengths, indicating that it has learned the stabilizing effect of chemical bonding. Additionally, the predicted interaction energies asymptotically approach zero at larger distances, reflecting the physical reality that atomic interactions diminish with separation. These findings suggest that the model has successfully captured meaningful interatomic interactions, by fitting to a dataset generated from

quantum-level calculations.



(a) (b)

Figure 6: Predicted Pair Energy (Hartree) vs. Distances in Angstrom for (a) C-H pair and (b) H-H pair. The trained model predicts a negative energy value that stabilizes molecule at around bond length, while also showing convergence to 0 Hartree at large distances.

7 Conclusions

This project successfully developed a C++ based multi-layer perceptron capable of predicting molecular conformer energies by learning pairwise atomic potentials directly from the ANI-1 dataset. The model achieved a Mean Absolute Error of 6.67 Hartree on the test set, significantly outperforming a traditional Lennard-Jones potential and demonstrating its ability to learn meaningful chemical interactions, such as bond stabilization and energy decay at larger interatomic distances. While the model's accuracy decreased for larger molecules with more heavy atoms, its interpretable nature and superior performance over classical methods highlight the potential of data-driven approaches for computationally efficient and accurate energy predictions in diverse chemical systems.

8 Future Work

Building upon the current model, several directions for future work could enhance its performance and applicability. Firstly, exploring more sophisticated network architectures, such as incorporating attention mechanisms or graph neural network layers, could better capture complex, long-range atomic interactions. Implementing batch processing and experimenting with advanced optimizers

like Adam or RMSprop could also lead to faster convergence and potentially improved local minima.

Secondly, expanding the training dataset to include a wider variety of molecular structures, sizes, and chemical compositions, possibly from other QM-derived datasets, would be crucial for improving the model's generalization capabilities, especially for larger and more diverse molecules where current performance is limited.

Finally, further developing methods for model interpretability would be beneficial. This could involve techniques like layer-wise relevance propagation or sensitivity analysis to more deeply understand which atomic features and interactions are most influential in the model's energy predictions, thereby increasing trust and providing more detailed chemical insights.

9 References

[1] Folmsbee, D., Hutchison, G. (2020, July 9). Assessing conformer energies using electronic structure and machine learning methods. Wiley Online Library. https://onlinelibrary.wiley.com/doi/10.1002/qua.26381

[2] Durrant, J. D., McCammon, J. A. (2011, October 28). Molecular dynamics simulations and drug discovery. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC3203851/

[3] Smith, J. S., Isayev, O., Roitburg, A. E. (2017, February 7). ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. ScienceDirect. https://www.sciencedirect.com

[4] Smith, J. S., Isayev, O., Roitberg, A. E. (2017, December 19). ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. Scientific Data. https://doi.org/10.1038/sdata.2017

[5] S Smith, J., Isayev, O., Roitberg, A. (2017). ANI-1 data set: 20M DFT energies for non-equilibrium small molecules (Version 1). figshare. https://doi.org/10.6084/m9.figshare.5287732.v1