Machine Learned Conformer Energy Prediction via Approximating Pairwise Potential

Carmen Matar & Curtis Wu

UC Berkeley

Background

Motivation

- Molecular conformer energy estimation is an important topic in material science, drug discovery etc.
- Classic force fields, such as Lennard-Jones (LJ), provide computational speed, but lack accuracy and generalizability to some chemically diverse molecular systems.
- A strictly quantum mechanical (QM) approach, , such as DFT, offers high accuracy, but are computationally expensive especially for large scale simulations.
- An ML approach is an efficient and accurate alternative close to a quantum mechanical approach. However often not explainable.

Goal

Develop a neural force field model in an explainable way to evaluate if a learned force field structure can match or outperform traditional Lennard-Jones potentials in approximating molecular energies.

Objectives

- Develop a neural network model, implemented entirely in C++ using low-level libraries, to predict molecular conformer energies from learned pairwise atomic interactions.

- Train and validate model using ANI-1 dataset, which provides accurate reference energies derived from Density Functional Theory (DFT) with the ωB97x functional and the 6-31G(d) basis set.

- Assess and quantify whether our learned neural force field will match or surpass the accuracy of classical potentials (LJ), while preserving computational efficiency and interpretability.

Data

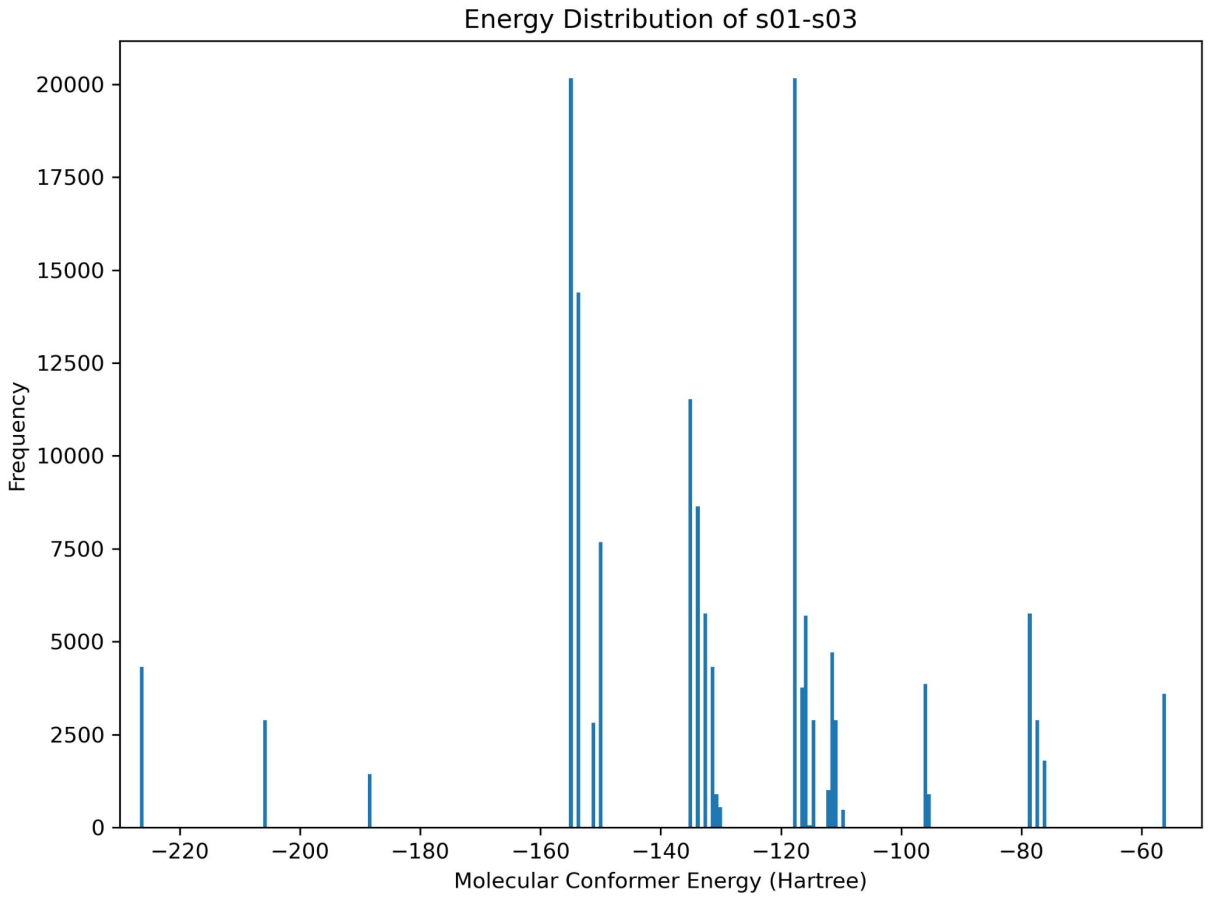
Dataset: ANI-1 (Accurate Neural Network Interatomic Potentials

- DFT calculated molecular energies for diverse small organic molecules.
- Organized into subsets with varying numbers of heavy atoms

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

Data

- Randomly selected 100 conformers for each molecule in s01-s04 ->
- 5900 conformers total
- Randomly selected one unique molecule with different number of heavy atoms as test set
- Training Set: 5600 conformers;
- Test Set: 300 conformers



UC Berkeley

Methodology

Preprocessing Steps

- Convert ANI-1 *.h5* data into CSV for training
- Extracted pairwise atomic data (atomic types, distances) and total conformer-level energies
- Normalized interatomic distances for stable neural network training

Input Representation

- Encode each atomic pair as:
  - One-hot vectors for atom types (C, H, O, N)
  - Normalized interatomic distance

Neural Force Field Model Architecture: MLP {9, 256, 128, 64, 1}

The MLP takes in an atomic pair, outputs an energy

Methodology

Training Procedure

- **Loss Calculation:** Mean Squared Error (MSE) of sum of predicted atomic pair energy and true energy is used as the loss function.
- **Backward Propagation:** Gradients for weights/biases calculated using derived equation.
- Trained by molecule (atom pairs), accumulated gradients are applied using a set learning rate.

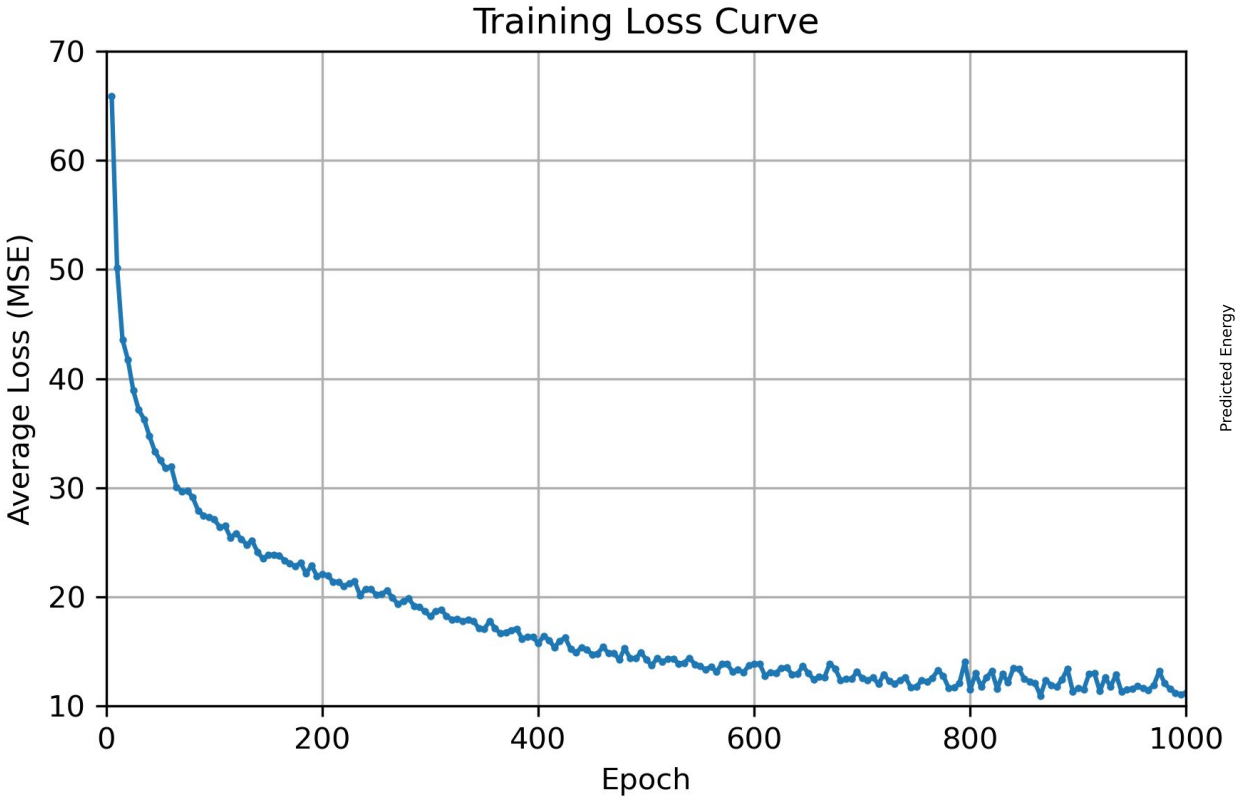
```
#include <iostream>
#include <fstream>
#include <sstream>
#include <vector>
#include <map>
#include <cmath>
#include <algorithm>
#include <random>
#include <set>
#include <string>
#include <iomanip>
```

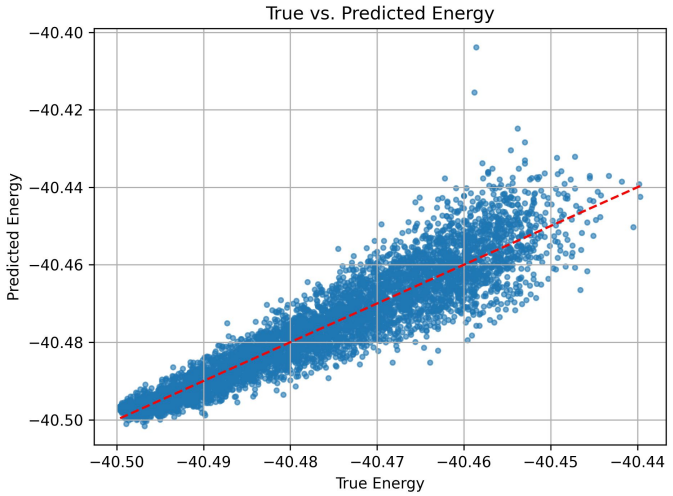
```
class MLP {
public:
    MLP(int input_dim, const std::vector<int>& hidden_layers, int output_dim, double lr);

    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);
};
```


UC Berkeley

Results and Discussion

Trained for 1000 epochs, at 0.0001 learning rate
Test Set (300 conformers) MAE: 6.67 Hartree
LJ MAE: 2039 Hartree



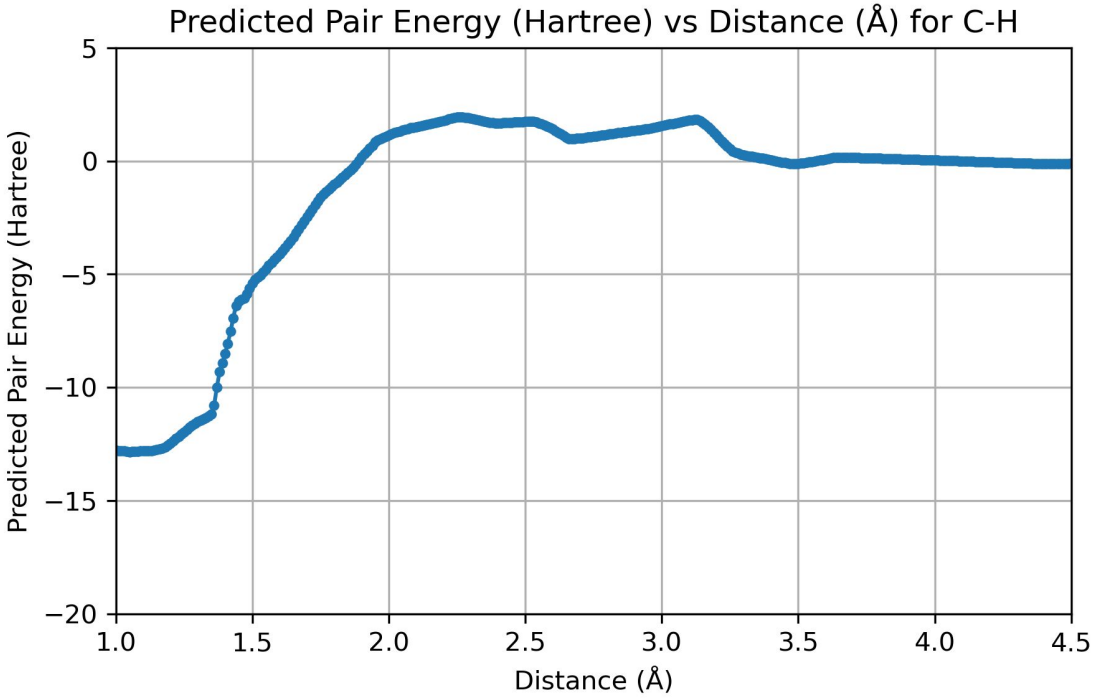


UC Berkeley

Results and Discussion

Observations

- Predicted pairwise energy stabilizes molecule at ~ bond length
- As distance increases, interaction energy converges to around 0
- Model learned pairwise atomic interactions, from the QM total molecular conformer energy target function



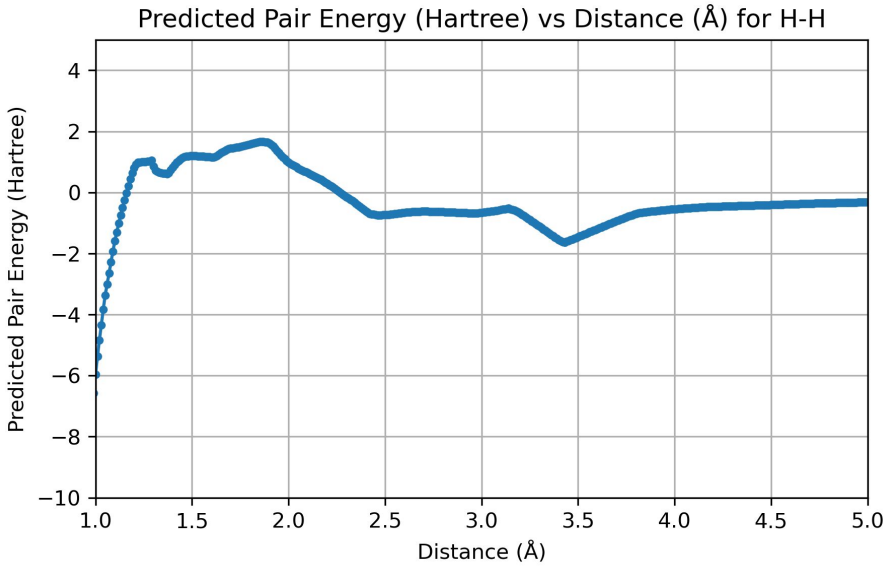
UC Berkeley

Future Work & Project Expansion

More sophisticated network architecture/batch/optimizer

Expansion of the dataset

Network Interpretability



UC Berkeley

Supporting Information & References

- J. S. Smith et al., ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost.
- J. S. Smith et al., ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules.

Thank you for your attention!

Thank you to Dr. Agrawal and Austin for their support throughout the semester!

UC Berkeley