# BIO ENG C142: Final Project Report

Qile Yang

May 2, 2025

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

## 1 Introduction

Understanding and predicting the behavior of molecules and materials at the atomic level is fundamental to advancements across various scientific disciplines, including drug discovery, materials science, and catalysis [1–3]. Accurate computational modeling allows researchers to explore chemical space, predict reaction outcomes, and design novel functional materials without the need for costly and time-consuming physical experiments.

At the heart of highly accurate computational methods lies quantum mechanics. Density Functional Theory (DFT) is a prominent approach that offers a favorable balance between accuracy and computational tractability for many systems [4]. Instead of solving the complex many-body Schrödinger equation directly, DFT cleverly recasts the problem in terms of the electron density, $\rho(\mathbf{r})$. The core idea is that the ground state energy and all other ground state properties are unique functionals of the ground state electron density, $E_0 = E[\rho_0]$. In practice, DFT often relies on solving the Kohn-Sham equations, a set of single-particle equations that yield the electron density of the interacting system.

Despite the successes of DFT, a significant challenge remains in balancing computational accuracy with efficiency. The computational cost of DFT calculations scales cubically, limiting its application to relatively small systems or short simulation timescales [4, 5]. Conversely, classical force fields, which use simplified, em-

pirically parameterized functions to describe interatomic interactions, offer computational efficiency suitable for large-scale simulations (millions of atoms) and long timescales. However, these force fields often lack the necessary accuracy and transferability, especially for systems involving chemical reactions, complex electronic effects, or environments significantly different from those used in their parameterization [6].

To address the computational drawbacks of full DFT calculations, many machine-learning based approaches offer massive speedups. One landmark approach encodes molecules with an Atomic Environment Vector (AEV) representation, developed as part of the ANI framework, specifically the ANI-1 potential described by Smith et al.. [7] The ANI-1 potential and its associated AEVs demonstrated the ability to achieve near-DFT accuracy for predicting molecular energies and forces but at a significantly reduced computational expense, comparable to traditional force fields. The AEVs provide a fixed-size, symmetry and permutation-invariant descriptor of an atom's local chemical environment, making them suitable inputs for simpler neural network models by essentially enforcing a hard prior that the local environment contains all relevant energy information. The approach has been shown to be effective and easily extensible for a wide range of molecular systems, including organic molecules and biomolecules, and has been eimproved for larger systems and applications. [8]

Here, I try to refine the existing ANI-1 potential to improve its performance on a specific class of molecules, namely, small organic molecules. I use the ANI potential as inspiration to adapt different model architectures to a subset of the GDB dataset [9] comprised only of small molecules with the atoms Hydrogens, Oxygens, Carbons, and Nitrogens. The goal is to achieve a more accurate and efficient model for predicting molecular properties, particularly for small organic molecules.

## 2 Methods

I explored several neural network architectures and training strategies to refine the ANI potential for small organic molecules. The primary goal was to enhance the model's ability to predict molecular properties with higher accuracy while maintaining computational efficiency. To achieve this, I experimented with a variety of architectural modifications and training methodologies. All models were trained using the exact same atomic environment vector (AEV) representation implemented with the torchani package. [10] The output of the model was a single scalar value representing the predicted molecular energy.

All architectures were built to be duplicated for use on each atom. The main ones tested included the incorporation of multi-head self-attention layers [11] before feedforward layers to capture complex and long-range interactions in the atomic environment. Residual connections were also evaluated to facilitate gradient flow and mitigate the vanishing gradient problem, which is particularly important for deeper networks. Various activation functions were also evaluated, including ReLU, GELU [12], and CELU, to determine their impact on model performance and convergence speed.

The dataset used for training and evaluation was split into 80% training, 10% validation, and 10% testing to ensure a fair evaluation of the model's performance. To further enhance the ro-

bustness of the evaluation process and mitigate overfitting, I employed K-Fold cross-validation with $k = 3$. This approach allowed the model to be trained and validated on multiple subsets of the data, providing a more comprehensive assessment of its generalization capabilities.

The primary evaluation metric for the models was the Mean Absolute Error (MAE) in kcal/mol, which directly reflects the accuracy of the predicted molecular energies. This metric was chosen because it provides an intuitive measure of the average deviation between predicted and true values, making it suitable for assessing the performance of regression models in this domain.

All training was conducted using the Adam optimizer, which is well-suited for handling sparse gradients and adaptive learning rates. Hyperparameters, including batch size, learning rate, number of epochs, and L2 regularization strength, were manually tuned through an iterative process to optimize the model's performance. Early stopping was employed based on the validation loss to prevent overfitting and ensure that the model did not continue training once its performance on unseen data began to degrade.

To accelerate computation and enable the training of more complex models, all computations were conducted on an NVIDIA RTX3050 laptop GPU through a windows subsystem for linux on the x86-64 architecture with 32GB of RAM. This setup provided sufficient computational power to handle the dataset and the computational demands of training neural networks of this. The final model was selected based on the lowest validation MAE observed during training. Once selected, the model was evaluated on the held-out test set to report its performance, ensuring that the reported results reflect its ability to generalize to unseen data.

# 3 Results

The best model architecture was a set of four identical but independent simple Feedforward networks (one for each species) with a single hidden layer of 128 weights that used a ReLU activation function, making up 197636 learnable parameters in total.

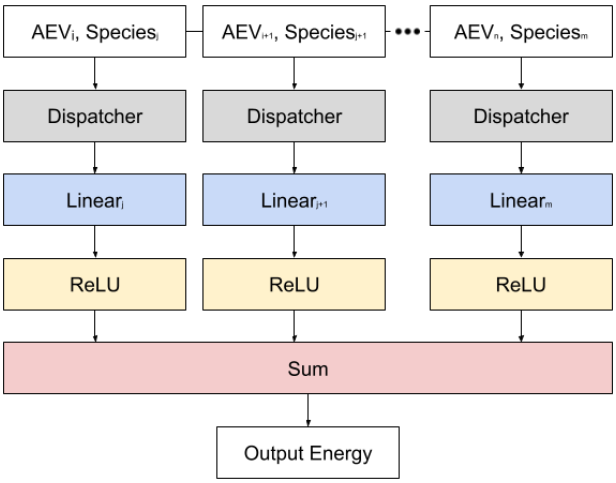


Figure 1: **Model Architecture.** The atomic species and coordinates are transformed into an unordered collection of AEVs and species Labels based on each individual atom in an input molecule. Each input is then "dispatched" based on the Species label to its corresponding feedforward network (one for each species) and an activated scalar value. The values are then summed together to get the final total energy of thge molecule.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa.

Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

## 4 Discussion

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

potential limitations and improvements: not investigated for label imbalance, not investigated for almost-leakage like dataset behaviour, many training adjustments could be made, could use different networks for each atom like they did in the schrod. paper that improved on ani.

## 5 Code Availability

All code, data, and pytorch models used in this project are publicly available on GitHub under the MIT license at `https://github.com/Qile0317/bioe142-final-project`.

# 6 References


1. Wang, Y., Chen, J. & Kang, Z. *In silico protein design promotes the rapid evolution of industrial enzymes* 2018.
2. Dominy, B. N. & Shakhnovich, E. I. Native atom types for knowledge-based potentials: application to binding energy prediction. *Journal of medicinal chemistry* **47,** 4538–4558 (2004).
3. Cicaloni, V., Trezza, A., Pettini, F. & Spiga, O. Applications of in silico methods for design and development of drugs targeting protein-protein interactions. *Current topics in medicinal chemistry* **19,** 534–554 (2019).
4. Engel, E. *Density functional theory* (Springer, 2011).
5. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. *Chemical reviews* **112,** 289–320 (2012).
6. Herbers, C. R., Li, C. & van der Vegt, N. F. Grand challenges in quantum-classical modeling of molecule–surface interactions. *Journal of computational chemistry* **34,** 1177–1188 (2013).
7. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chemical science* **8,** 3192–3203 (2017).
8. Devereux, C., Smith, J. S., Huddleston, K. K., Barros, K., Zubatyuk, R., Isayev, O. & Roitberg, A. E. Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *Journal of chemical theory and computation* **16,** 4192–4202 (2020).
9. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of chemical information and modeling* **52,** 2864–2875 (2012).
10. Gao, X., Ramezanghorbani, F., Isayev, O., Smith, J. S. & Roitberg, A. E. TorchANI: a free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. *Journal of chemical information and modeling* **60,** 3408–3415 (2020).
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
12. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415* (2016).