Use of an ANN to Predict Molecular Energies of Organic Compounds with DFT Accuracy

Introduction:

With the rise of Machine Learning (ML) usage in the Chemical and Biological sciences, both supervised and unsupervised learning methods have begun to play a large role in the analysis of computationally complex problems. Learning models, like Artificial Neural Networks (ANNs) have provided possible alternatives to high cost models such as Hybrid Density Functional Theory (DFT)[2,4] or Quantum Monte Carlo Simulation (QMC)[4,8]. These *ab initio* methods[4], though first-in-class in accuracy, require high computational loads; thus the development of low cost ANNs with similar accuracy has become highly valued.

Given the large number of datapoints available for molecular energies, energy inference through ML serves as a perfect candidate for an ANN. Through development of a layered neural network, and training on the large available datasets, an ML model could learn the patterns and motifs present in similar energy molecular geometries. By using this method, ANNs can achieve similar accuracy to complex models at fractional computational costs.

The largest issue in the implementation of a chemical ML model is the creation of a valid and fully encompassing computational input. In prior works, Behler introduced new formats for the molecular representation of molecules by applying varied symmetry functions to develop new environments[5]. In a more recent expansion on this paper, a new, highly modified version of these symmetry functions has allowed for the training and validation of ANNs to predict potential energies of organic molecules of various sizes[1]. For an ANN, the model is trained on an element's Atomic Environment Vector (AEV) as its input. Applying layers of matrix multiplications and non-linear activation functions to the input, the ANN predicts the contribution of the energy for that atom. This energy contribution is summed over all atoms in the molecule, which is then compared to the ground truth.

Fig. 1:

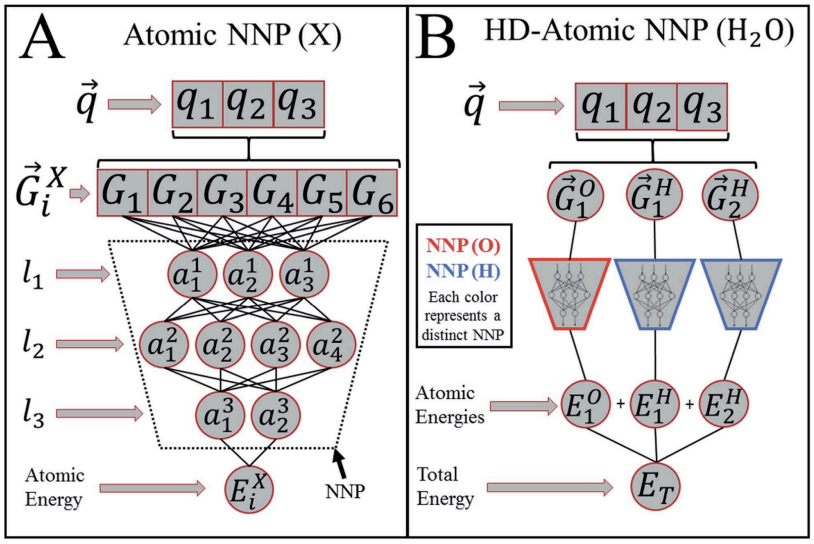

Fig. 1: A: Behler's Original Symmetry Equation Model[5]; B: Smith's New Atomic Environment Vector Model based on Behler's work.[1]

Applying the work Behler has done, I developed a new ANN trained on the ANI-1 dataset to predict organic molecular energies from AEVs. This model was further refined, with deep model architectural adjustments, hyperparameter tuning, and regularization techniques to avoid overfitting.

Methods:

For the training of the ANN, the ANI GDB-11 dataset[10] was used. This dataset includes atomic coordinates and molecular energies for organic molecules made of solely H, C, N, and O atoms with a maximum of 4 heavy atoms per molecule. In order to provide good training for the model while still allowing good conclusions to be made from validation and test data, a 75/10/15 train-val-test split was used.

For the architecture of the ANN, it was decided upon to use a model with 1 residual layer and 3 linear hidden layers for each atom. The residual layer was settled upon as a result of vanishing gradients past 3 hidden layers. For this same reason activation functions were adjusted to Leaky ReLU over traditional ReLU[9]. The addition of deeper layers was unnecessary as creating a high parameter count model would not avoid the high computational cost we are trying to improve on from other quantum computational models.

Fig. 2:

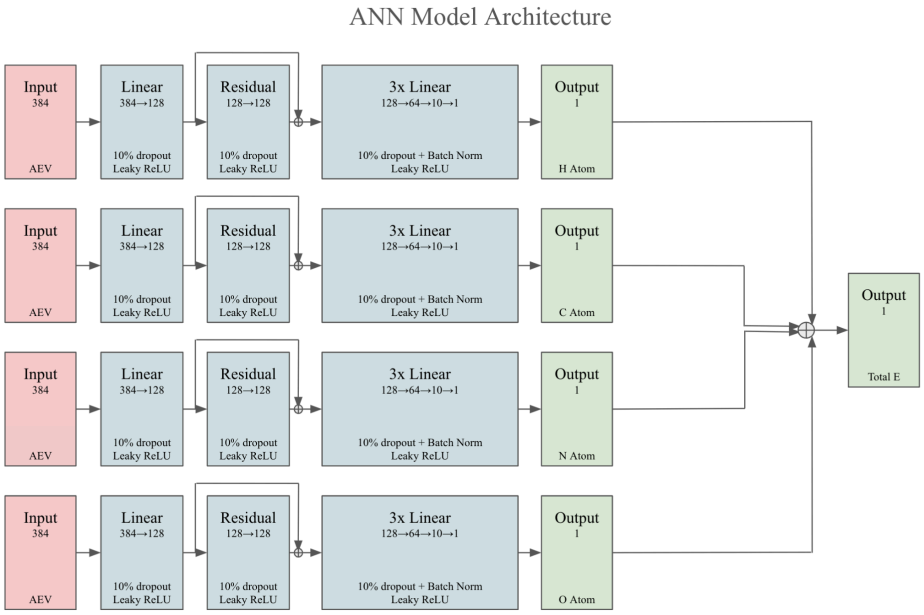

Fig. 2: Each atom's environment descriptor is processed by a shared feedforward network with four hidden layers (128, 128, 64, 10 dimensions) using Leaky ReLU activations. A residual connection links layers 1 and 2. Per-atom energy predictions are summed to yield the total molecular energy.

The model developed, takes in batches of molecules, and feeds input tensors forward through the layers listed in the architecture above. Output tensors are then evaluated against dataset ground truth values and the calculated MSE is back propagated. The Adam optimizer[6] was used to adapt weights for the model each epoch in the training cycle.

Hyperparameter tuning was attempted through an automated Bayesian optimization[3]; however, due to long training times only a few iterations were run, the “converged” hyperparameters were then manually adjusted. Through manual hyperparameter tuning, the optimal model parameters were found to be a learning rate of 1e-4, batch size of 2048 molecules, 200 epochs, and l2 regularization of 1e-5.

Fig. 3:

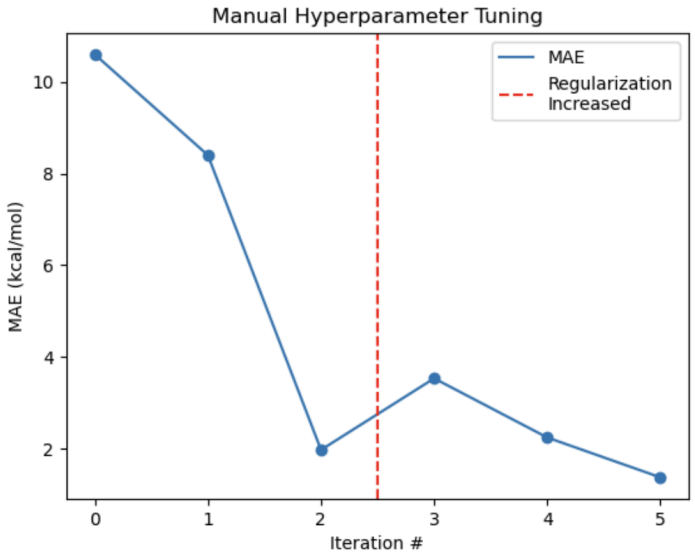

Fig. 3: Iter 0: lr=1e-3, 100 epochs, 4096 batch, l2=1e=-5; Iter 1: lr=1e-3, 100 epochs, 4096 batch, l2=1e=-7; Iter 2: lr=1e-4, 100 epochs, 2048 batch, l2=1e=-5; Iter 3: lr=1e-4, 100 epochs, 2048 batch, l2=1e=-5; Iter 4: lr=1e-4, 150 epochs, 2048 batch, l2=1e=-5; Iter 5: lr=1e-4, 200 epochs, 2048 batch, l2=1e=-5; Note: batch normalization and dropout was added after iteration 2.

Results and Discussion:

Training of the model was positive, resulting in an MSE below 1e-4 for train data and below 1e-5 for validation data across all folds of a 3-fold cross validation. The reason for the lower error in validation data is most likely due to the dropout used during the training mode of the model.

Fig. 4:

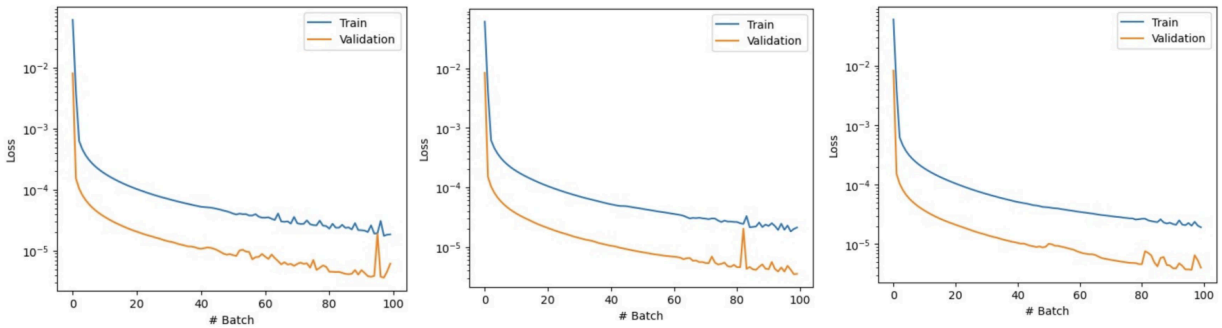

Fig. 4: 3-fold cross validation with consistent training and test loss below 1e-4 MSE. Only 100 epochs used for cross validation due to computational constraints.

Final training over the total 70% train split resulted in an MSE of approximately 1e-5 for train data and 1e-6 for validation data. Using this trained model on the test data split, an MAE of 1.48 kcal/mol and an RMSE of 2.04 kcal/mol was achieved. This is extremely similar to the values achieved in the Smith paper (RMSE: 1.81 kcal/mol)[1].

Fig. 5:

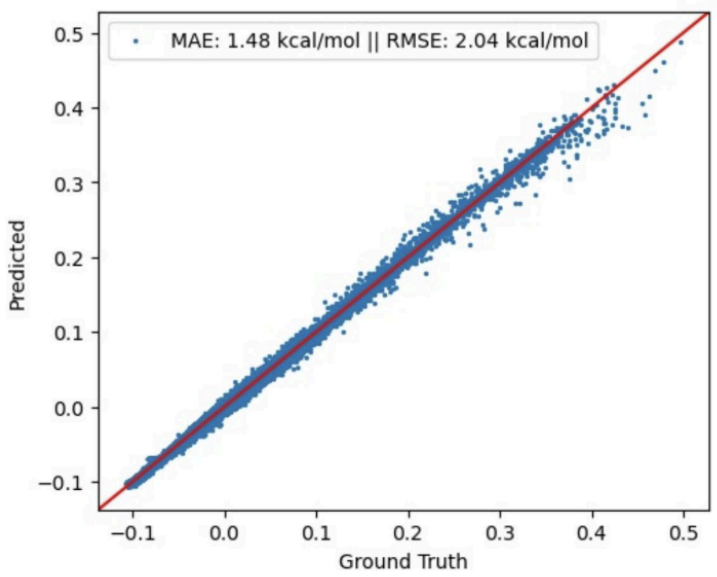

Fig. 5: Predicted energy values vs. Ground truth plot. MAE: 1.48 kcal/mol, RMSE: 2.04 kcal/mol.

Though the error achieved on the test dataset is slightly worse than the Smith results, the ANN developed in this experiment was not able to fully converge to its lowest error. Computational capabilities limited the training to 200 epochs; however, the training error curves seem to show that error has not yet fully converged. Thus, if additional time or computational power was available, a test error lower than Smith's value could be realistically achieved.

Though this error was not achieved yet, an acceptable prediction error is observed. The 1.48 kcal/mol error is in the same range as quantum chemical models like DFT and much better than the error of classical force fields[4] with nearly the same computational burden. Overall, this ANN provides a low cost alternative to DFT with room for improvement over models like Smith's ANI–1 model[1]. And with larger datasets and slight architecture adjustments, it seems possible to extend this model to more complex organic structures incorporating more atom types (S, P, B, I, Br, Cl, F, etc.) and higher amounts of heavy atoms.

References:

(1) Smith, Justin S., et al. “The Ani-1ccx and Ani-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules.” *Nature News*, Nature Publishing Group, 1 May 2020, www.nature.com/articles/s41597-020-0473-z.

(2) *Density-Functional Theory of Atoms and Molecules | Oxford Academic, academic.oup.com/book/41995. Accessed 1 May 2025.*

(3) *Frazier, Peter I. “A Tutorial on Bayesian Optimization.” arXiv.Org, 8 July 2018, arxiv.org/abs/1807.02811.*

(4) *Friesner, Richard A. “Ab Initio Quantum Chemistry: Methodology and Applications.” Proceedings of the National Academy of Sciences of the United States of America, U.S. National Library of Medicine, 10 May 2005, www.ncbi.nlm.nih.gov/pmc/articles/PMC1100737.*

(5) *Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces | Phys. Rev. Lett., journals.aps.org/prl/abstract/10.1103/PhysRevLett.98.146401. Accessed 2 May 2025.*

(6) *Kingma, Diederik P., and Jimmy Ba. “Adam: A Method for Stochastic Optimization.” arXiv.Org, 30 Jan. 2017, arxiv.org/abs/1412.6980.*

(7) *Low-Cost Autonomous Perceptron Neural Network Inspired by Quantum Computation | AIP Conference Proceedings | AIP Publishing, pubs.aip.org/aip/acp/article-abstract/1905/1/020005/1020896/Low-cost-autonomous-perceptron-neural-network?redirectedFrom=fulltext. Accessed 2 May 2025.*

(8) *Monte Carlo Methods in Ab Initio Quantum Chemistry | World Scientific Lecture and Course Notes in Chemistry, www.worldscientific.com/worldscibooks/10.1142/1170. Accessed 1 May 2025.*

(9) *Olamendy, Juan C. “Understanding Relu, Leakyrelu, and Prelu: A Comprehensive Guide.” Medium, Medium, 4 Dec. 2023, medium.com/@juanc.olamendy/understanding-relu-leakyrelu-and-prelu-a-comprehensive-guide-20f2775d3d64.*

(10) *Smith, Justin S., et al. “The Ani-1ccx and Ani-1x Data Sets, Coupled-Cluster and Density Functional Theory Properties for Molecules.” Nature News, Nature Publishing Group, 1 May 2020, www.nature.com/articles/s41597-020-0473-z.*