Evaluating the Performance of a Universal Electron Force Field

Austyn Masuno[1], Marivi Fernández-Serra[23], Anthony Mannino[23], Jose Soler[4]

[1]Claremont McKenna College, Claremont, CA, 91711, USA
[2]Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY, 11794, USA
[3]Institute for Advanced Computational Science, Stony Brook University, Stony Brook, NY 11794, USA
[4]Departamento e Instituto de Física de la Materia Condensada, Universidad Autónoma de Madrid, E-28049 Madrid, Spain

Motivation

The time evolution of many-body systems can be described using Molecular Dynamics (MD) simulations, which are split into two principal approaches. *Classical force fields* are suitable for large systems with long time scales, but do not consider electronic structure and the parametrization can be a complex and arduous process. *Ab initio molecular dynamics (AIMD)* simulations consider electronic structure, but are computationally demanding and limited to smaller systems with shorter time scales. Su and Goddard proposed an intermediate approach called *electron force field* (*EFF*) [1], combining classical force fields with electronic structure which was able to solve systems too large for AIMD simulations. Ultimately, they faced challenges predicting equilibrium geometries for simple molecules but nonetheless provided valuable insights towards the development of new EFFs.

Methods

Mol2vec through Uniform Manifold Approximation Projection (UMAP)

- A corpus is generated from 20M molecules from the ZINC database to train the Mol2vec model [2]. The model learns embeddings for each unique Morgan identifier which are 'words' (molecules are sentences). UMAP is applied to the transformed data.

CC(=O)OC1=CC=CC=C1C(=O)O
↓
['2246728737', '3545365497', ..., '3217380708']

EFF vs DFT

- The input of the EFF is the atomic geometries and species, and a simplified description of electronic structure called *e-balls*.

$$E_{tot} = E_k + E_x + E_c + E_{ee} + E_{ne} + E_{nn}$$

$$E_k = \sum_i \frac{c_1}{w_i^2} + \sum_{i<j} \frac{c_2}{w_{ij}^2} \exp\left[-c_3\left(\frac{r_{ij}}{w_{ij}} - 1\right) - c_4 \log^2\left(\frac{w_i}{w_j}\right)\right]$$

$$E_x = -\sum_{i<j} \frac{8c_5}{\sqrt{2\pi}} \delta_{\sigma_i\sigma_j} \frac{w_i^2 w_j^2}{w_{ij}^5} \exp\left[-\frac{r_{ij}^2}{2w_{ij}^2}\right]$$

$$E_c = -\sum_{i<j} c_6\left(1 - \delta_{\sigma_i\sigma_j}\right)\left[2\pi\left(w_{ij}^2 + w_{min}^2\right)\right]^{-3/2} \exp\left[-\frac{r_{ij}^2}{2w_{ij}^2}\right]$$

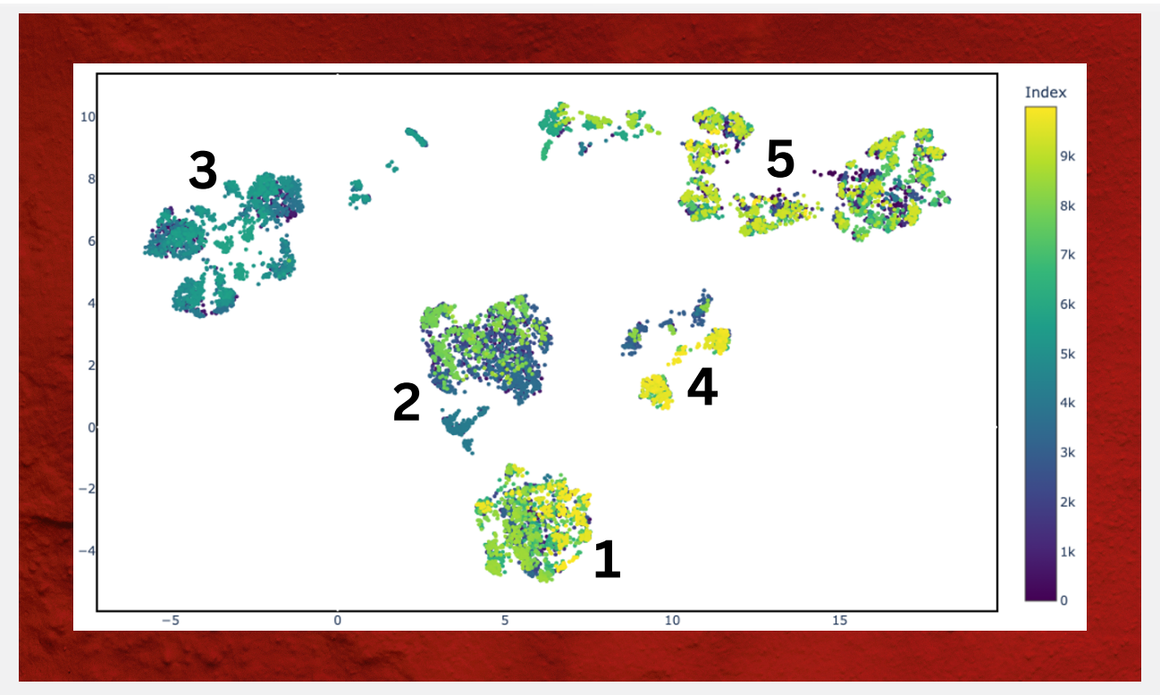
$$E_{ee} = \sum_{i<j} \frac{1}{r_{ij}} \mathrm{erf}\left[\frac{r_{ij}}{w_{ij}}\right]$$

$$E_{ne} = -\sum_{i,J} \frac{Z_J}{r_{iJ}} \mathrm{erf}\left[\frac{r_{iJ}}{w_i}\right]$$

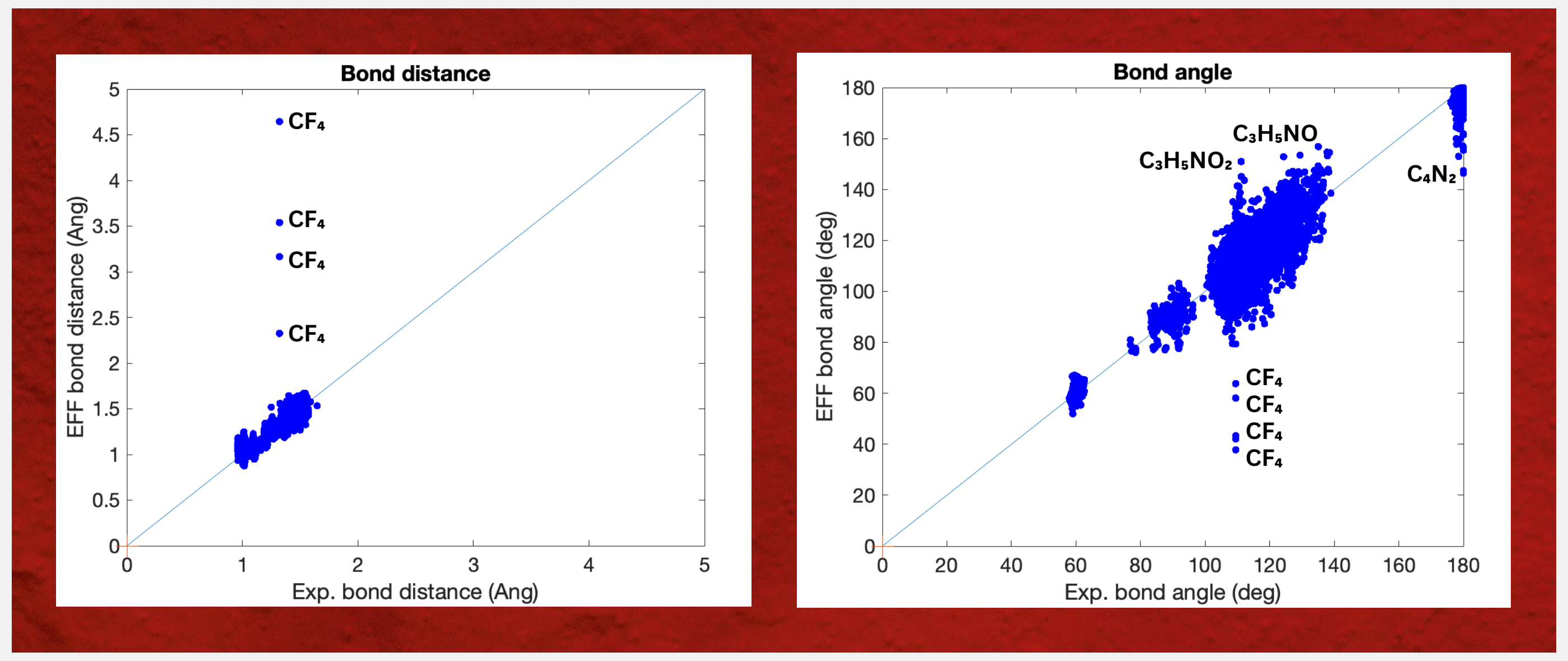
$$E_{nn} = \sum_{I<J} \frac{Z_I Z_J}{r_{IJ}}$$

- DFT calculations were performed with an interface between SIESTA, an efficient electronic structure code , and Atomic Simulation Environment (ASE), a set of tools and Python modules for atomistic simulations.

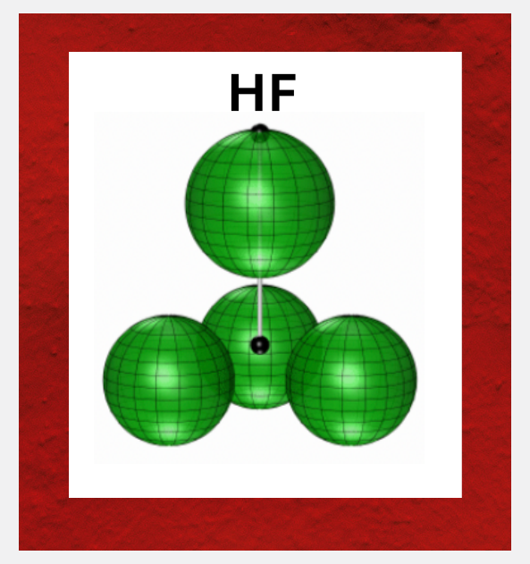
Results



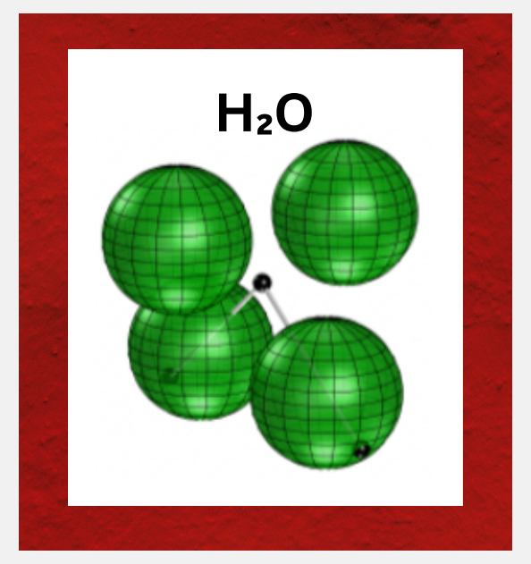
UMAP visualization for the first 10,000 molecules in the QM9 dataset. UMAP is a dimensionality reduction technique that is effective in preserving the local structure of data in a lower-dimensional space.

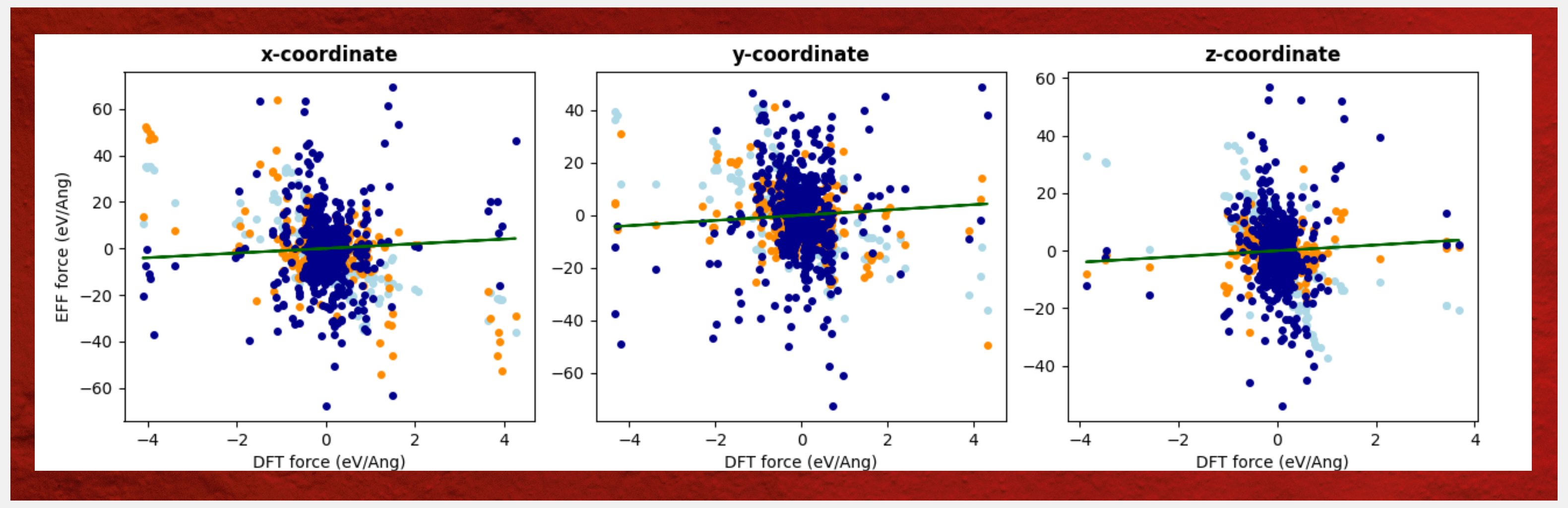


Comparison of bond distances (left) and angles (right) calculated with a version of the EFF with 2120 parameters to experimental values.

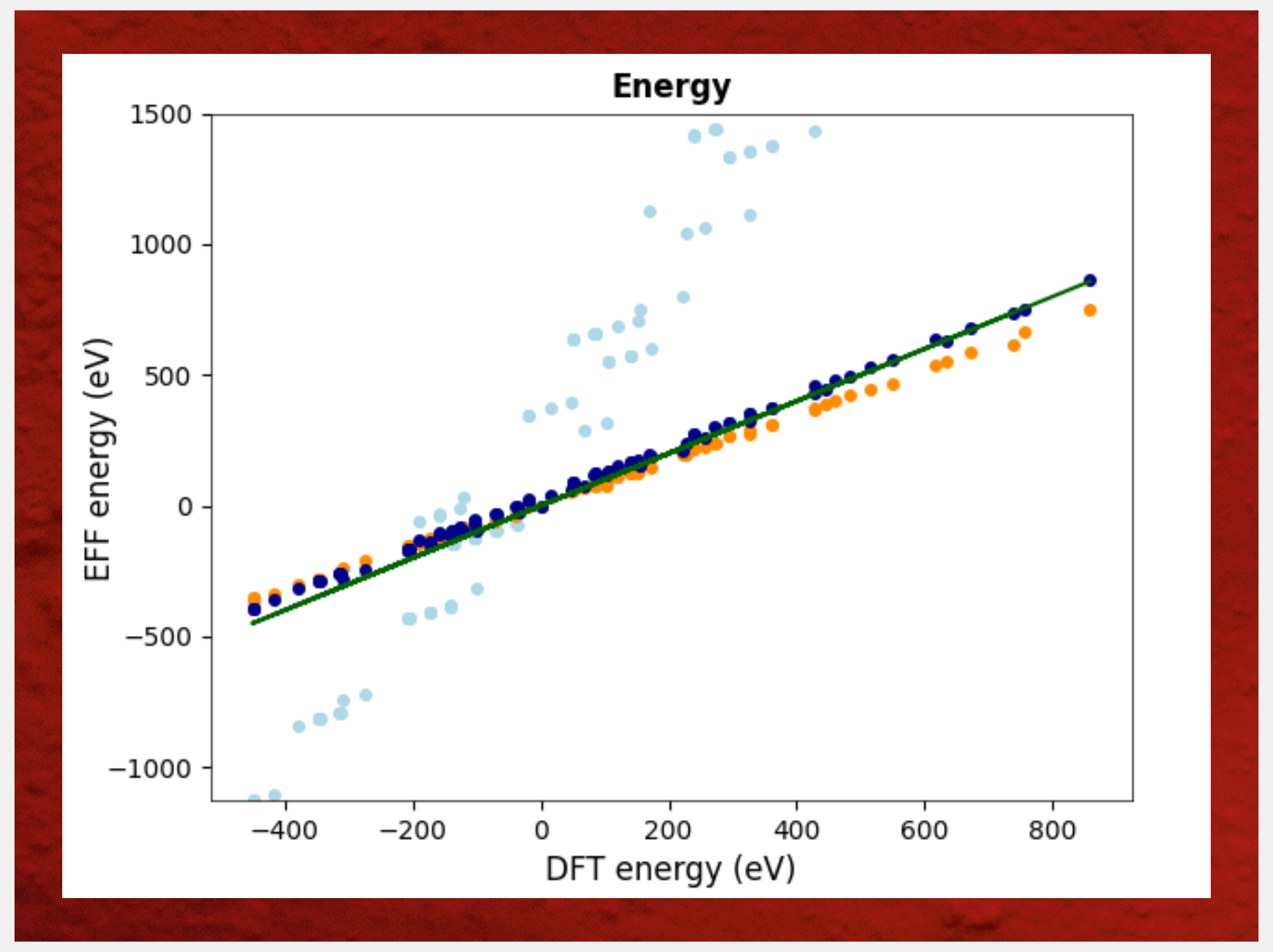


Geometries and e-ball configurations for a linear molecule (HF) and a nonlinear molecule (H_2O). Each green ball is a coincident spin up-down e-ball pair.





Comparison of forces (top) and energies (bottom) calculated with multiple versions of our EFF to those calculated with DFT.



# of parameters	
●	18
●	3708
●	2120

Challenges

1. UMAP analysis output distinct clusters of molecules, but no meaningful difference in the performance of the EFF was found based on which of these subsets the model was trained on.
2. All EFFs seem to give forces that are not only much larger than, but uncorrelated with what DFT predicts. Even with the addition of over over 2000 fitting parameters, the EFF is still not producing accurate forces for most molecules.
3. Comparing the forces by species, we find that hydrogen (H) and carbon (C) atoms perform better than oxygen (O), nitrogen (N), and flourine (F) atoms. Atoms with higher polarizability are more prone to electron density fluctuations induced by neighboring atoms and thus experience stronger van der Waals forces. The current version of the EFF doesn't incorporate this effect, which could be one explanation why the O, N, and F atoms are predicted less accurately.

Conclusion

We have developed an EFF that demonstrates moderate accuracy in reproducing experimental results, especially concerning equilibrium geometries and energies for various molecules. According to UMAP analysis, varying the data the model was trained on did not have a significant effect. However, a substantial improvement in performance was achieved by increasing the number of parameters. Despite these enhancements, the EFF still predicts forces on atoms that are too large for most molecules.

Future Work

1. It could be meaningful to compare UMAP results to other dimensionality reduction techniques such as PCA.
2. The current version of the EFF lacks accurate polarizability effects. An extension of this project could involve implementing a simplified version of van der Waals interactions.
3. Our current model only considers pairwise interactions, however considering many-body effects could be important for bonding in certain materials.

References + Acknowledgements

[1] J. Su & W. Goddard. Excited electron dynamics modeling of warm dense matter. Phys. Rev. Lett. 2007, 99, 185003
[2] S. Jaeger, S. Fulle, S. Turk. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. J. Chem. Inf. Model. 2018, 58, 27-35

This research was supported by the Stony Brook University Physics REU program, which is funded by the National Science Foundation under Grant PHY-2243856.