EXPANDING THE EXTRAPOLATION LIMITS OF NEURAL NETWORK FORCE FIELDS USING PHYSICS-BASED DATA AUGMENTATION
WORKSHOP SUBMISSIONS

Anonymous authors
Workshop paper under double-blind review

ABSTRACT

Even though machine learning force fields are quite accurate in the prediction of forces and energies in the sampled region, they fail to extrapolate, which results in the unphysical behavior of the system during molecular dynamics simulations. We propose to overcome this problem by performing data augmentation. To expand the original dataset random perturbations of atoms were performed. The corresponding increase in the energy of the system was calculated under the assumption of harmonicity. The required spring constants were obtained from the original dataset by fitting a gaussian mixture model to the bond lengths distribution. The resulting force field performance was improved in the regions far from training data.

1 INTRODUCTION

Molecular dynamics (MD) simulation is a powerful tool to study systems' physical and chemical properties involving millions of atoms. Force field prediction is an essential part of MD simulations. There are different ways one can predict forces acting on each atom of the system that vary in their computational complexity and the resulting accuracy.

Machine learning-based force-fields (ML-FFs) fill the gap between computationally efficient but less accurate classical force fields and computationally expensive *ab initio* calculations, which resolve quantum molecular effects. The inputs to ML-FFs are the atomic positions and types, with outputs being the forces acting on each atom and the total energy of the system. The double and single bonds, used in classical FFs, are only models of reality and do not capture quantum effects. ML-FFs thus seek to efficiently model the enormous range of chemical possibilities arising from *ab initio* calculations, obviating preconceived notions and knowledge of fixed bonds. Being a universal function approximation, only training data limits ML-FFs (Unke et al., 2021).

Data augmentation has been employed successfully in a vast array of image classification problems, summarized in (Shorten & Khoshgoftaar). In essence, training images can have a variety of manipulations employed but retain the same label. More recent work has explored data augmentation in natural language processing (Shorten et al.), where linguistic constraints can be applied to reduce the overfitting effect. We seek to continue this paradigm by using physics-informed data augmentation to enhance ML-FFs.

Even though ML-FFs are quite accurate in the prediction of forces and energies in the sampled region, they fail to extrapolate (Fig. 1). This results in unphysical behavior of the system during MD simulations. As we can see in Fig. 1 there appears an unphysical energy barrier for the SchNet model followed by a constant region compared to *ab initio* reference that has a steady increase before going to the constant region. Besides that, when the distance between two atoms goes to zero, the energy should go to infinity. However, for the ML-FF, the energy can actually experience a decrease in that region. We want to solve the preceding problems by explicitly incorporating into the model the increase in energy while moving away from the sampled data. By preventing the energy decrease, we hope to avoid going over the barrier and the atoms colliding with each other.

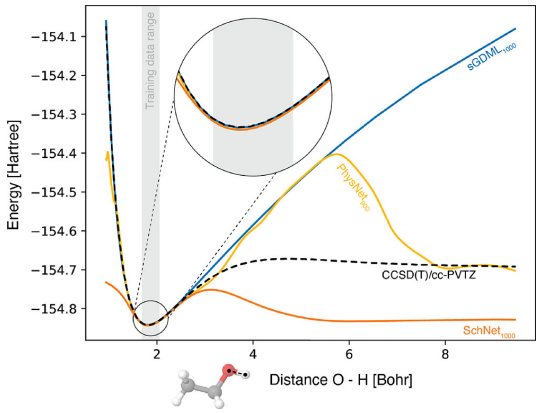

Figure 1: One-dimensional cut through the potential energy surface of ethanol along the O–H bond distance for different ML-FFs (solid blue, yellow, and orange lines) compared to *ab initio* reference data (dashed black line). The typical training region is highlighted grayUnke et al. (2021).

2 METHODS

2.1 SCHNET MODEL

We employ the SchNet architecture (Schütt et al.) for ML-FFs, which encodes physical constraints such as requiring smooth energy and force prediction as atom locations vary, and invariance in atom indexing and translation. Specifically, the forces and energies are respectively equivariant and invariant under rotations.

We developed a private fork of (Axelrod et al.), which is an implementation of SchNet in the PyTorch (Paszke et al., 2019) framework. Our fork includes streamlined routines for the generation of augmented data, and we plan to incorporate our functionality at some point in the future. The hyperparameters used for our network are provided in the appendix.

Our data came from a 500 kelvin simulation of an ethanol molecule using *ab initio* molecular dynamics, meaning that nuclei trajectories were integrated over time using force fields calculated by solving for the electronic wavefunction and nuclei charges. The data is open and may be found in the git repository (Axelrod et al.).

2.2 DATA AUGMENTATION

In order to incorporate the energy increase into the SchNet model, we augment the original data set by including global statistical information into individual data points. The input data sets to train SchNet-like models almost always come from *ab initio* molecular dynamics simulations (Unke et al., 2021), and our technique relies on this fact. Without having the training data sampled from an *ab initio* molecular dynamics calculation, our assumptions are incorrect, so attention must be paid to the source of the training data.

Because the input training data to SchNet is a bonded molecule in equilibrium, the atoms tend to cluster in energy troughs which are approximately governed by quadratic behavior for small variations from the equilibrium position. From thermodynamics we know that if the energy trough is characterized as $E(\mathbf{r}) = E_0 + \frac{1}{2}k||\mathbf{r}||^2$, then the distribution of time-averaged atom locations follows a Gaussian distribution with probability density function proportional to $\exp\left(-\frac{k}{2k_BT}||\mathbf{r}||^2\right)$, where k is known as a spring constant in kcal mol^{-1} $\mathring{A}^{-2}$, $||\mathbf{r}||$ is the change in distance between the original equilibrium and new positions of the atoms in $\mathring{A}$, k_B is Boltzmann's constant in kcal mol^{-1} K^{-1}, and T is the temperature in K. The spring constant in this equation will be the basis of our data augmentation technique, as for each unique pair of atom types, we will be able to estimate this spring constant by fitting the data (e.g. to a Gaussian) for which we can empirically estimate

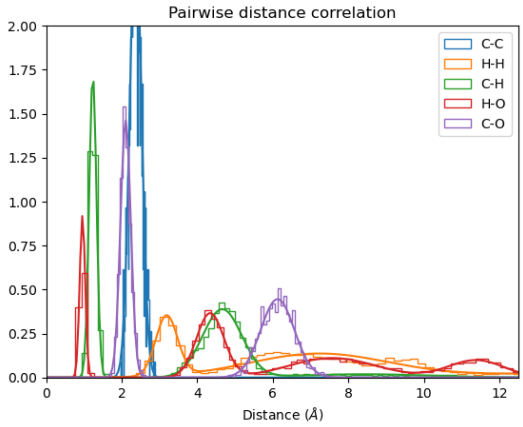

Figure 2: Pairwise distance correlations for the 500 kelvin ethanol molecule data set.

its parameters and thus find an estimate for k, since T is fixed. This simple model serves as our conceptual basis.

However, in a larger system, atoms interact with multiple neighbors, so a simple model with few parameters (e.g. a regular Gaussian) does not necessarily characterize the system. For these atoms in larger molecules, a model with more parameters (e.g. a Gaussian mixture model) will be required to accurately fit the data.

We introduce the methodology for which we estimate the values of k, which first begins by preprocessing a set of data for which the spring constants can be estimated. To preprocess a data set, we loop over each point and compute pairwise distances from each atom to all other types of atoms. Careful attention must be paid to avoid double-counting. From there, a model for the pairwise distance correlation can be developed, where our spring constant will depend on not only the types of atoms in each pair but also the distance between the atoms.

Our exploration of these data sets has revealed that Gaussian mixture models (GMMs) model multibody interactions very well. Indeed, Fig. 2 illustrates this: GMMs match the histograms on a data set consisting of 1000 configurations of an ethanol molecule. The selection of the number of Gaussians in each mixture is a hyperparameter for our data augmentation technique, which may be important for modeling atom species with many different spring constants dependent on their distance.

We produced an augmented data set by including new data where the positions of the original atoms are perturbed by a small random vector $||\Delta\mathbf{r_i}|| \in (0, \epsilon]$ over n randomly chosen atoms in the geometry for $1 \leq i \leq n$. The perturbation vector is different for each atom, noting that both ϵ and n are hyperparameters of the model. However, since the original data set is large and duplicating the entire data set is expensive, we randomly choose l geometries for which to do the augmentation. Note that l is also a hyperparameter of the model.

The corresponding change in energy is then approximated by the spring constant belonging to the nearest Gaussian in terms of interatomic spacing. This allows the global statistical behavior of the molecule to be reflected in individual augmented data points. The change in the energy gradient due to the spring constant is similarly evaluated as $k\Delta\mathbf{r_i}$.

3 Results

3.1 Spring Constants

Fig. 2 depicts the fit of the Gaussian mixture models to each of the pairwise distance correlations shown as histograms. By our reasoning, the inverse variances are equal to $k/(k_BT)$., where the data was generated at 500 K, and $k_B = 1.987 \cdot 10^{-3}$ kcal/mol. In order to determine which value of k to use for a pair of atoms, we interpolate by selecting the nearest spring constant. This is representative of the mean and variance of each component of the GMM, without weights.

3.2 IMPROVEMENT OF POTENTIAL ENERGY SURFACE FAR FROM SAMPLING REGION

For this experiment, we used 200 geometries for augmentation in addition to 1000 from the original data. From there, we vary the maximum displacement position of the perturbed atoms across the values $\{0.01, 0.03, 0.05, 0.07, 0.09, 0.11, 0.2, 0.5\}\mathring{A}$. After the model was fitted for each set of data points, its physical behavior was confirmed via the energy conservation during molecular dynamics runs in the NVE ensemble. To see if our model performance was improved we have calculated the potential curve with respect to O-H distance (Fig. 3). We can see that the *ab initio* reference calculation follows a steep potential energy gradient far from the atoms' equilibrium positions. The SchNet model, when trained on raw data, fails to capture that behavior. The figure shows that at the cost of slight biases in the equilibrium potential energy value, the potential energy prediction far from the trough can be greatly improved. There is little correlation between the value of maximum displacement and the introduced bias as well as the increase of potential energy far from equilibrium. Therefore, the expected improvement of the potential curve is subject to hyperparameter optimization. Overall, because of the increase in global accuracy, we expect our model will provide better results when employed in molecular dynamics simulations leveraging force fields calculated by our model.

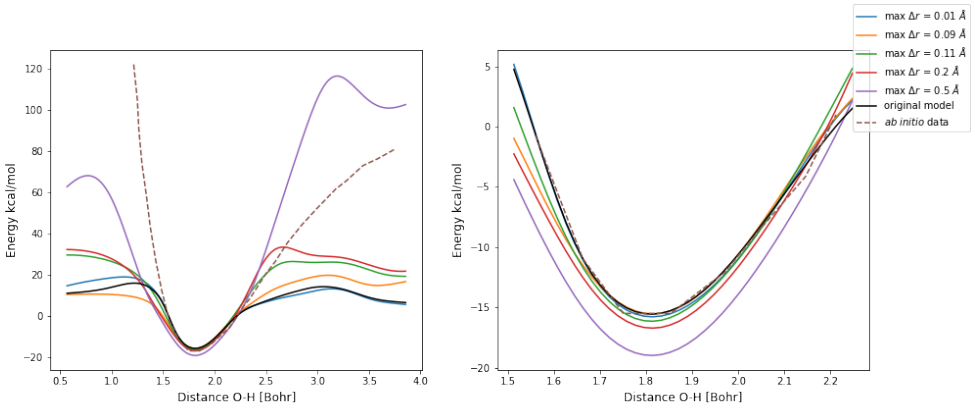

Figure 3: Potential energy curves generated from augmented datasets for a few values of our augmentation technique's hyperparameters.

4 DISCUSSION

Our employed interpolation technique for the spring constants uses the analogy to a single Boltzmann distribution, so we employed the nearest interpolation scheme as a function of distance for each type of atom pair. Future work could investigate linear interpolation schemes or other techniques taking advantage of the structure of the Gaussian mixture model in that local spring constants could be defined as an average weighted by each Gaussian. We suspect the latter approach would provide better results due to faster transitions from one spring constant to another.

One weakness of our approach is that training data is assumed to always be at or near equilibrium. As the temperature rises, the training data will include more points farther from equilibrium. In this case, the energy is just as likely to decrease as it is to increase when moving from the original point. Our approach, however, always assumes that an increase in potential energy is encountered. Thus, the current approach can be improved by incorporating information on the energy gradient of the augmented geometry. For instance, if the gradient aligns with the displacement, we decrement energy and vice versa.

REFERENCES

Simon Axelrod, Daniel Schwalbe-Koda, Wang Wujie, Ang Shijun, Rafael Gomez-Bombarelli, and Ryan Solaski. Neuralforcefield. URL https://www.github.com/

learningmatter-mit/NeuralForceField.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Kristof T. Schütt, Farhad Arbabzadah, Stefan Chmiela, Klaus R. Müller, and Alexandre Tkatchenko. Quantum-chemical insights from deep tensor neural networks. 8(1):13890. ISSN 2041-1723. doi: 10.1038/ncomms13890. URL https://www.nature.com/articles/ncomms13890.

Connor Shorten and Taghi M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. 6(1):60. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://doi.org/10.1186/s40537-019-0197-0.

Connor Shorten, Taghi M. Khoshgoftaar, and Borko Furht. Text Data Augmentation for Deep Learning. 8(1):101. ISSN 2196-1115. doi: 10.1186/s40537-021-00492-0. URL https://doi.org/10.1186/s40537-021-00492-0.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111. URL https://doi.org/10.1021/acs.chemrev.0c01111. PMID: 33705118.

A APPENDIX

Table 1: Hyperparameters for SchNet model trained on 500K ethanol dataset.

Parameter	Value
n_atom_basis	256
n_filters	256
n_gaussians	32
cutoff	5.0
trainable_gauss	true
dropout_rate	0.2

Table 2: Calculated GMM spring constants for 500 K ethanol data set.

Atom Pair	**Distance** ($\mathring{A}$)	k (**kcal mol**$^{-1}$ $\mathring{A}^{-2}$)
H-H	1.79	129
	2.48	27.7
	3.00	25.0
	3.69	14.0
H-C	1.11	533
	2.16	64.5
	2.91	14.2
H-O	0.97	737
	2.08	125
	2.76	85.6
	3.38	23.5
C-C	1.54	388
C-O	1.45	455
	2.47	122