PyITA: A Taylor Expansion-Based Data Augmentation Program for ANN Potentials Applied to TiO_2

Zhengbo Xiang
American School in Japan; Tokyo, Japan
Email: 21xiangz@asij.ac.jp

Abstract – The efficiency of Artificial Neural Network (ANN) potentials enables the modeling of materials at scales that are too computationally expensive for conventional first-principles approaches. However, the force and energy-prediction accuracy of ANNs are generally limited by the availability of training data and training hours. The enlistment of more efficient training methods can partially mitigate this limitation. In this paper, I demonstrate the capabilities of my new Python program, PyITA, which executes a recently demonstrated Taylor expansion-based data augmentation technique[11]. Using my program, I was able to evaluate the powerful methodology by constructing and comparing ANN potentials for the chemical species titania (TiO_2). I compared the error distributions of the augmented potentials with that of the non-augmented potentials. Potentials were trained on both a large (7815 structures) and a small (500 structures) dataset. I ultimately found insufficient evidence to confirm that the data augmentation method is effective for increasing the accuracy of either force predictions or energy predictions on either dataset.

Key Words – Materials Informatics, Artificial Neural Networks, Machine Learning

INTRODUCTION

I. Background

Atomistic simulations are crucial to an informatics approach to materials science. Thanks to their extraordinary accuracy, *first principles*-based quantum mechanical modelling is often used to yield potentials that are crucial to the study of material characteristics[1,2]. For instance, Density Functional Theory (DFT) calculations rely on a very accurate approximation of the many-body Schrodinger equations[3]. DFT has been applied to the study of superconductivity[4], magnetism[5], and phase-change materials[6]. Owing to its flexibility and accuracy, it is a powerful tool.

However, even DFT can be too computationally expensive and impractical for the study of certain complex materials and interfaces[7]. Increasingly, ANNs trained on reliable data (such as DFT calculations) are used to predict material properties instead[7,8,9,10]. Critically, the difficulty of ANN potential training scales linearly with the number of atoms in the training set. In contrast, the difficulty of DFT calculations often scales cubically, quartically, or even quintically with structure size. As such, ANN potentials can be applied to systems that are too large to be practically modelled by DFT calculations.

Despite their advantages over first principles-based methods, the training of sufficiently accurate ANN potentials is often hampered by the limited availability of sufficiently large datasets. Data augmentation allows us to enhance limited training datasets, improving the data-efficiency/effectiveness of ANN training models. Advances in data augmentation techniques will reduce the research cost of future studies in computational materials science.

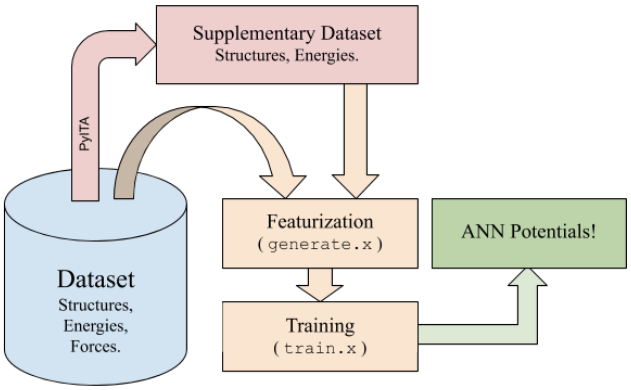

FIGURE 1: This flowchart demonstrates how ANN potentials are constructed with *PyITA*. *PyITA* neither modifies nor replaces preexisting featurization and training methods. Instead, it creates additional input data from the original dataset by modifying copies of pre-existing structures. The supplementary dataset contains all data generated by *PyITA*. In this paper, any reference to an "enhanced", "extended", or "augmented" dataset refers to the set of all structures in the relevant parent and supplementary datasets (ie. the original dataset combined with the supplementary dataset).

II. *PyITA & Aenet*

Aenet[7] is an open source software package developed by Dr. Artrith and Dr. Urban. In this paper, Aenet was used to perform dataset featurization, conduct neural network training, and produce energy/force predictions.

Aenet contains three notable configuration files, conventionally named `generate.in` (for featurization), `train.in` (for training), and `predict.in` (for prediction). `generate.in` contains the locations (/path/to/file) of every structure file in the training dataset. The training dataset is a collection of XCrySDen-format (`.xsf`) structure files that state the atomic symbols and Cartesian positional coordinates of constituent atoms, the forces acting on each constituent atom (in the +x, +y, and +z directions respectively), and the total structural energies.

Cooper *et al.* demonstrated the incorporation of force information in training via a Taylor expansion as an incredibly new and promising method for increasing the effectiveness of ANN training[11]. The Taylor expansion itself allows us to computationally-efficiently estimate approximate energies of new structures. I independently implemented this method in a now open-source Python program: *PyITA* (Python Implementation of Taylor-expansion for Aenet)[12]. *PyITA* functions as a pre-featurization external data augmentation program (Figure 1).

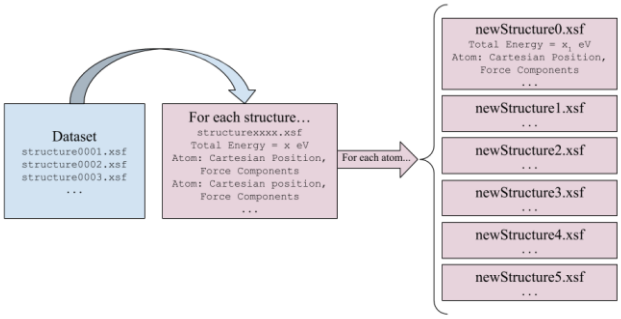

FIGURE 2: Flowchart demonstrating *PyITA*'s creation of additional structures. Six additional structures are created for each atom in each structure in the original dataset until the supplementary dataset reaches a sufficient size (based on user-configured *a*-value).

Force information from structures in the original dataset are used to generate additional approximate structure data (coordinates and energies) near the positions of existing structures. For each atom in any parent structure, *PyITA* can create six additional approximate structures (Figure 2).

After prompting the user for an augmentation-factor *a* (also referred to as the *a*-value in this paper) and a user-configurable displacement constant δ, *PyITA* reads every existing structure (`.xsf`) file in the provided training (parent) dataset into a list. A new structure is created by displacing any atom in any parent structure by δ in one of six directions (+x, +y, +z, -x, -y, -z). *PyITA* does so iteratively, starting from the first atom of the first existing structure and stopping when the number of created structures is equal to the product of *a* and the number of structures in the original dataset. The new structures are written as `.xsf` files in a user-designated directory, effectively forming the supplementary dataset. The file locations of all structures in the supplementary dataset are written to a new `generate.in` file. The original `generate.in` file remains untouched so that additional *PyITA* structures based on the same parent configuration file can be easily created.

Crucially, based on the first-order Taylor expansion demonstrated in Equations 10 and 11 of the *Cooper et al.* paper[11], we approximate the energy of each new structure as the energy of its parent structure subtracted by the product of δ and the component force experienced by the original structure in the direction of the displacement. Algorithmically, atomic displacement in any negative direction is treated as a displacement of equal magnitude in the corresponding positive direction such that $\delta = -\delta_0$.

As this method is essentially a form of data augmentation, I refer to datasets enhanced by *PyITA* as "augmented" in my paper.

III. *Approach & Promise*

Using *PyITA*, I test the feasibility of the Taylor expansion approach for the efficient training of ANN potentials on TiO_2. The study of TiO_2 is especially valuable because it is a powerful model material with applications in optics, nanotubes research, and photocatalysis. Recently, Xiang *et al.*[12] utilized TiO_2 to demonstrate the application of ANN potentials to the study of the flexoelectric effect.[13] In this paper, I utilize a training dataset containing 7815 TiO_2 structures. Dr. Artrith and Dr. Urban graciously provided this dataset to the public[7].

I hope to aid future research in the field of materials informatics by evaluating a new and potentially more effective method for the training of TiO_2 potentials.

METHODS

Initially, I trained five ANN potentials on a dataset consisting of 500 structures randomly chosen from a 7815-structure dataset. I also trained five additional ANN potentials on a *PyITA*-augmented version of the same 500-structure dataset (δ=0.03 and *a*=6). I assigned Trial Numbers (one through five) to the ANN potentials in each category. I trained each ANN potential to 15 iterations, randomly choosing 10% of each dataset as an independent test set. I configured the ANN architecture to include two hidden layers per atomic species, each with 15 nodes. I used the Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (LM-BFGS)[14] training method, as implemented in Aenet.

I then created two additional potentials. I trained one potential on the entirety of the 7815-structure TiO_2 dataset and another potential on a *PyITA*-augmented version of the full 7815-structure dataset (δ=0.03 and *a*=4). This time, I selected an augmentation factor lower than the earlier *a*=6

value due to computational constraints. However, I preserved the neural network architecture and training configurations of the previously constructed 500-structure potentials.

Finally, the most crucial outcome of the Cooper *et al.* Taylor expansion method is an improvement in the accuracy of force predictions. To test this, I created ten additional ANN potentials: Five trained on the previously discussed 500-structure dataset, and another five trained on an augmented version of said 500-structure dataset. Again, I held the neural network architecture constant.

RESULTS AND DISCUSSION

I computed the relative energy errors for each of the original 500-structure potentials based on data from the independent test sets. I calculated the median relative energy error for the five non-augmented ANN potentials to be approximately 3.27 meV, while the median relative energy error for the five augmented ANN potentials was approximately 4.22 meV. Further, the range of errors on the augmented potential was greater than the range of errors on the non-augmented potential (Figure 3). It appears that the augmentation technique as performed using the current approach failed to improve energy predictions based on the 500-structure datasets.

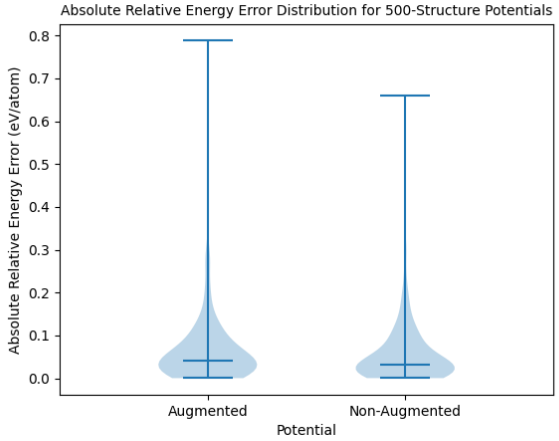

FIGURE 3: The violin plot represents a combination of all five respective potentials of the augmented and non-augmented datasets, as the distributions were computed with the combined validation data (across trials) for each dataset (*e.g.* The "augmented" distribution is constructed from the absolute relative energy errors of predictions by all five potentials based on the augmented dataset). The absolute errors in relative energy were computed with the independent test set. They are defined as the absolute values of the differences between the relative (per-atom) energies of the reference structure (computed by DFT) and the ANN-computed relative energies.

I initially hypothesized that since 500-structure datasets are unusually small for ANN potentials characterizing materials as complex as TiO_2, there may simply not have been a sufficiently diverse set of training data for the potential to yield consistently accurate predictions, regardless of the application of data augmentation techniques. If the structures represented in the training dataset are very similar, the neural network may overfit.

However, augmentation remains apparently ineffective in the 7815-structure potentials. The enhanced 7815-structure potential had a higher median and mean absolute relative energy error than its non-augmented counterpart. However, in contrast with our findings for the 500-structure potentials, the augmented 7815-structure potential yielded absolute relative energy errors with a much smaller range than the non-augmented 7815-structure potential (Figure 4). If future research identifies a decrease in the range of relative energy errors as a consistent feature of our data augmentation technique, *PyITA* could help achieve more robust extrapolation than what is possible with energy-only training.

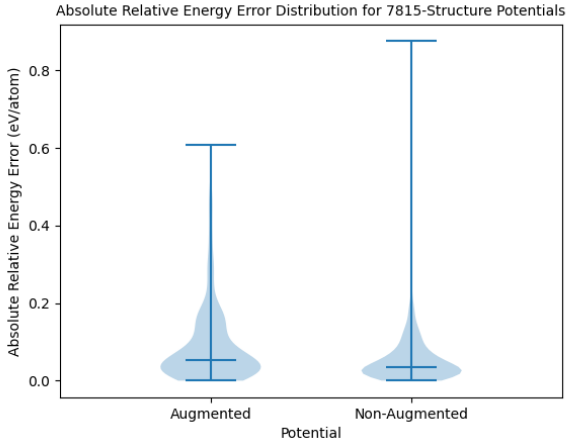

FIGURE 4: This plot demonstrates the distributions of absolute errors in relative energy for the respective augmented and non-augmented potentials trained on the full 7815-structure dataset (Left: Augmented, Right: Non-Augmented). The median absolute error in relative energy for the augmented potential was 5.4 meV, which is noticeably higher than the non-augmented potential's median absolute error of 3.5 meV. However, the augmented potential's error distribution has a noticeably smaller range.

I observed that distribution-features across trials varied greatly for both augmented and non-augmented 500-structure potentials. This variability may also exist in 7815-structure potentials. More trials (potentially with a larger iteration count and a larger independent test set) must be performed to concretely determine the impact of our augmentation technique on the distribution of absolute relative energy errors.

Finally, I calculated the frequency of errors in absolute force for each atom in each structure of the independent test set by subtracting the predicted interatomic forces from the reference (DFT-calculated) interatomic forces in each Cartesian direction. The distribution of absolute errors in force predictions are compared in Figure 5.

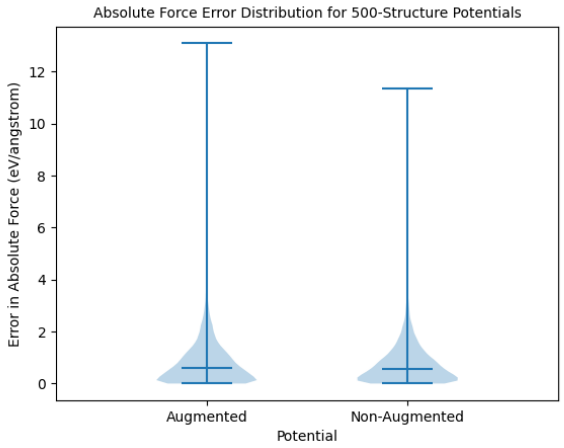

FIGURE 5: This plot demonstrates the distributions of absolute errors in force predictions for the augmented and non-augmented 7815-structure potentials. The median absolute error in force predictions for the augmented potential was marginally higher than that of the non-augmented potential (0.597 eV/Angstrom vs. 0.574 eV/Angstrom). The range of the errors in the augmented potential was also noticeably larger.

CONCLUSION AND FUTURE STUDIES

I have not found sufficient evidence to demonstrate that the Taylor expansion-based data augmentation technique is effective in increasing the accuracy of either force prediction or energy prediction in TiO_2. This does not, however, mean that the technique is not suitable for the study of TiO_2.

After I thoroughly tested my error calculation methodology and closely reviewed the outputs of my error calculation software, I suspect that the problem could be explained by the large variations in accuracy I observed across trials. In other words, due to the constraints presented by my limited access to computational resources, the per-trial variations in the distributions of both energy and force errors were large enough to make it difficult to draw meaningful conclusions.

I strongly suspect that the specific neural network architecture used in my study—especially due to the relatively low number of iterations/epochs—prevented the augmentation technique from being effective. Training to a greater number of iterations leads to better-fitting models, which could result in changes to the size of the training dataset being much more impactful.

It is also possible that the relatively low *a*-values used in my calculations have played a factor. With more time and computational resources, I would be able to determine an attainable "sweet spot": a computationally achievable configuration which noticeably improves my potentials' predictive capabilities.

Finally, my findings may hint to a limitation in the Cooper *et al.* methodology. That is, the first order Taylor expansion-based data augmentation strategy might consistently fail for LM-BFGS ANN potentials that are not trained to a sufficient number of iterations. It is possible that preserving the Cooper *et al.* Taylor expansion to a greater order would be appropriate when training resources are limited, as it would reduce the signal-to-noise ratio of the supplementary dataset. However, this modification might also greatly increase the computational intensiveness of the data augmentation process. Further study would be necessary to confirm this hypothesis.

Materials informatics is a field in its infancy. I eagerly anticipate exciting future research on data augmentation for the development of ANN-potentials.

AKNOWLEDGEMENTS

I would like to extend special thanks to my mentor Dr. Hieu Nguyen for providing me with guidance and affirmation throughout the duration of my research. Further, I wish to especially acknowledge my friend Kylie for helping me debug *PyITA* and its accompanying back-end scripts through many sleepless nights. I express my deepest gratitude to the Aenet team for providing the public with excellent, well-documented, and free software which will continue to greatly accelerate the development of the field of materials informatics. Finally, I appreciate the help I received from Dr. Nongnuch Artrith and the Aenet community through the Aenet mailing list.

REFERENCES

[1] Ward, L., & Wolverton, C. (2017). Atomistic calculations and materials informatics: A review. Current Opinion in Solid State and Materials Science, 21(3), 167–176.

[2] Pokluda, J., Černý, M., Šob, M., & Umeno, Y. (2015). Ab initio calculations of mechanical properties: Methods and applications. Progress in Materials Science, 73, 127–158.

[3] Kryachko, E. S., & Ludeña, E. V. (2014). Density functional theory: Foundations reviewed. Physics Reports, 544(2), 123–239.

[4] Uzunok, H. Y., Zafer, T., Tütüncü, H. M., Karaca, E., Bağcı, S., & Srivastava, G. P. (2020). Probing physical properties and superconductivity of noncentrosymmetric superconductors LaPtGe and LaPtGe3: A first-principles study. Computational Materials Science, 185, 109949.

[5] Pandey, S., Demaske, B., Ejegbavwo, O. A., Berseneva, A. A., Setyawan, W., Shustova, N., & Phillpot, S. R. (2020). Electronic structures and magnetism of Zr-, Th-, and U-based metal-organic frameworks (MOFs) by density functional theory. Computational Materials Science, 184, 109903.

[6] Zhang, W., Deringer, V. L., Dronskowski, R., Mazzarello, R., Ma, E., & Wuttig, M. (2015). Density-functional theory guided advances in phase-change materials and memories. MRS Bulletin / Materials Research Society, 40(10), 856–869.

[7] Artrith, N., & Urban, A. (2016). An implementation of artificial neural-network potentials for atomistic materials simulations: Performance for TiO2. Computational Materials Science, 114, 135–150.

[8] Artrith, N., & Kolpak, A. M. (2014). Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. Nano Letters, 14(5), 2670–2676.

[9] Lacivita, V., Artrith, N., & Ceder, G. (2018). Structural and Compositional Factors That Control the Li-Ion Conductivity in LiPON Electrolytes. Chemistry of Materials: A Publication of the American Chemical Society, 30(20), 7077–7090.

[10] Sosso, G. C., Miceli, G., Caravati, S., Behler, J., & Bernasconi, M. (2012). Neural network interatomic potential for the phase change material GeTe. Physical Review. B, Condensed Matter, 85(17), 174103.

[11] Cooper, A. M., Kästner, J., Urban, A., & Artrith, N. (2020). Efficient training of ANN potentials by including atomic forces via Taylor expansion and application to water and a transition-metal oxide. Npj Computational Materials, 6(1), 54.

[12] Xiang, Z. (2020). pyITA (Version 0.1) [Computer software]. Github. https://github.com/alanatransrights/pyITA

[13] Xiang, L., Zeng, X., Huang, X., & Li, G. (2020). The application of artificial neural-network potentials for flexoelectricity: Performance for anatase-type TiO2. Physics Letters. A, 384(10), 126217.

[14] Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal of Scientific Computing*, *16*(5), 1190–1208.