Using Supervised Learning to Calculate the Surface Energy of Water Dimers

Genwei Zhang and Delano Usiukiewicz
Department of Computer Science, University of Oklahoma, Norman, Oklahoma
Email: genweizhang@ou.edu, delanou@ou.edu

Abstract

The task of computing the surface energy of a molecule is very time intensive if using conventional quantum chemistry simulation methods. A typical simulation problem involves calculating potential energy of a water dimer. Using supervised learning and the distances between the atoms in the molecule we aim to accurately reproduce water-dimer potential energy. By implementing this approach, we expect advancement in problem-solving capabilities in the field of quantum chemistry.

Introduction

In molecular quantum mechanics, a frequent task is simulating the potential energy surface (PES) of a molecule, an essential research area that can be used to explore some properties of molecular structures. Possible applications include: discovering the minimum energy shape of a molecule or calculating the chemical reaction rates. In this project our goal is, using the supervised learning (SL), to train a PES-Predictive model that can predict the potential energy of a water-dimer as accurately as the classical quantum simulation functions. The establishment of such a model can facilitate calculating water-dimer PES more efficiently when compared with conventional methods.

Our dataset consisted of 10,000 different molecular configurations where each instance consists of 15 different atomic distances and an output of potential energy.

To determine which representation would work best for this application we used prebuilt supervised-learning packages in R: Radom Forest, and Neural Net. After finding the optimal parameters for each representation, we found the Neural Net to result in a smaller RMSD on our testing dataset.

Once the representation was chosen, we implemented our own neural net and performed feature selection. To reduce the "curse of dimensionality" we were able to eliminate the intra-molecular distances between atoms and only retained the inter-molecular distances.

Related Work

In the work reported by Morawietz *et al*., this team used a full dimensional neural network for predicting water dimer potential energies based on environment-dependent DFT energies, atomic forces, and charges. They successfully demonstrated that the neural network could simulate or predict the MD trajectories with the accuracy of essentially the same quality of those trajectories generated from AIMD simulations. We initially planned to reproduce their work using similar neural network methods. However, we think there are two places that can be further improved. First, the training data in this work are: DFT energies, forces and charges, but since they are all functions of atom positions, we could only use atom coordinates as the training input. This way we can save a lot time on preparing the input data and theoretically obtain similar prediction accuracy. Additionally, this paper mentioned the many-body function, but for water-dimer we think two-body function might be better.

Methods and Results

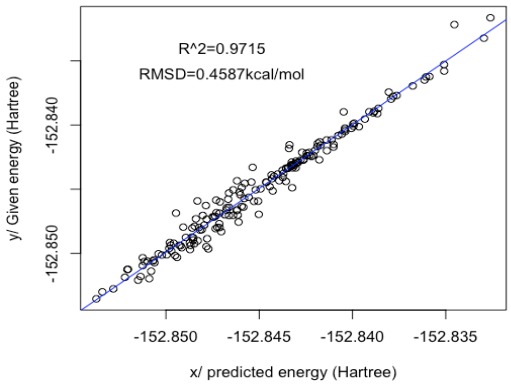

Figure 1. Random forest result

When initially comparing pre-existing supervised learning packages in R we compared random forest to neural net for the RMSD value and plotted the results. The first graph is from the Random Forest while the second is from the Neural Net. As evident in the graph, the lower RMSD value comes from the Neural Net.

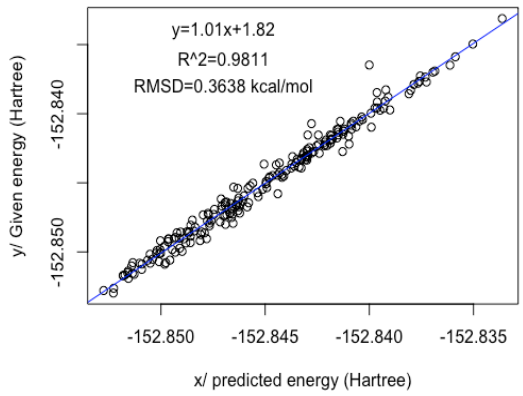

Figure 2. Neural net result

After deciding upon the supervised learning algorithm, the next step was to reduce the dimensions and perform

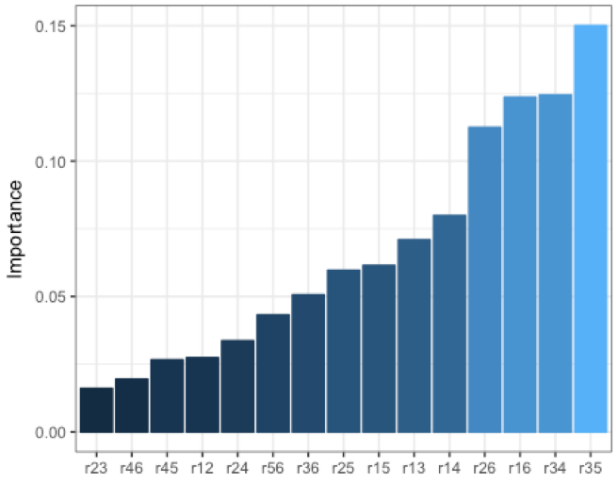

Figure 3. Importance of variables

feature extraction. The results are given in figure 3. After setting a threshold of 5% we were able to eliminate the intramolecular distances between atoms.

The last step was to compare the performance of our custom neural net using different number of neurons in the hidden layer as well as different number of iterations. We found that a 1 million iterations using 50 neurons yielded a lower value at 0.225 kcal/mol. When moving to 10 million steps there was no significant different between the two with 50 neurons yielding 0.194 kcal/mol and 10 being 0.001 kcal/mol higher.

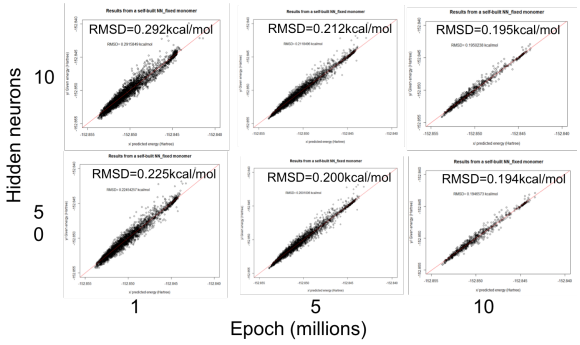

Figure 4. Custom nerual net results

Analysis and Future Work

Neural networks work better than random forest on reproducing water dimer potential energy. We developed a neural network script that can be used to train predictive models with larger amount of input data (~10k) than pre-existing neural net packages in R. Supervised Learning can accurately reproduce water dimer potential energy with RMSD < 1 kcal/mol.

In the future we would: perform bagging of all trained models to improve the prediction performance, run even longer iterations to check and solve the over-fitting issues, and apply stacking to further make the prediction more accurate through combining distinct types of methods (i.e. NN, regression, RF, SVM).

References

Handley, C.M., G.I. Hawe, D.B. Kell, and P.L. Popelier, Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning. Phys. Chem. Chem. Phys., 2009. 11(30): p. 6365-76.

Morawietz, T., V. Sharma, and J. Behler, A neural network potential-energy surface for the water dimer based on environment-dependent atomic energies and charges. J. Chem. Phys., 2012. 136(6): p. 064103.

Gillan, M.J., D. Alfe, A.P. Bartok, and G. Csanyi, First-principles energetics of water clusters and ice: a many-body analysis. J. Chem. Phys., 2013. 139(24): p. 244504.

Yao, K., J.E. Herr, and J. Parkhill, The many-body expansion combined with neural networks. J. Chem. Phys., 2017. 146(1): p. 014106.

Che, D., Q. Liu, K. Rasheed, and X. Tao, Decision tree and ensemble learning algorithms with their applications in bioinformatics. Adv. Exp. Med. Biol., 2011. 696: p. 191-9.

https://www.kaggle.com/arathee2/random-forest-vs-xgboost-vs-deep-neural-network.