UNIVERSITY OF MIAMI
DEPARTMENT of COMPUTER SCIENCE

Development of an Enhanced Machine Learning Based QM/MM Method

Courtney Sever[1,2], Caroline Velez[3], Xiang Zhong[3], Orlando Acevedo[3]
[1]Department of Computer Science, Computing for Structure REU, [2]Florida State University, Tallahassee, Florida, 32304, United States
[3]Department of Chemistry, University of Miami, Coral Gables, Florida 33146, United States

Introduction

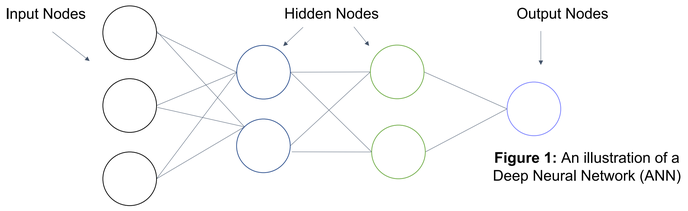


Figure 1: An illustration of a Deep Neural Network (ANN)

Mixed quantum mechanical and molecular mechanical (QM/MM) simulations coupled to free energy perturbation theory and Monte Carlo sampling (FEP/MC) is a highly accurate method for computing free energy profiles of chemical reactions in solution. However, these methods often rely on lower level or semiempirical QM methods due to their associated high computational costs, which may introduce error. Machine learning (ML) and the use of Artificial Neural Networks (ANN) for computing the energies of molecular structures provide the ability to retain high-level QM accuracy at or below the cost of semiempirical methods. ANNs require training of large data sets of chemical reaction structures representing the path along the free energy profile, e.g., reactants, transition states, and products. This research effort developed a systematic method for deriving the structures, computing the energies at the desired QM theory level, and updating the BOSS-Gaussian software package to incorporate AENET machine learning. As a test of the methodology, the Henry reaction between nitromethane and formaldehyde was computed in water using the enhanced ML-QM/MM methodology. The reaction was computed at the MP2/6-31+G(d) theory level for the solute and utilized the TIP4P water model for bulk solvent. Hundreds of thousands of structures were generated and trained in the ANN using our automated method. To compute the total free energy profile for the making/breaking C-C bond of the Henry reaction spanning a distance of 1.50–3.50 Å required 6-7 million QM calculations.

What is the Henry Reaction?

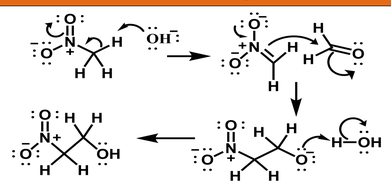


Figure 2: Mechanism of Henry Reaction

Discovered in 1895, the Henry reaction is an extremely useful reaction in the formation of carbon-carbon bonds. More specifically, the reaction is a combination of a nitroalkane and either an aldehyde or ketone in the presence of a base to form β-nitro alcohols. The main drawback of using this reaction is the low selectivity and unwanted side products that come with it. In recent years, it has been of great interest in finding selective methods for synthesis, as well as finding ways to reduce the environmental impact. One way to reduce the impact on the environment is to perform the reaction in water, however a small set of data supports the Henry reaction occurring in water.

Overall Problem and Objective

Molecular behavior is best modeled using quantum mechanics(QM). However, QM calculations are computationally costly, and thus cannot be applied to thousands of molecules at once. Therefore alternative methods such as the quantum mechanical and molecular mechanical (QM/MM) method, as well as the free energy perturbation theory and Monte Carlo sampling (FEP/MC) can be utilized. The purpose of this project is to train the ANNs well enough that they are able to produce an activation energy, $\Delta G^{\ddagger}$, with little error. Once this is completed, this reaction can be run in other solvents than water.

$$i\hbar\frac{\partial}{\partial t}\Psi = H\Psi$$

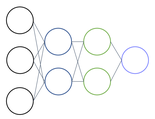
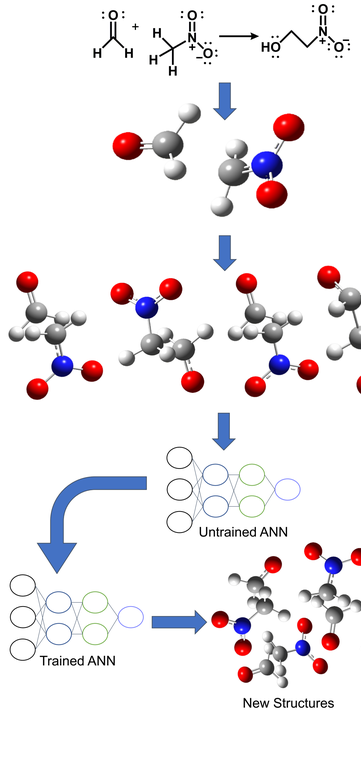


Figure 3: QM equation with an ANN

Method

Figure 4: Step by Step of the Methodology.



1. Using Gaussian09, both reactants of the Henry reaction, formaldehyde and nitromethane, are constructed and optimized using MP2/6-31+G(d) in order to find a transition state (TS).
2. Once the optimization is complete, frequency calculations are run, as well as forward and reverse intrinsic reaction coordinates (IRC) are also run.
3. Once the IRC calculations are done, both the forward and reverse structures created using the IRC's are run through the AENET generate program, which helps fit the ANN's appropriately.
4. Using the structures from the forward and reverse IRC, cutoff's are optimized by adjusting the angular and radial values using a range from 3-9 Å with increments of 0.2 and was then further refined from 4.4-4.8 Å with increments of 0.1. The most accurate results were found using an angular cutoff of 4.8 Å and a radial cutoff of 4.6 Å.
5. Once that is done, a training is run for 1000 iterations, using 80% of the structures for training and 20% for testing. Once that is done the quality of the ANN's is then verified by having the machine predict on structures with energies known already so that the root-mean-squared-error(RMSE) can be found.
6. QM/MM/MC/FEP calculations are then run using a custom version of BOSS. The reaction is run within 740 water molecules (TIP4P) for any C-C distance between 1.50-3.50 Å with small increments of 0.25 Å. This is done using the AENET machine learning. From this, 50,000 structures are generated.
7. Using the newly generated structures, many different training sets are run with by using 50,000 structures each time, letting the ANN compute energies, and if they are within an error of 2 kcal/mol, the structures are then removed from the training.

Results

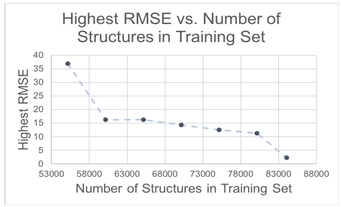


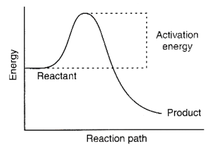
Figure 5: Graph of most recent training data with respective RMSE values.

Cutoff Values (kcal/mol)	Training Size	Highest RMSE
2.00	49814	6.2
2.00	72545	4.5
1.00	84053	2.3

Figure 6: The last iteration of each training mixture run with their respective cutoffs, training size and highest RMSE.

Figures 5 & 6 show the most up to date results on the training of the ANN's for the Henry reaction. Figure 5 shows the results from the most recent training mixture, showing that with only 7 iterations, the ANN was able to reduce the RMSE from almost 40.0 kcal/mol to slightly more than 2.0 kcal/mol. Figure 6 shows the results from 3 different training mixtures, detailing the differences in the cutoff criteria and training size. Figure 6 represents the results from three different runs, each with added structures made using QM/MM/MC/FEP calculations. If the ANN is able to produce a computed energy within a certain cutoff criteria, as outline in the figure, the structure is accepted and subsequently removed from the training set. This figure shows that in very few runs, the cutoff value was already lowered due to the minimized RMSE values. The goal is to eventually decrease the cutoff value to 0.1 kcal/mol.

Future Direction



Once this is completed, the next step would be to take the Henry reaction and apply it to different solvents, such as acetonitrile. This step would widen the possibilities for the Henry reaction ANN, maybe to even create ANN's in the same way for different reactions, such as a Diels-Alder reaction, and even with different theory levels, showing exactly how limitless machine learning is in chemistry.

Figure 7: Activation Energy($\Delta G^{\ddagger}$) Graph

Discussion

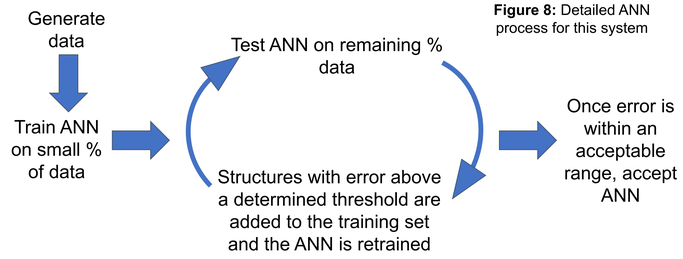


Figure 8: Detailed ANN process for this system

The next step is to further refine the generation of ANN training data. In the method portrayed above, rather than using the bulk of the data generated to train the network, only a small percentage of that data goes into training the ANN's, and the rest are used to test them. A certain percentage of structures with error above a predetermined threshold are added to the training set, as these are structures that greatly differ from anything the network has been already exposed to. The ANN is then retrained and retested, and the process repeats until all test data fall below an acceptable error range. The structures derive from QM/MM/MC/FEP calculations, which can be dramatically different from structures already in the existing training set, and therefore high RMSE errors can be initially detected, signifying the need to add more unique structures. Another possible source of error is the overfitting of the network to similar structures. If the ANN receives too many structures of similar orientation, it will develop a bias towards that output and will become less accurate for structures that differ drastically. The main goal is to be able to calculate a $\Delta G^{\ddagger}$ of the Henry reaction to within 1 kcal/mol error as compared to full QM calculations.

Conclusion

- Artificial Neural Networks are highly adaptable to large data sets, and show great potential in predicting chemical properties.
- Currently, the ANN can predict energy of the Henry reaction to a RMSE value of less than 2.3 kcal/mol.
- QM/MM/MC/FEP calculations using this enhanced ML significantly increases the speed of molecular simulations by an order of magnitude.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1659144.
Gratitude is expressed to the Center for Computational Sciences (CCS) support of this work.

Citations

1. Kostal J, Voutchkova AM, Jorgensen WL. Investigation of solvent effects on the rate and stereoselectivity of the Henry reaction. Org Lett. 2012;14(1):260–263. doi:10.1021/ol2030394
2. Vilseck, J. Z., Kostal, J., Tirado-Rives, J., Jorgensen, W. L., J. Comput. Chem. 2015, 36, 2064–2074. DOI: 10.1002/jcc.24045
3. GaussView, Version 5.0.9, Roy Dennington, Todd A. Keith, and John M. Millam, Semichem Inc., Shawnee Mission, KS, 2016.
4. Gaussian 09, Revision B.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.
5. Adapted from previous poster by T. Bestwick