Automatic Determination of Chemical Reaction Mechanisms with Neural Networks

Daniel S. Levine
Schrödinger, Inc.
120 West 45th St
New York, NY 10036
`daniel.levine@schrodinger.com`

Abstract

We present an automatic protocol for the determination of complex reaction mechanisms and reaction networks. Starting from a root node, new nodes are spawned recursively by adding an artificial force between two or more atoms (attractive for unbonded, repulsive for bonded) and carrying out a quantum mechanical optimization as described by Maeda and coworkers to generate new products that are kinetically accessible. A graph convolutional neural network is used to select the atoms which are subject to these forces, rapidly pruning unpromising search directions in the generation of child nodes. Paths in the resulting network are then automatically passed to transition state calculations which determine precise barriers for all elementary steps. The complete reaction network can then be used to simulate expected product ratios and selectivities.

1 Introduction

The rational design and improvement of chemical reactions greatly benefits from a comprehensive understanding of the underlying mechanism of the reaction of interest. These reactions are usually complicated, multi-step transformation, often involving catalysts or other auxiliary reagents, and a wide variety of products may result[1]. Complicating matters further, multiple competing mechanisms may be possible, while side-reactions (unwanted products) and thermodynamic sinks may prevent the desired reaction from being the dominant pathway[2]. The ability to design efficient and selective chemical reactions would allow for the rapid development of new materials, alternative green energy storage methods, preparation of commodity chemicals from renewable feedstocks, and cheaper fine chemical and pharmaceutical manufacturing.

Exploration of the chemical space of a collection of molecules, the space of stable molecules that can be obtained by a sequence of chemically reasonable transformations, in a balanced (that is, unbiased) and automated way can be used to generate reaction networks –graphs of the local chemical space– in which nodes represent reaction products and the edges connecting the nodes are elementary chemistry steps with a characteristic associated reaction barrier[3]. Such a network can be used to predict which pathways are operative for a certain set of reagents, and therefore also predict product ratios and selectivities.

Unfortunately, the chemical space of a set of molecules grows very rapidly with the size and complexity of the reactant molecules. Effectively sampling this space requires efficient tools to prune unproductive search directions[1]. It remains an open question as to how this sampling method can be accomplished in an automated manner. In this work, we describe our use of a neural network to effectively provide this critical, automatic pruning step, in the greater context of automatically and efficiently discovering reaction mechanisms without any user interference.

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

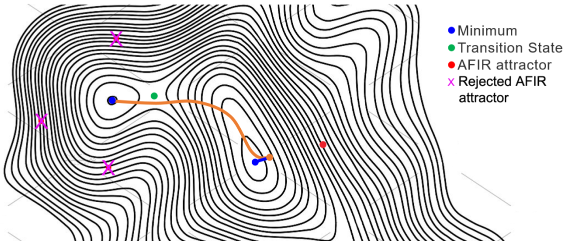


Figure 1: Potential energy surface with points of interest labeled. The orange path represents the path of the AFIR trajectory from the first minimum (blue dot) toward the AFIR attractor (red dot). The short blue path represents a relaxation to the second minimum, i.e. the product. The crossed-out (purple) AFIR attractors are of the type eliminated by the neural network screen. Note that there are no minima located near these attractors except for the starting point.

2 Methods

The Artificial Force Induced Reaction (AFIR) method of Maeda is a conceptually simple way of inducing relatively low-barrier chemical reactions [4]. In this scheme, a quantum mechanical-based optimization of two molecules is carried out in which an artificial constant force has been added between two atoms in the molecules, pushing them together (and dragging the rest of the molecule around with it, though it is free to adopt whatever conformation is necessary to lower the energy). This constant force serves to pull reactants over any activation barriers to generate products. The main challenge in this scheme is selecting which pairs of atoms should be pushed together. A conceptually related idea is the ZStruct2 method of Zimmerman, which seeks to find "driving coordinates" which correspond to interesting elementary steps, such as bond-making/breaking[5]. In both cases, the selection of these search directions is the key step to efficient exploration of chemical space. Once these search directions are determined, we employ the AFIR method to generate the reaction network. The exact transition states are determined by use of an automated TS search workflow [6].

The AFIR method can be thought of as creating attractors on the potential energy surface which pull the system from one potential well to another. These attractors are not in general located at actual minima because this would require knowing the structure of the product. However, since most of the degrees of freedom of the molecules are unconstrained, as long as the one constrained bond distance is reasonable, the attractor will be located fairly close to a minimum. If the constraint is unreasonable, there will not be a nearby minimum and hence we call such a constraint "unproductive" (since it does not lead to a productive reaction). Figure 1 illustrates this situation schematically. The main result reported here is a means of identifying unproductive AFIR attractors (purple X's in Figure 1) from productive ones (red dots).

Graph-convolutional neural networks trained on previously reported experimental observations have recently seen great success as predicting overall reaction transformations[7]. The ability of these networks to represent the whole molecule and detect the effects of remote functional groups on local reactivity prompted our investigation of these networks for pruning the search space during reaction exploration. A graph-convolutional classifier which used the molecular connectivity, atom identities, and basic hybridization information to generate the convolution along with the atoms being pushed together was trained on data obtained from previous AFIR trajectories, where the classification was whether or not a reaction (of any type) took place. As the goal of a reaction network is to determine an unbiased understanding of the available reaction space for an ensemble, we do not distinguish what reaction is occurring; we seek all reactions.

For a concrete chemical example, consider the molecules highlighted in blue in Figure 2. All possible pairs of atoms in the two molecules could be considered as possible reactions. However, the vast majority of these will not lead to a productive reaction. Since the quantum mechanical optimizations to determine if such a reaction is reasonable or not is fairly expensive, we wish to minimize such calculations. The reaction proposals marked in red could be instantly dismissed by an expert chemist as unlikely to be productive. We use a graph-convolutional neural network to be able to make the same sorts of predictions, dramatically reducing the AFIR algorithm's cost.

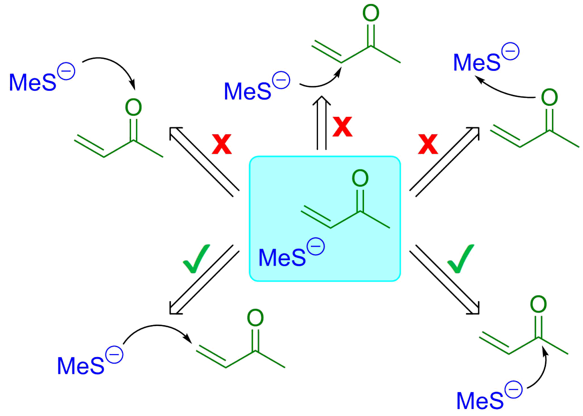
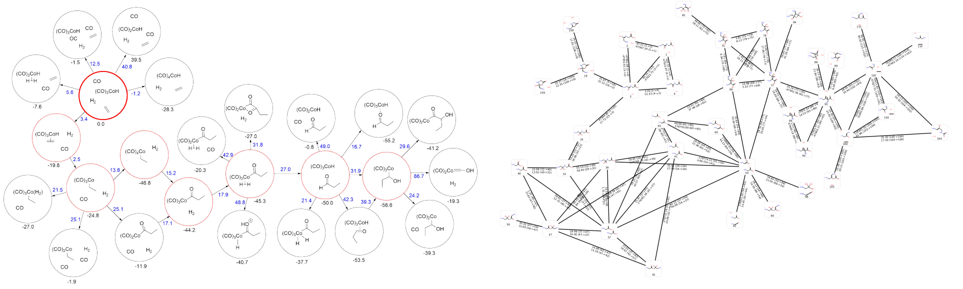


Figure 2: Examples of AFIR trajectories in terms of real chemistry. An expert chemist could quickly dismiss the reaction proposals marked with an X; the neural network classifier can as well.



(a) Hydroformylation reaction network with catalytic cycle highlighted in red.

(b) Reactions of dipeptides with water in acidic conditions.

Figure 3: Reaction networks generated with AFIR scheme

In order to obtain an active learning method, a committee of five networks is used to determine uncertainty in the predicted classification. If the uncertainty is deemed to be too high, the AFIR trajectory is performed and the resulting data is added to the training set of the model.

3 Results

The graph-convolutional neural network was trained on 6785 AFIR trajectories and obtained ROC AUC scores of about 0.75 on out-of-sample data. Using the network in conjunction with the AFIR scheme to build reaction networks, thresholds were selected so that about 1% of AFIR trajectories which were predicted to produce a non-trivial result (i.e. a reaction) did not actually result in a reaction (the false positive rate), while conversely about 4% of AFIR trajectories which were predicted to not produce a reaction actually did (the false negative rate). The remaining 95% of data points were correctly classified. This is extremely high enrichment over not using the neural network to screen (i.e. considering all pairs of non-bonded atoms) for potential reactivity. In such cases without the network, upwards of 90% of AFIR trajectories produce no result and are wasted compute time.

This neural network/AFIR scheme was able to predict the mechanisms of a number of complex organic and inorganic transformations without interference from users, including peptide hydrolysis in acidic and neutral conditions and the Co-catalyzed hydroformylation of ethylene. Further investigations are currently in progress, as well as considerable augmentation of the training data set.

References

[1] Amanda L. Dewyer, Alonso J. Argüelles, and Paul M. Zimmerman. Methods for exploring reaction space in molecular systems. *WIREs Comput. Mol. Sci.*, 8(2):e1354, March 2018.

[2] Paul M. Zimmerman. Automated discovery of chemically reasonable elementary reaction steps. *J. Comput. Chem.*, 34(16):1385–1392, June 2013.

[3] Dmitrij Rappoport, Cooper J. Galvin, Dmitry Yu. Zubarev, and Alán Aspuru-Guzik. Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry. *J. Chem. Theory Comput.*, 10(3):897–907, March 2014.

[4] Satoshi Maeda, Yu Harabuchi, Makito Takagi, Tetsuya Taketsugu, and Keiji Morokuma. Artificial Force Induced Reaction (AFIR) Method for Exploring Quantum Chemical Potential Energy Surfaces. *Chem. Rec.*, 16(5):2232–2248, October 2016.

[5] Amanda L. Dewyer and Paul M. Zimmerman. Finding reaction mechanisms, intuitive or otherwise. *Org. Biomol. Chem.*, 15(3):501–504, January 2017.

[6] Leif D. Jacobson, Art D. Bochevarov, Mark A. Watson, Thomas F. Hughes, David Rinaldo, Stephan Ehrlich, Thomas B. Steinbrecher, S. Vaitheeswaran, Dean M. Philipp, Mathew D. Halls, and Richard A. Friesner. Automated Transition State Search and Its Application to Diverse Types of Organic Reactions. *J. Chem. Theory Comput.*, 13(11):5780–5797, November 2017.

[7] Connor W. Coley, Wengong Jin, Luke Rogers, Timothy F. Jamison, Tommi S. Jaakkola, William H. Green, Regina Barzilay, and Klavs F. Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science*, 10(2):370–377, 2019.