Learning Scoring Functions for Chemical Expert Systems

Chloé-Agathe Azencott, Matthew A. Kayala, and Pierre Baldi

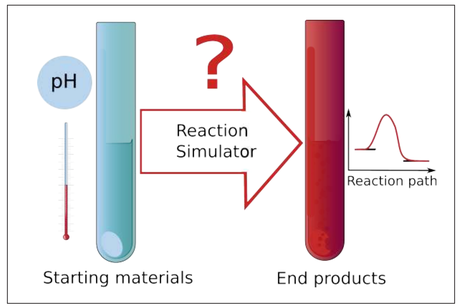
UCIrvine

Institute for Genomics and Bioinformatics
Bren School of Information and Computer Sciences

Background

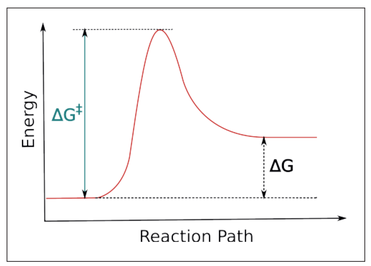
Reaction Simulator

Being able to computationally simulate reactions would augment organist chemists' problem-solving abilities and enable chemical space exploration.



Energy of Activation

Activation energy of an elementary reaction step is a core component of our reaction simulator. It is a measure of reaction favorability and is directly related to reaction rates.



Computing the energy of activation of a reaction is generally done through quantum mechanical or experimental methods, which are expensive and time consuming on a large scale. We propose to use machine learning techniques to infer the activation energy function.

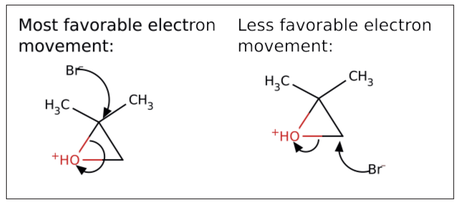
Very little quantitative data is available for machine learning training; however, chemists can tell which of many possible reactions is more favorable based solely on qualitative knowledge of trends in molecule stability and reaction rates. We therefore propose a method to generate qualitative data for use in a machine learning framework.

Data Generation

The Reaction Explorer `http://www.reactionexplorer.org` is an expert system that covers basic undergraduate chemistry. It is based on 1500+ rules covering elementary reaction steps, defined as simple electron movements from source to sink orbitals, and uses 2000+ full multistep reaction test cases.

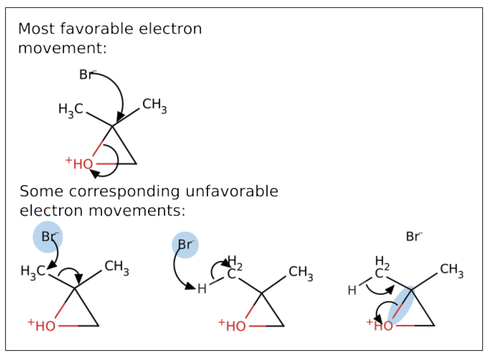
Method 1: Priority Ordering

The rules are ordered by priority. We can use this information to create ordered pairs of potential elementary steps.



Method 2: Implicitly Unfavored Reactions

Given what simple elementary movement is favored, we can infer that movements starting from the same orbital but ending in a different one (or conversely) are less likely to happen.



Limiting ourselves to ionic reactions on molecules composed exclusively of C, H, N, O and halides, we generate:

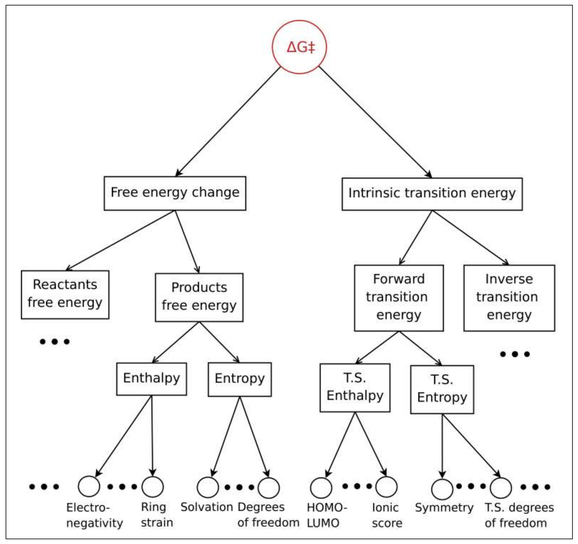
Using Method 1: 3,457 ordered pairs.

Using Method 2: 475,684 ordered pairs (corresponding to 16,806 most favorable reactions).

Statistical Machine Learning

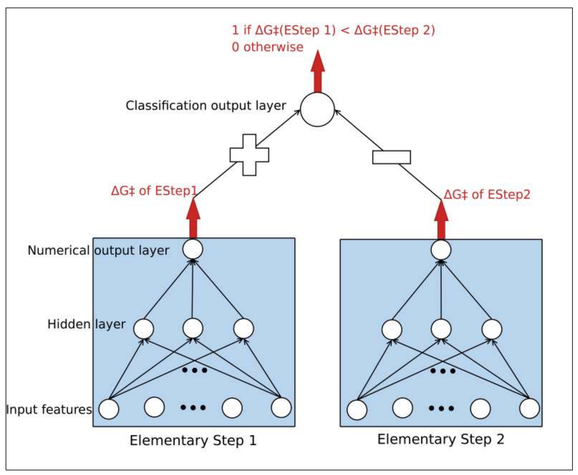
Data Representation

We derive a feature representation of elementary reaction steps in a principled way.



Neural Network Architecture

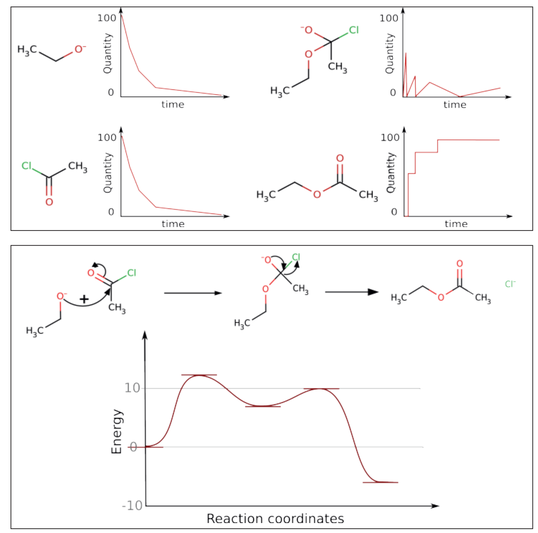
We machine learn the energy of activation using the following architecture.



Results

With 15 hidden nodes, we obtain cross-validated classification accuracies of 93.7 % over the data generated by priority ordering (Method 1) and 99.0 % over the data generated by considering implicitly unfavored reactions (Method 2).

Reaction Simulator



Conclusion

We proposed a method to compensate for the lack of quantitative data by leveraging qualitative knowledge. When applied to the prediction of energies of activation for ionic reactions of molecules composed of C, H, N, O atoms and halides, it reproduces the qualitative trends and ranking order of reactivities necessary to solve relevant chemistry problems.

Differentiating the relevance of data generated from each method, as the first dataset pairs relatively likely reactions while the second one includes highly unlikely reactions, is one of our priorities.

In the near future we plan to validate the reaction simulator on external reaction databases (e.g. SPRESI).

We also plan to expand the coverage of our predictor to include for instance third row elements, organo-metallic chemistry, pericyclic reactions, and free radical chemistry.

Related Work

J. H. Chen and P. Baldi. Reaction mechanism prediction by transformation rules and general principles. ACS National Meeting, Fall 2008. Philadelphia, PA.

J. H. Chen and P. Baldi. Reaction simulation expert system for synthetic organic chemistry. ACS National Meeting, Spring 2009. Salt Lake City, UT.

Further Information

Contact: `cazencot@ics.uci.edu`

Talk on Tuesday, March 24th at 4:40pm
Salt Palace Convention Center - 254 A CINF - Adv. Scoring Functions

Reaction Simulator: Jonathan H. Chen
Talk on Thursday, March 26th at 10:30am
Salt Palace Convention Center - 254 A CINF - Cal. Papers

Acknowledgments

Matthew A. Kayala and Jonathan H. Chen
Prof. Pierre Baldi

OpenEye and ChemAxon Academic Software Licences

NIH/NLM Biomedical Informatics Training Grant
NSF Grants 0321390 and 0513376
The Dreyfus Foundation Special Grant Program in Chemical Sciences