

Comparison of Computer Simulation Methods for Predicting Chemical Reactions

Graham Gibson

IBM, Austin TX, 78723, USA.

(Dated: January 2, 2016)

In this paper we compare two primary methods of predicting basic organic chemistry reaction predictions. We analyze two types of models, an NLP based Neural Network and an agent based model. We compare and contrast the complexity, accuracy, and generalizability of both models as applied to predicting organic chemistry reactions. Two basic reaction mechanisms are explored, Elimination and Addition reactions. These two mechanisms are simple but fundamental to the set of organic reactions, as many more complicated reactions use these mechanisms as intermediates. We first verify the model on an alkene halogen addition reaction and then investigate the models generalizability to the elimination reaction.

PACS numbers:

I. INTRODUCTION

The problem of reliably predicting organic chemical reactions is vital in computation chemistry and biology. A model that is able to describe 3-dimension conformations of reactions would go a long way towards solving the more complicated protein folding problem.[?] Finding such a solution would have extremely beneficial consequences in the area of targeted drug development and computational medicine. Such a solution would also drastically reduce drug development time and help eliminate clinical trial risk. [?]

Traditional computational chemistry approaches have involved analyzing the potential energy surface of the atoms in the n-molecule system. The energy surface contains an array of parameters from 3-d cartesian coordinates, to electronegativity, to electron-electron repulsion. Finding the most likely product is equivalent to finding minima in the energy surface using traditional techniques like gradient descent. This method is often termed molecular dynamics. [?] A second approach that is frequently taken is to go deeper into the reaction and look at the driving quantum mechanical principles that are at play. These methods use a combination of approximation techniques to guess the Hamiltonian of an atom and analytically solve for the orbital energy. This allows one to identify reactive orbitals very precisely. The two primary methods are called Hartree-Fock and Density Functional Theory. [?] [?] Although these approaches are invaluable in their description of the energy of chemical reactions, they are both quite complicated and computationally involved. The quantum mechanical approaches are extremely resource intensive as they involve the approximation of complicated integrals and matrix multiplications. [?] A simpler and potentially less computationally intensive approach would be to consider chemistry at a more macroscopic level. The reasoning behind this is that chemistry students begin studying reaction mechanisms and gain the ability to predict basic reactions long before they learn about solving Hamiltonian equations. Organic chemistry textbooks focus on what we may call the "grammar" of organic chemistry: mechanisms and properties that allow chemistry students to predict chemical reactions without the use of quantum mechanics. Using this view we can define two separate higher level approaches to predicting chemical reactions.

A. First Approach: Language Based Modeling

Continuing with the idea of modeling organic reactions as language over the alphabet of atoms, predicting a chemical reaction has a natural analogy between verifying if a sentence is valid in a given language. The problem of verifying a sentence is in a language has been studied in depth in the field of computer science and has been most successfully modeled by using recursive neural networks (RNNs) such as the Long Short Term Memory network architecture. [?] This architecture maintains state of the neural network by copying the hidden state H_{t-1} to the hidden state H_t using a series of memory gates. LSTM networks have gained a lot of popularity in the computer science community for their ability to remember context from many time steps previous to the current because of this copying of hidden layer state. [?] A common LSTM network approach to the problem of machine translation is depicted below as in [?]

This figure shows the input sequence being fed into the LSTM network one character at a time. In the chemistry analog we are feeding one atom in the textual representation of a molecule into the LSTM at a time step t . This textual representation is obtained by translating chemical reactions to their SMILES representation. [?] As targets we feed the known product of a chemical reaction. Formally, we have:

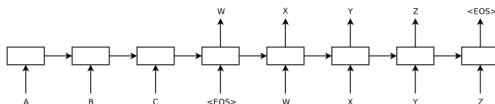


FIG. 1: LSTM network feeding in an input sequence and target sequence.

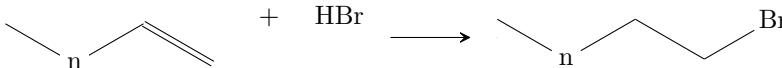
$$I = (i_1, i_2, \dots, i_n)$$

$$O = (o_1, o_2, \dots, o_n)$$

The LSTM is able to learn a function $f : I \rightarrow O$ using standard stochastic gradient descent by first mapping the input and output character sets to one-hot encodings. If we examine a function f produced by the neural network after an error threshold e is reached we can investigate the semantic meaning of the mapping. The map f takes input sequences of SMILES encoded reactants and outputs SMILES encoded products. For example



If we apply this function to molecules with variable length numbers of carbon atoms attached to the double bond we learn something about the meaning of the function mapping.



We see that the function f has in a sense learned properties of the reaction. First, it knows that reactive site in the addition reaction is the alkene $C = C$ bond. Second, the network is able to understand that the number of C atoms is invariant under the mapping.

Results

TABLE I: Addition Reaction Learning after 50,000 Training Iterations

Input	Output	Target	Correct In Training Data	
$CCCCC = C \text{ HB}$	$CCCCCCB$	$CCCCCCB$	Y	Y
$CC = C \text{ HI}$	$CCCI$	$CCCI$	Y	
$CCCCCCCCC = C \text{ HF}$	$CCCCCCCCCF$	$CCCCCCCCCF$	Y	Y
$CC = C \text{ HF}$	$CCCF$	<i>none</i>	Y	N
$CC = C \text{ HCl}$	$CCCF$	<i>none</i>	N	N

As can be seen from *Table 1* the map f is able to learn that the number of C atoms in the reactants and the products must be the same. However, the LSTM network was not able to abstract to halogens it has not seen before to understand that they behave in the same way. Unfortunately, the last example highlights one of the main issues with neural network based approach for predicting chemical reactions. The RNN is only able to learn lexicographic similarities between inputs and targets and not chemical similarities between training examples. The only way the network would have been able to correctly predict the last reaction is if it had previously seen the halogen in context of another reaction. In this way the neural network is learning the grammar of the chemical alphabet instead of the grammar of chemistry.

Complexity

Another issue of NLP neural network techniques is the size of the sample space. If we consider a reaction of two reactants whose strings are length k_1 and k_2 over an alphabet of size n a randomly initialized neural network will produce strings with length between $(0, n^{k_1+k_2})$. This space is much larger than the space derived from first principles which has size $\binom{m}{2}$ where m is the total number of atoms in the system: since we need only consider pairwise interactions of atoms.

B. Second Approach: Agent Based Modeling

Considering the limitations of the NLP approach described above we turned next to agent based modeling. Agent based models are comprised of multiple instantiations of classes of objects that are given objective functions. They are allowed to interact in a defined environment for a given number of time steps to improve their objective function. [?] Agent based models have been used with a great deal of success to model complex behaviors that arise from a simple set of heuristics. There is also a very natural encoding from the chemical world to the ABM world. We define an agent as follows

$$A_i = (x, y, z, e, b, n)$$

Where x, y, z are the standard cartesian coordinates, e is the experimentally determined electronegativity of the element, b is the maximal bond order for that given element, and n is the set of agent linked to A_i . For example a carbon atom in methane molecule at the origin of the agent space is defined by,

$$C_i = (0, 0, 0, 2.55, 4, 4)$$

By defining an agent this way we can use two simple heuristics to update agent A_i and agent B_i . Our global objective function is to maximize energetic stability of the system. We first let agents randomly move around a 3 dimensional cartesian coordinate and if the euclidean distance between agents A_i and B_i is within a sphere of radius r they are allowed to interact. The two heuristics used to update agents are as follows:

```

if  $|A_i(n)| < A_i(b)$  and  $|B_i(n)| < B_i(b)$  then add  $B_i$  to  $A_i(n)$  and add  $A_i$  to  $B_i(n)$ 
end if
if  $|A_i(n)| < A_i(b)$  and  $|B_i(n)| = B_i(b)$  and  $B_i(e) < A_i(e)$  then set  $B_i(n) = A_i$ 
end if

```

This is equivalent to the two chemical statements: If agent A_i can accept more bonds, it will, and if agent B_i can accept more bonds it similarly will. The second statement says that if a agent A_i encounters an agent B_i that has fulfilled its bond order, agent B_i may still break its links to bond with agent A_i if agent A_i 's electronegativity is greater. This is because bonds to elements with higher electronegativities are more stable. [?] These two heuristics define the bonding interaction between two agents but do not describe the geometric relationship between agents. This is handled by a separate heuristic determined by the principle of electron-electron repulsion. If we have agent A_i with $|A_i(n)| > 2$ then we need to define new (x, y, z) coordinates for the neighbors that minimizes electron-electron repulsion. This can be done by a simple randomized algorithm :

```

let  $x, y, z$  be random in  $\mathbb{R}$ 
for agent in  $A_i(n)$  do
  if  $Volume(A_i(x) + x, A_i(y) + y, A_i(z) + z) > Volume(A_i(x), A_i(y), A_i(z))$  then
     $A_i(x) \leftarrow A_i(x) + x$ 
     $A_i(y) \leftarrow A_i(y) + y$ 
     $A_i(z) \leftarrow A_i(z) + z$ 
  end if
end for

```

Below is the result of running the algorithm on a chiral center, C , of bond order 4 for 1000 time steps.

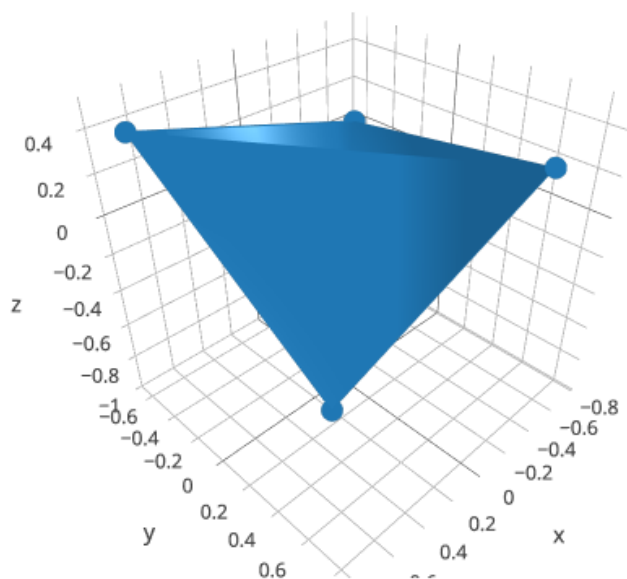
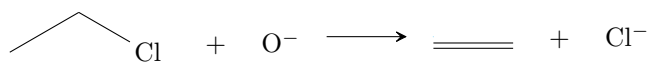
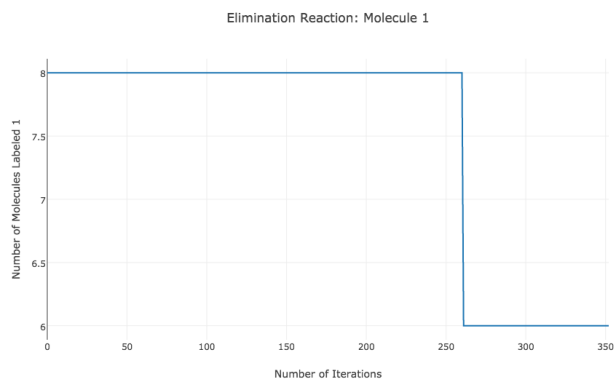


FIG. 2: Result of Stochastic Arrangement Algorithm of Carbon

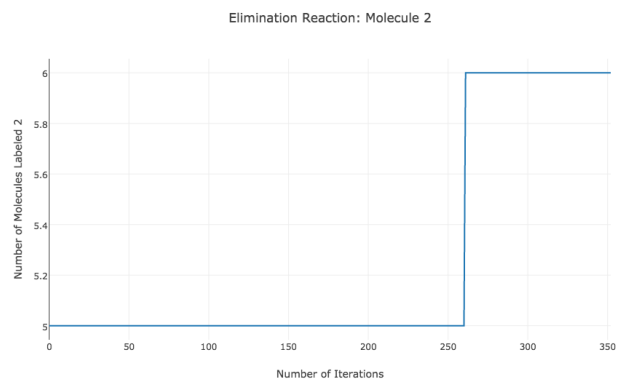
Results

As can be seen from the figures below the ABM based on these two simple heuristics is able to predict elimination reactions that employ the E1 Mechanism. The following result is from the ABM run on the reactants below.

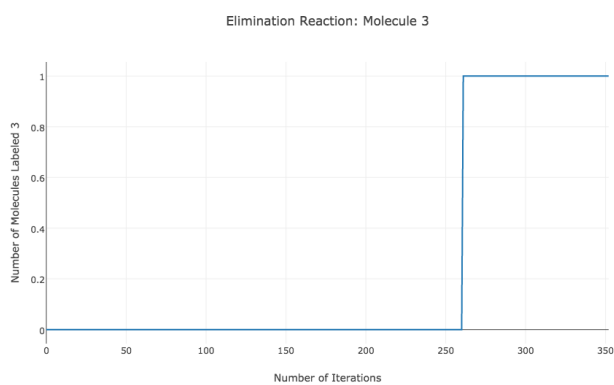




(a) Molecule 1

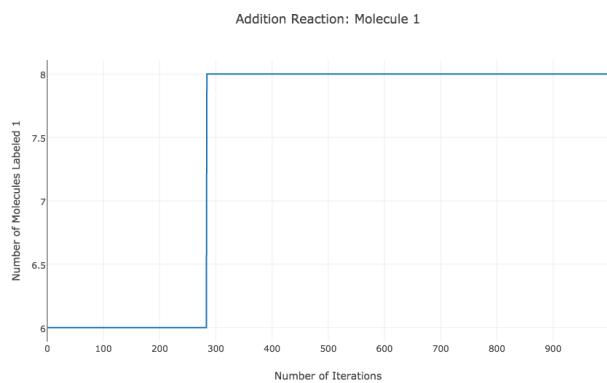


(b) Molecule 2

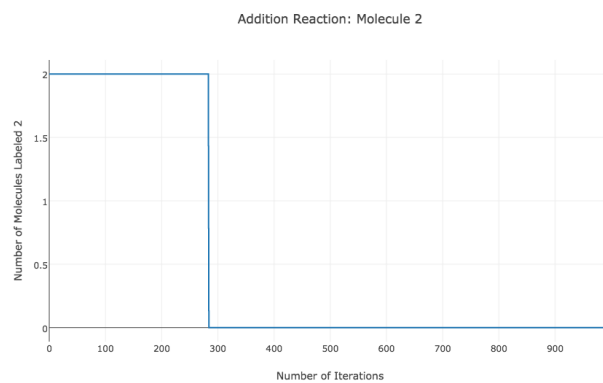


(c) Molecule 3

The ABM was also correctly able to predict addition reactions when given the following reactants.



(a) Molecule 1



(b) Molecule 2

II. CONCLUSIONS

Two models were presented in an attempt to correctly predict addition and elimination reaction mechanisms in organic chemistry. The NLP based LSTM network was correctly able to learn that the double bond of alkenes is the most reactive site and will react with halogens to add across the double bond. However, two main limitations exist to this model. First, the search space is too large to abstract to other mechanisms, such as elimination, without an extremely large amount of time spent training. Second, using the language model causes the network to lose chemical information by only creating lexographic relationships between atoms. Improving the training of such a network would be very profitable because the main advantage of the model is the simplicity of training data required. One would be able to predict chemical reactions without any heuristic assumptions, i.e. chemical knowledge, by simply feeding the network known reactions.

The ABM was able to learn more mechanisms, both addition and elimination, and converge in a smaller number of time steps due to the smaller search space. However, unlike the LSTM, certain chemical heuristics must be assumed to be universally successful in predicting reactions. As the model is tested on more and more reactions, new heuristics may have to be added to account for boundary conditions which may lead to a messy solution of a long chain of logical if statements. Further research on both models is required to see how well they are able to abstract to the rich language of organic chemistry.

-
- [1] Jack Cazes. Encyclopedia of Chromatography -Vol 2. Pg. 1505. CRC Press, 2005
 - [2] Uri Wilensky, William Rand. An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo. April 10, 2015
 - [3] Daylight Chemical Information Systems, Inc. SMILES - A Simplified Chemical Language. 2008 [Online access: November 20, 2015]
 - [4] Andrej Karpathy. The Unreasonable Effectiveness of Recurrent Neural Networks. May 21, 2015. [Online access: December 16, 2015]
 - [5] A Chemist's Guide to Density Functional Theory, W. Koch and M.C. Holthausen, WILEY-VCH (2001).
 - [6] Errol G. Lewars. Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics. Springer; 2011 pg 9
 - [7] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. J. Sequence to Sequence Learning with Neural Networks. *arXiv:1409.3215 [cs.CL]* [Online access: December 18, 2015]
 - [8] L. BORTOLUSSI, A. DOVIER AND F. FOGOLARI. Agent-based protein structure prediction. Multiagent and Grid Systems, 3(2), 2007, pp. 183-197.
 - [9] Morra G, Meli M, Colombo G. Molecular dynamics simulations of proteins and peptides: from folding to drug design. Curr Protein Pept Sci. 2008 Apr;9(2):181-96.
 - [10] C. David Sherrill. An Introduction to Hartree-Fock Molecular Orbital Theory. School of Chemistry and Biochemistry, Georgia Institute of Technology. June 2000. <http://vergil.chemistry.gatech.edu/notes/hf-intro/hf-intro.pdf> [Online access: December 20, 2015]
 - [11] James Daniel Whitfield, Peter John Loved, Alan Aspuru-Guzik. Computational Complexity in Electronic Structure. arXiv:1208.3334v1 [physics.chem-ph] 16 Aug 2012. [Online access: December 22, 2015]