UNIVERSITY OF CALIFORNIA,
IRVINE

Computational Tools for Chemical Reactions: Simulation & Prediction

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

David Fooshee

Dissertation Committee:
Professor Pierre Baldi, Chair
Professor David Van Vranken
Professor Daniel Hirschberg

2017

DEDICATION

To:

Mom and Dad, for always encouraging my curiosity
Timmie, for unwavering support
and Maxwell, for keeping things interesting

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to acknowledge financial support from the National Institutes of Health, the National Science Foundation, and DARPA.

This dissertation contains reprinted material appearing in the *Journal of Chemical Information and Modeling*, the *Journal of Cheminformatics*, and *Environmental Science & Technology*. I would like to thank the coauthors of those works for their contributions. I would also like to thank Jordan Hayes and Yuzo Kanomata for computing support.

Finally, I would like to thank my adviser for his support, inspiration, and guidance. This work would not have been possible without him.

CURRICULUM VITAE

David Fooshee

EDUCATION

Doctor of Philosophy in Computer Science **2017**
University of California Irvine *Irvine, CA*

Masters in Computer Science **2012**
University of California Irvine *Irvine, CA*

Bachelor of Science in Biomedical Engineering **2005**
Washington University in St. Louis *St. Louis, MO*

RESEARCH EXPERIENCE

Graduate Research Assistant **2009–2017**
University of California Irvine *Irvine, CA*

Research Intern **2006–2009**
M. D. Anderson Cancer Center *Houston, TX*

Researcher I **2004–2005**
Washington University School of Medicine *St. Louis, MO*

TEACHING EXPERIENCE

Teaching Assistant **2009–2011**
University of California Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Deep learning for reaction prediction and product identification **2017**
Under submission

Synergies between quantum mechanics and machine learning in reaction prediction **2016**
Journal of Chemical Information and Modeling

Accurate and efficient target prediction using a potency-sensitive influence-relevance voter **2015**
Journal of Cheminformatics

Atmospheric oxidation of squalene: Molecular study using COBRA modeling and high-resolution mass spectrometry **2015**
Environmental Science & Technology

ReactionMap: An efficient atom-mapping algorithm for chemical reactions **2013**
Journal of Chemical Information and Modeling

COBRA: A computational brewing application for predicting the molecular composition of organic aerosols **2012**
Environmental Science & Technology

SOFTWARE

Reaction Predictor `http://cdb.ics.uci.edu`
Search for unknown structures by predicting sequences of elementary chemical reactions.

Reaction Map `http://cdb.ics.uci.edu`
Predict atom-mappings for chemical reactions.

ABSTRACT OF THE DISSERTATION

Computational Tools for Chemical Reactions: Simulation & Prediction

By

David Fooshee

Doctor of Philosophy in Computer Science

University of California, Irvine, 2017

Professor Pierre Baldi, Chair

Achieving human-level performance at predicting chemical reactions remains an open problem with broad potential applications. Here we describe a deep learning-based tool for chemical reaction prediction and product identification. Significant efforts were made to curate and refine a new, high-quality data set of hand-selected chemical reactions written at the level of elementary electron movements. Using deep artificial neural networks trained on this data, we demonstrate a high degree of accuracy at predicting real-world reactions. Because predictions are made at the elementary step level, they can be chained together to form multi-step reaction pathway searches, to help identify unknown side products.

We also present a computational brewing application, COBRA, capable of simulating complex chemical mixtures. We demonstrate its efficacy at modeling both the photooxidation of isoprene, and the oxidation of squalene in the presence of ozone, by comparing predicted results with results obtained from high-resolution mass spectrometry.

In addition, we address the problem of atom-mapping for chemical reactions, by designing a new atom-mapping algorithm that can be used to annotate unmapped reactions.

Chapter 1

Introduction

Predicting, modeling, and understanding chemical reactions are tasks that present many interesting challenges – and opportunities – for computational approaches. Presented in this dissertation are several computational tools and algorithms designed to address these challenges.

1.1 Dissertation Outline and Contributions

1.1.1 Atom-Mapping for Chemical Reactions

The purpose of this work is to take a significant step towards solving the atom-mapping problem. Plainly stated, in a balanced chemical reaction where all the atoms are accounted for, the atom-mapping problem is the problem of constructing a one-to-one map between all atoms on the left-hand (reactants) side and the right-hand (products) side. Our approach uses a combination of maximum common chemical subgraph search, and minimization of an assignment cost function derived empirically from training data. On large test sets

with thousands of reactions randomly sampled from the SPRESI commercial database, we demonstrate up to 36% greater accuracy compared to existing methods.

1.1.2 Computational Chemical Reaction Simulation

Atmospheric organic aerosols (OA) represent a significant fraction of airborne particulate matter and can impact climate, visibility, and human health. These mixtures are difficult to characterize experimentally due to the enormous complexity and dynamic nature of their chemical composition. We introduce a novel Computational Brewing Application (COBRA) and apply it to modeling oligomerization chemistry stemming from condensation and addition reactions of monomers pertinent to secondary organic aerosol (SOA) formed by photooxidation of isoprene.

1.1.3 COBRA Modeling of Atmopheric Squalene Oxidation

We apply our Computational Brewing Application (COBRA) to simulate the oxidation of squalene in the presence of ozone, and compare predicted results with those observed by high-resolution mass spectrometry experiments. COBRA predicts over one billion molecular structures between 0-1450 Da, which correspond to about 27,000 distinct elemental formulas. Over 83% of the squalene oxidation products inferred from the mass spectrometry data are matched by the simulation. Simulation indicates a prevalence of peroxy groups, with hydroxyl and ether groups being the second-most important O-containing functional groups formed during squalene oxidation.

1.1.4 Deep Learning for Chemical Reaction Prediction

In this work, we apply deep learning to the prediction and ranking of elementary chemical reactions. By chaining sequences of elementary reactions, we can further predict multi-step reaction transformations. Furthermore, given a set of reactants, and a set of unknown mass targets, we can search for sequences of elementary reactions, i.e. reaction pathways, leading to the formation of those targets. We demonstrate significantly greater prediction accuracy on a benchmark data set, when compared with shallow neural network architectures. Finally, we demonstrate a new approach to predicting reactive sites using recurrent neural networks, specifically long short-term memory (LSTM) architectures.[57, 50]

Chapter 2

Atom-Mapping for Chemical Reactions

Large databases of chemical reactions provide new data-mining opportunities and challenges. Key challenges result from the imperfect quality of the data, and the fact that many of these reactions are not properly balanced or atom-mapped. Here we describe ReactionMap, an efficient atom mapping algorithm. Our approach uses a combination of maximum common chemical subgraph search, and minimization of an assignment cost function derived empirically from training data. We use a set of over 259,000 balanced, atom-mapped reactions from the SPRESI commercial database to train the system, and we validate it on random sets of 1,000 and 17,996 reactions sampled from this pool. These large test sets represent a broad range of chemical reaction types, and ReactionMap correctly maps about 99% of the atoms, and about 96% of the reactions, with a mean time per mapping of two seconds. Most correctly mapped reactions are mapped with high confidence. Mapping accuracy compares favorably with ChemAxon's AutoMapper versions 5 and 6.1, and the DREAM web tool. These approaches correctly map 60.7%, 86.5%, and 90.3% of the reactions, respectively, on the same data set. A ReactionMap server is available on the ChemDB web portal at

http://cdb.ics.uci.edu.

2.1 Introduction

Large databases of chemical reactions, such as Beilstein (Reaxys/Elsevier) and SPRESI (InfoChem), create several new data-mining opportunities and challenges. One major opportunity is that by mining these databases, properly configured machine learning algorithms should be able to learn a theory of chemistry and chemical reactivity, and be capable of generalizing it to predict the outcome of arbitrary reactions, with a host of possible applications.

Before such a formidable task can be attempted, several other problems must first be solved. One fundamental problem, not addressed here, is that these databases are inherently commercial and not readily available to the academic research community for mining purposes. A second problem has to do with the inconsistent quality of the data entered into these databases over the years. Most of the reactions are not balanced and not atom-mapped. This alone creates significant problems for automated machine understanding of chemical reactions and reactivity.

The purpose of this work is to take a significant step towards solving the atom mapping problem. Plainly stated, in a balanced chemical reaction where all the atoms are accounted for, the atom mapping problem is the problem of constructing a one-to-one map between all atoms on the left-hand (reactants) side and the right-hand (products) side. When a reaction is unbalanced, it is still possible to consider a partial atom mapping by trying to identify a maximal subset of atoms that can be mapped in one-to-one fashion between the left- and right-hand sides. Most of the time there is a unique solution. In cases involving symmetries, there may be several symmetrically equivalent mappings.

Balancing reactions is likely to require a fairly deep understanding of chemical reactivity. It is possible that a complete solution for the atom mapping problem may also require a fairly deep understanding of chemical reactivity. Our intuition, however, is that this is not the case, for the majority of situations. Accordingly, the goal here is to develop an algorithm that can address the atom mapping problem on the broadest possible scale without requiring any deep knowledge of chemistry, resorting primarily to topological properties and correlations between the molecular graphs of the reactants and products.

The atom mapping problem can be formulated in several ways. Since molecules are typically represented by graphs, the atom mapping problem is, at its core, a question of graph matching. A reaction's reactants, R, and products, P, represent two undirected, potentially disconnected graphs. Nodes in R are labeled, while nodes in P are not. We seek a mapping from R to P such that each node in R is assigned to the correct node in P; that is, the assignment must reflect the underlying chemical rearrangement that occurs in the reaction. Although less natural, the atom mapping problem can obviously be defined also in the reverse direction, from products to reactants.

Graph matching is a common task in computer vision and automated object recognition. Prior work addressing the problem includes that of Rangarajan and Mjolsness, who described a Lagrangian relaxation network approach [110]. They define a distance measure between adjacency matrices for two graphs, then search for a permutation matrix under constraints defined in the framework of deterministic annealing such that the distance between graphs is minimized. Another approach, described by Taskar et al.[126], is a convex optimization problem formulation wherein the loss function is a Hamming distance between the target mapping and a candidate mapping, which simply counts the number of differing variables between the target and candidate solution. Yet another approach involves the use of spectral methods for the permutation group [59].

Prior work on the atom mapping problem generally falls under two categories: common

substructure-based methods, and optimization-based methods [30]. Common substructure methods generate an atom mapping by correlating matching subgraphs between reactants and products, followed by additional computation to correlate any remaining atoms which were not part of the common substructure [93, 11]. Optimization-based methods attempt to minimize the number of bonds broken and formed during a reaction, or to minimize some other cost function over the possible atom mappings for a given reaction [5, 33, 55, 82, 42].

Here we combine elements of both approaches. First we identify maximum common subgraphs between the left and right sides of a reaction. Then we search for an assignment of the remaining atoms that minimizes a cost function derived empirically from chemical reaction data.

2.2 Data

To obtain training and testing data, we first removed from the 2005 version of the SPRESI database all reactions that were not balanced and/or not fully atom-mapped. This resulted in a set of 259,595 reactions (SP05_set). From this set we extracted two subsets: SP05_test and SP05_training. SP05_test contained 1,000 reactions randomly extracted from SP05_set, whereas SP05_training contained the remaining 258,595 reactions. Finally, we also extracted a set, SP09_test, containing 17,996 new balanced and atom-mapped reactions. These reactions were present in the 2009 version of SPRESI but not in the 2005 version. Figure 1 shows the distribution of the number of reactants per reaction in SP05_set.

We used SP05_training to extract first and second neighbors reaction rules as described below. This resulted in 87,371 first neighbors, and 668,684 second neighbors reaction rules. Finally, we tested our approach on both SP05_test and SP09_test.

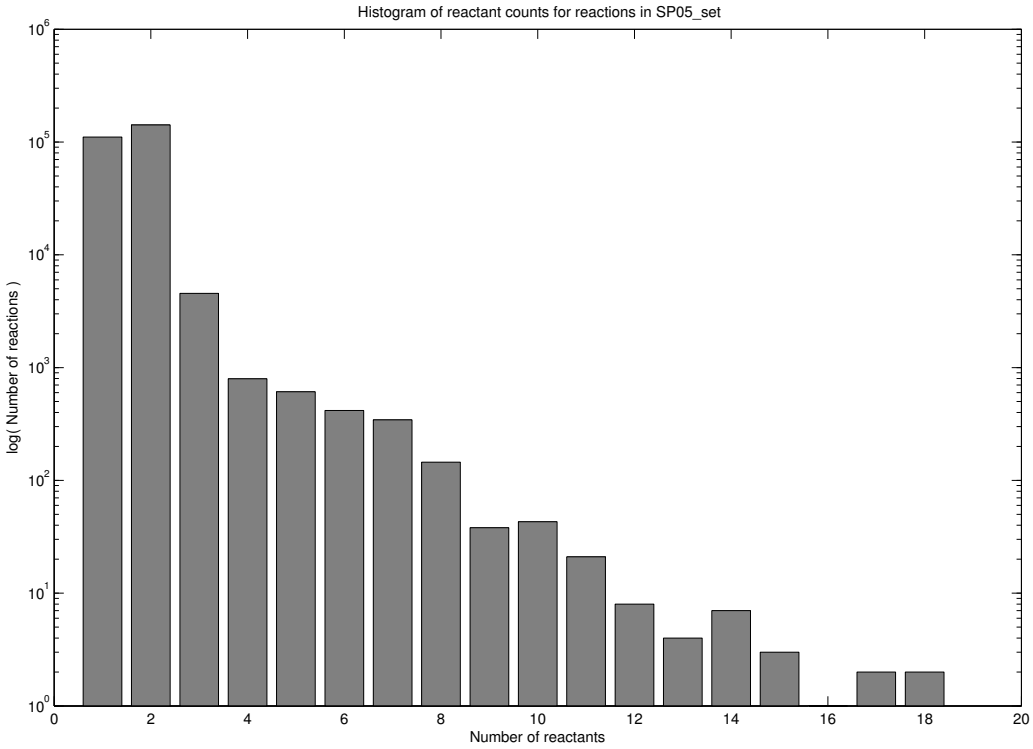


Figure 2.1: Histogram, using a logarithmic scale on the y-axis, of the number of reactants per reaction computed on the 259,595 reactions from SP05_set. Most reactions have at most two reactants.

2.2.1 Database of reaction rules

The database of reaction rules is built using the SP05_training data as follows. For each reaction, we extract the lists of atom "state keys" $k^{(1)}(u_i)$ and $k^{(2)}(u_i)$ for reactants, $k^{(1)}(v_j)$ and $k^{(2)}(v_j)$ for products. These are SMILES strings representing the atom u_i with its first neighbor ($k^{(1)}(u_i)$), and first and second neighbors ($k^{(2)}(u_i)$). We then compare the two lists $k^{(1)}(\underline{u})$ and $k^{(1)}(\underline{v})$, and define the first neighbors reaction rules by looking at the differences between the two. Second neighbors reaction rules are similarly defined, using $k^{(2)}(\underline{u})$ and $k^{(2)}(\underline{v})$. For example, the simple addition reaction [CH2:2]=[CH2:1].[BrH:3] >> [CH3:2][CH2:1][Br:3], shown in Figure 2A, results in the following first neighbors reaction rules:

- C=C >> C[CH2]; rule for the carbon atom with map index 2.
- C=C >> CCBr; rule for the carbon atom with map index 1.
- Br >> [CH2]Br; rule for the bromine atom.

2.3 Atom mapping algorithm

Although the algorithm we describe here is not dependent on the particular way a reaction is represented in a computer program, we will use SMILES strings [137, 139, 138]. SMILES is a widely used language for representing molecules and reactions as simple text strings. It also provides the foundation for SMARTS and SMIRKS [1], which allow the efficient representation of molecular patterns (SMARTS) and reaction mechanisms (SMIRKS).

Following the SMILES convention, we will assume that hydrogen atoms not involved in the reaction are implicitly represented, i.e. they are not found in the SMILES string representing the reaction. Moreover, for now we will only consider balanced reactions – that is, the reactants side will contain the same number and same type of atoms as the products side.

With this in mind, the method we use to construct a one-to-one mapping between reactant and product atoms can be broken down into three steps:

1. Perform maximum common substructure (MCS) search and map the atoms belonging to the common substructures that are found.
2. For all atoms that were not mapped in the previous step, apply a bipartite matching with cost function.
3. Return the predicted mapping, and any equivalent mappings if requested.

2.3.1 Maximum common substructure (MCS) search (Phase 1)

The reactants and the products of a reaction can be considered as graphs where the atoms and bonds represent vertices and edges, respectively. An MCS algorithm can be used to find all the maximal common subgraphs between reactants and products. Corresponding atoms in the common subgraphs are then assigned equal map indices, the assumption being that these portions of the molecules remain unchanged during the reaction.

The maximum common subgraph problem is NP-hard in general and thus the application of exact MCS algorithms can be problematic when the molecular graphs contain many nodes. Moreover, not all the atoms in a reaction can be mapped in this way, as the product molecules are the result of structural changes that occur during the reaction (bonds formed and broken). Therefore an MCS algorithm should be supported by other strategies in order to map any remaining atoms. An example of this approach can be found in the article by Leber et al.[83].

In our case, the "exact" MCS algorithm provided by the OEChem library [2] (OpenEye Scientific Software) is attempted for a limited time interval (five seconds). This algorithm attempts to find the global maximum common substructure between left- and right-hand sides. If the MCS algorithm returns a match, we continue with phase 2 of the ReactionMap algorithm. Alternatively, if the time interval has elapsed before the exact MCS algorithm completes, it is interrupted, and a second attempt at MCS detection is made using the "approximate" MCS search function. During either of these steps, we may iterate several times in order to recursively find smaller and smaller common substructures in order to match as much of the left- and right-hand sides as possible. Additionally, it is often helpful to relax the specificity of the graph matching employed by the MCS algorithm. That is, we often want to ignore differences in bond order or formal charge in favor of having better total graph coverage from the MCS algorithm. Thus we configure the MCS algorithm to

discern atoms solely based on atomic number (ignoring ring membership, aromaticity, formal charge, and so on), and we configure it to ignore bond order. Once we have a result from the MCS step, we continue with the second phase of the algorithm, in which the remaining atoms must be mapped.

2.3.2 Bipartite matching (Phase 2)

Ideally, the MCS search will return the largest common substructure between reactants and products. The atoms and bonds for which no match is found generally represent "what happened" during the reaction, i.e. the bond rearrangements and consequent atom displacements specific to that particular reaction. On average, after the MCS step, we are left with about 5% of atoms that remain to be atom-mapped.

The approach we use for assigning these atoms is a bipartite matching with cost function. If u_i (v_j) is an atom in the reactants (products) side, we assign a cost $c_{i,j}$ to each of the pairs (u_i, v_j). The problem is then to find the matching that minimizes the total cost of the mapping. In other words, each u_i is assigned to a v_i such that the global cost of the mapping M, $c(M) = \sum_{(i,j)} c_{i,j}$, is the minimum possible. This is the well known "assignment problem" or "stable marriage problem", and can be efficiently solved with combinatorial optimization techniques. Specifically, we employ the Munkres algorithm [96], sometimes called the Hungarian method. For this algorithm, the number of operations scales as $O(n^3)$ (n being the number of vertices of the bipartite graph), which results in a polynomial running time. To further increase the speed of the matching, instead of assigning an infinite cost to mapping different atom types (e.g. mapping a carbon atom to a nitrogen), we apply the Munkres algorithm for each atom type separately.

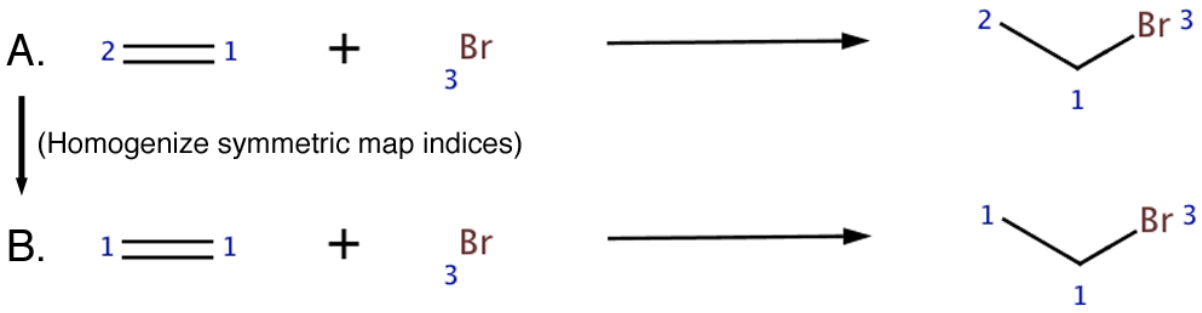


Figure 2.2: Example of symmetric ambiguity in atom mapping. In A, carbon 1 and 2 are symmetrically equivalent. In B, a homogenization step gives them the same map index, indicating they are interchangeable in the final mapping.

2.3.3 Assignment of mapping costs

Mapping costs $c_{i,j}$ are assigned as follows, based on the database of reaction rules described above.

For each reactant atom u_i, we create two state keys $k^{(1)}(u_i)$ and $k^{(2)}(u_i)$. This procedure is repeated for the product atoms v_j to produce $k^{(1)}(v_j)$ and $k^{(2)}(v_j)$. For each pair of atoms (u_i, v_j) we then create the reaction keys (SMILES strings) $r^{(m)}(u_i, v_j) = k^{(m)}(u_i) >> k^{(m)}(v_j)$ for $m = 1, 2$. Finally, we assign the following mapping costs:

1. If $r^{(2)}(u_i, v_j)$ is in our database of reaction rules, then $c_{i,j} = 1$.
2. If $r^{(2)}(u_i, v_j)$ is not in our database of reaction rules, but $r^{(1)}(u_i, v_j)$ is in it, then $c_{i,j} = 5$.
3. If neither $r^{(2)}(u_i, v_j)$ nor $r^{(1)}(u_i, v_j)$ is found in the database, but $k^{(1)}(u_i) = k^{(1)}(v_j)$, then $c_{i,j} = 10$.
4. If none of the above conditions applies, then $c_{i,j} = 100$.

In other words, if during training we observed the atom u_i mapped to atom v_j with a corresponding reaction key $r^{(2)}(u_i, v_j)$, then we assign the lowest possible cost (=1). We

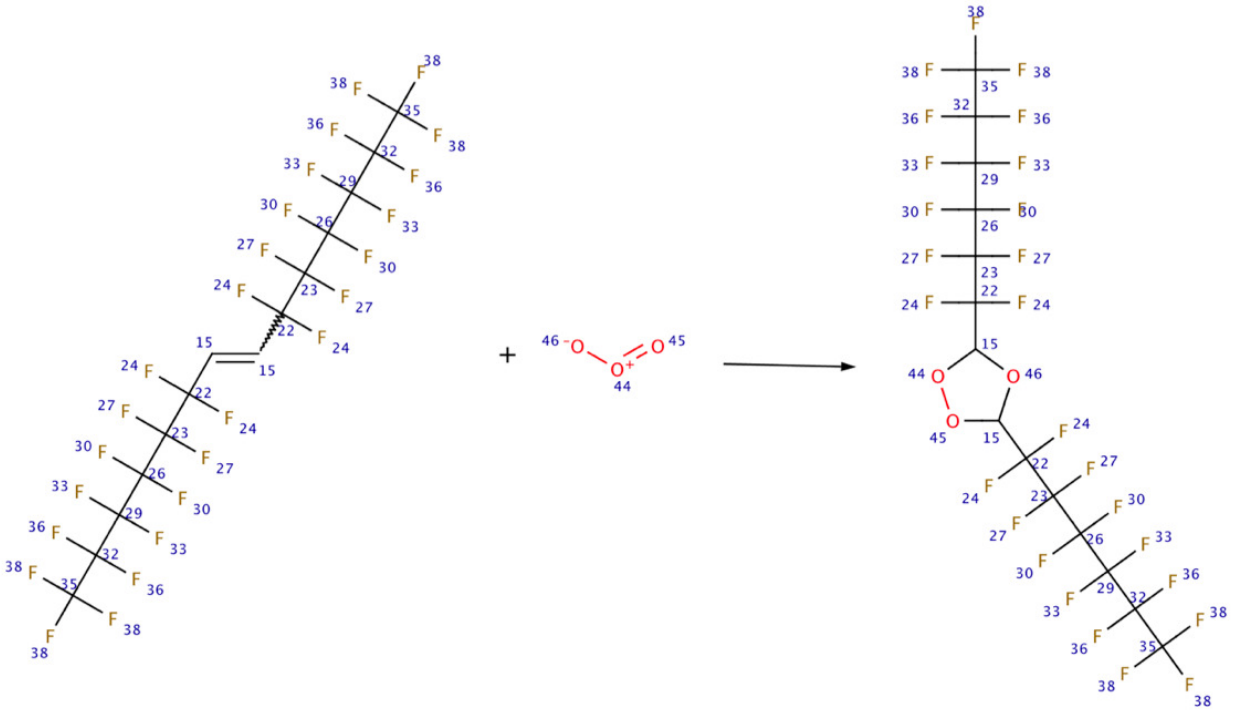


Figure 2.3: Example of a highly symmetric reaction and its homogenized map indices.

assign a slightly higher cost (=5) if $r^{(1)}(u_i, v_j)$ was observed but $r^{(2)}(u_i, v_j)$ was not. A moderate cost (=10) is assigned if neither $r^{(2)}(u_i, v_j)$ nor $r^{(1)}(u_i, v_j)$ was observed but u_i and v_j are the same atom and have the same first neighbors. The highest cost (=100) is assigned for all other cases. Many different cost values were tested (data not shown) and these were found to work best. In this way, we account for the chemical knowledge extracted during the training phase, while allowing a certain degree of freedom for assigning mappings that were never observed.

2.3.4 Symmetrically equivalent mappings

Instead of enumerating and reporting numerous symmetrically equivalent mappings for a given reaction, our approach is to "homogenize" the map indices for symmetrically equivalent atoms. That is, reactant atoms which are symmetrically equivalent are assigned the same map index to indicate their equivalence (Fig. 2). After this homogenization step, the

Table 2.1: Percentage of correctly mapped reactions when using different algorithm components.

Algorithm	**SP05_test (1,000 reactions)**	**SP09_test (17,996 reactions)**
MCS only	78.2%	89.7%
Bipartite matching only	6.2%	2.0%
Combined	96.2%	95.7%

atom mapping proceeds normally. In the final proposed mapping one can see, by locating the homogenized atom map indices, each possible location where members of a symmetrically equivalent group of atoms could be mapped. Thus our simple example from above would be reported as [CH2:1]=[CH2:1].[BrH:3] >> [CH3:1][CH2:1][Br:3]. Map index 2 is omitted, as it was homogenized by assigning an index of 1 to equivalent atoms. This transformation is illustrated in Figure 2, which shows both the original mapping (Fig. 2A), and the homogenized version (Fig. 2B). We show a more complex example of symmetric mapping in Figure 3, which illustrates an actual mapping from our test set.

If multiple mappings are requested, we can enumerate the symmetrically equivalent outcomes of an atom mapping. Considering our example again, we would in this case return both:

[CH2:1]=[CH2:2].[BrH:3] >> [CH3:1][CH2:2][Br:3]

and

[CH2:1]=[CH2:2].[BrH:3] >> [CH3:2][CH2:1][Br:3],

which are the two equivalent mappings for this reaction.

2.4 Results

In the previous sections we outlined the ReactionMap algorithm, and the data used to train and test that algorithm. Here we describe the results obtained from our methods.

Table 2.2: Percentage of correctly mapped atoms when using different algorithm components.

Algorithm	**SP05_test (1,000 reactions)**	**SP09_test (17,996 reactions)**
MCS only	86.3%	93.7%
Bipartite matching only	49.8%	44.9%
Combined	99.4%	98.8%

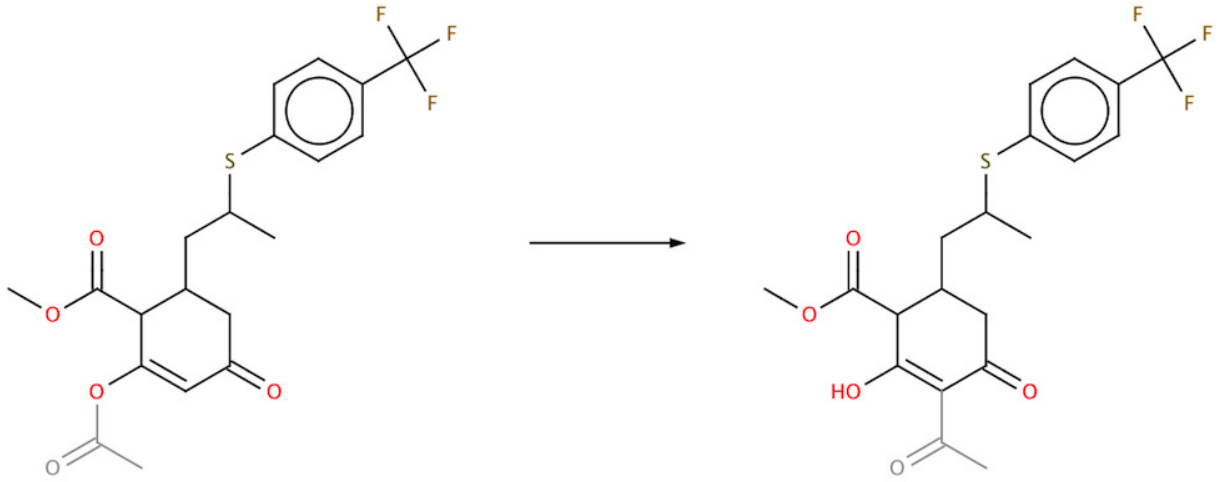


Figure 2.4: Correctly mapped reaction illustrating both maximum common substructure (MCS) search and bipartite matching steps. The grey region indicates atoms that were not mapped during the MCS step, but were correctly mapped during the bipartite matching step.

2.4.1 Accuracy

Mapping results are summarized in Tables 1 and 2. The full algorithm is able to correctly map all the atoms for 96.2% of the reactions in the test SP05_test set, and 95.7% of the reactions in the SP09_test with a mean time per mapping of two seconds. For the remaining reactions, a subset of the atoms is still mapped correctly, on average about 80%. Thus the percentage of correctly mapped atoms across all the reactions is 99.4% for the SP05_test set and 98.8% for the SP09_test set. These percentages are probably a slight underestimate of what can be achieved by our methods since there are probably some errors in SPRESI affecting both the training and testing procedures.

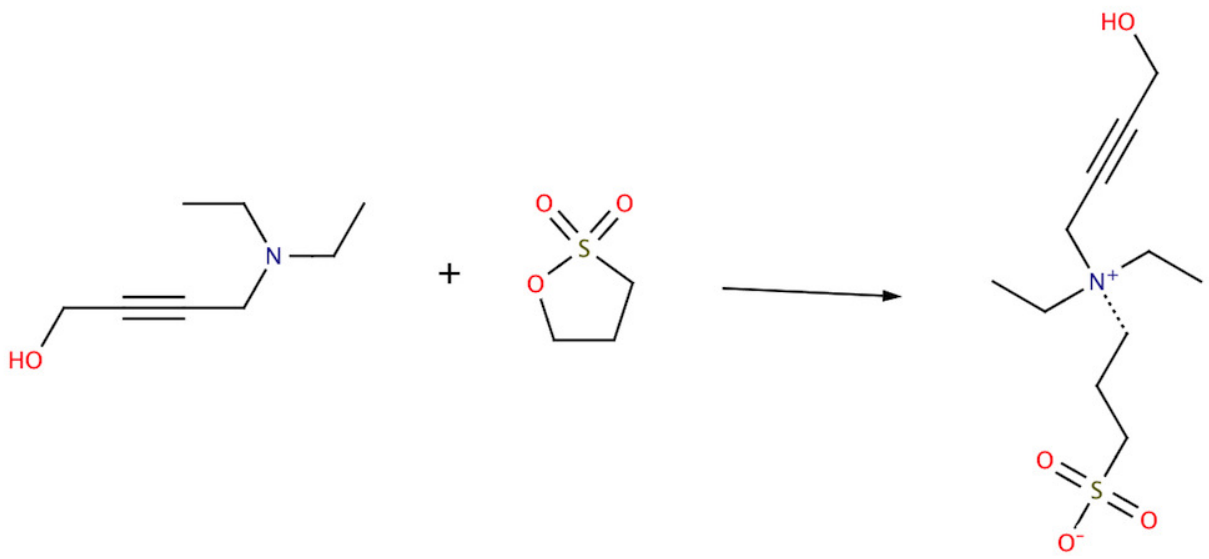


Figure 2.5: Correctly mapped reaction that requires only the MCS step to map all atoms. The dotted line indicates the new bond formed between the two reactant molecules.

The tables also show the effect of each component of the algorithm on overall performance, and why each step is necessary. MCS alone, for instance, can only map about 78.2% of the reactions in the SP05_test set, and the bipartite matching alone can only map 6.2% of the reactions in the same set. Within each component, we also observe a greater variability across datasets. For instance, MCS alone can map 89.7% of the reactions contained in the SP09_test set. The origin of this variability may be in part explained by fluctuations in the composition of these two datasets. For example, the percentages of reactions with exactly one reactant and exactly two reactants are 42% and 54% respectively in the SP05_test set, whereas these percentages are equal to 50% and 45.5% respectively in the SP09_test set. In any case, the MCS component is responsible for most of the accuracy, with a small but significant additional improvement from the bipartite matching component.

ReactionMap succeeds at mapping a wide variety of reactions, including many involving large, complex molecules and structural rearrangements. Figure 4 illustrates a successful mapping which requires both the MCS step and the bipartite matching of the remaining atoms. While many reaction mappings require a combination of MCS and bipartite matching,

Table 2.3: Percentage of reactions correctly mapped in reverse.

SP05_test (1,000 reactions)	**SP09_test (17,996 reactions)**
94.4%	95.0%

Table 2.4: Percentage of atoms in agreement between forward and reverse mapped reactions.

Result	**SP05_test (1,000 reactions)**	**SP09_test (17,996 reactions)**
Correct mapping	99.6%	99.7%
Incorrect mapping	83%	76.2%

some can be performed using MCS alone. Figure 5 illustrates one such situation. Here the two reactants bond together to form a single product. Since there is no structural rearrangement of either reactant, this is an ideal situation for MCS to map the full reaction.

Additionally, we can perform the mapping in the reverse direction, matching products to reactants. For this we use first and second neighbor rules extracted as described above, but in the reverse direction. Table 3 summarizes the percentage of reactions mapped correctly in reverse. We tried many different ways of combining the additional information from the reverse mappings to complement forward mappings, but none led to a robust performance improvement. Nonetheless, performing the reverse mappings is still useful, as we can check the agreement between the two mappings and gain an idea of our confidence in the final mapping. Table 4 summarizes the average percentage of atoms in agreement for correct and incorrect mappings. Specifically, incorrect mappings have, on average, 83% of atoms matching between the suggested forward and reverse mappings. For correct mappings, this agreement averages 99.6%. Figure 6 shows the average mapping success rate versus the percent agreement between forward and reverse mappings for SP09_test. We find that 99% of mappings are correct when agreement is above 95%. For lower levels of agreement, the likelihood of a correct mapping declines rapidly. Additionally, of the 96.2% (SP05_test) and 95.7% (SP09_test) of reactions mapped correctly, 98.1% and 98.9% of these, respectively, have greater than 95% forward-reverse agreement. Thus the majority of correctly mapped reactions are mapped with high confidence.

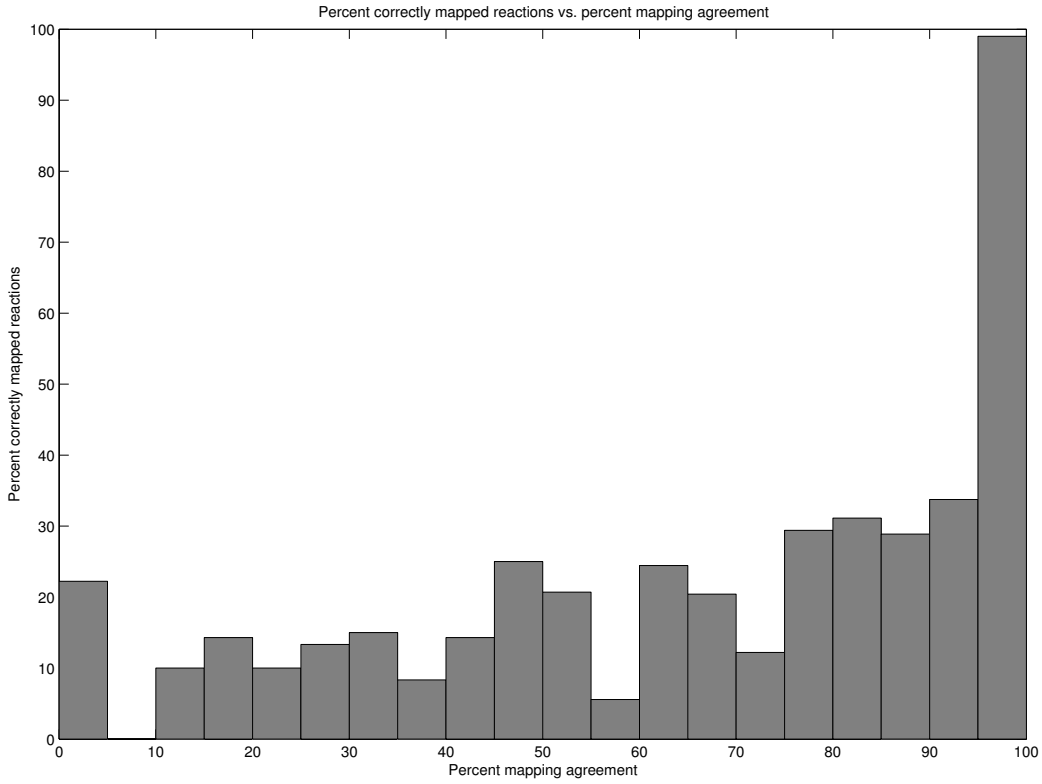


Figure 2.6: Histogram showing the percentage of correct mappings for various degrees of forward-reverse agreement. For agreement greater than 95%, 99% of mappings are correct, while lower levels of agreement have significantly fewer correct mappings on average.

2.4.2 Comparison with other predictors

We compared our results on SP05_test with the performance of three publicly available atom mapping tools on that same test set (Table 5). The software tested included ChemAxon's AutoMapper [3], versions 5 and 6.1, and the DREAM web tool described by First et al.[42]. AutoMapper uses an MCS approach, while DREAM takes an optimization-based approach, using linear programming to minimize bonds broken and formed. AutoMapper versions 5 and 6.1 mapped 60.7% and 86.5% of reactions correctly, respectively, while DREAM mapped 90.3% of the reactions correctly. ReactionMap correctly mapped 96.2% of the reactions.

ReactionMap's average mapping speed on SP05_test was two seconds per reaction. Au-

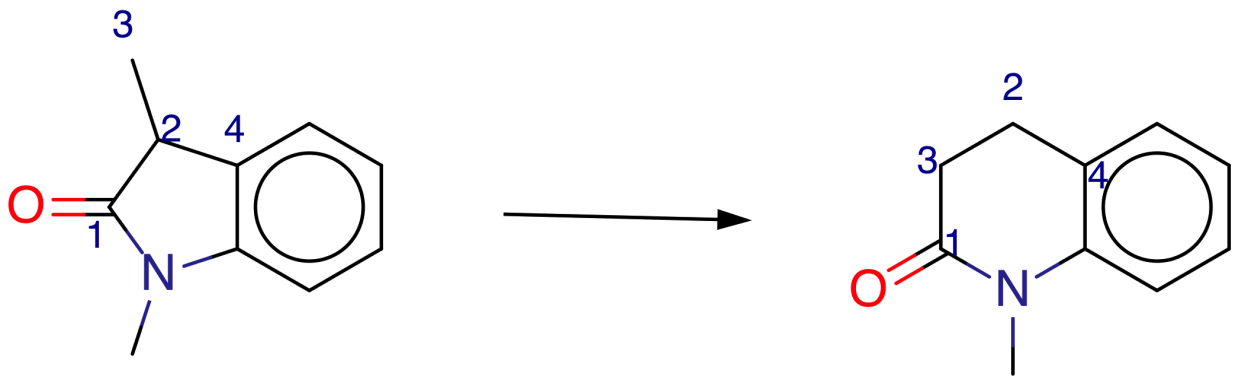


Figure 2.7: Example of an incorrectly mapped reaction due to an error during the MCS step. The correct mapping involves breaking the bond between carbons 1 and 2, followed by bond formation between carbons 1 and 3. Instead, the MCS step returned a mapping that broke bond 2-4 and joined carbons 3 and 4.

Table 2.5: Comparison with other predictors.

Algorithm	**Time per mapping**	**SP05_test (1,000 reactions)**
ReactionMap	2s	96.2%
AutoMapper 5	0.03s	60.7%
AutoMapper 6.1	0.02s	86.5%
DREAM	< 2s	90.3%

toMapper appears to sacrifice mapping accuracy for speed, with average mapping times of 0.02-0.03s per reaction. Since we could only interact with DREAM through a web interface, rather than running tests locally, a direct speed comparison is not available. However, we estimate the mapping time to be slightly less than two seconds per reaction on this data set, based on turnaround time from the DREAM web service.

2.4.3 Error types

Errors made by the algorithm are a function of two factors: how many atoms the MCS search maps between reactants and products, and how well the bipartite matching performs on the remaining unmapped atoms. Errors can occur during either step. Figure 7 shows the case where an error occurs during the MCS step. Here, bond 1-2 breaks, and a new bond forms between carbons 1 and 3. The MCS step is able to find a common substructure

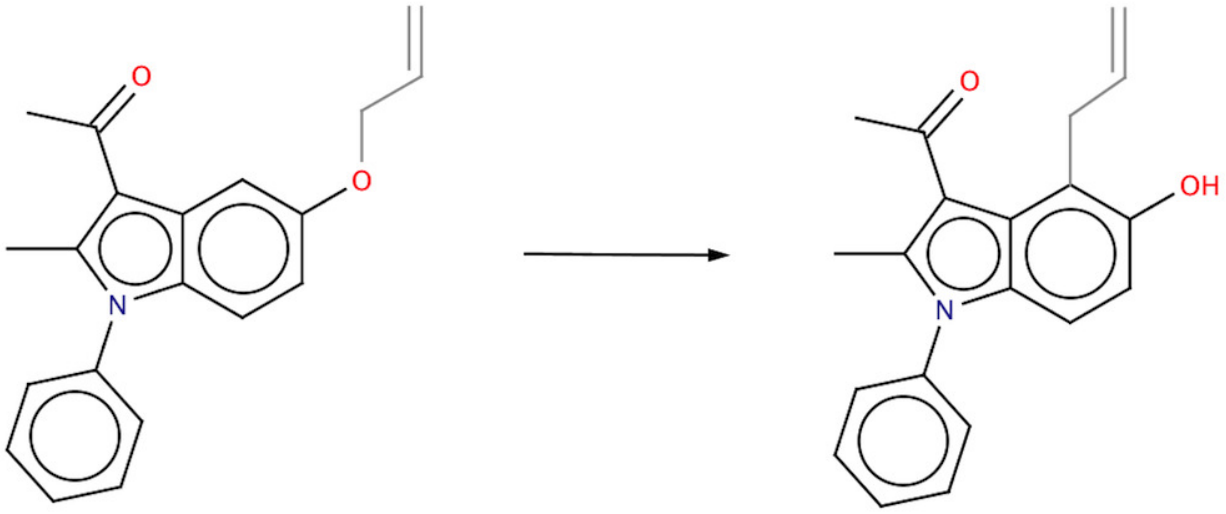


Figure 2.8: Example of an incorrectly mapped reaction due to an error during the bipartite matching step. The grey region indicates atoms that were not mapped during the MCS step. Bipartite matching of these atoms returned the incorrect (flipped) mapping order.

which includes all atoms and all bonds except bond 2-4. Thus it returns a mapping that breaks bond 2-4, and joins carbons 3 and 4. The second possibility – an error during the bipartite matching step – is shown in Figure 8. In this example, a rearranged propene group is mapped with its indices flipped. Figures 9 and 10 show examples of reactions that can be problematic because of interdependent or sequential bond breaking and formation. In Figure 9, the bond between carbons 5 and 6 must be broken to allow formation of bonds 1-3, 8-5, and 4-6. Figure 10 shows a similar situation – the bond between sulfur 1 and carbon 2 must be broken, with the cyclopentadiene rejoining the two intermediates. When given the three reactants present after bond 1-2 is broken (Figure 11), ReactionMap produces the correct mapping.

Though a handful of errors do occur as outlined above, ReactionMap's mapping success rate of about 96% at the reaction level and 99% at the atom level is encouraging considering the huge variety and complexity of reactions tested, and possible directions for future improvements, such as using larger training sets. In future work, mapping accuracy could be improved by developing additional heuristics, or by harnessing additional empirical chemical

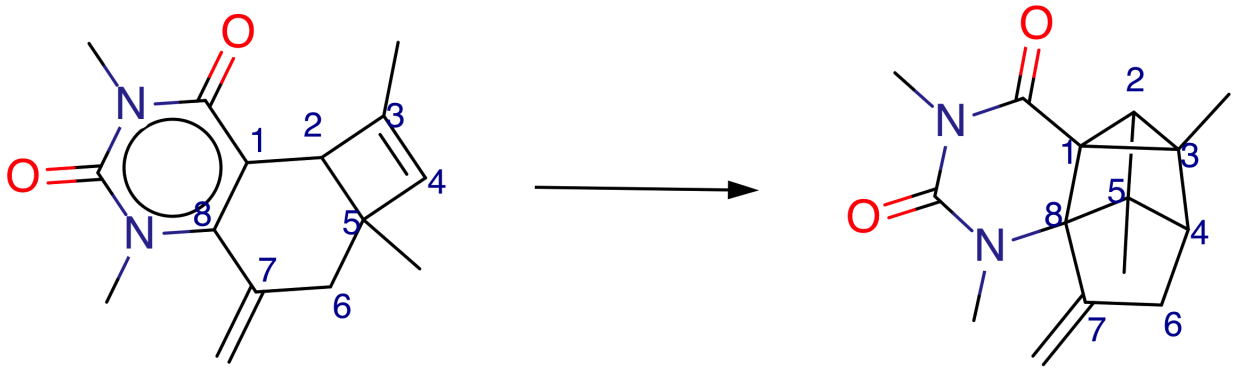


Figure 2.9: Example of a problematic reaction that was mapped incorrectly. The challenge stems from interdependent or sequential bond breaking and formation. Here, bond 5-6 must first be broken to allow formation of bonds 1-3, 8-5, and 4-6.

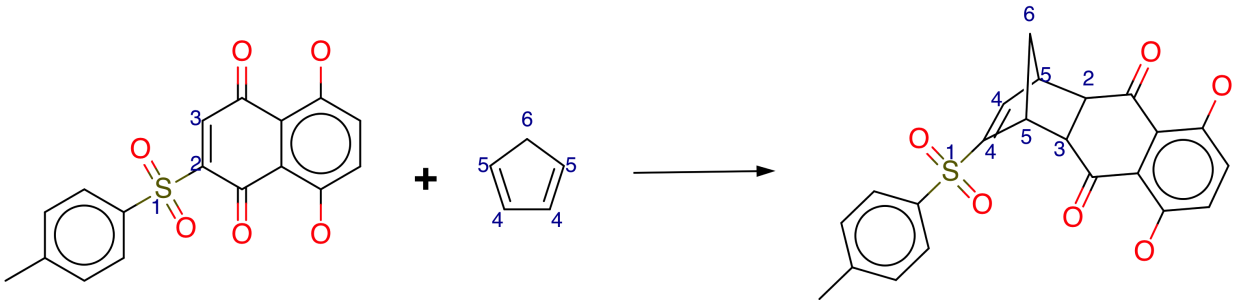


Figure 2.10: Example of another problematic reaction that was mapped incorrectly. Bond 1-2 is broken, yielding two fragments, which are rejoined by the cyclopentadiene. If the algorithm is asked to map the three reactants after bond 1-2 is broken, the result is correct.

data, to address error cases like those described above. The algorithm could also be further optimized for unbalanced reactions and thus should become useful in time as a tool for filtering and cleaning large reaction databases, although these remain largely unavailable for mining purposes. In any case, ReactionMap represents a strong tool for solving the atom mapping problem and opening the door for using more completely atom-mapped chemical datasets for future machine learning and other chemoinformatics endeavors. A ReactionMap server is available on the ChemDB web portal at http://cdb.ics.uci.edu.

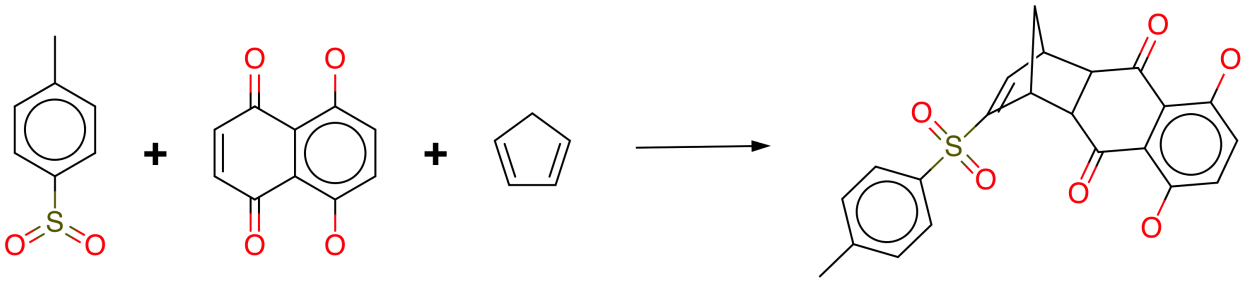


Figure 2.11: Reaction after breaking bond 1-2 in Figure 10. From this point, the reaction is mapped correctly, illustrating the challenge of interdependent bond breaking and bond formation.

Chapter 3

Computational Chemical Reaction Simulation

Atmospheric organic aerosols (OA) represent a significant fraction of airborne particulate matter and can impact climate, visibility, and human health. These mixtures are difficult to characterize experimentally due to their complex and dynamic chemical composition. We introduce a novel Computational Brewing Application (COBRA) and apply it to modeling oligomerization chemistry stemming from condensation and addition reactions in OA formed by photooxidation of isoprene. COBRA uses two lists as input: a list of chemical structures comprising the molecular starting pool and a list of rules defining potential reactions between molecules. Reactions are performed iteratively, with products of all previous iterations serving as reactants for the next. The simulation generated thousands of structures in the mass range of 120500 Da and correctly predicted ˜70% of the individual OA constituents observed by high-resolution mass spectrometry. Select predicted structures were confirmed with tandem mass spectrometry. Esterification was shown to play the most significant role in oligomer formation, with hemiacetal formation less important, and aldol condensation insignificant. COBRA is not limited to atmospheric aerosol chemistry; it

should be applicable to the prediction of reaction products in other complex mixtures for which reasonable reaction mechanisms and seed molecules can be supplied by experimental or theoretical methods.

3.1 Introduction

Atmospheric organic aerosols (OA), airborne particles comprised primarily of organic material, are estimated to contribute up to 50% of the total particulate matter mass at continental midlatitudes and up to 90% at forested areas[65]. OA impact climate and visibility by interacting with solar radiation, atmospheric oxidants, and water vapor[118] and are associated with adverse effects on human health as discussed in Mauderly and Chow[92] and references therein. The climate, visibility, and health effects of OA are dependent on their chemical composition[109, 37, 9]. However, the diversity of OA formation and growth mechanisms makes detailed molecular chemical composition of OA difficult to predict by traditional modeling or characterize by experimental methods. During their residence time in the atmosphere, OA undergo chemical aging processes[113, 64] that further enhance molecular complexity through, for example, the formation of nitrogen-containing organic compounds (NOC)[81, 23, 35]. Recent advances in high-resolution mass spectrometry (HR-MS) have enabled simultaneous detection of hundreds of individual molecules in OA[81, 99, 80, 117, 15]. HR-MS tools are useful in providing the molecular formulas for OA constituents. Ion fragmentation patterns observed in tandem mass spectrometry (MS^n) experiments provide additional information about the structures. However, interpretation of MS^n data is complicated by the presence of multiple structural isomers and lack of sufficiently diagnostic fragmentation patterns. In previous work, reliable structural information could only be extracted for the low molecular weight (MW) species[66, 6] and homologous oligomers[121, 26, 98] present in OA, leaving the structures of the majority of complex

oligomers uncharacterized. HR-MS experiments suggest that in many cases, oligomeric compounds in atmospheric OA are produced from repetitive reactions between the OA constituents[6, 98, 111, 97]. There is strong evidence that condensation reactions such as esterification and aldol condensation[121, 98, 127, 16, 17, 52, 27, 148] and addition reactions such as hemiacetal formation[46, 62, 63] are quite common in both organic aerosols and in aqueous solutions of OA. Advanced computational approaches are needed for predicting the OA composition using a bottom-up approach, in which the starting low-molecular weight compounds are known from experiments, and reaction rules for combining these compounds into oligomers can be unambiguously defined. This work describes the first application of a Computational Brewing Application (COBRA) to modeling oligomerization products observed in the detailed composition of OA. Figure 3.1 shows a diagrammatic representation of COBRA, which is a customizable simulation engine for chemical reactions. Prior work from our group related to predicting the mechanisms of organic chemical reactions has focused on a rule-based approach with broad knowledge of general organic chemistry[28] or a machine learning approach to reaction prediction based on ranking mechanistic steps by productivity[67]. COBRA differs from these approaches in that it is optimized to simulate the evolution of a complex mixture by combinatorial computation of many thousands of chemical reactions between mixture constituents, based on a chosen pool of starting (seed) compounds and specific reaction mechanisms. For the case study presented in this work, the formation of a large set of experimentally observed high-MW oligomeric products in OA derived from the photooxidation of isoprene[98, 97] is modeled by considering chemical transformations of a basic set of monomers through the following oligomerization reactions: esterification, aldol condensation, and hemiacetal formation. We demonstrate that COBRA succeeds at modeling relevant oligomerization chemistry, predicting and visualizing high-MW compound structures, and predicting unique reaction products in OA. More broadly, the COBRA approach is applicable to studying the evolution of a wide range of organic mixtures as long as the starting components and reaction mechanisms can be suggested.

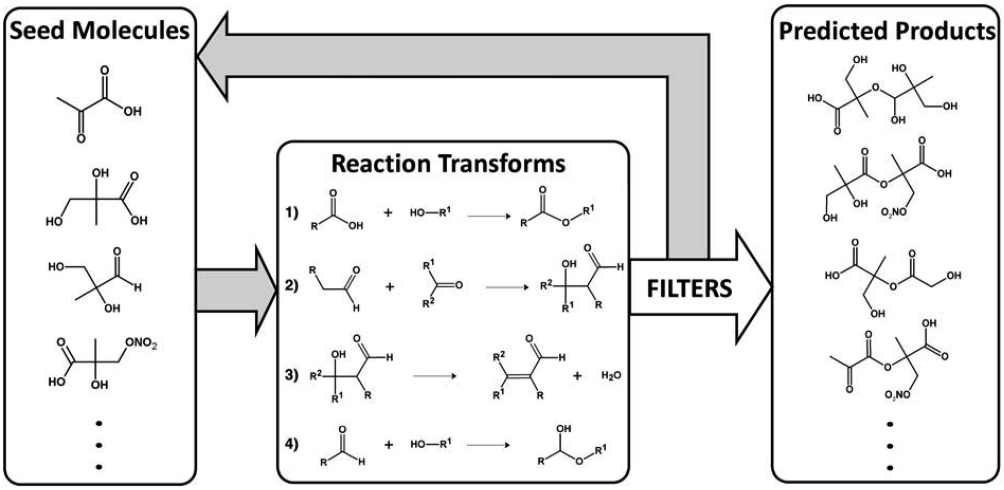


Figure 3.1: A schematic representation of COBRA. COBRA converts a set of seed molecules into a set of predicted products using predefined chemical reaction rules. This process is repeated for several iterations, with predicted products going back into the pool of reacting molecules after each step. The four reactions used in this work are shown in the center panel, and the full list of 27 seed molecules is shown in Table 2.1.

3.2 Materials & methods

3.2.1 COBRA

Chemical structures are input into COBRA using the widely used chemical string representation SMILES[136]. Reaction transforms are defined using the reaction transform language SMIRKS. The SMIRKS language is a superset of SMILES and can be used to define reaction transforms to an arbitrary degree of specificity. We leverage the programming language Python in conjunction with OpenEyes OEChem library[2] to process the input SMILES and SMIRKS into predicted products. Standard valence rules are explicitly programmed into the reaction transforms in order to predict chemically meaningful structures. Specifically, R-groups in the reactions shown in Figure 1 were allowed to be alkyl or acyl groups but not hydrogen alone. Filters can be imposed to prevent compounds with certain properties from participating in reactions or from being included in the pool of final products. The simulation includes the following iterative steps. 1) For each molecule and each pair of molecules

Table 3.1: Seed molecules used for the COBRA simulations described in this work.

Formula	Name	Structure
$C_2H_4O_2$	glycolaldehyde	
$C_2H_4O_2$	acetic acid	
$C_2H_4O_3$	glycolic acid	
$C_3H_4O_2$	methylglyoxal	
$C_3H_4O_2$	acrylic acid	
$C_3H_6O_2$	hydroxyacetone	
$C_3H_6O_2$	propanoic acid	
$C_3H_4O_3$	pyruvic acid	
$C_3H_6O_3$	lactic acid	
$C_3H_4O_4$	3-hydroxypyruvic acid	
$C_3H_4O_4$	2-hydroxy-3-oxopropanoic acid	
$C_4H_6O_2$	methacrylic acid	
$C_4H_6O_3$	2-methyloxopropanoic acid	
$C_4H_8O_3$	2-methylglyceraldehyde	

Formula	Name	Structure
$C_4H_8O_4$	2-methylglyceric acid	
$C_4H_8O_4$	erythrose/threose	
$C_4H_6O_4$	methylmalonic acid	
$C_4H_6O_4$	2-hydroxy-2-methyl-3-oxopropanoic acid	
$C_4H_6O_5$	hydroxy(methyl) propanedioic acid	
$C_4H_7O_6N$	2-methylglyceric acid 3-nitrate	
$C_4H_8O_5$	2,3,3-trihydroxy-2-methylpropanoic acid	
$C_5H_8O_3$	2-hydroxy-2-methylbutanedial	
$C_5H_8O_4$	methylsuccinic acid	
$C_5H_8O_4$	2-hydroxy-2-Me-4-oxobutanoic acid	
$C_5H_8O_5$	2-hydroxy-2-methylbutanedioicacid	
$C_5H_{10}O_3$	2,4-dihydroxy-2-methylbutanal	
$C_5H_{10}O_5$	2,3,4-trihydroxy-2-Mebutanoic acid	

in the reactant pool, all unimolecular reaction transforms and all bimolecular transforms, respectively, are applied. Identical reactions already performed in the previous iterations are recognized and skipped. 2) The product molecules are filtered. For this simulation, we automatically filter out any molecule with greater than 40 heavy (C, O, N) atoms. This filter restricts the molecular weights of the products below approximately 500 Da, the limit at which molecular assignments made with 0.1 mDa accuracy can be uniquely determined for $C_cH_hO_oN_nS_s$ species[72]. This restriction significantly decreased the model run time while still capturing the chemistry of the most abundant compounds in isoprene high-NOx SOA[97]. 3) Molecules that were not filtered out during the previous step are added back into the reactant pool, and the system returns to step one. The simulation is complete either after the requested number of iterations has passed or after a specified number of unique reactions has been simulated (30,000 reactions for the full simulation and the exclusion simulations, in this study). Results can then be conveniently visualized and searched for specific chemical structures matching any specified structural criteria. Additionally, we can compute the sequence of reactions that led to a given products formation.

3.2.2 Experimental Methods

Isoprene OA was photochemically generated in a 5 m3 Teflon chamber, as described previously[98]. Samples were generated in dry air, under high-NOx (VOC:NOx <1) conditions and in the absence of inorganic seed particles. H2O2 was used as an OH precursor. The initial mixing ratio of isoprene was 250 ppbv (parts per billion by volume). Blank samples were produced in an identical manner as OA samples, without UV radiation. Samples were collected on Teflon filters (0.2 m pore size, Millipore), vacuum sealed, and frozen prior to analysis. Negative ion mode direct injection ESI (tip voltage 4 kV) was used as an ionization method for solvent-extracted OA samples. The solvents used for analysis were water and acetonitrile (both HPLC grade, Fluka) at a 1:1 volume ratio. Mass analysis was done with a high-resolution linear ion trap (LTQ-) Orbitrap (Thermo Corp.) at Pacific Northwest National Lab (PNNL) Environmental Molecular Science Laboratory facility (EMSL), with a mass resolution of 60,000 m/m at m/z 400. MS^n studies were performed in the LTQ, with mass selection in the 0.5 m/z range and collision-induced-dissociation energies of 2040 energy units. Product ions of MS^n were analyzed in the Orbitrap.

3.3 Results & discussion

Table 3.1 lists the seed molecules for the simulation. These low-MW species have been identified as important building blocks in the formation of isoprene photooxidation OA[97]. The majority of these molecules have been detected in isoprene OA[25] and are primarily multifunctional carbonyl, alcohol, and carboxyl compounds derived from the oxidation of isoprene with the hydroxyl radical (OH)[121, 97, 40, 51, 120, 74, 32, 103, 123, 124]. Some of these molecules have sufficiently high vapor pressure to exist primarily in the gas phase. For example, glycolaldehyde has a room temperature vapor pressure of 0.028 Torr[105] and may participate in aerosol mass-growth reactions initiated in the gas- or heterogeneous phase.

Note that glyoxal is not included among the seed molecules used in these simulations because it was not found among the monomer units inferred from mass spectrometry of isoprene oxidation products as described in ref 22. The 3-nitrate ester of 2-methylglyceric acid (2MGA)[121, 98, 27, 123] represents the sole NOC in the list of seed molecules. Available information about NOC monomers is limited due to the small number of studies of NOC composition in the literature[97].

COBRA was used to simulate the evolution of a virtual OA mixture by applying the four reaction transforms shown in Figure 1 to the set of 27 starting molecules in Table 3.1. The resulting product pool contained 135,107 predicted structures after performing 30,000 unique reactions, representing a total of 758 unique elemental formulas of the type $C_cH_hO_oN_n$ with MW <500 Da. The 2 orders of magnitude difference between the number of structures and elemental formulas reflects a large number of predicted structural isomers. The experimental HR-MS data contains 464 neutral elemental formulas, and 323 of these formulas (70%) were predicted by COBRA. Figure 3.2 shows the experimental HR-MS spectrum separated into two panels. The first panel shows all HR-MS peaks in the experimental spectrum that were predicted by COBRA (Figure 2A), and the second one shows peaks that did not appear in the COBRA output (Figure 2B). As Figure 2 demonstrates, the 70% fraction of the experimentally observed peaks predicted by COBRA represents the most abundant compounds with MW <500 Da. The remaining 30% fraction has average peak intensities that are more than an order of magnitude smaller compared to the average peak intensities associated with the successfully predicted compounds, suggesting that COBRA captures the essential chemistry producing oligomers in isoprene high-NOx OA. The high degree of overlap between the predicted and observed formulas confirms that the oxygenated hydrocarbon seed molecules proposed in our previous work are the relevant oligomer building blocks in this type of OA[97]. The fact that a relatively small set of modeled reactions was sufficient to account for roughly 70% of the HR-MS peaks suggests that the oligomerization chemistry in isoprene OA is well constrained.

The fraction of the experimentally observed compounds that are correctly predicted by the simulation can be a misleading metric of success in some cases. For example, if the simulation were significantly overdefined, and generated every possible $C_cH_hO_oN_n$ formula (˜105) allowed by the valence rules, this fraction would become 100%. Therefore the reverse comparison of the fraction of predicted peaks that show up in the experiment is just as important. In the present case, 43% of predicted formulas correspond to compounds detected by HR-MS, and the remaining 57% are not observed. This level of agreement can be viewed as good, considering that the simulation does not use any kinetics restrictions and predicts the formation of compounds that may be below the limit of detection of HR-MS.

It is remarkable that a total of 62 NOC molecular formulas were predicted stemming from 2MGA-3-nitrate alone, representing 41% of the total NOC observed in the isoprene OA sample in the m/z 120500 range (Figure 2A). Note that this particular simulations ability to predict NOC compounds is limited because only a single NOC seed molecule is included. Nonetheless, the data confirm that 2MGA-3-nitrate is a prolific oligomer building block in isoprene OA that produces a variety of products through oligomerization in the condensed phase[121, 98, 27, 123].

Since structural complexity increases with molecular size, the number of structural and stereo isomers that can be produced by oligomerization reactions also grows exponentially with increasing mass of the individual products (Figure 2C). For the 323 experimentally observed molecular formulas, COBRA predicts 102,650 unique structures, with the number of isomers ranging from 1 to nearly 3,000. The apparent decrease in the number of hits as MW approaches 500 Da is a result of filtering; we artificially limit structures in the product pool to those with less than 40 heavy atoms and do not consider structures with molecular weights in excess of 500 Da. Despite the large number of predicted isomers, the simulation results help constrain possible structures, especially for lower-MW compounds. For example, over 1,000 unique molecular structures associated with a molecular mass of 142.063 Da, assigned

to $C_7H_{10}O_3$, can be retrieved from Internet-based chemical inventories (e.g., SciFinder). In contrast, COBRA predicts only two structures of C7H10O3, thereby dramatically reducing the complexity (Table 2). This represents more than a one hundred-fold increase in the level of confidence in assigning molecular structures to a given formula obtained by HR-MS. Even at high masses, COBRA predictions remain useful because the structural information for high-MW compounds is not readily available from chemical databases, and this type of modeling is a practical step toward understanding the possible structures of large oligomers.

A few published studies used MS^n fragmentation data to determine structures of the oligomers in isoprene high-NOx SOA[121, 26, 97]. All molecular structures described in the previous studies match structures predicted by COBRA. To further test the validity of the COBRA predictions, we made a comparison to MS^n studies previously performed by our group of NOC oligomers observed in SOA generated from isoprene photooxidation[97]. Figure 3.3 shows the predicted structures of two NOC oligomers, $C_8H_{13}NO_9$ (267.059 Da) and $C_{14}H_{20}NO_{13}$ (411.101 Da), that are most consistent with MS^n experiments. The MS^n experiments were performed using the negative ion mode; therefore, the molecular masses of the detected compounds are those of deprotonated molecules. Oligomer fragmentation is often characteristic of the functional groups and monomer units present within the compound. For example, HNO_3 and CH_3NO_3 are characteristic MS^n losses from organic nitrates, and $C_4H_6O_3$ is a characteristic loss from the carbonyl ester unit of 2MGA[121, 97].

For a given molecular formula, the proposed structures based on MS^n experiments described in Nguyen et al. match at least one of the structures predicted by COBRA. We note that structural isomers may not have distinct fragmentation patterns, and therefore it is possible that the experimental MS^n data represents several structural isomers predicted by COBRA. COBRA predicted four isomeric structures for $C_8H_{13}NO_9$, and at least one predicted structure is consistent with the fragmentation pattern produced by MS^n studies. For $C_{14}H_{21}NO_{13}$, at least one out of 44 predicted structures is consistent with the MS^n fragmentation pattern.

Therefore, it is reasonable to hypothesize that other structures predicted by COBRA may be present in the OA, perhaps but not necessarily, in lower quantity than the most abundant isomer.

To elucidate which seed molecules were the most important as oligomer building blocks for isoprene OA, we performed 27 exclusion simulations, each with a different seed molecule omitted. Each exclusion simulation was limited to the same number of reactions (30,000). The total number of experimental molecular formulas recovered in each case was compared against a simulation in which no seed molecules were omitted. The exclusion percent was calculated as a percent reduction in the number of predicted formulas matching experimental formulas. The highest exclusion percent (19%) was for the 3-nitrate ester of 2-methylglyceric acid, consistent with its special role of being the sole NOC precursor. Specifically, the removal of 2MGA-3-nitrate eliminated all NOC oligomers from the product pool. Removal of other seed molecules resulted in smaller variation in the number of predicted formulas, ranging from 6 to 6%. Note that because of the fixed number of reactions in these simulations, an exclusion of a less significant contributor to chemistry can actually lead to a slight increase in the total number of predicted formulas, yielding a negative exclusion percent. None of the nitrogen-free compounds appeared to stand out from the rest, suggesting some redundancy in the pool of seed molecules. The redundancy in seed molecules lowered the percent overlap between simulation and experiment. These results highlight the need for future work in determining missing precursors and reaction mechanisms. In general, this type of exclusion analysis is useful for assessing the contribution of a single molecule to the total product pool, e.g. the importance of glyoxal in heterogeneous reaction with various OA constituents,[54, 88, 89, 150] or the contribution of first vs second generation VOC oxidation products to producing condensable organic compounds.

We performed additional exclusion simulations, in which one of the four reactions shown in Figure 3.1 was disabled. The results of the simulations based on the three remaining

Table 3.2: Results from an exclusion analysis, in which one of the four reactions shown in Figure 3.1 is disabled.

reaction transform	reaction type	percent reduction in experimentally observed formulas	percent reduction in theoretically predicted formulas	percent reduction in theoretically predicted structures
1	esterification	40.9	12.8	37.8
2	aldol condensation	0.3	0.7	5.6
3	elimination of water from aldol condensates	0	1.5	0.01
4	hemiacetal formation	10.2	−0.1	46.7

reactions were compared to the results of the full simulation. Table 3.2 lists the percent reduction in the number of experimentally observed molecular formulas recovered by the simulation, the percent reduction in the number of theoretically predicted molecular formulas, and the percent reduction in the number of theoretically predicted structures generated by the simulation. The esterification reaction (Reaction 1) played the most significant role in recovering experimentally observed formulas, with an exclusion percent of 40.9%, followed by the hemiacetal formation reaction (Reaction 4) with an exclusion percent of 10.2%. The esterification and hemiacetal formation reactions also made the largest contributions to the total number of theoretically predicted structures. Exclusion of esterification from the simulation reduced the number of structures in the COBRA output by 37.8%, while exclusion of hemiacetal formation resulted in a 46.7% decrease. Although hemiacetal formation is prolific at generating theoretical structures, the results demonstrate that esterification is the most important reaction in promoting oligomerization in isoprene OA, while hemiacetal formation is of secondary importance. These results are consistent with a previous study, wherein reducing aerosol liquid water content changed the composition of isoprene high-NOx SOA most drastically by hindering the esterification reaction[98]. They are also consistent with the fast and efficient reactions of OA compounds containing carboxylic acid groups and carbonyls with alcohols observed in Bateman et al.[18]. In contrast, aldol condensation (Reactions 2 and 3) does not appear to play a significant role in forming the oligomers. Removal of either Reaction 2 or Reaction 3 resulted in a small reduction in the number of

theoretically produced structures and recovered HR-MS formulas.

Once a pool of COBRA-generated simulation products is available, it can easily be mined for desired products or classes of products. For example, we are interested in the potential for isoprene photooxidation to generate ,-unsaturated aldehydes, a class of genotoxic compounds found in cigarette smoke and air pollution[146, 147, 21, 128, 79, 8]. The most common atmospheric ,-unsaturated aldehydes with adverse health effects are gas-phase species such as acrolein (C_3H_4O)[70]. However, adverse health effects are also strongly linked to inhalation of particulate matter[92, 112, 36, 38, 108, 116], which contains more complex and poorly characterized organics. COBRA predicted 6,131 structures corresponding to ,-unsaturated aldehydes in the isoprene photooxidation SOA data set. This large set of structures can be filtered further, by mass, O/C ratio, or any other property required to yield a focused set of predicted structures. Applying a computational tool like COBRA along with the molecular-level experimental characterization of products is a powerful first step to identifying condensed-phase atmospheric toxins.

One limitation of COBRA is that it does not currently include kinetics information and therefore cannot model the relative abundances of individual products. This limitation may be important in cases where the product branching ratios from a particular reaction are dissimilar or dependent on atmospheric conditions, in which case the results from COBRA will artificially enhance the importance of noncompetitive products. For example, in the absence of nitrogen oxides (NOx) the same OH-initiated oxidation chemistry produces more hydroperoxides and fewer carbonyl products[12]. However, a scaling factor can be incorporated in the transformation rules if the branching ratios are known or can be estimated with ab initio methods. For the comparison with the ESI-based MS data, this is not a major limitation because there are no simple relationships between the detection efficiency in ESI and the concentration of the analyte species[125].

Applying COBRA to the simulation of oligomerization in isoprene OA is a good example of

the utility of computational tools for understanding complex natural mixtures and verifying experimental data. An important strength of COBRA is its ability to handle a complex simulation and generate a large number of predicted compounds. While the combinatorial explosion of generated structures means computation time is currently on the order of several days, this time can be significantly reduced in future work by using parallel computing algorithms. If pertinent experimental data are available, direct comparison between the observed and predicted compounds can be made to better constrain chemistry and chemical structures. Furthermore, as COBRA reduces the number of possible structures available for a given chemical formula, the predicted structures may be useful for discriminating isobaric species in a mass spectrum. Applying a small set of transformation rules to predict an experimentally determined high-resolution mass spectrum may become a powerful tool for better understanding the chemistry of complex systems comprised of hundreds or thousands of individual compounds.

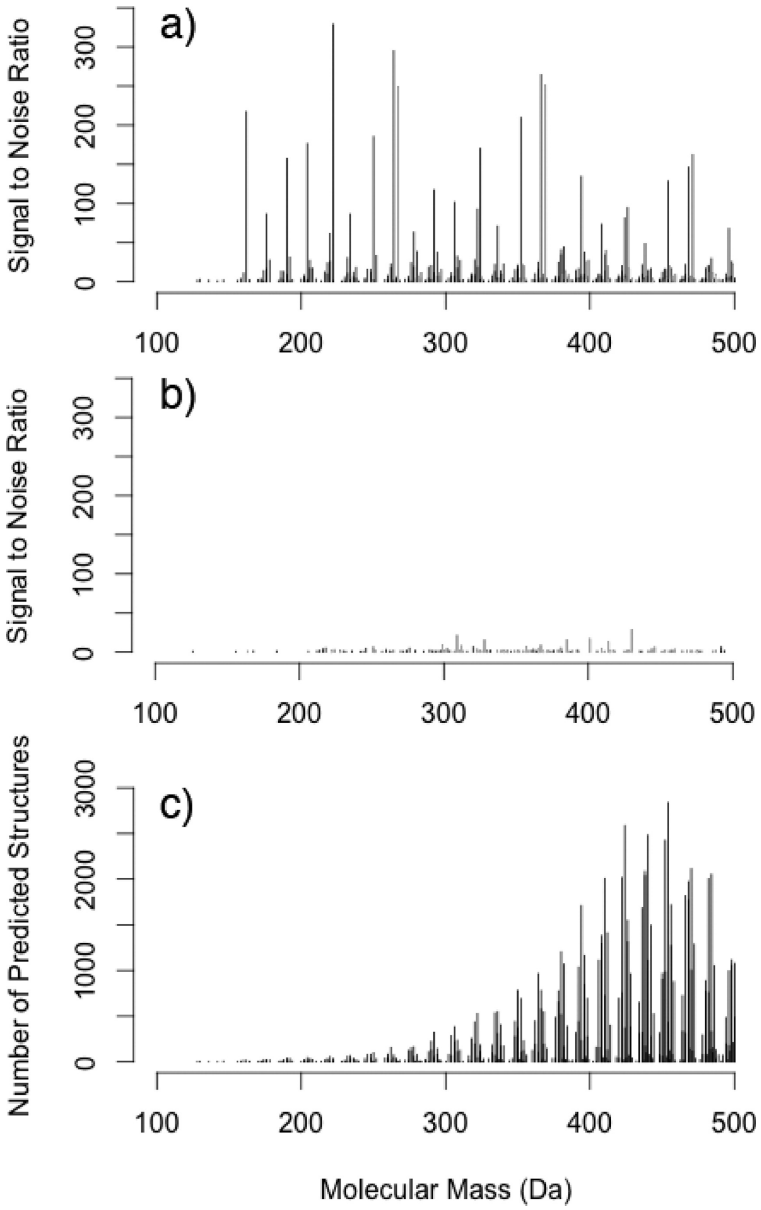


Figure 3.2: Comparison between the HR-MS experiment and COBRA predictions. Panel (a) shows the HR-MS peaks that are predicted by COBRA (70%). Panel (b) shows the remaining 30% of HR-MS peaks that are not predicted by our simulation. In both panels, NOC peaks are colored green. Panel (c) indicates the number of isomeric structures predicted at each observed molecular mass.

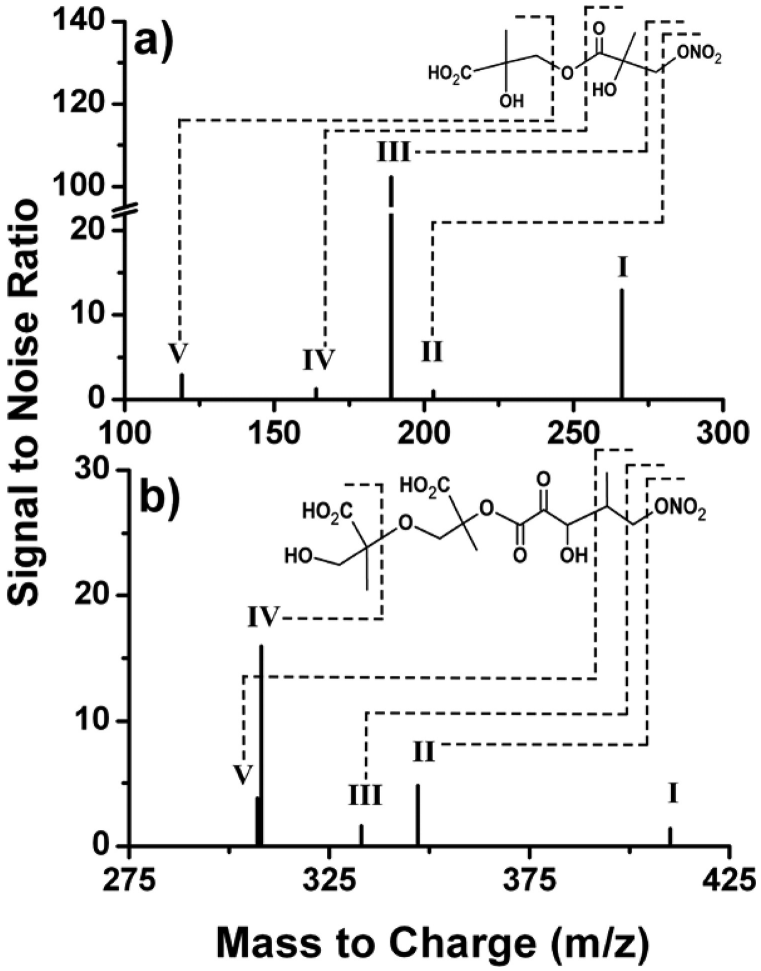


Figure 3.3: Experimental MS^2 fragmentation spectra for precursor (I) ions (a) $C_8H_{12}NO_9$ and (b) $C_{14}H_{20}NO_{13}$. Two structures predicted by COBRA that are consistent with the fragmentation patterns are overlaid on the spectra. Product ions are obtained from fragmentation at the dashed bonds (after losses of neutral molecules). Product ions for (a) are as follows: II. m/z 203.056 (I-HNO_3); III. m/z 189.040 (I-CH_3NO_3); IV. m/z 164.020 (I-$C_4H_6O_3$); V. m/z 119.035 (I-$C_4H_5NO_5$). Product ions for (b) are as follows: II. m/z 347.098 (I-HNO_3); III. m/z 333.082 (I-CH_3NO_3); IV. m/z 308.062 (I-$C_4H_6O_3$); V. m/z 307.082 (I-$C_3H_5NO_3$).

Chapter 4

COBRA Modeling of Atmopheric Squalene Oxidation

Squalene is a major component of skin and plant surface lipids and is known to be present at high concentrations in indoor dust. Its high reactivity toward ozone makes it an important ozone sink and a natural protectant against atmospheric oxidizing agents. While the volatile products of squalene ozonolysis are known, the condensed-phase products have not been characterized. We present an analysis of condensed-phase products resulting from an extensive oxidation of squalene by ozone probed by electrospray ionization (ESI) high-resolution mass spectrometry (HR-MS). A complex distribution of nearly 1300 peaks assignable to molecular formulas is observed in direct infusion positive ion mode ESI mass spectra. The distribution of peaks in the mass spectra suggests that there are extensive cross-coupling reactions between hydroxy-carbonyl products of squalene ozonolysis. To get additional insights into the mechanism, we apply a Computational Brewing Application (COBRA) to simulate the oxidation of squalene in the presence of ozone, and compare predicted results with those observed by the HR-MS experiments. The system predicts over one billion molecular structures between 0 and 1450 Da, which correspond to about 27000 distinct elemental formulas.

Over 83% of the squalene oxidation products inferred from the mass spectrometry data are matched by the simulation. The simulation indicates a prevalence of peroxy groups, with hydroxyl and ether groups being the second-most important O-containing functional groups formed during squalene oxidation. These highly oxidized products of squalene ozonolysis may accumulate on indoor dust and surfaces and contribute to their redox capacity.

4.1 Introduction

Squalene (see Figure 4.1 is a naturally occurring product found in human and plant lipids[44]. This unsaturated, nonvolatile triterpene reacts readily with ozone, and is the most abundant ozone-reactive constituent of human sebum[133, 141]. Squalene accounts for 10-12% of adult skin lipids, making it one of the major lipids found on the surface of the human skin[143]. This contributes to squalenes presence in human skin flakes, which are a major component of indoor dust. Lipids in skin and plant oils, such as squalene, are a natural form of protection against oxidizing agents found in air, with squalene shown to account for about 40% of ozone removal by human skin and hair[143, 144].

Squalene is highly reactive toward ozone due to its six unconjugated carbon-carbon double bonds[44, 140]. It reacts with ozone based on the Criegee mechanism, whereby ozone attacks a double bond to form a primary ozonide (POZ), which then decomposes into a carbonyl and a carbonyl oxide, also known as a Criegee intermediate[144, 34]. Carbonyls, carboxyls, and -hydroxy ketones are the main functional groups in squalene ozonolysis products, largely arising from the isomerization and decomposition reactions of its carbonyl oxides, RCHOO and RC(CH3)OO[144].

In view of its importance as an ozone sink in human occupied environments, the reaction between squalene and ozone has been increasingly studied in recent years[133, 141, 143, 144,

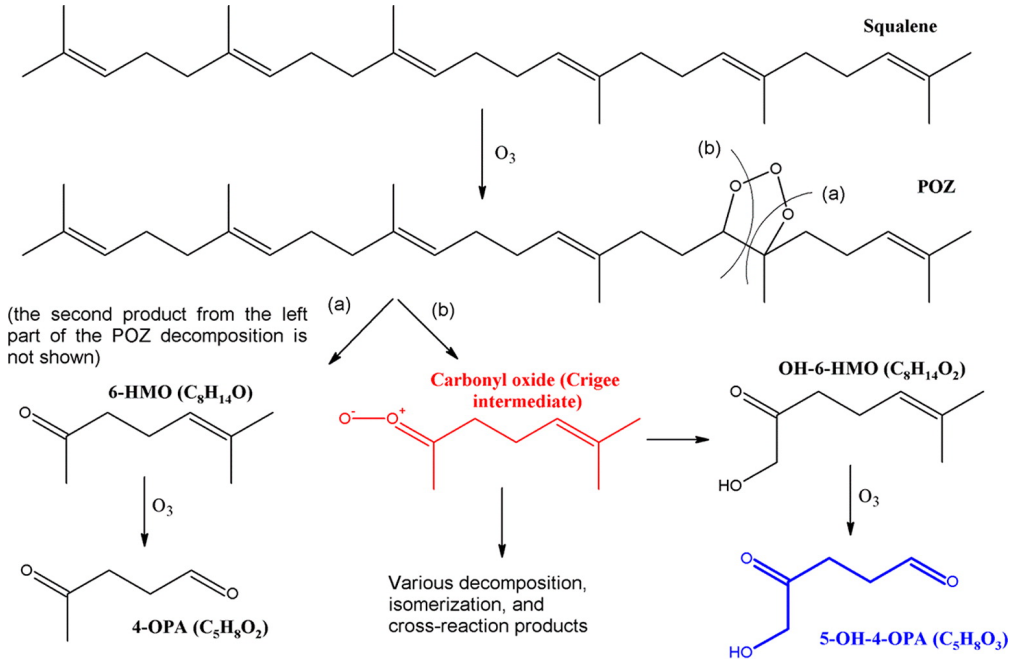


Figure 4.1: Structures of squalene, one of the six possible primary ozonides (POZ) squalene can form upon reaction with ozone, and selected volatile products resulting from the decomposition of the unstable POZ.

140, 106, 45, 134, 142, 145]. Human skin lipids reacting with ozone have been examined in real and simulated environments, and results indicate that humans are a significant sink for ozone present in such environments, accounting for up to half of ozone removal[133, 144, 142, 145]. For example, simulated office and aircraft cabin studies have examined the products of squalene ozonolysis due to human presence[144, 142, 145]. In one simulated aircraft cabin study, it was found that ozone reacting with squalene on human skin and clothes was responsible for a 40-60 ppb increase in the concentration of certain acids, aldehydes, and ketones and >55% ozone removal[142].

The reaction of squalene with ozone forms several primary products, including acetone, hydroxyacetone, 6-methyl-5-hepten-2-one (6-MHO), 1-hydroxy-6-methyl-5-hepten-1-one, and geranyl acetone, as well as a number of secondary products, including 4-oxopentanal (4-OPA), 5-hydroxy-4-oxopentanal (5-OH-4-OPA), 1,4-butanedial (succinic dialdehyde), 4-methyl-8-oxo-4-nonenal (4-MON), 4-methyl-4-octene-1,8-dial (4-MOD), 4-oxopentanoic acid, and

4-oxobutanoic acid[44, 143, 144, 90, 91]. In studies where squalene directly from human skin was scrubbed onto glass wool and exposed to 100 ppb ozone, acetone (˜17 ppb) and 6-MHO (˜14 ppb) represented the most abundant first-generation products, while 4-MON (˜4 ppb) and 4-OPA (˜3.5 ppb) represented the most abundant second-generation products[144]. Some of these products are represented in Figure 4.1.

Many of these products have also been detected in forest air and are believed to result from reactions of ozone with plant surfaces containing squalene and related terpenes[90, 91]. Squalenes high degree of unsaturation allows for the formation of tertiary and higher-order reaction products, such as ethanedial (glyoxal)[144, 140]. The volatility of these poorly characterized oxidation products is low, and they may be present on surfaces of indoor dust particles contributing to their redox activity and potential for irritancy[7].

In this study, squalene is exposed to relatively high levels of ozone to simulate its prolonged oxidation, yielding a highly complex mixture of condensed-phase oxidized products. We utilize high-resolution mass spectrometry (HR-MS) to examine these oxidation products in detail and discuss the underlying mechanism. To further aid in understanding this complex system, we use a Computational Brewing Application (COBRA) to model the products of squalene oxidation by ozone[43]. COBRA is optimized to model the evolution of a complex chemical mixture by simulating millions of reactions between compounds in the mixture, after sets of starting products and chemical reaction mechanisms are specified. By studying these computationally predicted products and comparing them with the experimental data, we can better understand the molecular and structural characteristics of squalene oxidation products present on aged indoor surfaces and dust.

4.2 Materials & methods

4.2.1 Experimental Methods

Approximately 50 L of pure squalene (Sigma-Aldrich, >98% purity) was uniformly distributed over the surface of a 47 mm polyvinylidene fluoride (PVDF) filter (Millipore HVLP04700, 0.45 m pore size). The filter was sealed inside a stainless steel filter holder (Millipore XX4404700) and exposed to a 2 SLM (standard liters per minute) flow of oxygen containing 50 ppm ozone for a period of time, typically 1 h. Such a high concentration was chosen to simulate the effects of long-term exposure of squalene to ambient levels of ozone. In terms of integrated exposure, it is roughly equivalent to 1000 h of exposure to 0.050 ppm ozone; however, we cannot exclude the possibility of nonlinear effects on the mechanism associated with the very high ozone mixing ratio used in this study. In the first set of control experiments, a blank filter was exposed to the same flow of ozone. In the second set of control experiments, squalene was applied to the filter but not exposed to ozone. All experiments were done under dry conditions and in the absence of other copollutants, such as NOx. In actual indoor environments, water and NOx could potentially affect the reaction mechanism and product distribution. Therefore, comparisons of the observed and predicted products of oxidation of squalene with compounds found indoors should be done with care.

The reaction products were extracted from the filter with 5 mL acetonitrile containing 50 M NaCl to help improve ionization efficiency in the positive ion mode. The resulting solutions were analyzed with a high-resolution (m/m = 105 at m/z 450) linear-ion-trap (LTQ) Orbitrap mass spectrometer (Thermo Corp.) using an electrospray ionization (ESI) source in the positive ion mode. The compounds were detected as sodiated [M + Na]+ and/or protonated [M + H]+ species, with the strong predominance of the former. The resulting mass spectra were calibrated with respect to peaks in an Ultramark LTQ ESI positive ion calibration solution (Thermo Scientific) and with respect to the observed prominent

peaks in the oxidized squalene mass spectra, such as [C30H50OnNa]+ with n = 0-6, and [C20H34O12(C5H8O3)nNa]+ with n = 0-7. The data analysis was carried out as discussed in Nizkorodov et al.[100]. Specifically, peaks corresponding to 13C isotopes were removed; peaks that appeared in the oxidized filter control sample were discarded; formula assignments were limited to C1-80H2-140O0-50Na0-1+ with 0.0012 m/z tolerance; the H/C and O/C ratios were constrained to 0.5-2.2 and 0-1.2; and only closed-shell ions (no ion-radicals) were considered. Multiple formula assignments were sometimes possible for peaks with high m/z values; in these cases, preference was given to formulas that continued prominent CH_2 Kendrick series of unambiguously assigned peaks. All formulas discussed in the remainder of this paper correspond to the neutral compounds for the convenience of comparison with the COBRA simulation.

4.2.2 Computational Methods (COBRA)

COBRA is a computational brewing application that can be customized to simulate highly complex chemical systems[43]. It uses a bottom-up approach to simulation, wherein the starting compounds are selected and reaction rules describing the chemical transformations allowed within the system are defined based on experiments or prior chemical knowledge. Starting compounds are input using the commonly used SMILES language, and reaction rules are written using the related SMIRKS language[60, 1, 2]. SMIRKS is a reaction transform language used to define chemical reactions to an arbitrary degree of specificity. Each SMIRKS reaction rule represents one possible chemical transformation. Once the set of starting molecules and list of reaction rules is defined, COBRA simulates the evolution of the chemical system by exhaustively applying reaction rules to the starting pool of reactants, thereby generating a set of first-iteration products, which is mixed back into the pool of reactants for future iterations. The simulation continues in this way for an arbitrary number of iterations, resulting in a comprehensive list of all possible products of reactions.

In this study, COBRA was applied to modeling the oxidation of squalene in the presence of ozone. Two starting compounds were used for the simulation: squalene and ozone. Twelve chemical rules based on the current understanding of the mechanism of ozonolysis of alkenes were written to model the chemistry of the system (Table 4.1)[41]. Some of the rules produce Criegee intermediates (CIs) or other free radicals, which react further to form more stable products. To avoid modeling these unstable intermediates explicitly within the product pool, we combined CI- and radical-forming rules with later rules which consume CIs and radicals to form the final set of 12 SMIRKS-encoded rules used by the COBRA simulation.

Specifically, Rule 1 describes the formation of a Criegee intermediate (CI) and carbonyl from an alkene, as shown in Figure 4.1. Rule 2 corresponds to the formation of a secondary ozonide (SOZ) from the CI and a carbonyl, which is known to be efficient in condensed phases[13]. Rules 3-5 correspond to possible decomposition pathways of CI into stable products[41]. Rule 6 encompasses a series of reactions starting from an OH loss from CI, followed by the addition of OH to a double bond in a neighboring molecule, and an addition of molecular oxygen to all the alkyl radical sites to form peroxyl radicals, RO2[131]. The extent of OH loss from the CI in the condensed phase is uncertain, and may be low based on pressure-dependence studies, but we include it since it is one of the most important reactions of CIs in the gas phase[77, 39]. The RO2 radicals are relatively stable, and are expected to decay by cross-reactions resulting in carbonyl, alcohol, and peroxide products (Rules 7-9). Our first set of simulations included only rules 1-9 in Table 4.1, but they could not reproduce some of the major peaks in the experimental mass spectrum. Thus, the rules were amended to include three additional pertinent rules: hemiacetal formation (Rule 10), which appeared to occur in the products based on the experimental data; peroxyhemiacetal formation (Rule 11); and autoxidation of double bonds to carbonyls (Rule 12), which appeared to occur readily in squalene exposed to air (see Figure S1)[45].

We applied a filter such that the maximum mass of the generated structures was limited

to 100 heavy atoms (C and O), or approximately 1450 Da, because this was expected to cover the range of peaks detected by mass spectrometry. No other filters were applied. The simulation was allowed to proceed for four iterations, producing over 1 billion unique molecular structures after running for one month on an AMD Opteron 6274 2.2 GHz system. Ideally, the simulations would have run for six iterations to allow all six of squalenes double bonds to be oxidized, but extending the simulation was prohibitively expensive in terms of computer resources.

4.3 Results & discussion

4.3.1 Experimental Mass Spectrum

The background-corrected ESI mass spectrum of oxidized squalene contained more than 1500 peaks, and nearly 1300 of them could be unambiguously assigned to elemental formulas within the constraints described above. The reconstructed spectrum of the assigned neutral compounds is shown in Figure 4.2. ESI is regarded as a soft ionization technique, so it is assumed that each neutral compound should produce just one peak in the mass spectrum. In practice, some amount of ion fragmentation or cluster formation may occur even under soft ESI conditions. But even considering this complication, it is clear that ozonolysis of squalene generates a complex mixture of products. This can be contrasted with ozonolysis of lipids that have just one double bond, such as oleic acid or undecylenic acid, which produce a simpler product distribution[132, 149, 49]. For example, the undecylenic acid ozonolysis in Gomez et al.[49] produced a factor of 20 fewer peaks compared to the number of assigned peaks in the squalene system in this work.

In direct-infusion ESI-MS of complex mixtures, the relative intensities of different peaks are not necessarily indicative of the corresponding concentrations because of the selectivity in

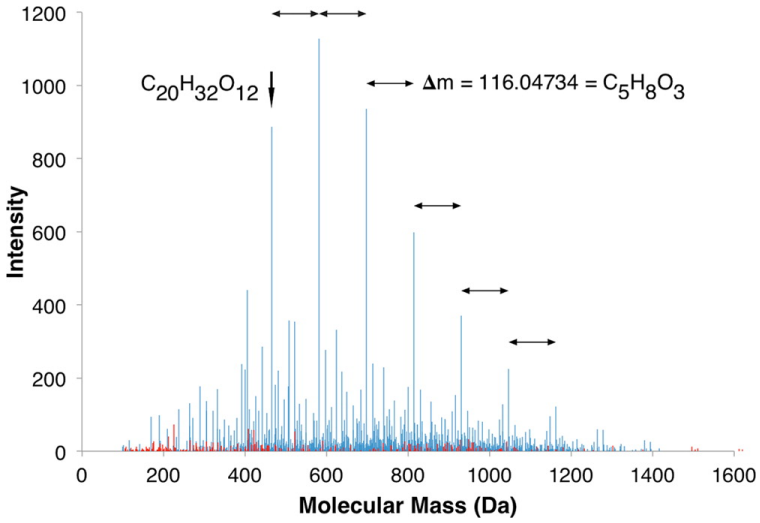


Figure 4.2: HR-MS experimental data and COBRA predictions of squalene ozonolysis products. HR-MS spectrum peaks that were predicted by the simulation are shown in blue. The remaining experimental peaks that were not predicted are shown in red. Within the simulation mass range of 0-1450 Da, over 83% of HR-MS peaks were predicted, including all major peaks as shown. Also noted is the series of strong peaks with m = $C_5H_8O_3$.

the ionization process. For example, the spectrum of unoxidized squalene, in addition to the major $C_{30}H_{50}$ squalene peak, contains peaks of $C_{30}H_{50}O_n$ with n = 1-6 corresponding to the products of squalene oxidation (all of them detected as Na adducts). These species are likely present at small levels in squalene, but they are enhanced in the mass spectrum because sodium ion affinity increases with the polarity of the analyte. Despite the selectivity in ionization, we can reasonably expect that all highly oxidized products should show up in the mass spectrum. Therefore, we can draw qualitative conclusions from the occurrence and relative intensity distribution of structurally related peaks.

One series of formulas that especially stands out in the experimental spectrum is a progression of $C_{20}H_{34}O_{12}(C_5H_8O_3)_n$ with n = 0-7 (Figure 4.2). The oligomer building block $C_5H_8O_3$ could represent a combination of isoprene (C_5H_8), from which squalene is built, and ozone (O_3). However, it is also possible that $C_5H_8O_3$ corresponds to an actual product of squalene ozonolysis, such as 5-hydroxy-4-oxopentanal (5-OH-4-OPA) shown in Figure 4.1. The molecules possessing both hydroxyl and carbonyl groups can cross-couple by means of hemi-

acetal formation reactions, and 5-OH-4-OPA should be especially active in such a process because aldehydes are more reactive than ketones. In fact, the COBRA simulations described below did not predict some members of the $C_{20}H_{34}O_{12}(C_5H_8O_3)_n$ series when the hemiacetal formation rule (Rule 10) was disabled, further supporting that reactions importance.

To help identify other series of related compounds in the mass spectrum, we carried out an analysis of the frequency of occurrence of mass differences between the peaks, similar to the one described in Kunenkov et al.[78]. According to the results, the most common mass differences correspond to $C_5H_8O_3$ (possibly 5-OH-4-OPA), $C_3H_6O_2$ (possibly hydroxyacetone), C_5H_8 (isoprene), O_3 (ozone), oxygen atom, and various combinations of these species. It is significant that all of these differences are chemically meaningful. For example, hydroxyacetone is a known product of squalene oxidation and an easily oligomerizable molecule[144]. Isoprene is a basic building block of squalene, so it is a naturally repeating unit for the squalene oxidation products. The reaction of ozone with alkenes produces SOZ compounds in high yields (Rule 2 in Table 4.1), with the formula of SOZ derived from the formula of the initial alkene with the addition of 3 oxygen atoms. Finally, the oxygen atom can be added to the molecule via the autoxidation reaction (Rule 12).

4.3.2 COBRA Simulations

COBRA simulations generated over 1 billion unique structures representing 26899 elemental formulas after four iterations of computational brewing (where the first iteration is the initial reaction of ozone with one of the double bonds in squalene). Overlap with experimental HR-MS results was good, with the majority of high-intensity peaks being predicted (Figure 4.2). Specifically, when considering HR-MS observed peaks within the COBRA simulation range, which was limited to a maximum of 100 heavy atoms, or roughly 1450 Da, we predicted 1045 of the 1258 experimentally observed peaks (83.1%). As Figure 4.2 demonstrates, the

peaks predicted by COBRA represent the most abundant compounds detected within this mass range. This suggests that the reaction rules defined in Table 4.1 and used by COBRA to generate predicted products reasonably capture the essential chemistry of the squalene ozone oxidation system.

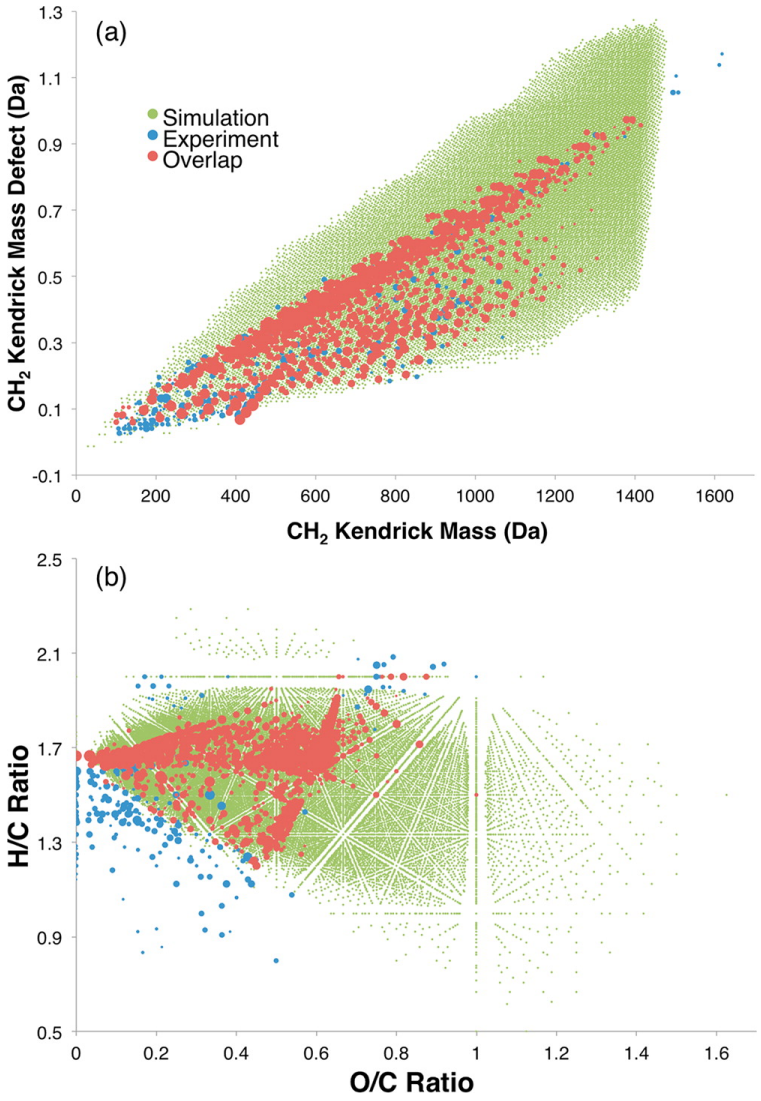


Figure 4.3: (a) Kendrick plot and (b) Van Krevelen diagram comparing the COBRA results (small green dots) with experimental results (blue circles with diameters proportional to the log of the peak abundance). Red circles are those for which simulated and experimental data overlap. The COBRA simulation produces products that are more oxidized on average (have higher O/C ratios).

Despite the good overlap between the observed and predicted formulas, the fact that 17% of the observed formulas were not reproduced by the simulation suggests that some reaction

mechanisms are still missing. Specifically, the simulation appears to miss experimentally observed compounds with low H/C and low O/C ratios, which tend to be outliers on the Van Krevelen diagram in Figure 4.3b. To further compare the group of observed compounds that is predicted and the group that is not predicted, we computed the average molecular formula and average double bond equivalent (DBE) for the two groups. For observed compounds that are predicted, the average formula is $C_{37.3}H_{61.4}O_{15.2}$ (DBE 7.6), while the average formula for observed compounds that are not predicted is $C_{32.6}H_{49.3}O_{8.3}$ (DBE 9.0). These formulas indicate that the simulation predicts observed compounds that are, on average, larger and more oxidized compared to the ones it does not predict. Interestingly, observed compounds that are not predicted tend to be less saturated on average (have a higher DBE).

We should note that the overlap of the predicted formulas and observed ones is not a definite proof that the chemistry rules we have chosen are correct and comprehensive. We tested the overlap between the formulas predicted for squalene ozonolysis in this work, and formulas observed in ozonolysis of limonene by Bateman et al.[19], and found a high degree of overlap between them (90% of the observed formulas appeared in the simulation output). Limonene is a monoterpene, which, like squalene, is built from isoprene units. Therefore, limonene undergoes ozonolysis by a similar mechanism, so the overlap in this case is not too surprising. We also tested the overlap between formulas predicted in this work and those observed in two other experiments: an isoprene photooxidation SOA[43], and a naphthalene photooxidation SOA[85]. Overlap between our squalene simulation from this work and the observed formulas in isoprene and naphthalene SOA was 62% and 23%, respectively. The lower degree of overlap with these two systems is consistent with the mechanistic differences between OH-driven oxidation for isoprene and naphthalene cases, and O3-driven oxidation in the squalene and limonene cases.

To provide further context for the aforementioned comparisons, we also tested the overlap between experimental results from this work and experimental results from the other works.

When comparing experimentally observed formulas for squalene ozonolysis in this work and formulas observed in the other works, the overlap (the percentage of formulas from the other studies that were also observed experimentally in this work) was as follows: 22% for the limonene ozonolysis system[19], 10% for the isoprene photooxidation system[43], and 7% for the naphthalene photooxidation system[85]. All of these comparison results are summarized in Table 4.2.

Figure 4.3 offers another perspective on the overlap between HR-MS and COBRA results. The top panel shows a CH_2 Kendrick plot for the experimentally observed and simulated peaks. The predicted structures have similar values of the Kendrick mass defects compared to the experimental ones. The Kendrick mass defect is calculated as the difference between the nominal mass and Kendrick mass with respect to the $^{12}CH_2$ group[100]. The second panel shows a Van Krevelen diagram, wherein the H/C and O/C ratios for all the compounds are plotted against each other[71]. This diagram has a wide range on the O/C axis, which is expected since the chemistry strongly favors oxygen atom addition to the molecules. The simulated compounds are on average more oxidized than the experimentally observed ones. This is not surprising, as the simulation can proceed without kinetic constraints, which is of course not true in reality. First of all, as the products build up, the product mixture likely becomes more viscous, constraining access of ozone and oxygen to the mixture components[130]. Second, it has been shown for the OH-driven oxidation of lipids that oxidation reactions adding oxygen atoms to the products are efficient at the early stages of oxidation, but at later stages they are replaced by fragmentation reactions, which our simulation does not take into account. For OH oxidation of squalane (the fully saturated version of squalene) the fragmentation processes took over when O/C ratio reached about 0.4.(34) On the basis of Figure 4.3b, the fragmentation may become significant in the ozone-squalene system once the O/C ratio reaches ˜0.7 since not many products are observed with O/C in excess of this value, while there is clearly a potential for accommodating larger amounts of oxygen according to the simulation.

We performed 13 exclusion simulations to examine which reaction rules were contributing most to the prediction of experimentally observed products. For each exclusion simulation, a single reaction rule was omitted, and the simulation was run until 500000 reactions were computed. Then the number of experimental molecular formulas recovered was compared against an analogous control simulation in which all rules were included. The exclusion percent is the percent reduction in the number of correctly predicted formulas when a given rule is omitted. That is, a higher exclusion percent indicates a more important rule in terms of contribution to the product distribution. With an exclusion percent of 16.3%, the most important rule in recovering experimentally observed formulas was the sequence of reactions 1 + 6 + 7 from Table 4.1, which involved OH loss from a Criegee intermediate (CI) with subsequent addition of O2 to the resulting radical, followed by carbonyl formation from the resulting RO2 group. The second most important rule, with an exclusion percent of 15%, was the sequence of reactions 1 + 6 + 9, analogous to the previous rule, but with peroxide formation from RO2 as the last step. The third most important, at 12.2%, appeared to be reaction 2 from Table 4.1, the formation of a trioxolane ring (SOZ) from CI and a carbonyl. Finally, formation of hemiacetals and peroxyhemiacetals (reactions 10 and 11 from Table 4.1) had an exclusion percent of 5.8% and appeared to be required to reproduce the most abundant peaks in the experimental spectrum. Some rules had an exclusion percent of 0%, implying that they either produce strictly a subset of products generated by other rules, or that they do not produce experimentally observed structures within the number of reactions computed in an exclusion simulation.

We also compared the table of major primary and secondary products of squalene ozonolysis listed in Weschler[141]. to both our experimental HR-MS data, and the simulated data. All of these products were predicted by the simulation, while half of them were detected in the HR-MS data. We note that experiments in Weschler[141] predominantly detected early products of oxidation generated at low ozone exposures, most of which are volatile species, whereas the experiments reported here focused on later-generation products remaining in

the condensed phase. It is therefore expected that some of the readily oxidizable compounds observed under the milder oxidation conditions in Weschler[141] were not observed here under stronger oxidation conditions. The simulation captures both early- and late-generation products of oxidation, regardless of whether they are volatile or not, hence it agrees with both sets of experiments.

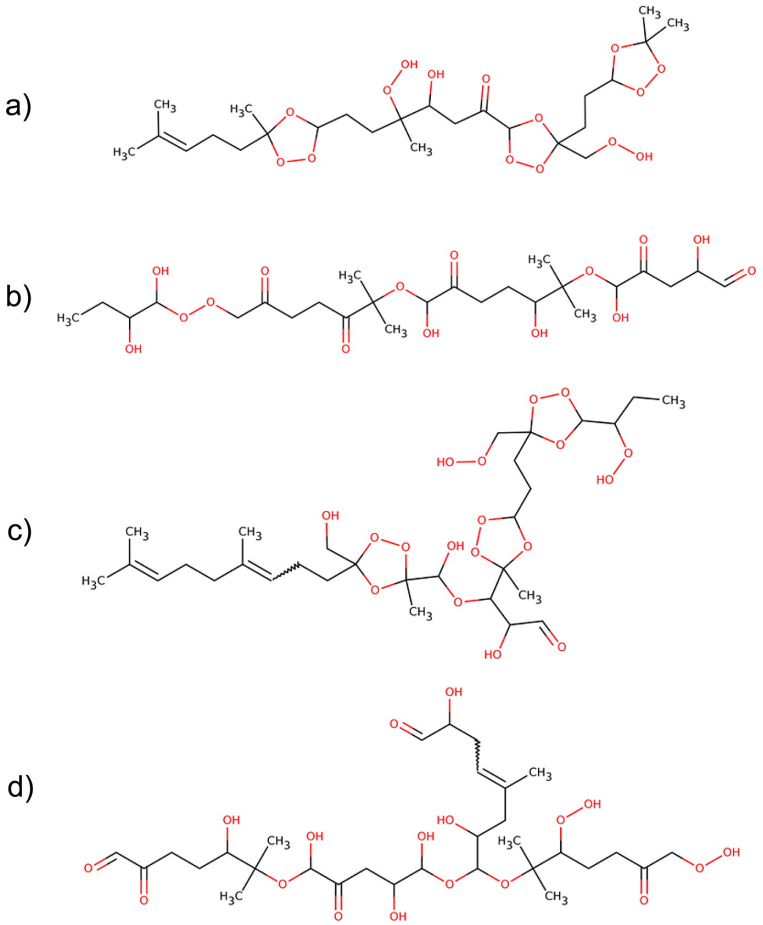


Figure 4.4: Examples of predicted structures from the two most abundant HR-MS peaks. Parts (a) and (b) are from the most abundant peak, with the formula $C_{25}H_{42}O_{15}$. Parts (c) and (d) are from the second most abundant peak, with the formula $C_{30}H_{50}O_{18}$. COBRA generated 9804 and 32527 structures for these two peaks, respectively, and the examples above were randomly selected from those sets.

One of the benefits of the simulation is that it has information on molecular structures of every product produced by the simulation. Figure 4.4 shows examples of COBRA-predicted structures with masses corresponding to the two most abundant HR-MS peaks. The struc-

tures in Figure 4.4A and 4.4B correspond to the highest intensity peak, $C_{25}H_{42}O_{15}$, and Figure 4.4C and 4.4D are structures selected from the second highest peak, $C_{30}H_{50}O_{18}$. We should emphasize that these structures are just randomly selected structural isomers out of thousands of possibilities. For example, there were as many as 9804 structures generated for the aforementioned C_{25} peak, and 32527 structures generated for the C_{30} peak. These numbers indicate the broad range of structural isomers generated by the system for a given molecular mass. For some molecular formulas, particularly at higher masses, there were hundreds of thousands of structures predicted. Of course, in reality, not all of these isomers will be accessible because of various kinetic and energetic constraints, which are not included in the simulation. Nevertheless, the implication of these results for the interpretation of the experimental mass spectrum (which cannot resolve isobaric compounds) is that the mixture of products is likely even more complex than the number of distinct peaks in the direct-infusion ESI mass spectrum would suggest.

Another advantage of the simulation is that it can provide statistical information on the distribution of important functional groups in the products (once again, we are neglecting possible kinetic and energetic constraints in the actual reaction). Therefore, we calculated the percentage of oxygen atoms in simulated products by functional group, to better understand the chemical composition of the complete set of predicted structures (Table 4.3). On the basis of the simulation results, peroxy groups are the most prevalent functional group containing oxygen atoms, accounting for 42.5% of oxygen atoms in the products. The next most prevalent oxygen-containing functional groups were hydroxyl, ether, and ketone groups, representing 20.9%, 18.0%, and 9.7% of all oxygen atoms in the simulated formulas, respectively. Figure 4.5 shows examples of predicted structures containing a high fraction of O atoms in peroxy groups. The two structures were selected from the pool of predicted products for the two most abundant HR-MS peaks, as in Figure 4.4. They represent the maximum fraction of oxygen atoms in peroxy groups among structures predicted at these masses and are shown for illustrative purposes. We note that while the simulation does not

consider possible kinetic and energetic constraints, these statistics and examples are still of use for understanding the composition and characteristics of the simulated product pool.

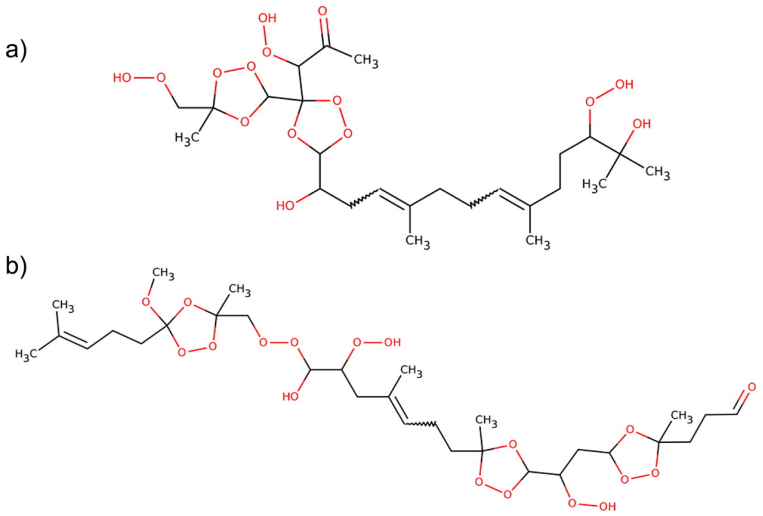


Figure 4.5: Examples of structures with a high number of peroxy functional groups. Peroxy groups were the most common functional group for O atoms in the predicted products (Table 4.3). The structures above were predicted for experimental peaks $C_{25}H_{42}O_{15}$ (a) and $C_{30}H_{50}O_{18}$ (b). These structures represent the highest fraction of O atoms in peroxy groups for structures predicted with these molecular formulas.

Simulating the oxidation of squalene in the presence of ozone using COBRA is a good example of the utility of computational tools for modeling and understanding complex natural systems. COBRA can aid in validating experimental data, generating proposed structures for compounds observed in mass spectrometry experiments when combined with electronic structure calculations, and generating possible structures for other experimentally observed chemical formulas. By defining a relatively simple system of chemical transformation rules and starting compounds, this computational approach can be a powerful tool for better understanding the complexity of a natural chemical system composed of tens or hundreds of thousands of compounds. These results are atmospherically relevant because they may help understand the nature of compounds found in highly aged indoor dust resulting from skin flakes and highly soiled indoor surfaces, such as airplane cabins and heavily populated office spaces.

Table 4.1: List of chemical rules used to define the squalene oxidation system.

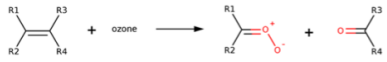
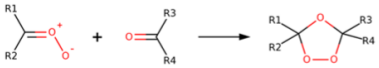
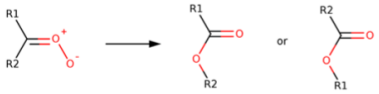
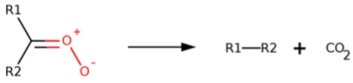
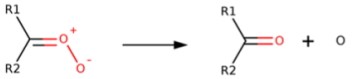
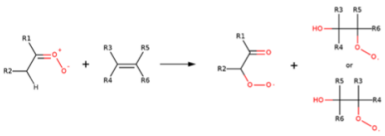
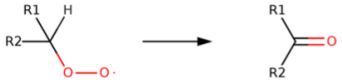
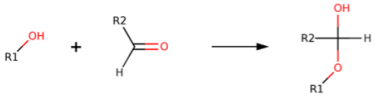
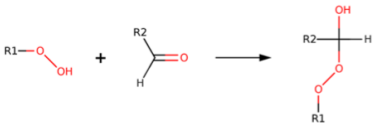
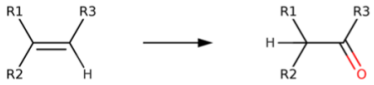
No.	Description	Chemical Rule
1	Ozone + alkene	
2	CI + carbonyl	
3	CI isomerization into acid or ester	
4	CO_2 elimination from CI	
5	O atom elimination from CI	
6	OH loss from CI + OH addition to an alkene + addition of O_2 to the resulting radicals	
7	Carbonyl formation from RO_2	
8	Alcohol formation from RO_2	
9	Peroxide formation from RO_2	
10	Hemiacetal formation	
11	Peroxyhemiacetal formation	
12	Double bond oxidation	

Table 4.2: Comparison of experimental and simulated results from this work with experimental results from several prior studies.

system	number of assigned peaks in system	percentage of assigned peaks matched by squalene + O_3 simulation	percentage of assigned peaks matched by squalene + O_3 experimental system
squalene + O_3 products	1258	83%	100%
limonene + O_3 SOA	924	90%	22%
isoprene + OH/NO_x SOA	463	62%	10%
naphthalene + OH/NO_x SOA	242	23%	7%

Table 4.3: Percentage of O atoms by functional group for all predicted products.

functional group	percentage of O atoms
peroxy	42.5%
hydroxyl	20.9%
ether	18.0%
ketone	9.7%
aldehyde	5.8%
carboxyl	2.0%
ester	1.1%

Chapter 5

Deep Learning for Chemical Reaction Prediction

Reaction Predictor is an application for predicting chemical reactions and reaction pathways. It uses deep learning to predict and rank elementary reactions by first predicting electron sources and sinks, pairing those sources and sinks to generate proposed reactions, and finally ranking the reactions by favorability. We demonstrate significantly improved accuracy and speed while predicting advanced organic chemistry. We describe multi-target pathway search functionality than can aid the identification of unknown masses observed by mass spectrometry. Finally, we discuss an alternative approach to predicting electron sources and sinks using recurrent neural networks, specifically long short-term memory (LSTM) architectures, operating directly on reactant SMILES strings. This approach has shown promising preliminary results.

5.1 Introduction

Achieving human-level performance at predicting chemical reactions remains an open problem with broad potential applications. Historically there have been three major categories of approaches to reaction prediction: rule-based expert systems, quantum-mechanical simulations, and machine learning-based systems.

Rule-based approaches to reaction prediction[47, 22, 29] are fast, but the requisite systems of manually-implemented rules and exceptions require painstaking maintenance. While they may provide good results within a limited chemical domain, rule-based systems are constrained by the extent to which a human expert has defined the underlying rules. Such systems do not scale well over the long term as new areas of chemistry are added. Furthermore, these systems typically make predictions at the level of overall chemical transformations. Multi-step reactions are condensed into a single transformation, and information about the elementary arrow-pushing steps comprising the multi-step reaction is not available. Yet these elementary steps are the building blocks for predicting novel multi-step global reactions and identifying side products.

Quantum-mechanical (QM) approaches may theoretically yield accurate results based on physical first principles, but in practice are highly sensitive to operator set-up, and are computationally expensive. Many recent studies involving QM-based prediction of reaction pathways are narrowly limited to a single chemical system[86, 107, 4, 10, 119]. A clear benefit of these methods, when successful, is their ability to quantitatively predict important reaction parameters such as free energies, energy barriers, transition states, and reaction rates. Still, they require considerable human intervention and are not suitable for making high-throughput predictions.

Machine learning (ML) approaches[69, 68, 135] are fast and scalable, but ideally require large data sets from which to learn. Obtaining such chemical data sets is a significant challenge, as

many are proprietary and not readily available for academic use. Furthermore, reactions may be unbalanced or not atom-mapped, complicating attempts at statistical learning. Nonetheless, the ML approach to reaction prediction remains promising, and recent advances in deep learning have opened the door to further performance gains.

We also note that these three approaches can be complementary to one another. Rule-based systems have been used to provide training examples for ML-based systems[69]. Similarly, ML-based systems can benefit from the implementation of a small number of carefully selected rules. In the case of Reaction Predictor, such rules are designed to address corner cases or express strong priors about specific products or resonance structures that may be problematic or redundant. Other work has explored using ML-based approaches to predict traditional QM observables like molecular atomization energies and electronic structure properties[114, 53, 61]. Recent work in our group demonstrated synergies between the QM- and ML-based approaches[115]. We showed that results derived by QM modeling can be used to derive new training examples for ML algorithms, creating a closed and automated positive feedback loop. In this loop, the ML system helps prioritize QM computations on elementary reaction steps, and the QM-computed Hartree-Fock activation energy is used to generate additional training examples to directly improve the ML system.

Reaction Predictor is a deep learning-based approach to reaction prediction that operates at the level of elementary reactions. Each predicted reaction involves the movement of electrons from an electron source to an electron sink. Chemical inputs are represented using SMILES and SMIRKS strings, and the OEChem toolkit[2, 60] enables additional chemical computation. In brief, the Reaction Predictor pipeline operates in the following multi-step fashion, given a set of input reactants:

1. Enumerate all possible electron sources and electron sinks within the input reactant molecules

2. Filter the list of candidate sources and sinks, predicting a smaller list containing only the most reactive sources and sinks

3. Propose reactions by enumerating all combinations of source-sink pairings

4. Rank the proposed reactions by favorability

Reaction Predictor features an offline pathway search tool that can help identify unknown products by searching for experimentally observed masses. The user submits a search job by entering a set of reactants and one or more mass targets, and the system performs the search in the background, emailing results when complete. If matches were found, the system provides not only proposed structures for the search targets, but also includes the full reaction pathway that yielded the products.

Here we describe the ML design and methodology underpinning Reaction Predictor's ML-based predictions. We test its performance on a benchmark data set of challenging real-world reactions, and demonstrate that it notably surpasses the performance of an early prototype in both accuracy and speed. We also demonstrate that an alternative approach to predicting electron sources and sinks simply by examining SMILES strings, using a long short-term memory (LSTM) architecture[57, 50], shows promising results.

5.2 Materials & methods

5.2.1 Data

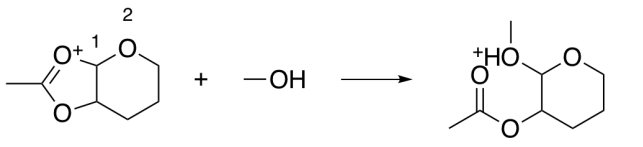
The unit of our data set is an "elementary reaction". Each elementary reaction represents an energetically favorable elementary mechanistic step, with a single labeled electron source and a single labeled electron sink. The early prototype used a data set that was a combination of

undergraduate organic reactions extracted from Reaction Explorer rules[29], plus a set of 368 hand-selected reactions from graduate-level organic chemistry texts[58, 87]. In total, there were 5,551 elementary reactions in the original data set. We note that, internally, Reaction Predictor has three distinct chemical prediction modes: one for polar, one for radical, and one for pericyclic reactions. Each of these uses its own separate underlying data set and trained predictive models. Here we will discuss the polar data set and models, as they represent the most developed prediction mode for Reaction Predictor.

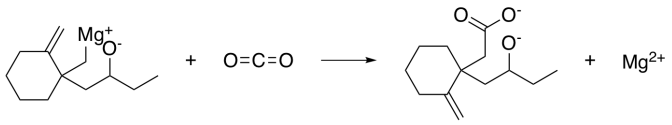
We used a benchmark data set of 289 single-step reactions taken from challenging multi-step transformations to test the performance of our system. These reactions were hand-curated to cover a broad range of advanced organic chemistry intended to test the system's ability to generalize on real-world reactions. They are not a subset of the training data.

5.2.2 Data set improvement

One significant limitation of using Reaction Explorer rules to generate the bulk of the prototype's original training reactions is the inherent bias towards undergraduate chemistry. Undergraduate texts often make simplifying assumptions or omit complicating details in order to more clearly present fundamental concepts. A more accurate understanding of the chemical reality is left to be clarified during advanced study. This presents a problem for learning algorithms, which can only learn and generalize based on what they are shown during training. To address this, we reviewed the existing data set in its entirety. Manual inspection of all elementary reactions led to the removal of 884 problematic reactions. Reactions were removed for any of several reasons. First, inspection showed that the S_N2 mechanism was overemphasized, and a number of redundant S_N2 reactions were removed. Second, some reactions were removed as they were unfavorable or less favorable than an alternative elementary mechanism. Third, some of the reactions simply contained errors.



(a)



(b)

Figure 5.1: Examples of problematic reactions. In (a), the problematic mechanism was an S_N2 attack of carbon "1" by ethanol. The correct mechanism should use an oxygen "2" lone pair to push out O^+, followed by ethanol attack. In (b), this Grignard reaction erroneously contained Mg^+ instead of MgBr.

Figure 5.1 shows examples of problematic reactions identified in the original data set. After removing these problematic reactions, we were left with a cleaned data set of 4667 polar elementary reactions.

Using this cleaned data set as our starting point, we have since added 6,361 high quality hand-curated reactions. These were selected by our coauthors Van Vranken, Gutman, and Mood, and cover a broad range of advanced organic chemistry that was not represented in the original data set. Some examples of notable new reaction types added to the data set are shown in Figure 5.2. In Figure 5.2a, a 12-step synthesis reported by Tu et al. involves an MPV reduction. Figure 5.2b represents a seven-step synthesis reported by Mulzer et al., which involves a Mitsunobu reaction. Figure 5.2c shows an eight-step synthesis by Baumann et al. that depends upon a Stetter reaction. For each of these examples, critical steps in the reaction mechanism required elementary reactions that were not represented in the original data set, but have since been added.

At time of writing, the total number of elementary reactions in our data set is 11,028 – more than double the original data set.

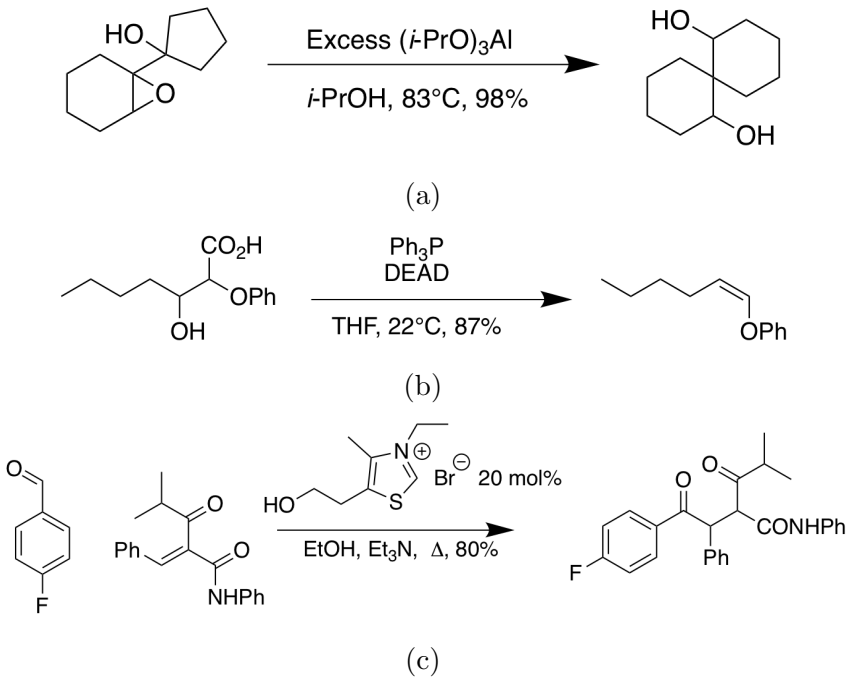


Figure 5.2: Examples of multi-step syntheses requiring chemistry that is now represented in the data set. Hand-curated elementary reactions were added to the training data to enable these reaction types, and many more, greatly improving predictive capability. Reaction (a) is a 12-step synthesis requiring an MPV reduction[129]; (b) is a seven-step synthesis involving a Mitsunobu reaction[95]; and (c) is an eight-step synthesis involving a Stetter reaction[20].

5.2.3 Combinatorial reaction generation

We experimented with using software to automatically generate thousands of additional elementary reactions for our training data set, using the following methodology. First, for a given reaction mechanism, we identified the core molecular template and the appropriate electron movement. We then systematically varied the substituents of the template reaction within realistic chemical constraints to generate all combinations of substitutions. In this way we generated tens of thousands of elementary reactions covering a range of fundamental reaction classes. In order to not overwhelm the existing data with biases towards the combinatorial reactions, we randomly sampled subsets on the order of a thousand reactions for each mechanism. Figure 5.3 illustrates this process.

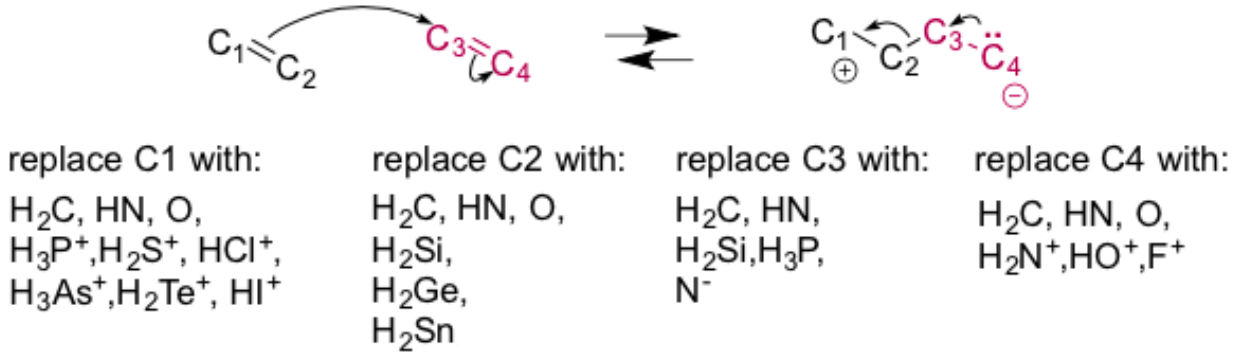


Figure 5.3: Illustration of combinatorial reaction generation, including reaction template (top) and substitution constraints (bottom). Thousands of reactions are produced by generating all combinations of allowed substitutions for C_1, C_2, C_3, and C_4.

5.2.4 Applying deep learning

Predicting reactive electron sources and sinks is a crucial step in the Reaction Predictor pipeline. If the best source or best sink is rejected during the source/sink filtering step, the desired reaction cannot be reproduced. The early prototype of Reaction Predictor placed great emphasis on *recall*, with little consideration for *precision*. That is, the system was biased towards predicting many potential sources and sinks, to avoid missing any, even

if most were false positives. Yet false positives have negative performance implications: they significantly increase computation time, which quickly adds up for pathway searches, wherein multiple single-step predictions are chained together to predict products of multi-step reactions. We developed source/sink filtering models focused on both precision and recall, as described below.

For source/sink filtering, we use a single 380,000 parameter, fully-connected feedforward neural network (also called a multilayer perceptron, or MLP) with 1,500 inputs, three hidden layers of 200 rectified linear units, and two independent sigmoid output units corresponding to a source prediction and a sink prediction. To reduce overfitting, 50% dropout was applied to each hidden layer[56, 14]. To train the model, weights were initialized as described in Glorot and Bengio, and updated using the Adam optimizer[73] on mini-batches of 64 examples. An exponentially decaying learning rate, and early stopping based on a validation set of 10% of the training set were used. Training was performed on an NVIDIA Titan X GPU.

We built the training data for our source/sink filtering network as follows. For each elementary reaction in the database, we extracted four training examples: (1) the labeled source, (2) the labeled sink, and (3) two randomly sampled non-source, non-sink examples. There are two advantages to this method compared with the original method of extracting two positively labeled examples (true source and true sink), plus all remaining negative examples. First, we avoid the significant imbalance inherent in the data, as we observe approximately 22-times more negative examples than positive examples. Previously this was addressed by oversampling the positive examples. The second advantage is we avoid adding potentially misleading examples to our training data. Specifically, we avoid negatively-labeled examples that should actually be considered secondary sources or sinks. For a given elementary reaction, the set of atoms not explicitly labeled source or sink contains mostly poor sources and poor sinks. However, some could be considered "second-tier" sources/sinks, either on their own or within a different molecular context. By randomly selecting from these atoms to

generate our negative examples, we gain a representative sample of the non-source, non-sink examples in our data, while also avoiding labeling all potential second-tier sources/sinks as negative examples. Extracting the data as described above, our training set for source/sink filtering consisted of 23,850 examples, half of which were positive examples.

After identifying sources and sinks, and pairing them together, we must rank the resulting set of proposed reactions. To do so, we train a deep Siamese architecture neural network[101, 24] to compute a reaction favorability score. Figure 5.4 illustrates this architecture. Training examples consist of ordered reaction pairs ($R_{favorable}$, $R_{unfavorable}$), where the favorable reaction is always presented to the left instance of the shared-weight neural network. Fixed weights of +1 and -1 for left and right outputs are connected to a final sigmoid unit. Thus the final output y approaches 1 if the left reaction was scored higher than the right reaction, and 0 otherwise. After model training, we use one instance of the shared-weight network to compute favorability scores for all reactions, and rank them based on those scores.

For the shared-weight network, we used two hidden layers of 300 tanh units, and a sigmoid output. Initialization and training proceeded as described above for the source/sink models. We generated training examples ($R_{favorable}$, $R_{unfavorable}$) as follows. For each elementary reaction in the data set, we have one reaction, $R_{favorable}$, formed by pairing the labeled electron source with the labeled electron sink. We can propose many additional *un*favorable reactions, by pairing the labeled source with all non-sinks, and all non-sources with the labeled sink, within the constraints of chemical feasibility. We use this set of unfavorable reactions to create additional training pairs ($R_{favorable}$, $R_{unfavorable}$), yielding 387,744 total training examples.

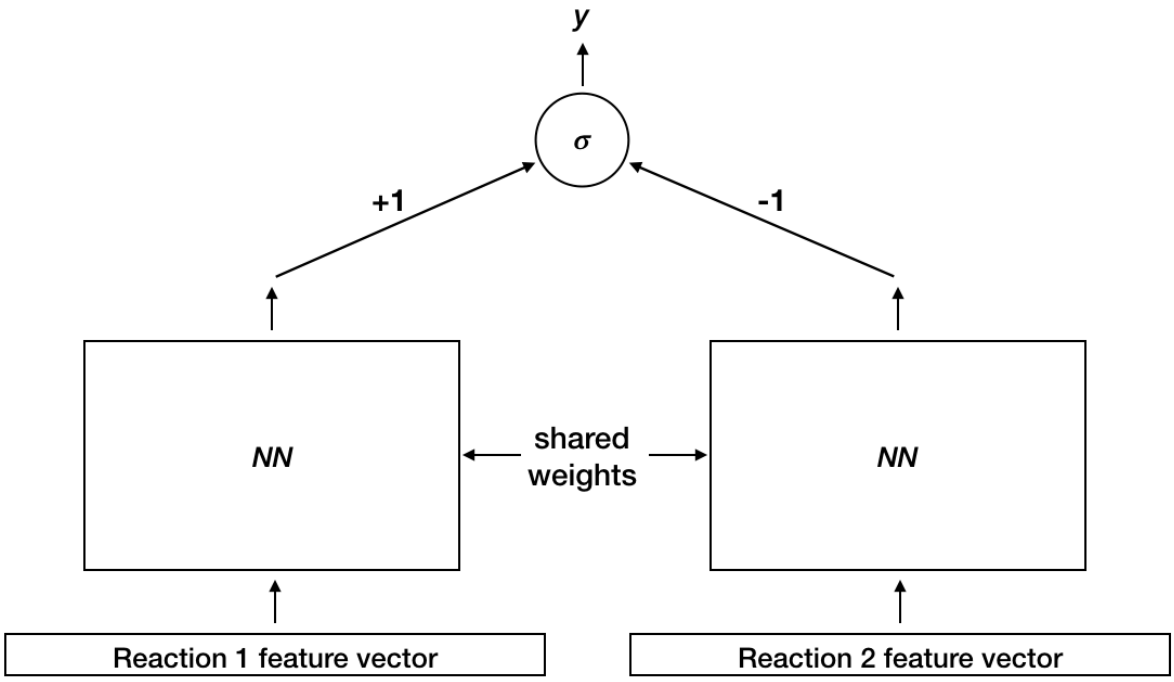


Figure 5.4: Siamese architecture for reaction ranking. Outputs from the left and right instances of the neural network NN have fixed weights of +1 and -1 into a final sigmoid unit. Thus the final output y approaches 1 if Reaction 1 scores higher than Reaction 2, and 0 otherwise. We can think of NN as computing a reaction favorability score that is learned by training on many examples of ($R_{favorable}$, $R_{unfavorable}$) reaction pairs.

5.2.5 Feature representation & selection

To make accurate source/sink predictions, we must extract relevant chemical information about each potential source/sink site within a reactant molecule. The features we use to capture this information fall into two categories: physicochemical, and graph-topological. Physicochemical properties are extracted at the atom level. Examples include partial and formal charge, presence and type of filled/unfilled orbitals, presence of lone pairs, and a steric coefficient. Graph-topological features capture properties about the atom and bond connectivity of the molecular graph. These are based on a variation of chemical fingerprints[122] and are extracted by enumerating paths and trees over a small neighborhood around a particular atom. Specifically, we allow paths up to depth 6 if the atoms along that path are heteroatoms or are part of a conjugated pi system, otherwise the maximum depth is 3.

Reaction-level features include (1) a combination of concatenated source and sink atom-level features, (2) features describing the type of orbitals involved, and (3) net change features, which are created by computing molecular fingerprints for both the reactants and products, then subtracting the former from the latter to capture "what changed" during the reaction.

Extracting the features described above for all examples in our data set yields 293,046 atom-level features, and 62,560 reaction-level features. We select the top 1,500 atom-level and top 2,000 reaction-level features with the highest mutual information and use those features as the input representations for our source/sink filtering and ranking models[76, 104].

5.2.6 Spectator molecules

A spectator molecule is present on both the reactant and product sides of an elementary reaction, but does not participate in the reaction mechanism. There is information to be gleaned, however, from a spectator molecule's *inaction*. Each spectator molecule can provide negative source/sink training examples for the filtering model, and unproductive elementary reaction examples for the ranking model. We added the ability to capture and consider these spectator molecules in our training data.

5.2.7 Offline pathway search

Reaction Predictor's Pathway Search feature allows the user to input starting materials and a target molecule to search for multi-step reaction pathways that yield the target. The prototype version had several limitations. First, it required a specific molecular structure for the target. Often a chemist may have only a molecular mass, not a structure, for a target of interest. Another limitation was that the searches had to be run "online", while the user waited for a result to be returned to the web application. For large, complicated pathway

searches, this often led to the web browser timing out while waiting for a server response. Another limitation was that the pathway searches were all running on a single machine. Multiple searches executing in parallel had to compete with each other for resources, slowing down the whole machine, including single-step Reaction Predictor queries.

To address the issues described above, we implemented new Pathway Search functionality. Most importantly, there is an offline search mode that allows the user to enter an email address in addition to the search parameters, and receive match results by email when they are ready. Search jobs are intelligently distributed across our compute cluster with hundreds of available CPU cores, so they do not slow down the main Reaction Predictor website or each other. Another important feature is multi-target search. Targets can be specific molecules, masses, or a mixture of both. Users can enter as many search targets as desired for a given set of reactants. This makes Pathway Search a powerful tool for suggesting alternative reaction pathways and identifying unknown products observed in mass spectrometry data.

5.2.8 Additional features

While implementing the major features described above, many other improvements were made, including new features and improved chemical capabilities. Here we describe a non-exhaustive selection of these additions.

Intramolecular reactions Reaction Predictor considers intramolecular reactions in cases where more than a single reactant molecule is present. In the earlier prototype, if two or more reactants were present, only *inter*molecular reactions were considered. This change is especially important for Pathway Search, where multiple reactants will either be directly entered or inevitably formed during the course of the search, and to prevent intramolecular reactions at all subsequent steps is to omit a broad swath of potential reactions. Multi-step

reaction mechanisms frequently have steps involving intramolecular reactions, and they must be considered, even in the presence of many reactant molecules, in order to achieve high-quality multi-step predictions. Some examples of intramolecular reactions, including a Grob fragmentation, a Neber rearrangement, and an intramolecular Prins reaction, are shown in Figure 5.5[102, 31, 75].

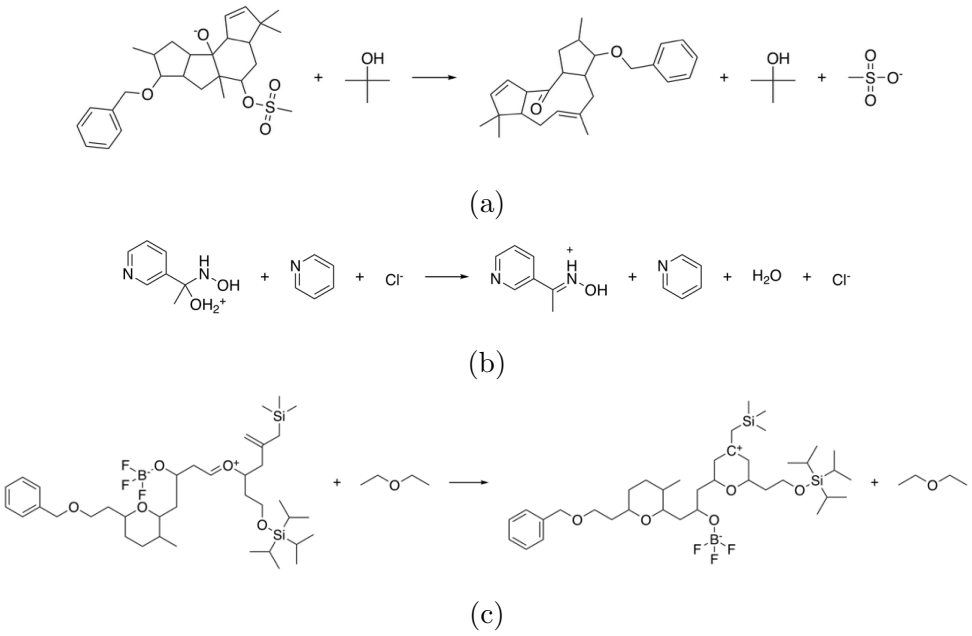


Figure 5.5: Examples of intramolecular reactions that can be predicted by the system. A Grob fragmentation (a), Neber rearrangement (b), and intramolecular Prins reaction (c) are shown. Enabling the prediction of intramolecular reactions when multiple reactants are present significantly increases the number of proposed reactions, but is necessary for the prediction of many reaction pathways including these.

"Continue reacting" button for single-step predictions Users can manually explore a multi-step reaction pathway in a guided stepwise fashion. After submitting reactants and generating a list of predicted products, users can click a button next to any of the results to use those products as the reactants for a new single-step prediction.

Improved modeling of alkali metals The alkali metals Li, Na, and K, are now modeled with a more sophisticated understanding of their bonding potential. Originally, these metals

were only allowed to form a single bond, while in reality they can form up to six. This is now reflected in reaction predictions involving these atoms.

Additional atom types Reactants containing Sc, Ti, Zn, As, or Se atoms were not accepted in the early prototype. We have improved the underlying chemical model to be capable of handling these elements.

Multiple equivalents for reaction pathways Multi-step reaction pathways often rely upon the availability of multiple equivalents of a starting material. Users can now specify that arbitrarily many equivalents of a given reactant should be present for a reaction pathway search. Reaction Predictor will then keep track of how many times the reactant has been consumed, and add the reactant back into the reactant pool as many times as the user-specified number of equivalents allows.

ML features We carefully evaluated the physicochemical and graph-topological features used by our ML algorithms. Some were identified that were either irrelevant or counter-productive for making source/sink predictions. For example, we observed instances where increasing the molecular mass of a molecule, by adding a side-chain to a distant part of the molecule, would cause a source or sink that had at first been predicted correctly, to suddenly be misclassified. We removed this feature, as the reactivity of a candidate source or sink should not be a function of the mass of its parent molecule. We also removed a feature designed to capture whether a source/sink was located centrally or peripherally within the molecular graph, as it too was chemically irrelevant for determining the quality of a source or sink.

5.3 Results & discussion

5.3.1 Single-step performance

We assessed single-step reaction prediction performance using the benchmark data set of 289 reactions described above. Table 5.1 shows the results achieved by Reaction Predictor, compared with results from the early prototype.

Table 5.1: Single-step reaction prediction performance for Reaction Predictor (RP), compared with the early prototype, using a benchmark data set of 289 reactions.

	RP	**prototype**
Products recovered	83.0%	58.1%
Products ranked in top-5	79.2%	76.8%
Mean time per reaction (sec)	8.5	91.8
Total reactions proposed	22,283	92,158

First we note that the early prototype failed while attempting to run the benchmark data set, due to its inability to handle certain atom types, as described above. We incorporated the necessary updated library into that version, so that it could complete the test.

We observe that Reaction Predictor significantly outperforms the early prototype, recovering 83.0% of the expected products, vs. 58.1% for the prototype. This difference is particularly striking when we consider that the prototype version is biased towards predicting more false positive sources/sinks, in exchange for, ideally, improved recall. We see this reflected in the total number of reactions proposed by each version, with the prototype proposing 92,158 reactions, while our current version proposes 22,283. Yet, even though the prototype predicted so many more reactions, it still only recovered 58.1% of the expected products. That Reaction Predictor can predict nearly 70,000 fewer reactions while recovering significantly more of the products (83.0%) indicates a dramatic improvement in source/sink filtering capability. We believe this is a result of two factors: the improved breadth and quality of the expanded training data, and the application of deep learning to better learn from that data.

Reaction Predictor also outperforms the early prototype in terms of how many correct products are ranked within the top-5 of the ordered reaction list, but the margin is less pronounced. This result is not wholly unexpected, as the ranker has historically been the strongest ML component in the system. Being able to generate hundreds of thousands of training examples for the ranking model, even with only several thousand elementary reactions in the training set, seems to convey enough information to yield a highly accurate trained ranking model. Our observation that doubling the number of elementary reactions in the training set led to a dramatic improvement in source/sink identification, but a relatively minor improvement in ranking performance, seems to imply that the task of source/sink prediction is more difficult than the ranking task. How much of this apparent imbalance is a reflection of chemical reality, versus how much may be an artifact of our particular computational approach and ML design, is unclear. From the perspective of a human chemist, it seems surprising that the source/sink prediction task should be significantly more challenging. We suspect that ultimately the problem is the scarcity of available training data. As more data becomes available, we expect the source/sink predictions to become increasingly accurate.

We also note the dramatic difference in average runtime per reaction. For the early prototype, each prediction took 91.8 seconds on average, while the current version took less than one tenth as long, at 8.5 seconds per reaction. This performance improvement is a major boon for offline pathway searches, which benefit greatly from faster prediction speed.

5.3.2 Combinatorial data

We tested the effect of including combinatorially-generated reactions in our training set by running a direct comparison between a version trained on the standard training set and a version trained on a combinatorially-augmented training set. For this test, the standard

training set contained 10,052 elementary reactions, while the combinatorial version contained 36,902 elementary reactions. Decision thresholds were tuned to achieve an equivalent false positive rate for each version, and performance was measured on our benchmark data set of 289 reactions. The results indicated decreased accuracy for the version trained with combinatorial data. Specifically, only 70.2% of the correct products were recovered, and only 70.9% of those recovered were ranked in the top-5, compared with 80.3% and 77.6%, respectively, for the non-combinatorial version. Thus we did not include combinatorial data in the latest Reaction Predictor training set.

We hypothesize that the homogeneity of the combinatorially-generated reactions introduced biases that degraded predictive performance. We attempted to counteract this by randomly sampling a small fraction of the total reactions generated from each of our combinatorial reaction templates and including only those in the training set. Nonetheless, we observed degraded performance even when the smaller samples of reactions were used in the training set. We believe there is likely value in using combinatorial reaction generation, if done in such a way as to closely mimic the molecular variety observed in real-world reactions. However, the time required to carefully design and validate the necessary templates and constraints, while also aiming to simulate a realistic variety of molecular contexts, is by our estimation better spent simply writing high-quality elementary reactions by hand. For now, we leave combinatorial reaction generation aside for potential future work.

5.3.3 Pathway search results

Offline pathway search allows chemists to submit many jobs at once, each one searching for arbitrarily many targets, and have them run in parallel in the background, until results are found and returned via email. The improved performance of our single-step predictions, both in terms of accuracy and speed, make pathway search a powerful tool for understanding

unknown masses observed by mass spectrometry, proposing alternative synthesis pathways, or identifying side products.

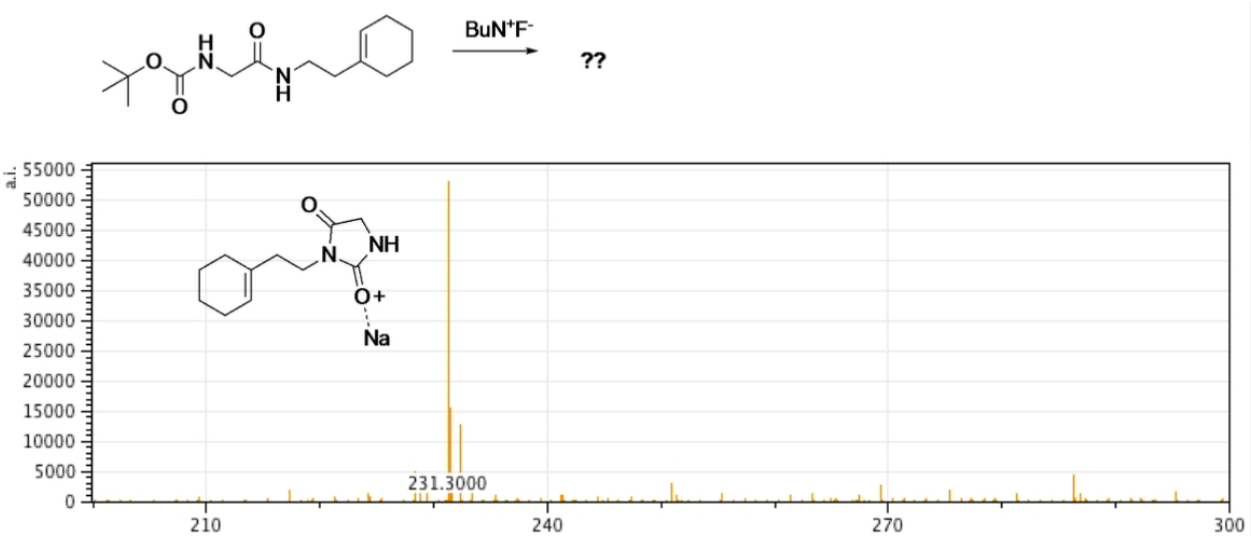


Figure 5.6: A fluoride deprotection afforded an unexpected product with m/z 231. Pathway search found 23 product pathways for the de-sodiated mass, including the highly plausible imidazolidine-2,4-dione shown here.

We submitted a number of pathway search jobs based on actual mass spectrometry data to test the feature's ability to identify unknown peaks. Figures 5.6 and 5.7 illustrate typical instances where pathway search suggested plausible product pathways or structures for unidentified target masses. In Figure 5.6, an unexpected product was found after fluoride deprotection[94]. Pathway search suggested 23 pathways to generate this mass beginning from the actual starting materials used, including the highly plausible imidazolidine-2,4-dione shown. Figure 5.7 illustrates a malonate alkylation that generated the desired product in low yield. Pathway search identified plausible structures corresponding to over-alkylation of the reactant, in effect "troubleshooting" the reaction by suggesting an explanation for the low yield observed.

Other times, pathway search may not find a match for a target of interest, or it may find matches that are implausible or were arrived at via implausible mechanisms. These failure modes can be alleviated, to some extent, by adjusting the parameters of the pathway search,

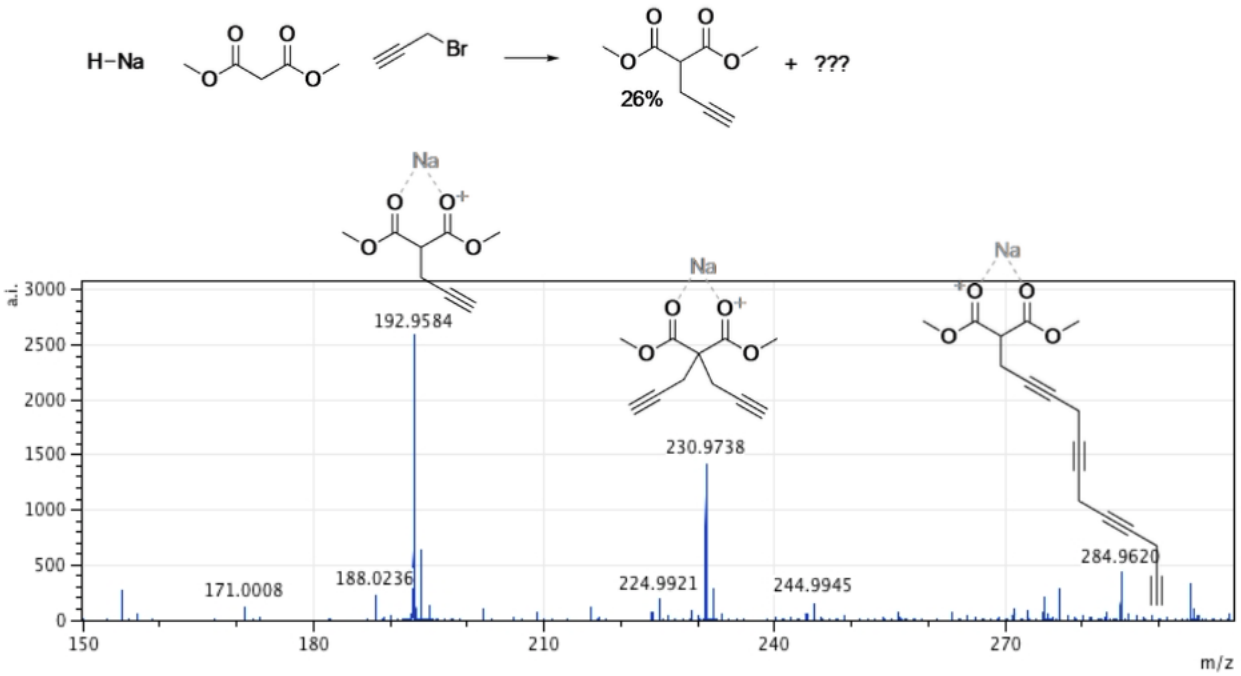


Figure 5.7: A malonate alkylation generated the desired product in low yield. Pathway search generated plausible structures corresponding to over-alkylation.

which are fully customizable. Users can control the branching factor – how many of the top-ranked reactions are pursued at each search step – as well as the maximum search depth. Even implausible structures matching a target mass may spark ideas or suggest paths to identifying plausible alternatives. Continued improvements to source/sink prediction and ranking accuracy are ultimately the most important factors contributing to pathway search efficacy.

5.4 LSTMs for source and sink prediction

Our MLP-based source and sink predictor has some inherent limitations. The input features only cover a limited amount of context for any given molecule, e.g. neighborhoods of atoms within six bonds. Furthermore, these features do not contain any information about other reactants in a given reaction, which could render the prediction task virtually impossible in

some cases, where a part of a molecule could act as a sink or a source depending on which other reactants are present.

We therefore explore a fundamentally different model to overcome these limitations. On a high level, this model is based on recurrent neural networks that operate on the canonicalized SMILES strings of all reactants. It is able to learn and use features that encode the context of the entire reaction when making predictions for locations of sinks or sources. This model is able to operate on an arbitrary number of reactants of arbitrary size/length[1] and is invariant to the ordering of reactants as presented. We achieve this invariance by operating on reactants in parallel, and extracting fixed-length feature vectors for all reactants, which are merged into groups of reaction-level features by averaging. A different part of our network uses these reaction-level, and order-invariant, features as context when it predicts sources and sinks for a given molecule.

A more technical description of the model and operating procedure is as follows: SMILES representations of reactants are canonicalized using an implicit representation of hydrogen atoms. We then remove all information that could trivialize the source/sink prediction task. For example, in our data set all potential hydrogen sinks (or sources) appear as the first element in the SMILES string. Thus we remove these and add a 'start-of-sequence' character to *every* SMILES string. The training objective of our model is to predict, for every character in the input SMILES strings of all reactants, whether it corresponds to an atom (or bond) that will act as a sink, source, or both. We further replace multi-letter atoms (e.g. 'Cu', 'Al') with new unique single characters.

The neural network model conceptually consists of eight elements, shown in Figure 5.8. The layers of this model operate on all reactants in parallel, and receive information about the operations on other reactants of the reaction only at the reaction-level-feature merging step (see step 4). The input consists of a variable number (number of reactants) of character

[1]However, we expect that the accuracy of predictions will degrade for reactions with very large molecules.

strings of variable length (length of individual reactant smiles). We therefore use zero-padding for mini-batches to speed up training.

1. The first layer of the neural network maps the input characters into a learned set of vectors (also called embedding), where the number of vectors equals the number of unique characters in all smiles.

2. This layer is followed by a one-dimensional convolutional neural network (CNN)[84], and then further by a recurrent bi-directional LSTM[57, 50].

3. The layers of the bi-directional LSTM traverse the sequences that correspond to a processed version of the original smiles strings in the forwards and backwards directions. By doing so they accumulate and compute a fixed-length vector representation of the individual reactants, i.e. learned fingerprint vectors for reactions.

4. These representation vectors are then redistributed across all processing streams of all reactants: a given stream receives the fingerprint of its own molecule and the average of fingerprints of the other molecules/reactants. This way we achieve commutativity of reactants while not mixing contexts of "this" and "other" reactants.

5. These two context vectors are concatenated and replicated across the output of the convolutional layer (2). We thereby re-use the character-level representations that are produced by the convolutional network (for efficiency and to promote weight sharing), while also augmenting them with reaction-level information about all reactants.

6. A bi-directional LSTM, separate from and unrelated to (3), then operates on this augmented stream of vectors and produces one output for each vector. These vectors have a one-to-one correspondence to characters in the input smiles.

7. The LSTM outputs are then (optionally) processed by a CNN that can refine and sharpen predictions on a local level.

8. A final sequence-distributed MLP then computes the class-probability predictions for each element of a given reactant. These are the character-level predictions of sources/sinks for every atom (or non-atom symbol) in the input smiles strings, computed for all reactants in parallel.

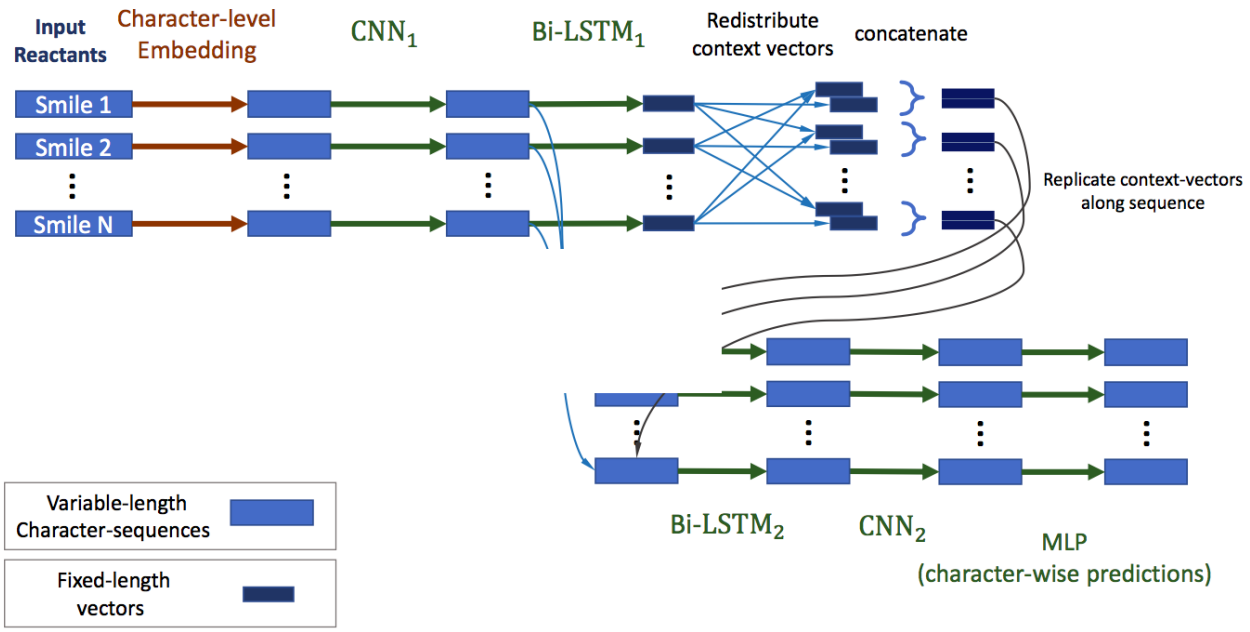


Figure 5.8: Schematic of the LSTM architecture used for source/sink prediction. This approach operates directly on SMILES strings representing all reactant molecules, and is able to make source/sink predictions using context from the entire set of reactants.

This model can predict multiple sources/sinks for one reaction or even reactant. We therefore post-process the model's predictions for all reactions by ranking the predicted probabilities, and only using the single most confident source and sink prediction for any given reaction.

The LSTM architecture uses rectified linear units for its hidden layers, and sigmoid outputs. It was trained for 60 epochs on a set of 10,052 labeled SMILES strings. Weights were initialized according to Glorot and Bengio, and updated using the Adam optimizer, with an exponentially decaying learning rate. Training was performed on an NVIDIA Titan X GPU.

To gauge the performance of the LSTM on challenging reactions, we checked its source/sink prediction accuracy on the benchmark data set of 289 reactions, and compared those results

with the MLP-based predictor's performance. Table 5.2 summarizes these results. When we considered only the top-1 highest-scored source and sink, the MLP recovered both correctly in 31.5% of the reactions, while the LSTM recovered both correctly for 47.0% of the reactions. If we consider the top-5 predictions, the MLP recovers both the correct source and the correct sink for 90.0% of the reactions, and the LSTM for 81.4% of them. Similarly, top-10 accuracy for the MLP was 96.9%, and 89.6% for the LSTM. Finally, top-20 accuracy for the MLP was 99.0%, versus 95.7% for the LSTM.

Table 5.2: MLP and LSTM source/sink prediction accuracy on the benchmark data set of 289 reactions. Predictions were considered correct only if both the true source and the true sink were identified within the top-ranking 1, 5, 10, or 20 source/sink predictions.

Top-N accuracy (both source & sink recovered)	**MLP**	**LSTM**
Top-1	31.5%	47.0%
Top-5	90.0%	81.4%
Top-10	96.9%	89.6%
Top-20	99.0%	95.7%

The top-1 score of 47.0% for the LSTM is a very promising result, and is particularly impressive as it outperforms the MLP's top-1 performance. But this result is tempered by the LSTM's lower performance at recovering the labeled sources/sinks within top-5 and top-10 constraints. The crucial job of the source/sink filtering model is to recover all reactive electron sources and sinks, while minimizing the prediction of false positives. Table 5.2 shows that while the MLP is less accurate at predicting a single best source and sink, it is still able to recover more of the correct sources/sinks, with fewer false positives, when multiple source/sink predictions are allowed. In reality, we expect that a complex chemical reactant will have multiple potentially-reactive sources and sinks, and the MLP recovers those more efficiently than the LSTM. In fact, even if we allow the LSTM to pick its top-20 proposed sources and sinks, its reaction-level accuracy of 95.7% is still slightly lower than the top-*10* MLP result of 96.9%. Nonetheless, these results are very promising and indicate potential for using an LSTM architecture for effective source/sink prediction. We emphasize that one advantage of our LSTM approach is its ability to consider the entire context of the reactants

when making its source/sink predictions. This capability fundamentally shifts our expectations of what a good source/sink filter should do. Whereas the MLP predicts a set of potentially many reactive sources and sinks, but is necessarily less accurate in choosing the single best ones because it lacks contextual information, the LSTM can see the entire reaction context, and, at least in theory, predict the best source/sink pair given that complete information.

5.5 Conclusion

Reaction Predictor is a unique and powerful tool for predicting chemical reactions at the level of elementary mechanistic steps. Deep learning coupled with a curated and expanded set of training data has yielded significant advances in both speed and predictive accuracy. Pathway Search takes advantage of these performance gains to aid in the identification of unknown products by searching in the background and emailing results to the user. Finally, we demonstrate a promising LSTM-based approach to predicting reactive sites based solely on SMILES strings. This could be used in future work to complement and improve the existing MLP-based source/sink filters. Ultimately we expect Reaction Predictor will continue to improve over time as new opportunities for refinement are identified, and as more training data becomes available.

Bibliography

[1] Daylight. URL `www.daylight.com`.

[2] *OEChem, version 1.7.4.* OpenEye Scientific Software, Inc., Santa Fe, NM, USA, 2012 www.eyesopen.com.

[3] *Marvin Beans, version 6.1.* ChemAxon Ltd., www.chemaxon.com. 2013.

[4] Enrique Abad, Roland K. Zenn, and Johannes Kästner. Reaction mechanism of monoamine oxidase from qm/mm calculations. *The Journal of Physical Chemistry B*, 117(46):14238–14246, 2013. doi: 10.1021/jp4061522. URL `http://dx.doi.org/10.1021/jp4061522`. PMID: 24164690.

[5] Tatsuya Akutsu. Efficient extraction of mapping rules of atoms from enzymatic reaction data. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, RECOMB '03, pages 1–8, New York, NY, USA, 2003. ACM. ISBN 1-58113-635-8. doi: 10.1145/640075.640076. URL `http://doi.acm.org/10.1145/640075.640076`.

[6] K. E. Altieri, S. P. Seitzinger, A. G. Carlton, B. J. Turpin, G. C. Klein, and A. G. Marshall. Oligomers formed through in-cloud methylglyoxal reactions: Chemical composition, properties, and mechanisms investigated by ultra-high resolution ft-icr mass spectrometry. *Atmos. Environ.*, 42:1476, 2008.

[7] S. E. Anderson, J. Franko, L. G. Jackson, J. R. Wells, J. E. Ham, and B. J. Meade. Irritancy and allergic responses induced by exposure to the indoor air chemical 4-oxopentanal. *Toxicol. Sci.*, 127:371, 2012.

[8] E. Andre, B. Campi, S. Materazzi, M. Trevisani, S. Amadesi, D. Massi, C. Creminon, N. Vaksman, R. Nassini, M. Civelli, P. G. Baraldi, D. P. Poole, N. W. Bunnett, P. Geppetti, and R. Patacchini. Cigarette smoke-induced neurogenic inflammation is mediated byalpha,beta-unsaturated aldehydes and the trpa1 receptor in rodents. *J. Clin. Invest.*, 118:2574, 2008.

[9] M. O. Andreae and D. Rosenfeld. Aerosol-cloud-precipitation interactions. part 1. the nature and sources of cloud-active aerosols. *Earth-Sci. Rev.*, 89:13, 2008.

[10] Milica Andrejić and Ricardo A. Mata. Local hybrid qm/qm calculations of reaction pathways in metallobiosites. *Journal of Chemical Theory and Computation*, 10

(12):5397–5404, 2014. doi: 10.1021/ct5008313. URL http://dx.doi.org/10.1021/ct5008313. PMID: 26583223.

[11] Joannis Apostolakis, Oliver Sacher, Robert Körner, and Johann Gasteiger. Automatic determination of reaction mappings and reaction center cnformation. 2. validation on a biochemical reaction database. *J. Chem. Inf. Model.*, 48(6):1190–1198, 2008. doi: 10.1021/ci700433d. URL http://pubs.acs.org/doi/abs/10.1021/ci700433d. PMID: 18533714.

[12] R. Atkinson. Atmospheric chemistry of vocs and nox. *Atmos. Environ.*, 34:2063, 2000.

[13] Philip S Bailey. *Ozonation in organic chemistry V1: Olefinic compounds*. Elsevier, 1978.

[14] Pierre Baldi and Peter Sadowski. The dropout learning algorithm. *Artificial Intelligence*, 210:78 – 122, 2014. ISSN 0004-3702. doi: http://dx.doi.org/10.1016/j.artint.2014.02.004. URL http://www.sciencedirect.com/science/article/pii/S0004370214000216.

[15] U. Baltensperger, R. Chirico, P. F. DeCarlo, J. Dommen, K. Gaeggeler, M. F. Heringa, M. Li, A. S. Prvt, M. R. Alfarra, D. S. Gross, and M. Kalberer. Recent developments in the mass spectrometry of atmospheric aerosols. *Eur. J. Mass. Spectrom.*, 16:389, 2010.

[16] K. C. Barsanti and J. F. Pankow. Thermodynamics of the formation of atmospheric organic particulate matter by accretion reactions - 2. dialdehydes, methylglyoxal, and diketones. *Atmos. Environ.*, 39:6597, 2005.

[17] K. C. Barsanti and J. F. Pankow. Thermodynamics of the formation of atmospheric organic particulate matter by accretion reactions - part 3: Carboxylic and dicarboxylic acids. *Atmos. Environ.*, 40:6676, 2006.

[18] A. P. Bateman, M. L. Walser, Y. Desyaterik, J. Laskin, A. Laskin, and S. A. Nizkorodov. The effect of solvent on the analysis of secondary organic aerosol using electrospray ionization mass spectrometry. *Environ. Sci. Technol.*, 42:7341, 2008.

[19] A. P. Bateman, S. A. Nizkorodov, J. Laskin, and A. Laskin. Time-resolved molecular characterization of limonene/ozone aerosol using high-resolution electrospray ionization mass spectrometry. *Phys. Chem. Chem. Phys.*, 11:7931, 2009.

[20] Kelvin L. Baumann, Donald E. Butler, Carl F. Deering, Kenneth E. Mennen, Alan Millar, Thomas N. Nanninga, Charles W. Palmer, and Bruce D. Roth. The convergent synthesis of ci-981, an optically active, highly potent, tissue selective inhibitor of hmg-coa reductase. *Tetrahedron Letters*, 33(17):2283 – 2284, 1992. ISSN 0040-4039. doi: http://dx.doi.org/10.1016/S0040-4039(00)74190-6. URL http://www.sciencedirect.com/science/article/pii/S0040403900741906.

[21] K. Berhane, M. Widersten, A. Engstrom, J. W. Kozarich, and B. Mannervik. Detoxication of base propenals and other alpha, beta-unsaturated aldehyde products of radical reactions and lipid peroxidation by human glutathione transferases. *Proc. Natl. Acad. Sci. U.S.A.*, 91:1480, 1994.

[22] Edward S Blurock. Reaction: system for modeling chemical reactions. *Journal of chemical information and computer sciences*, 35(3):607–616, 1995.

[23] D. L. Bones, D. K. Henricksen, S. A. Mang, M. Gonsior, A. P. Bateman, T. B. Nguyen, W. J. Cooper, and S. A. Nizkorodov. Appearance of strong absorbers and fluorophores in limonene-o3 secondary organic aerosol due to nh4+-mediated chemical aging over long time scales. *J. Geophys. Res.*, 115:D05203, 2010.

[24] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, pages 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc. URL `http://dl.acm.org/citation.cfm?id=2987189.2987282`.

[25] A. G. Carlton, C. Wiedinmyer, and J. H. Kroll. A review of secondary organic aerosol formation from isoprene. *Atmos. Chem. Phys.*, 9:4987, 2009.

[26] A. W. H. Chan, M. N. Chan, J. D. Surratt, P. S. Chhabra, C. L. Loza, J. D. Crounse, L. D. Yee, R. C. Flagan, P. O. Wennberg, and J. H. Seinfeld. Role of aldehyde chemistry and nox concentrations in secondary organic aerosol formation. *Atmos. Chem. Phys.*, 10:7169, 2010.

[27] M. N. Chan, J. D. Surratt, M. Claeys, E. S. Edgerton, R. L. Tanner, S. L. Shaw, M. Zheng, E. M. Knipping, N. C. Eddingsaas, P. O. Wennberg, and J. H. Seinfeld. Characterization and quantification of isoprene-derived epoxydiols in ambient aerosol in the southeastern united states. *Environ. Sci. Technol.*, 44:4590, 2010.

[28] J. H. Chen and P. Baldi. No electron left behind: A rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.*, 49:2034, 2009.

[29] Jonathan H. Chen and Pierre Baldi. No electron left behind: A rule-based expert system to predict chemical reactions and reaction mechanisms. *Journal of Chemical Information and Modeling*, 49(9):2034–2043, 2009. doi: 10.1021/ci900157k. URL `http://dx.doi.org/10.1021/ci900157k`. PMID: 19719121.

[30] William Lingran Chen, David Z. Chen, and Keith T. Taylor. Automatic reaction mapping and reaction center detection. *WIREs Comput. Mol. Sci.*, 00:1–30, 2013. ISSN 1759-0884. doi: 10.1002/wcms.1140. URL `http://dx.doi.org/10.1002/wcms.1140`.

[31] John Y.L Chung, Guo-Jie Ho, Michel Chartrain, Chris Roberge, Dalian Zhao, John Leazer, Roger Farr, Michael Robbins, Kateeta Emerson, David J Mathre, James M

McNamara, David L Hughes, Edward J.J Grabowski, and Paul J Reider. Practical chemoenzymatic synthesis of a 3-pyridylethanolamino 3 adrenergic receptor agonist. *Tetrahedron Letters*, 40(37):6739 – 6743, 1999. ISSN 0040-4039. doi: http://dx.doi.org/10.1016/S0040-4039(99)01353-2. URL `http://www.sciencedirect.com/science/article/pii/S0040403999013532`.

[32] M. Claeys, B. Graham, G. Vas, W. Wang, R. Vermeylen, V. Pashynska, J. Cafmeyer, P. Guyon, M. O. Andreae, P. Artaxo, and W. Maenhaut. Formation of secondary organic aerosols through photooxidation of isoprene. *Science*, 303:1173, 2004.

[33] John D. Crabtree and Dinesh P. Mehta. Automated reaction mapping. *J. Exp. Algorithmics*, 13:15:1.15–15:1.29, February 2009. ISSN 1084-6654. doi: 10.1145/1412228.1498697. URL `http://doi.acm.org/10.1145/1412228.1498697`.

[34] R. Criegee. Mechanism of ozonolysis. *Angew. Chem., Int. Ed. Engl.*, 14:745, 1975.

[35] D. O. De Haan, M. A. Tolbert, and J. L. Jimenez. Atmospheric condensed-phase reactions of glyoxal with methylamine. *Geophys. Res. Lett.*, 36:5, 2009.

[36] R. Delfino, J. Sioutas, and C. Malik. S., potential role of ultrafine particles in associations between airborne particle mass and cardiovascular health. *Environ. Health Perspect.*, 113:934, 2005.

[37] K. S. Docherty, E. A. Stone, I. M. Ulbrich, P. F. DeCarlo, D. C. Snyder, J. J. Schauer, R. E. Peltier, R. J. Weber, S. M. Murphy, J. H. Seinfeld, B. D. Grover, D. J. Eatough, and J. L. Jimenez. Apportionment of primary and secondary organic aerosols in southern california during the 2005 study of organic aerosols in riverside (soar-1). *Environ. Sci. Technol.*, 42:7655, 2008.

[38] D. W. Dockery, C. A. Pope, X. Xu, J. D. Spengler, J. H. Ware, M. E. Fay, B. G. Ferris, and F. E. Speizer. An association between air pollution and mortality in six u.s. cities. *N. Engl. J. Med.*, 329:1753, 1993.

[39] G. T. Drozd and N. M. Donahue. Pressure dependence of stabilized criegee intermediate formation from a sequence of alkenes. *J. Phys. Chem. A*, 115:4381, 2011.

[40] E. O. Edney, T. E. Kleindienst, M. Jaoui, M. Lewandowski, J. H. Offenberg, W. Wang, and M. Claeys. Formation of 2-methyl tetrols and 2-methylglyceric acid in secondary organic aerosol from laboratory irradiated isoprene/nox/so2/air mixtures and their detection in ambient pm2.5 samples collected in the eastern united states. *Atmos. Environ.*, 39:5281, 2005.

[41] Barbara J Finlayson-Pitts and James N Pitts Jr. *Chemistry of the upper and lower atmosphere: theory, experiments, and applications*. Academic press, 1999.

[42] Eric L. First, Chrysanthos E. Gounaris, and Christodoulos A. Floudas. Stereochemically consistent reaction mapping and identification of multiple reaction mechanisms through integer linear optimization. *J. Chem. Inf. Model.*, 52(1):84–92, 2012. doi: 10.1021/ci200351b. URL `http://pubs.acs.org/doi/abs/10.1021/ci200351b`.

[43] D. R. Fooshee, T. B. Nguyen, S. A. Nizkorodov, J. Laskin, A. Laskin, and P. Baldi. Cobra: A computational brewing application for predicting the molecular composition of organic aerosols. *Environ. Sci. Technol.*, 46:6048, 2012.

[44] P. Fruekilde, J. Hjorth, N. R. Jensen, D. Kotzias, and B. Larsen. Ozonolysis at vegetation surfaces: a source of acetone, 4-oxopentanal, 6-methyl-5-hepten-2-one, and geranyl acetone in the troposphere. *Atmos. Environ.*, 32:1893, 1998.

[45] D. Fu, C. Leng, J. Kelley, G. Zeng, Y. Zhang, and Y. Liu. Atr-ir study of ozone initiated heterogeneous oxidation of squalene in an indoor environment. *Environ. Sci. Technol.*, 47:10611, 2013.

[46] R. M. Garland, M. J. Elrod, K. Kincaid, M. R. Beaver, J. L. Jimenez, and M. A. Tolbert. Acid-catalyzed reactions of hexanal on sulfuric acid particles: identification of reaction products. *Atmos. Environ.*, 40:6863, 2006.

[47] Johann Gasteiger and Clemens Jochum. Eros a computer program for generating sequences of reactions. *Organic Compunds*, pages 93–126, 1978.

[48] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL `http://proceedings.mlr.press/v9/glorot10a.html`.

[49] A. L. Gomez, J. Park, M. L. Walser, A. Lin, and S. A. Nizkorodov. Uv photodissociation spectroscopy of oxidized undecylenic acid films. *J. Phys. Chem. A*, 110:3584, 2006.

[50] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005.

[51] D. Grosjean, E. L. Williams, and E. Grosjean. Atmospheric chemistry of isoprene and of its carbonyl products. *Environ. Sci. Technol.*, 27:830, 1993.

[52] J. F. Hamilton, A. C. Lewis, T. J. Carey, and J. C. Wenger. Characterization of polar compounds and oligomers in secondary organic aerosol using liquid chromatography coupled to mass spectrometry. *Anal. Chem.*, 80:474, 2008.

[53] Katja Hansen, Grégoire Montavon, Franziska Biegler, Siamac Fazli, Matthias Rupp, Matthias Scheffler, O. Anatole von Lilienfeld, Alexandre Tkatchenko, and Klaus-Robert Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9 (8):3404–3419, 2013. doi: 10.1021/ct400195d. URL `http://dx.doi.org/10.1021/ct400195d`. PMID: 26584096.

[54] R. M. Healy, J. C. Wenger, A. Metzger, J. Duplissy, M. Kalberer, and J. Dommen. Gas/particle partitioning of carbonyls in the photooxidation of isoprene and 1,3,5-trimethylbenzene. *Atmos. Chem. Phys.*, 8:3215, 2008.

[55] Markus Heinonen, Sampsa Lappalainen, Taneli Mielikäinen, and Juho Rousu. Computing atom mappings for biochemical reactions without subgraph isomorphism. *J. Comput. Biol.*, 18(1):43–58, 2011. URL `http://dblp.uni-trier.de/db/journals/jcb/jcb18.html#HeinonenLMR11`.

[56] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012. URL `http://arxiv.org/abs/1207.0580`.

[57] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[58] R. W. Holman. The art of writing reasonable organic reaction mechanisms, 2nd edition (grossman, robert b.). *Journal of Chemical Education*, 80(11):1259, 2003. doi: 10.1021/ed080p1259. URL `http://dx.doi.org/10.1021/ed080p1259`.

[59] Jonathan Huang, Carlos Guestrin, and Leonidas Guibas. Fourier theoretic probabilistic inference over permutations. *The Journal of Machine Learning Research*, 10:997–1070, 2009.

[60] CA James, D Weininger, and J Delany. Daylight theory manual daylight version 4.82. daylight chemical information systems, 2003.

[61] Jon Paul Janet and Heather J. Kulik. Predicting electronic structure properties of transition metal complexes with neural networks. *Chem. Sci.*, 8:5137–5152, 2017. doi: 10.1039/C7SC01247K. URL `http://dx.doi.org/10.1039/C7SC01247K`.

[62] M. Jang, B. Carroll, B. Chandramouli, and R. M. Kamens. Particle growth by acid-catalyzed heterogeneous reactions of organic carbonyls on preexisting aerosols. *Environ. Sci. Technol.*, 37:3828, 2003.

[63] M. S. Jang, N. M. Czoschke, and A. L. Northcross. Semiempirical model for organic aerosol growth by acid-catalyzed heterogeneous reactions of organic carbonyls. *Environ. Sci. Technol.*, 39:164, 2005.

[64] J. L. Jimenez, M. R. Canagaratna, N. M. Donahue, A. S. H. Prevot, Q. Zhang, J. H. Kroll, P. F. DeCarlo, J. D. Allan, H. Coe, N. L. Ng, A. C. Aiken, K. S. Docherty, I. M. Ulbrich, A. P. Grieshop, A. L. Robinson, J. Duplissy, J. D. Smith, K. R. Wilson, V. A. Lanz, C. Hueglin, Y. L. Sun, J. Tian, A. Laaksonen, T. Raatikainen, J. Rautiainen, P. Vaattovaara, M. Ehn, M. Kulmala, J. M. Tomlinson, D. R. Collins, M. J. Cubison, E. J. Dunlea, J. A. Huffman, T. B. Onasch, M. R. Alfarra, P. I. Williams, K. Bower, Y. Kondo, J. Schneider, F. Drewnick, S. Borrmann, S. Weimer, K. Demerjian, D. Salcedo, L. Cottrell, R. Griffin, A. Takami, T. Miyoshi, S. Hatakeyama, A. Shimono, J. Y.

Sun, Y. M. Zhang, K. Dzepina, J. R. Kimmel, D. Sueper, J. T. Jayne, S. C. Herndon, A. M. Trimborn, L. R. Williams, E. C. Wood, A. M. Middlebrook, C. E. Kolb, U. Baltensperger, and D. R. Worsnop. Evolution of organic aerosols in the atmosphere. *Science*, 326:1525, 2009.

[65] M. Kanakidou, J. H. Seinfeld, S. N. Pandis, I. Barnes, F. J. Dentener, M. C. Facchini, R. Van Dingenen, B. Ervens, A. Nenes, C. J. Nielsen, E. Swietlicki, J. P. Putaud, Y. Balkanski, S. Fuzzi, J. Horth, G. K. Moortgat, R. Winterhalter, C. E. L. Myhre, K. Tsigaridis, E. Vignati, E. G. Stephanou, and J. Wilson. Organic aerosol and global climate modelling: a review. *Atmos. Chem. Phys.*, 5:1053, 2005.

[66] K. E. Kautzman, J. D. Surratt, M. N. Chan, A. W. H. Chan, S. P. Hersey, P. S. Chhabra, N. F. Dalleska, P. O. Wennberg, R. C. Flagan, and J. H. Seinfeld. Chemical composition of gas- and aerosol-phase products from the photooxidation of naphthalene. *J. Phys. Chem. A*, 114:913, 2009.

[67] M. A. Kayala, C. A. Azencott, J. H. Chen, and P. Baldi. Learning to predict chemical reactions. *J. Chem. Inf. Model.*, 51:2209, 2011.

[68] Matthew A. Kayala and Pierre Baldi. Reactionpredictor: Prediction of complex chemical reactions at the mechanistic level using machine learning. *Journal of Chemical Information and Modeling*, 52(10):2526–2540, 2012. doi: 10.1021/ci3003039. URL `http://dx.doi.org/10.1021/ci3003039`. PMID: 22978639.

[69] Matthew A. Kayala, Chloé-Agathe Azencott, Jonathan H. Chen, and Pierre Baldi. Learning to predict chemical reactions. *Journal of Chemical Information and Modeling*, 51(9):2209–2222, 2011. doi: 10.1021/ci200207y. URL `http://dx.doi.org/10.1021/ci200207y`. PMID: 21819139.

[70] J. P. Kehrer and S. S. Biswal. The molecular effects of acrolein. *Toxicol. Sci.*, 57:6, 2000.

[71] S. Kim, R. W. Kramer, and P. G. Hatcher. Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van krevelen diagram. *Anal. Chem.*, 75:5336, 2003.

[72] S. Kim, R. P. Rodgers, and A. G. Marshall. Truly exact mass: Elemental composition can be determined uniquely from molecular mass measurement at ¡0.1 mda accuracy for molecules up to ¡500 da. *Int. J. Mass Spectrom.*, 251:260, 2006.

[73] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://arxiv.org/abs/1412.6980`.

[74] T. E. Kleindienst, M. Lewandowski, J. H. Offenberg, M. Jaoui, and E. O. Edney. The formation of secondary organic aerosol from the isoprene plus oh reaction in the absence of nox. *Atmos. Chem. Phys.*, 9:6541, 2009.

[75] David J. Kopecky and Scott D. Rychnovsky. Mukaiyama aldolprins cyclization cascade reaction: a formal total synthesis of leucascandrolide a. *Journal of the American Chemical Society*, 123(34):8420–8421, 2001. doi: 10.1021/ja011377n. URL http://dx.doi.org/10.1021/ja011377n. PMID: 11516301.

[76] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL https://link.aps.org/doi/10.1103/PhysRevE.69.066138.

[77] J. H. Kroll, J. S. Clarke, N. M. Donahue, J. G. Anderson, and K. L. Demerjian. Mechanism of hox formation in the gas-phase ozone-alkene reaction. 1. direct, pressure-dependent measurements of prompt oh yields. *J. Phys. Chem. A*, 105:1554, 2001.

[78] E. V. Kunenkov, A. S. Kononikhin, I. V. Perminova, N. Hertkorn, A. Gaspar, P. Schmitt-Kopplin, I. A. Popov, A. V. Garmash, and E. N. Nikolaev. Total mass difference statistics algorithm: a new approach to identification of high-mass building blocks in electrospray ionization fourier transform ion cyclotron mass spectrometry data of natural organic matter. *Anal. Chem.*, 81:10106, 2009.

[79] C. Lambert, J. McCue, M. Portas, Y. Ouyang, J. Li, T. G. Rosano, A. Lazis, and B. M. Freed. Acrolein in cigarette smoke inhibits t-cell responses. *J. Allergy Clin. Immunol.*, 116:916, 2005.

[80] A. Laskin, J. S. Smith, and J. Laskin. Molecular characterization of nitrogen-containing organic compounds in biomass burning aerosols using high-resolution mass spectrometry. *Environ. Sci. Technol.*, 43:3764, 2009.

[81] J. Laskin, A. Laskin, P. J. Roach, G. W. Slysz, G. A. Anderson, S. A. Nizkorodov, D. L. Bones, and L. Q. Nguyen. High-resolution desorption electrospray ionization mass spectrometry for chemical characterization of organic aerosols. *Anal. Chem.*, 82: 2048, 2010.

[82] Mario Latendresse, Jeremiah P. Malerich, Mike Travers, and Peter D. Karp. Accurate atom-mapping computation for biochemical reactions. *J. Chem. Inf. Model.*, 52(11): 2970–2982, 2012. doi: 10.1021/ci3002217. URL http://pubs.acs.org/doi/abs/10.1021/ci3002217.

[83] Markus Leber, Volker Egelhofer, Ida Schomburg, and Dietmar Schomburg. Automatic assignment of reaction operators to enzymatic reactions. *Bioinformatics*, 25(23):3135–3142, December 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp549. URL http://dx.doi.org/10.1093/bioinformatics/btp549.

[84] Yann LeCun and Yoshua Bengio. The handbook of brain theory and neural networks. chapter Convolutional Networks for Images, Speech, and Time Series, pages 255–258. MIT Press, Cambridge, MA, USA, 1998. ISBN 0-262-51102-9. URL http://dl.acm.org/citation.cfm?id=303568.303704.

[85] H. J. Lee, P. K. Aiona, A. Laskin, J. Laskin, and S. A. Nizkorodov. Effect of solar radiation on the optical properties and molecular composition of laboratory proxies of atmospheric brown carbon. *Environ. Sci. Technol.*, 48:10217, 2014.

[86] Rong-Zhen Liao and Walter Thiel. Comparison of qm-only and qm/mm models for the mechanism of tungsten-dependent acetylene hydratase. *Journal of Chemical Theory and Computation*, 8(10):3793–3803, 2012. doi: 10.1021/ct3000684. URL `http://dx.doi.org/10.1021/ct3000684`. PMID: 26593020.

[87] R. Daniel Libby. Advanced organic chemistry, part a: Structure and mechanism, 4th edition (carey, francis a.; sundberg, richard j.). *Journal of Chemical Education*, 78(3):314, 2001. doi: 10.1021/ed078p314.1. URL `http://dx.doi.org/10.1021/ed078p314.1`.

[88] J. Liggio, S. M. Li, and R. McLaren. Reactive uptake of glyoxal by particulate matter. *J. Geophys. Res., [Atmos.]*, 110:D10304, 2005.

[89] J. Liggio, S.-M. Li, and R. McLaren. Heterogeneous reactions of glyoxal on particulate matter: identification of acetals and sulfate esters. *Environ. Sci. Technol.*, 39:1532, 2005.

[90] S. Matsunaga, M. Mochida, and K. Kawamura. Variation on the atmospheric concentrations of biogenic carbonyl compounds and their removal processes in the northern forest at moshiri, hokkaido island in japan. *J. Geophys. Res.*, 109:D04302, 2004.

[91] S. Matsunaga, M. Mochida, and K. Kawamura. High abundance of gaseous and particulate 4-oxopentanal in the forestal atmosphere. *Chemosphere*, 55:1143, 2004.

[92] J. L. Mauderly and J. C. Chow. Health effects of organic aerosols. *Inhalation Toxicol.*, 20:257, 2008.

[93] James J. McGregor and Peter Willett. Use of a maximum common subgraph algorithm in the automatic identification of ostensible bond changes occurring in chemical reactions. *J. Chem. Inf. Comput. Sci.*, 21(3):137–140, 1981. doi: 10.1021/ci00031a005. URL `http://pubs.acs.org/doi/abs/10.1021/ci00031a005`.

[94] Aaron D. Mood, Ilandari Dewage Udara Anulal Premachandra, Stanley Hiew, Fuqiang Wang, Kevin A. Scott, Nathan J. Oldenhuis, Haoping Liu, and David L. Van Vranken. Potent antifungal synergy of phthalazinone and isoquinolones with azoles against candida albicans. *ACS Medicinal Chemistry Letters*, 8(2):168–173, 2017. doi: 10.1021/acsmedchemlett.6b00355. URL `http://dx.doi.org/10.1021/acsmedchemlett.6b00355`.

[95] Johann Mulzer, Andreas Pointner, Alexander Chucholowski, and Gisela Bruntrup. threo-3-hydroxycarboxylic acids as key intermediates in a highly stereoselective synthesis of (z)- and (e)-olefins and enol ethers. *J. Chem. Soc., Chem. Commun.*, pages 52–54, 1979. doi: 10.1039/C39790000052. URL `http://dx.doi.org/10.1039/C39790000052`.

[96] J. Munkres. Algorithms for the assignment and transportation problems. *J. Soc. Ind. Appl. Math.*, 5(1):32–38, March 1957.

[97] T. B. Nguyen, J. Laskin, A. Laskin, and S. A. Nizkorodov. Nitrogen containing organic compounds and oligomers in secondary organic aerosol formed by photooxidation of isoprene. *Environ. Sci. Technol.*, 45:6908, 2011.

[98] T. B. Nguyen, P. J. Roach, J. Laskin, A. Laskin, and S. A. Nizkorodov. Effect of humidity on the composition of isoprene photooxidation secondary organic aerosol. *Atmos. Chem. Phys.*, 11:6931, 2011.

[99] S. A. Nizkorodov, J. Laskin, and A. Laskin. Molecular chemistry of organic aerosols through the application of high resolution mass spectrometry. *Phys. Chem. Chem. Phys.*, 13:3612, 2011.

[100] S. A. Nizkorodov, J. Laskin, and A. Laskin. Molecular chemistry of organic aerosols through the application of high resolution mass spectrometry. *Phys. Chem. Chem. Phys.*, 13:3612, 2011.

[101] Y. Chauvin P. Baldi. Neural networks for fingerprint recognition. *Neural Computation*, 5(3):402–418, 1993.

[102] Leo A. Paquette, Jiong Yang, and Yun Oliver Long. Concerning the antileukemic agent jatrophatrione: the first total synthesis of a [5.9.5] tricyclic diterpene. *Journal of the American Chemical Society*, 124(23):6542–6543, 2002. doi: 10.1021/ja020292z. URL `http://dx.doi.org/10.1021/ja020292z`. PMID: 12047168.

[103] F. Paulot, J. D. Crounse, H. G. Kjaergaard, J. H. Kroll, J. H. Seinfeld, and P. O. Wennberg. Isoprene photooxidation: new insights into the production of acids and organic nitrates. *Atmos. Chem. Phys.*, 9:1479, 2009.

[104] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[105] M. Petitjean, E. Reyes-Perez, D. Perez, P. Mirabel, and S. Le Calve. Vapor pressure measurements of hydroxyacetaldehyde and hydroxyacetone in the temperature range (273 to 356) k. *J. Chem. Eng. Data*, 55:852, 2009.

[106] L. Petrick and Y. Dubowski. Heterogeneous oxidation of squalene film by ozone under various indoor conditions. *Indoor Air*, 19:381, 2009.

[107] Iakov Polyak, Manfred T. Reetz, and Walter Thiel. Quantum mechanical/molecular mechanical study on the mechanism of the enzymatic baeyer–villiger reaction. *Journal of the American Chemical Society*, 134(5):2732–2741, 2012. doi: 10.1021/ja2103839. URL `http://dx.doi.org/10.1021/ja2103839`. PMID: 22239272.

[108] C. A. Pope, R. T. Burnett, M. J. Thun, E. E. Calle, D. Krewski, K. Ito, and G. D. Thurston. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *J. Am. Med. Assoc.*, 287:1132, 2002.

[109] U. Poschl. Atmospheric aerosols: Composition, transformation, climate and health effects. *Angew. Chem., Int. Ed. Engl.*, 44:7520, 2005.

[110] A. Rangarajan and E. D. Mjolsness. A lagrangian relaxation network for graph matching. *Trans. Neur. Netw.*, 7(6):1365–1381, November 1996. ISSN 1045-9227. doi: 10.1109/72.548165. URL `http://dx.doi.org/10.1109/72.548165`.

[111] A. Reinhardt, C. Emmenegger, B. Gerrits, C. Panse, J. Dommen, U. Baltensperger, R. Zenobi, and M. Kalberer. Ultrahigh mass resolution and accurate mass measurements as a tool to characterize oligomers in secondary organic aerosols. *Anal. Chem.*, 79:4074, 2007.

[112] A. C. Rohr, S. A. Shore, and J. D. Spengler. Repeated exposure to isoprene oxidation products causes enhanced respiratory tract effects in multiple murine strains. *Inhalation Toxicol.*, 15:1191, 2003.

[113] Y. Rudich, N. M. Donahue, and T. F. Mentel. Aging of organic aerosol: Bridging the gap between laboratory and field studies. *Annu. Rev. Phys. Chem.*, 58:321, 2007.

[114] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.*, 108:058301, Jan 2012. doi: 10.1103/PhysRevLett.108.058301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.108.058301`.

[115] Peter Sadowski, David Fooshee, Niranjan Subrahmanya, and Pierre Baldi. Synergies between quantum mechanics and machine learning in reaction prediction. *Journal of Chemical Information and Modeling*, 56(11):2125–2128, 2016. doi: 10.1021/acs.jcim.6b00351. URL `http://dx.doi.org/10.1021/acs.jcim.6b00351`. PMID: 27749058.

[116] J. M. Samet, F. Dominici, F. C. Curriero, I. Coursac, and S. L. Zeger. Fine particulate air pollution and mortality in 20 u.s. cities, 19871994. *N. Engl. J. Med.*, 343:1742, 2000.

[117] P. Schmitt-Kopplin, A. Gelencsr, E. Dabek-Zlotorzynska, G. Kiss, N. Hertkorn, M. Harir, Y. Hong, and I. Gebefgi. Analysis of the unresolved organic fraction in atmospheric aerosols with ultrahigh-resolution mass spectrometry and nuclear magnetic resonance spectroscopy: Organosulfates as photochemical smog constituents. *Anal. Chem.*, 82:8017, 2010.

[118] J. H. Seinfeld and J. F. Pankow. Organic atmospheric particulate material. *Annu. Rev. Phys. Chem.*, 54:121, 2003.

[119] Mitsuo Shoji, Hiroshi Isobe, and Kizashi Yamaguchi. Qm/mm study of the s2 to s3 transition reaction in the oxygen-evolving complex of photosystem ii. *Chemical*

Physics Letters, 636:172 – 179, 2015. ISSN 0009-2614. doi: http://dx.doi.org/10.1016/j.cplett.2015.07.039. URL http://www.sciencedirect.com/science/article/pii/S0009261415005527.

[120] R. S. Spaulding, G. W. Schade, A. H. Goldstein, and M. J. Charles. Characterization of secondary atmospheric photooxidation products: Evidence for biogenic and anthropogenic sources. *J. Geophys. Res.*, 108:4247, 2003.

[121] J. D. Surratt, S. M. Murphy, J. H. Kroll, N. L. Ng, L. Hildebrandt, A. Sorooshian, R. Szmigielski, R. Vermeylen, W. Maenhaut, M. Claeys, R. C. Flagan, and J. H. Seinfeld. Chemical composition of secondary organic aerosol formed from the photooxidation of isoprene. *J. Phys. Chem. A*, 110:9665, 2006.

[122] S. Joshua Swamidass, Jonathan Chen, Jocelyne Bruand, Peter Phung, Liva Ralaivola, and Pierre Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl1):i359–i368, 2005. doi: 10.1093/bioinformatics/bti1055. URL +http://dx.doi.org/10.1093/bioinformatics/bti1055.

[123] R. Szmigielski, J. D. Surratt, R. Vermeylen, K. Szmigielska, J. H. Kroll, N. L. Ng, S. M. Murphy, A. Sorooshian, J. H. Seinfeld, and M. Claeys. Characterization of 2-methylglyceric acid oligomers in secondary organic aerosol formed from the photooxidation of isoprene using trimethylsilylation and gas chromatography/ion trap mass spectrometry. *J. Mass Spectrom.*, 42:101, 2007.

[124] Y. Tan, A. G. Carlton, S. P. Seitzinger, and B. J. Turpin. Soa from methylglyoxal in clouds and wet aerosols: Measurement and prediction of key products. *Atmos. Environ.*, 44:5218, 2010.

[125] K. Tang, J. S. Page, and R. D. Smith. Charge competition and the linear dynamic range of detection in electrospray ionization mass spectrometry. *J. Am. Soc. Mass Spectrom.*, 15:1416, 2004.

[126] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: a large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, ICML '05, pages 896–903, New York, NY, USA, 2005. ACM. ISBN 1-59593-180-5. doi: 10.1145/1102351.1102464. URL http://doi.acm.org/10.1145/1102351.1102464.

[127] M. P. Tolocka, M. Jang, J. M. Ginter, F. J. Cox, R. M. Kamens, and M. V. Johnston. Formation of oligomers in secondary organic aerosol. *Environ. Sci. Technol.*, 38:1428, 2004.

[128] D. Trombetta, A. Saija, G. Bisignano, S. Arena, S. Caruso, G. Mazzanti, N. Uccella, and F. Castelli. Study on the mechanisms of the antibacterial action of some plant alpha,beta-unsaturated aldehydes. *Lett. Appl. Microbiol.*, 35:285, 2002.

[129] Yong Qiang Tu, Li Ming Yang, and Yao Zu Chen. A facial one-step approach to stereospecific spirocyclic diols from -hydroxyepoxides. *Chemistry Letters*, 27(4):285–286, 1998. doi: 10.1246/cl.1998.285. URL https://doi.org/10.1246/cl.1998.285.

[130] C. Turek and F. C. Stintzing. Stability of essential oils: A review. *Compr. Rev. Food Sci. Food Saf.*, 12:40, 2013.

[131] C. Von Sonntag and H. P. Schuchmann. The elucidation of peroxyl radical reactions in aqueous solution with the help of radiation-chemical methods. *Angew. Chem., Int. Ed. Engl.*, 30:1229, 1991.

[132] M. L. Walser, Y. Desyaterik, J. Laskin, A. Laskin, and S. A. Nizkorodov. High-resolution mass spectrometric analysis of secondary organic aerosol produced by ozonation of limonene. *Phys. Chem. Chem. Phys.*, 10:1009, 2008.

[133] C. Wang and M. S. Waring. Secondary organic aerosol formation initiated from reactions between ozone and surface-sorbed squalene. *Atmos. Environ.*, 84:222, 2014.

[134] H. Wang, C. He, L. Morawska, P. McGarry, and G. Johnson. Ozone-initiated particle formation, particle aging, and precursors in a laser printer. *Environ. Sci. Technol.*, 46: 704, 2012.

[135] Jennifer N. Wei, David Duvenaud, and Alán Aspuru-Guzik. Neural networks for the prediction of organic chemistry reactions. *ACS Central Science*, 2(10):725–732, 2016. doi: 10.1021/acscentsci.6b00219. URL http://dx.doi.org/10.1021/acscentsci.6b00219. PMID: 27800555.

[136] D. Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31, 1988.

[137] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36, 1988.

[138] David Weininger. Smiles. 3. depict. graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.*, 30(3):237–243, July 1990. ISSN 0095-2338. doi: 10.1021/ci00067a005. URL http://dx.doi.org/10.1021/ci00067a005.

[139] David Weininger, Arthur Weininger, and Joseph L. Weininger. Smiles. 2. algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci.*, 29(2):97–101, 1989. doi: http://pubs.acs.org/cgi-bin/archive.cgi/jcisd8/1989/29/i02/f-pdf/f_ci00062a008.pdf.

[140] J. R. Wells, G. C. Morrison, B. K. Coleman, C. Spicer, and S. W. Dean. Kinetics and reaction products of ozone and surface-bound squalene. *J. ASTM Int.*, 5:1, 2008.

[141] C. J. Weschler. Roles of the human occupant in indoor chemistry. *Indoor Air*, 2015.

[142] C. J. Weschler, A. Wisthaler, S. Cowlin, G. Tams, P. Strm-Tejsen, A. T. Hodgson, and W. W. Nazaroff. Ozone-initiated chemistry in an occupied simulated aircraft cabin. *Environ. Sci. Technol.*, 41:6177, 2007.

[143] C. J. Weschler, S. Langer, A. Fischer, G. Bek, J. Toftum, and G. Clausen. Squalene and cholesterol in dust from danish homes and daycare centers. *Environ. Sci. Technol.*, 45:3872, 2011.

[144] A. Wisthaler and C. J. Weschler. Reactions of ozone with human skin lipids: Sources of carbonyls, dicarbonyls, and hydroxycarbonyls in indoor air. *Proc. Natl. Acad. Sci. U. S. A.*, 107:6568, 2010.

[145] A. Wisthaler, G. Tamas, D. P. Wyon, P. Strom-Tejsen, D. Space, J. Beauchamp, A. Hansel, T. D. Maerk, and C. J. Weschler. Products of ozone-initiated chemistry in a simulated aircraft environment. *Environ. Sci. Technol.*, 39:4823, 2005.

[146] G. Witz. Biological interactions of alpha,beta-unsaturated aldehydes. *Free Radical Biol. Med.*, 7:333, 1989.

[147] E. L. Wynder and D. Hoffmann. Tobacco and health. *N. Engl. J. Med.*, 300:894, 1979.

[148] F. Yasmeen, N. Sauret, J.-F. Gal, P.-C. Maria, L. Massi, W. Maenhaut, and M. Claeys. Characterization of oligomers from methylglyoxal under dark conditions: a pathway to produce secondary organic aerosol through cloud processing during nighttime. *Atmos. Chem. Phys.*, 10:3803, 2010.

[149] J. Zahardis and G. A. Petrucci. The oleic acid-ozone heterogeneous reaction system: products, kinetics, secondary chemistry, and atmospheric implications of a model system - a review. *Atmos. Chem. Phys.*, 7:1237, 2007.

[150] J. Zhao, N. P. Levitt, R. Zhang, and J. Chen. Heterogeneous reactions of methylglyoxal in acidic media: implications for secondary organic aerosol formation. *Environ. Sci. Technol.*, 40:7682, 2006.