Predictive Modeling of Polycyclic Aromatic Hydrocarbon Formation During Pyrolysis

by

Mengjie Liu

B.S. Chemical and Biomolecular Engineering,
Georgia Institute of Technology, 2014

M.S. Chemical Engineering Practice,
Massachusetts Institute of Technology, 2017

Submitted to the Department of Chemical Engineering
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2020

Author . . . .
Department of Chemical Engineering
December 19, 2019

Certified by . . . .
William H. Green
Hoyt C. Hottel Professor of Chemical Engineering
Thesis Supervisor

Accepted by . . . .
Patrick S. Doyle
Robert T. Haslam (1911) Professor of Chemical Engineering
Chairman, Committee for Graduate Students

Predictive Modeling of Polycyclic Aromatic Hydrocarbon Formation During Pyrolysis

by

Mengjie Liu

Submitted to the Department of Chemical Engineering
on December 19, 2019, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY IN CHEMICAL ENGINEERING

Abstract

Polycyclic aromatic hydrocarbons (PAHs), large molecules comprised of multiple aromatic rings like anthracene or pyrene, are a notable intermediate and byproduct in combustion or pyrolysis of hydrocarbon fuels. On their own, they have been shown to pose a significant health risk, with certain PAHs being linked to increased cancer risk in humans. In addition, PAHs are known to play an important role as building blocks towards larger particles, known as soot or black carbon, which contribute a significant fraction of atmospheric $PM_{2.5}$ pollution (particulate matter with diameters under $2.5\,\mu m$). These particulates pose additional health risks and can also contribute to global climate change via radiative forcing. This motivates interest in understanding the chemical pathways leading to the formation of these PAHs, which could inform better models to predict PAH emissions and optimize methods to reduce their formation.

This thesis presents methods to improve the capabilities of automatic mechanism generation software in modeling the complex chemistry involved in PAH formation. In particular, it focuses on the Reaction Mechanism Generator (RMG) software, an open-source package developed primarily in Python. RMG automatically identifies species and reactions which are relevant at conditions of interest to aid construction of detailed mechanisms, but it has not been previously applied for PAH chemistry. To do so, new algorithms were developed to improve treatment of aromaticity and chemical resonance to better reflect the true behavior of molecules within the limitations of programmatic representations. The effect of polycyclic ring strain on parameter estimation was also investigated, highlighting challenges in capturing 3D conformational effects using the existing estimation frameworks and methods to address them. These improvements to fundamental algorithms play an important role in how thermochemical and kinetic parameters are estimated. The combined utility of these developments is demonstrated by the generation of a detailed mechanism for modeling PAH formation up to pyrene in acetylene pyrolysis, which represents an important milestone in RMG capabilities. Analysis of the model provides insight into the relative contributions of various PAH formation pathways, revealing that hydrogen abstraction, acetylene addition pathways are the key contributors to PAH formation in this system.

Thesis Supervisor: William H. Green
Title: Hoyt C. Hottel Professor of Chemical Engineering

Acknowledgments

My Ph.D. would not have been possible without the help and support of numerous people throughout my time here at MIT.

First and foremost, I would like to express my gratitude to my thesis advisor, Professor William Green, for all of the support and feedback that he has given me over the course of my Ph.D. He trusted me with freedom to explore projects I was interested in, letting me switch to working on RMG full time from my original plan to also do experiments. Looking back, joining this group was definitely the right choice.

I would like to thank my thesis committee members, Professors Richard West, Heather Kulik, and Yuriy Román for providing me with valuable suggestions during our annual meetings, which have helped shape my thesis into its final form. I especially appreciate the various bits of coding advice that Richard has given me, which have definitely helped me become a better developer.

I am also very grateful to the numerous Green Group members who I have had the pleasure of working with during my time here. I would first like to thank the older RMG developers, Drs. Connie Gao, Nathan Yee, Kehang Han, and Nick Vandewiele, for their mentorship and advice. Connie taught me about RMG and RMG-website and gave me ideas which led to much of the work in Chapter 2. Nathan taught me about RMG-database and gave me my first introduction to git and bash. Kehang and Nick both helped me learn many aspects of software development and server administration, and I worked with Kehang to write the initial version of the RMG developer guidelines.

I also appreciate working with Dr. Mark Goldman, Dr. Alon Grinberg Dana, Ryan Gillis, Matt Johnson, and Mark Payne on various aspects of RMG. I enjoyed learning about RMG together with Mark G. when we joined the group, and our numerous discussions about reaction degeneracy and symmetry numbers. I want to thank Alon for all of his work on developing RMG, Arkane, and ARC, and also for helping to create a great software development culture in our group. Ryan has contributed significantly to sulfur chemistry in RMG as well as the amount of joy in our corner of the office. Matt's expertise in numerical methods has been very helpful in many discussions, as well as his work towards improving the RMG-database. Mark P. has also provided invaluable help in various aspects of the development process, and also as a fellow sysadmin on the RMG server. I would like to especially thank him for his help when we had to move the entire RMG server from its old server room to our lab with a week's notice.

I had great experiences working with many people on the various projects that I have been a part of. For the crude oil upgrading project, I would like to thank Drs. Caleb Class, Lawrence Lai, and Soumya Gudiyella. Though we did not actually overlap, I bothered Caleb extensively while working on the di-tert-butyl sulfide project. Lawrence has been a great cubicle-mate and friend. The office was always fun with him around, and I also want to thank him for starting the group tradition of going to Whitehead for lunch on Fridays. I also enjoyed working with Soumya, who always provided useful insights from her extensive experience and really helped the success of our project.

On aromatics formation, I am extremely grateful to Jim Chu, Dr. Zach Buras, and Dr. Mica Smith. While Jim, Zach, and Mica worked primarily in the laser lab, their work on modeling aromatics formation was closely intertwined with my own. While I focused on

RMG development, they were very prolific in gathering and generating data to improve their models. Without their help, I'm certain that the aromatics capabilities of RMG would not have progressed as far as it has.

On the jet fuel project, I would like to thank Drew Weibel and Dr. Ray Speth from AeroAstro, and Dr. Agnes Jocher. Drew and Ray taught me a lot about jet fuel and engines, and it was very interesting to interesting to see the broader impacts of fuel chemistry. Agnes arrived a critical time and provided useful knowledge on bridging the gap between detailed chemistry and soot formation modeling.

I would like to thank all of the quantum chemistry experts who have helped me over the years: Colin Grambow, Dr. Enoch Dames, Dr. Adeel Jamal, Dr. Phalgun Lohur, Dr. Yi-Pei Li, and Dr. Duminda Ranasinghe. Without them, I would have wasted so much time figuring out how to run quantum calculations and troubleshooting the cryptic error messages. I also want to thank Colin for teaching me many bash tricks, though I could never hope to reach his level of proficiency.

Of course, my time in the Green Group could not have gone as smoothly as it did without the support of our admin, Gwen Wilcox. From scheduling meetings, to planning travel, to organizing RMG workshop, Gwen is always there to provide advice from her extensive experience.

I am also very thankful for all of the friends I've made in the last few years. From our first-year study group, Stephanie Doong, Christy Chao, Mike Lee, Shuting Feng, Miao Wang, and German Parada have been wonderful friends. I really appreciate our birthday celebrations, secret Santa, and hiking trips. Also, I'd like to thank Jim Chu, Joey Gu, Albert Liu, Yinying Ren, Yiming Mo, Steven Liu, William Dong, and Kai-Jher Tan for our occasional get-togethers and making sure I don't forget how to speak Chinese. I also appreciate everyone from our practice school group: Stephanie, Jim, Yiming, Sarah Shapiro, Lisa Volpatti, Yinying Ren, Jonathan Chou, Yamini Krishnan, and Shankul Vartak. I'm really glad that we all got along so well, and I enjoyed our occasional reunions. I would also like to thank Abigail Halim, Ross Regitsky, and Eric Lee for their friendship and support.

Last but not least, I would like to thank my parents, Jianyu and Shaohong, for supporting me all my life and encouraging me to pursue my dreams. I want to thank my sister, Jessie, for our interesting conversations and keeping me up to date on the latest memes. I would also like to thank my grandparents for their long-distance support. Finally, I would like to thank my aunt, Dr. Ning Yang, for inspiring me to pursue chemical engineering and a Ph.D.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

There has been growing interest in recent years towards investigating formation of polycyclic aromatic hydrocarbons (PAHs) in hydrocarbon systems. This is in part due to the increasing immediacy of concerns like air pollution and global climate change. PAHs are known to have harmful effects on human health,[1] which was a primary motivator in the U.S. EPA's definition of 16 priority PAHs in 1976.[2] PAHs are also known to be an important building block towards the formation of black carbon, or more generally soot, which contributes to climate change via radiative forcing and is also harmful to human health.[3, 4] Total U.S. emissions of $PM_{2.5}$ (particulate matter with diameters $\leq 2.5\,\mathrm{um}$) in 2014 are estimated at over 6 million tons, of which over 0.44 million tons was black carbon (Figure 1-1).[5]

Black carbon emissions can be largely attributed to combustion processes, whether in the form of fires, engines, power generation, waste incineration, or residential heating. An aspect which is not fully understood is the mechanism by which the initial PAHs are formed, which ultimately lead to the formation of black carbon. Thus, one major focus of this work is investigating the chemical mechanisms of PAH formation in systems such as pyrolysis or combustion. A major challenge with this objective is the inherent size and complexity of these chemical systems in terms of the number of chemical species and reactions involved. This motivates the application of automatic mechanism generation software, which can streamline the process of building these complex chemical mechanisms.

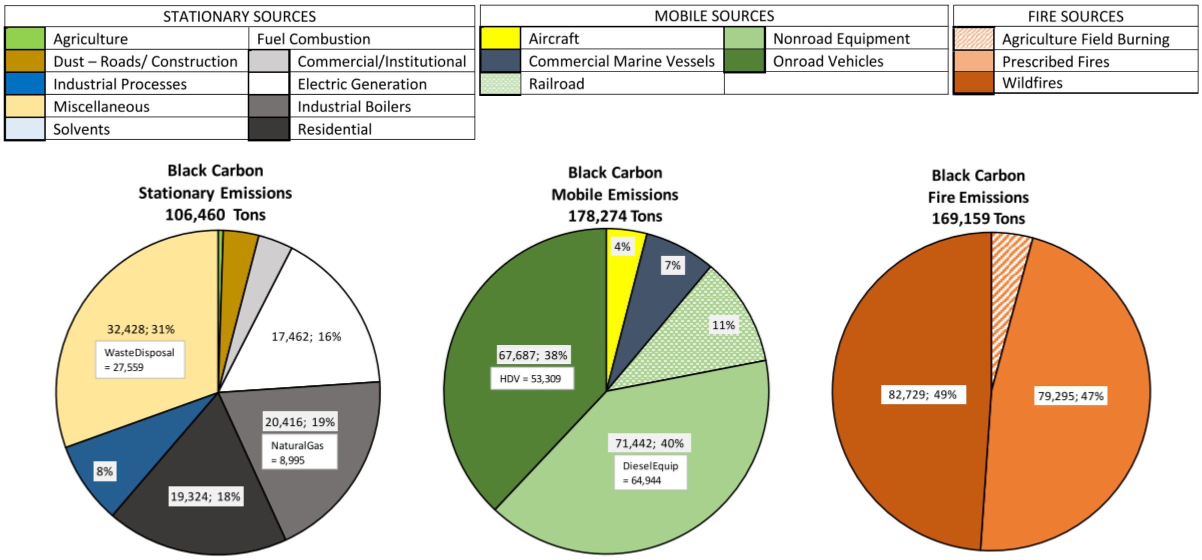


Figure 1-1: U.S. black carbon emissions in 2014, broken down by source. Adapted from 2014 National Emissions Inventory Version 1.0 profile.[5]

1.1 Automatic mechanism generation

Numerous automatic chemical mechanism generation software have been developed over the years to aid in the construction of detailed chemical mechanisms, including Genesys,[6] MAMOX+,[7] EXGAS,[8] and REACTION.[9] The software developed in our group is the Reaction Mechanism Generator (RMG),[10] which is the largest and most actively-developed open-source program in this category.

RMG employs an iterative model generation scheme built around a core/edge reaction model. The core contains the current model, which includes species and reactions which have been identified to be important at the user-defined conditions. The edge contains potential species and reactions which are obtained by reacting core species. Each iteration is comprised of a core expansion step where the mechanism is simulated and the species selection criteria is used to move an edge species into the core, and an edge expansion step where reactions are generated between core species to create new edge species. When no more edge species meet the species selection criteria, model generation is complete. This general cycle is shown in Figure 1-2.

Automatic mechanism generation tools have a clear advantage over manual mechanism creation in their ability to screen a more comprehensive set of potential reactions and

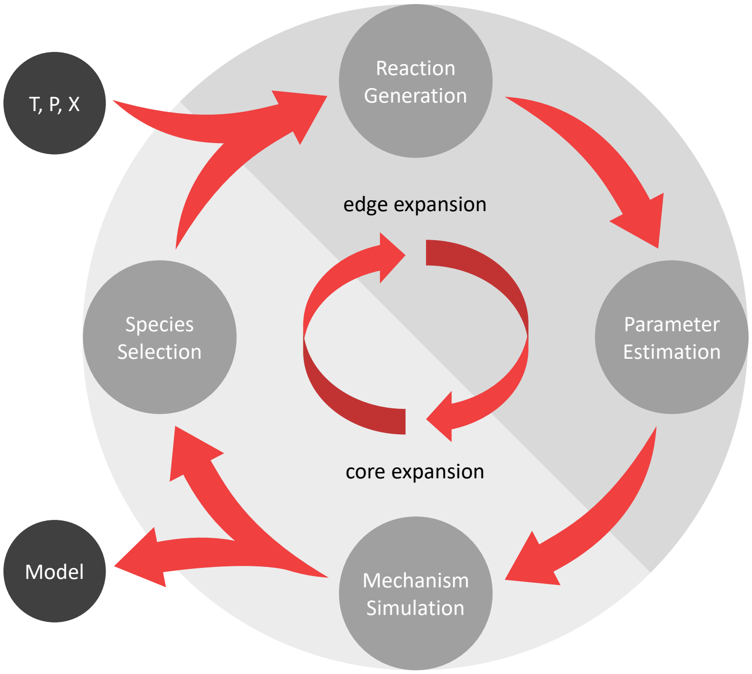


Figure 1-2: Scheme showing the iterative nature of mechanism generation in RMG. User specifies temperature, pressure, and composition, which are used to initiate the process.

automatically identify the important pathways. However, their accuracy is dependent on two key aspects which must work together: the fundamental algorithms for representing chemical concepts, and the data available for characterizing specific types of chemistry.

The software must be able to accurately represent molecules, determine potential reaction sites, and generate possible reactions. RMG uses resonance structures to capture the important chemical features of molecules and account for the effects of electron delocalization. Reaction generation depends on a set of pre-determined reaction families with associated reaction templates.

In addition, mechanism generation in generally is heavily dependent on the quality of the model parameters, primarily species thermochemistry and reaction rate constants. While RMG has the capability of estimating these parameters, the accuracy of these estimates depends strongly on the training data which is provided to the estimation algorithms.

1.2 Polycyclic aromatic hydrocarbons

Interest in uncovering the formation mechanisms of PAHs goes back many decades. A notable early work was done by Bittner and Howard, where they investigated species profiles in a benzene flame using molecular beam mass spectrometry.[11] They proposed a few reaction paths to explain the formation of species including naphthalene and pyrene, including acetylene and vinylacetylene addition. Acetylene addition in particular has received substantial attention as possibly the most important route for PAH growth, beginning with the work of Frenklach *et al.* in developing the hydrogen abstraction, C_2H_2 addition (HACA) pathway.[12] HACA pathways from benzene to larger PAHs have been thoroughly explored since.[13–15]

Aside from the HACA mechanism, many other notable pathways have been explored. Propargyl recombination has been shown to be an important pathway to formation of benzene, especially when considering chemically activated pathways.[16, 17] Cyclopentadienyl recombination has been shown to provide a route to naphthalene without needing to first form benzene.[18–20] Phenyl addition and cyclization has been proposed as a much more efficient pathway to quickly grow PAHs.[21, 22] Benzyne addition and fragmentation has also been explored as a efficient method of adding another aromatic ring.[23, 24]

Even though many pathways have been explored until now, either experimentally or using quantum chemistry methods, there is still too much data missing to build a complete picture of PAH formation. Many detailed chemical mechanisms have been published which include PAH formation pathways (e.g. [25–27]) but they commonly include "global reactions" which may include many elementary steps in reality, with estimated or fitted rate constants. The intrinsic complexity of PAH chemistry and sparsity of available data pose significant challenges to building an accurate mechanism. In this area, automatic mechanism generation has the potential to improve understanding by synthesizing and extrapolating from available knowledge.

1.3 Thesis overview

The overall goal of my work has been to improve the ability of RMG to accurately model PAH chemistry, in order to study how they form during hydrocarbon pyrolysis. Towards this end, I have made several contributions to improving the methods in RMG for handling aromatic and polycyclic species. Additionally, I have worked on expanding the RMG database with thermochemical and kinetic data relevant to PAH formation.

Chapter 2 describes major improvement in RMG for treatment of aromaticity and aromatic molecules. These improvements include new algorithms for generating resonance structures for aromatics, including the introduction of Clar structures as a better approach for representing aromaticity in PAHs. A new kekulization algorithm was also developed to enable reaction generation for aromatic bonds, which has significant implications on kinetics estimation for aromatic species.

Chapter 3 dives into more detail about chemical resonance in general, describing improvements which were made to the overall resonance structure generation algorithm in RMG. In particular, new resonance transformations were implemented for lone pairs, which are important when considering heteroatoms such as nitrogen and sulfur. Filtration of resonance structures was also implemented to identify only the most representative structures, therefore minimizing the total number of structure needed.

Chapter 4 discusses improvements to how polycyclic species are treated in RMG and how thermochemistry and kinetics estimation is affected by ring strain. Ring perception in graph theory is discussed briefly, motivating the implementation of new algorithms for determining the smallest set of smallest rings and the set of relevant cyles. For thermochemistry estimation, new ring strain corrections for strained polycyclic molecules have been calculated and added to the RMG group additivity database. For kinetics estimation, a ring membership attribute was implemented to encode 3D structure information in atom attributes to improve kinetics estimates in intramolecular addition families.

Chapter 5 highlights major developments which are available in the most recent RMG v3.0 release. One important change is the transition to Python 3, which was a necessary step for future-proofing RMG given the official end-of-life for Python 2 on January 1, 2020. RMG

3 also brings many new features, including surface mechanism generation, isotopic mechanism generation, uncertainty analysis, and a neural network thermochemistry estimator. There have also been many improvements to molecular representation, including new atom types and the resonance improvements discussed in Chapters 2 and 3. New kinetics families also enable RMG to generate models for a wider variety of chemical systems.

Chapter 6 brings together the developments discussed in the previous chapters in the automatic generation of a mechanism for PAH formation in acetylene pyrolysis. New thermochemistry and kinetics data for key PAH formation pathways were calculated and added to the RMG database. Using the new algorithms and data, a model was generated to capture pathways from acetylene up to pyrene. The model is validated against experimental data across a range of temperatures and compared to other acetylene pyrolysis models. The performance of this model demonstrates the success of the prior developments in advancing aromatic chemistry in RMG.

Finally, Chapter 7 discusses some ideas for future work in improving RMG's capabilities and further understanding of PAH formation chemistry. These ideas include improvements to 3D geometry considerations in RMG, promising quantum chemistry methods for studying PAH chemistry, and general ideas for improving RMG as a reliable software package.

References

(1) Kim, K.-H.; Jahan, S. A.; Kabir, E.; Brown, R. J. *Environ. Int.* **2013**, *60*, 71–80.

(2) Keith, L. H. *Polycycl. Aromat. Compd.* **2015**, *35*, 147–160.

(3) Jacobson, M. Z. *Nature* **2001**, *409*, 695–697.

(4) Dockery, D. W.; Pope III, C. A.; Xu, X.; Spengler, J. D.; Ware, J. H.; Fay, M. E.; Ferris, B. G.; Speizer, F. E. *N. Engl. J. Med.* **1993**, *329*, 1753–1759.

(5) U.S. Environmental Protection Agency *2014 National Emissions Inventory*; tech. rep.; 2017.

(6) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B. *Chem. Eng. J.* **2012**, *207-208*, 526–538.

(7) Ranzi, E.; Faravelli, T.; Gaffuri, P.; Garavaglia, E.; Goldaniga, A. *Ind. Eng. Chem. Res.* **1997**, *36*, 3336–3344.

(8) Warth, V.; Stef, N.; Glaude, P. A.; Battin-Leclerc, F.; Scacchi, G.; Côme, G. M. *Combust. Flame* **1998**, *114*, 81–102.

(9) Moréac, G.; Blurock, E. S.; Mauss, F. *Combust. Sci. Technol.* **2006**, *178*, 2025–2038.

(10) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(11) Bittner, J. D.; Howard, J. B. *Symp. Combust.* **1981**, *18*, 1105–1116.

(12) Frenklach, M.; Clary, D. W.; Gardiner, W. C.; Stein, S. E. *Symp. Combust.* **1985**, *20*, 887–901.

(13) Kislov, V. V.; Sadovnikov, A. I.; Mebel, A. M. *J. Phys. Chem. A* **2013**, *117*, 4794–816.

(14) Mebel, A. M.; Georgievskii, Y.; Jasper, A. W.; Klippenstein, S. J. *Proc. Combust. Inst.* **2017**, *36*, 919–926.

(15) Liu, P.; Li, Z.; Bennett, A.; Lin, H.; Sarathy, S. M.; Roberts, W. L. *Combust. Flame* **2019**, *199*, 54–68.

(16) Miller, J. A.; Klippenstein, S. J. *J. Phys. Chem. A* **2001**, *105*, 7254–7266.

(17) Miller, J. A.; Klippenstein, S. J. *J. Phys. Chem. A* **2003**, *107*, 7783–7799.

(18) Melius, C. F.; Colvin, M. E.; Marinov, N. M.; Pitz, W. J.; Senkan, S. M. *Symp. Combust.* **1996**, *26*, 685–692.

(19) Mebel, A. M.; Kislov, V. V. *J. Phys. Chem. A* **2009**, *113*, 9825–9833.

(20) Long, A. E.; Merchant, S. S.; Vandeputte, A. G.; Carstensen, H.-H.; Vervust, A. J.; Marin, G. B.; Van Geem, K. M.; Green, W. H. *Combust. Flame* **2018**, *187*, 247–256.

(21) Shukla, B.; Susa, A.; Miyoshi, A.; Koshi, M. *J. Phys. Chem.* **2008**, *112*, 2362–2369.

(22) Shukla, B.; Tsuchiya, K.; Koshi, M. *J. Phys. Chem. A* **2011**, *115*, 5284–5293.

(23) Comandini, A.; Brezinsky, K. *J. Phys. Chem. A* **2011**, *115*, 5547–5559.

(24) Comandini, A.; Abid, S.; Chaumeix, N. *J. Phys. Chem. A* **2017**, *121*, 5921–5931.

(25) Chernov, V.; Thomson, M. J.; Dworkin, S. B.; Slavinskaya, N. A.; Riedel, U. *Combust. Flame* **2014**, *161*, 592–601.

(26) Narayanaswamy, K.; Blanquart, G.; Pitsch, H. *Combust. Flame* **2010**, *157*, 1879–1898.

(27) Tao, H.; Wang, H.-Y.; Ren, W.; Lin, K. C. *Fuel* **2019**, *255*, 115796.

Chapter 2

Capturing aromaticity in automatic mechanism generation software

2.1 Introduction

Aromatic species are highly relevant in many processes related to petroleum fuels. In both combustion and refining, polycyclic aromatic hydrocarbons (PAHs) are the primary stepping stone towards carbon-rich aggregates such as soot or coke. In working towards understanding the chemical mechanisms involved in these processes, automatic mechanism generation has become an invaluable tool. However, the complexity of modeling chemical aromaticity has limited the extent to which these automatic algorithms can properly simulate PAH chemistry.

In this paper, we discuss improvements made to Reaction Mechanism Generator (RMG) that have improved the accuracy, stability, and efficiency of the algorithm when handling aromatic species. A brief overview of the existing aromaticity infrastructure in RMG is given, followed by details of the improvements made, including a novel application of Clar structure theory to cheminformatics. Finally, a test model for pyrolysis of iodonaphthalene and acetylene is presented to demonstrate the new capability of RMG to model PAHs.

2.2 Current treatment of aromaticity

The overall RMG algorithm has previously been described in detail,[1] so only specifics relating to aromaticity will be discussed here. The native molecular representation in RMG is through graph objects where the vertices are atoms and the edges are bonds. or aromatics, RMG uses a benzene bond type and two aromatic atom types: `Cb` for a standard benzene carbon with two benzene bonds, and `Cbf` for a fused benzene carbon with three benzene bonds, *e.g.*, the shared carbons in naphthalene. RMG currently does not perceive aromaticity for five-membered rings or rings containing heteroatoms, although it is being considered for future development.

In RMG, a chemical species is defined not by a single graph representation, but rather a collection of graphs corresponding to different resonance structures. Resonance structures are generally accepted as a way to represent the true structure of a species while being limited to discrete bond orders and fixed electron locations. As such, this is a vital step that enables proper enumeration of the possible reactions for a particular species as well as proper recognition of different structures as being the same species. For aromatic species, RMG currently uses two types of resonance structures: Kekulé structures and "Aromatic" structures. Kekulé structures are a standard representation of aromatics using single and double bonds (Fig. 2-1a). The other common representation where aromatic rings are denoted by an inscribed circle is referred to as the Aromatic structure in RMG, where all ring bonds are represented using the benzene bond type (Fig. 2-1b).

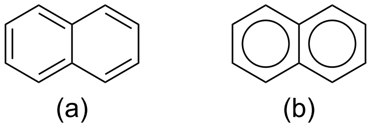


Figure 2-1: (a) One of three possible Kekulé structures for naphthalene. (b) The Aromatic structure for naphthalene, representing an average of the Kekulé structures.

To generate an Aromatic structure from a Kekulé structure, RMG uses RDKit[2] to identify aromatic bonds when RMG then converts to the benzene bond type. To generate Kekulé structures from an Aromatic structure, RMG currently uses RDKit to perform the kekulization, which is only able to generate a single, non-deterministic Kekulé structure.

Therefore, for a typical stable aromatic species, RMG will have two resonance structures: the Aromatic structure and a single Kekulé structure. For radical aromatic species, RMG currently attempts to delocalize the radical based on what it perceives to be allyl shifts. As a result, many incorrect resonance structures may be generated in cases where the radical cannot actually be delocalized, such as in the case of naphthyl, shown in Figure 2-2.

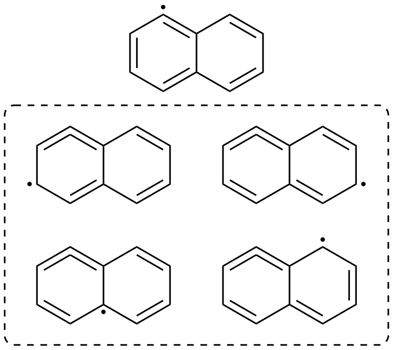


Figure 2-2: Aryl radicals such as the 1-naphthyl radical are essentially localized and are not conjugated with the aromatic π system. However, the current version of RMG will still attempt to delocalize it due to lack of 3D structure understanding.

As previously alluded to, RMG depends on resonance structures in order to generate proper reactions. However, RMG is currently unable to react a benzene bond. Instead, it uses the Kekulé structure to generate reactions for aromatic species when the aromatic bonds are affected. The most common reaction types where this is applicable for aromatics are inter- and intra-molecular radical addition to an aromatic ring. Because RMG uses the Kekulé structure to generate the reaction, the rate estimation is also done using that structure, which often results in the application of a rate rule originally intended for an alkene. This leads to frequent overestimation of the rates for these kinds of reactions.

For thermochemistry, RMG relies primarily on group additivity which works well with the existing Aromatic structures. Since base group additivity does not account for polycyclic ring strain, RMG applies additional corrections using a bicyclic decomposition method where bicyclic corrections are combined using a heuristic to get corrections for larger polycyclics.[3] Since group contributions for aromatic carbons were fitted using molecular representations where all aromatic atoms are identified as such, using the Aromatic structure will give the best prediction, which is currently done. Therefore, the improvements discussed hereafter will focus on kinetics.

2.3 Algorithm improvements

2.3.1 Resonance structures

The search for a more representative and concise alternative to using Kekulé structures for PAHs led to the idea of using Clar structures. To more accurately represent the aromatic character of multi-ring aromatics, Clar formulated a method that has become known as the π-sextet rule.[4] Essentially, this rule states that the structure with the largest number of disjoint aromatic π-sextets is the most important in characterizing the properties of the compound. Figure 2-3 shows two possible π-sextets assignments for phenanthrene: a single sextet in the center ring, or two sextets in the outer rings. According to the the heuristic, the structure with two sextets is the Clar structure, which is most representative of the true aromatic character. This structure indicates that the two outer rings are more aromatic than the center ring, which has been observed computationally.[5]

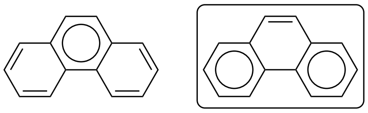


Figure 2-3: Two possible π-sextet assignments for phenanthrene. The boxed structure with two π-sextets is the Clar structure.

The Clar structure for a species can also give a good indication of its reactivity. A notable example is Clar's work on the difference in reactivity between two isomers of tribenzoperylene.[6] While five sextets could be assigned in one of the isomers with no double bonds remaining, only four could be assigned to the other isomer. Experimentally, Clar observed the isomer with only four sextets to be much more reactive to addition by maleic anhydride. Compounds which do not have any double bond assignments after assigning sextets are known as fully benzenoid or fully aromatic, and are generally more stable and unreactive than compounds which have additional double bonds.

The three rules prescribed by Clar for drawing these structures are as follows:

1. Sextets cannot be placed in adjacent rings.

2. Any π-electrons remaining after sextet assignment should be assigned as double bonds.

3. The structure with the most π-sextets is most representative of aromatic character and reactivity.

The first two rules are simply to satisfy the valence of each atom, while the last rule is a key point which limits the number of Clar structures and distinguishes them from Kekulé structures.

To automatically generate Clar structures, an integer linear programming (ILP) approach described by Hansen and Zheng was implemented.[7] Their algorithm is an intuitive adaptation of the rules for drawing Clar structures, formulated as follows:

$$\begin{aligned} &\text{maximize} && z = \sum_{r \in M} y_r \\ &\text{s.t.} && \sum_{b \in N(a)} x_b + \sum_{r \in R(a)} y_r = 1 \ \forall \ a \in A(M) \\ &&& x_b, y_r \in \{0, 1\} \ \forall \ b \in B(M), r \in M \end{aligned} \tag{2.1}$$

where M is the full molecule, $A(M)$ is the set of atoms in the molecule, $B(M)$ is the set of bonds, $N(a)$ is the set of bonds involving atom a, $R(a)$ is the set of rings involving atom a, b is any bond, and r is any ring. The variable y_r is 1 if the ring contains a π-sextet and 0 otherwise, and the variable x_b is 1 if the bond is a double bond and 0 otherwise. Essentially, the algorithm maximizes the number of π-sextet assignments subject to the constraint of satisfying atom valencies.

The algorithm was implemented in RMG by extracting connectivity information from the native Molecule object into the appropriate vectors and matrices, then solving the ILP using the open source mixed-integer linear programming solver, lpsolve.[8] Additionally, a recursive strategy was used to enumerate all Clar structures for molecules which have more than one.

After the addition of this new resonance structure type, the resonance algorithm in RMG was rewritten to give special treatment to aromatic species, summarized in Table 2.1. While Kekulé structure are still being generated for monocyclic aromatics for back-compatibility, they have been replaced by Clar structures for polycyclic aromatics. The Aromatic structure is still kept for accurate thermo estimation using Benson group additivity. Additionally, aryl radicals are no longer erroneously delocalized into the aromatic system.

Table 2.1: Resonance structures generated for various molecule categories.

All Species	**Aromatic Species Only**
Radical species (except aryl radicals)	Monocyclic aromatic species
electron delocalized structures	Aromatic structure
Lone pair containing species	Kekulé structures
lone pair/radical resonance	Polycyclic aromatic species
Nitrogen containing species	Aromatic structure
single/triple bond to double/double bond resonance	Clar structures

This process also uncovered issues in RDKit's aromaticity perception algorithm, which uses an atom-centered π-electron counting method based on Hückel's Rule. However, the algorithm fails in two important cases: rings with exocyclic double bonds and rings with delocalized radicals. Rings which have two double bonds which are connected to, but are not part of the ring are considered aromatic by π-electron counting despite the fact that they do not have the cyclic electron delocalization characteristic of aromaticity. In the other case, if a radical is delocalized into an aromatic ring, the π-electron counting method in RDKit fails to identify the radical as being in a π orbital and participating in aromatic stabilization.

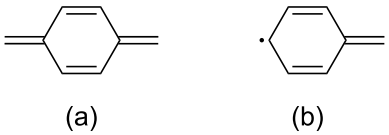


Figure 2-4: (a) Molecule which RDKit incorrectly identifies as aromatic. (b) Molecule which RDKit incorrectly identifies as non-aromatic.

In the first case, as in Figure 2-4a, RMG can identify the false positive because the bonds in the ring cannot be converted to aromatic bonds because the carbons with exocyclic double bonds would be hypervalent. In the second case, Figure 2-4b being an example, there was not a simple fix, but the aromatic resonance structure method was rewritten to maximize the number of aromatic rings identified by shifting delocalized radicals around the molecule. That enabled proper identification of radical aromatics and resulted in much more deterministic output from that method, which was important in making the overall resonance algorithm more robust.

To summarize, the implementation of Clar structures as a replacement for Kekulé structures allows

- capturing differences in aromaticity
- reducing the total number of structures

and the subsequent changes to the resonance algorithm enabled

- preventing incorrect delocalization of aryl radicals
- catching errors in aromaticity detection by RDKit

2.3.2 Reaction generation

As mentioned before, benzene bonds were previously non-reactive in RMG and were only relevant for thermochemistry generation and for describing the environment around a reacting bond. To provide more flexibility and precision in defining rate rules, changes were implemented to enable benzene bonds to react.

The difficulty with benzene bond reactivity is that changing an aromatic bond requires changes to other bonds in the ring as well, making it difficult to do with the reaction recipe approach used in RMG, which can only implement fixed, local changes.

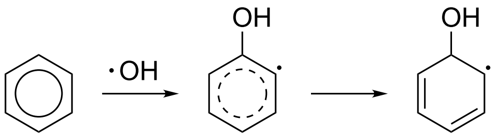


Figure 2-5: Addition of hydrogen radical to benzene with aromatic bonds. A second step is necessary to convert the aromatic bonds to single/double bonds.

To solve this problem, a new kekulization method that can reconcile the bonds in the intermediate structure is implemented in RMG. The new algorithm builds on concepts from both RDKit and OpenBabel's [9] kekulization algorithms with the addition of degree of freedom (DOF) analysis. The strategy is to analyze each aromatic bond in the molecule and try to determine its proper bond order, either single or double. The determination is made based on atom valency and the number of available electrons for bonding. Simultaneously, DOF analysis is used to count the number of DOFs associated with a ring or bond, which indicates whether or not an assumption can be made about the bond order. As bonds are fixed, the DOFs are reduced, allowing more bonds to be determined. This approach reduces the amount of trial-and-error as compared to path-exploration type methods such as in RDKit. Having this algorithm lets RMG treat aromatic bonds like normal bonds while applying the reaction recipe and then reconcile the bond orders later in the post-processing step (Fig. 2-5).

Enabling benzene bond reactivity highlighted an issue with how RMG calculates reaction degeneracy. Previously, the degeneracy of any reaction in RMG was set equal to the number of subgraph isomorphic matches between top level groups in reaction family trees and the reactant species. Most notably, if a reaction results in products which are resonantly equivalent, such as in Figure 2-6, the degeneracy will be overestimated. This issue is particularly pronounced with aromatics, since they have more resonance structures on average.

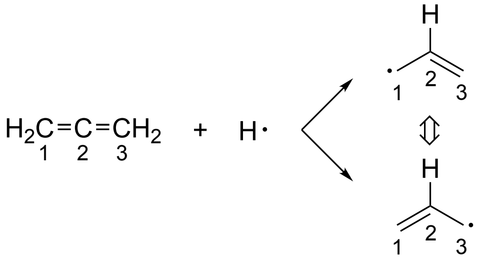


Figure 2-6: Addition of an H atom to the center carbon in allene, which gives two resonantly equivalent products. Current RMG counts these as separate reactions, so it assigns an incorrect degeneracy of 2 when calculating the rate coefficient. Using atom IDs enables RMG to identify both reactions as being identical, giving a degeneracy of 1.

To address this, a new method of distinguishing atoms in molecules by using unique IDs is implemented. With atom IDs, it is now possible to track the movement of atoms as a result of a reaction. By comparing the atom IDs in the product structures, it is possible to determine if they were formed via the same atom movements or not. Molecules which match and have the same IDs do not contribute to degeneracy, while molecules which match but have different IDs do contribute, such as in Figure 2-7. This change prevents programming artifacts from affecting the computed reaction degeneracy.

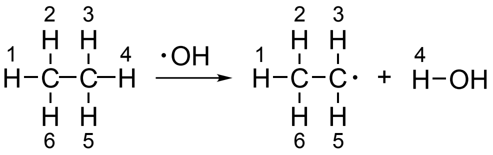


Figure 2-7: Hydrogen abstraction from ethane. Each hydrogen which could be abstracted would result in the same product but with different labeling, and therefore increase the degeneracy. This gives the correct degeneracy of 6, which is unchanged from the current method.

2.4 Model construction and results

To demonstrate that with these modifications, RMG can now be used to model aromatic compounds, an exploratory modeling study was done for co-pyrolysis of iodonaphthalene and acetylene. This system is of interest for investigating PAH formation, and many experimental and computational studies exist as a result. The model presented here was generated directly using RMG with minimal system-specific adjustments, described presently. Shock-tube data by Lifshitz *et al.* is used for comparison.[10]

The RMG input conditions were a temperature range of 800-1200 K, pressure of 1 atm, and initial composition of 20% naphthalene and 80% acetylene, without pressure dependence. The mechanism construction was run for four days, giving a final mechanism with 349 species and 3721 reactions. The RMG input file and final mechanism can be found in the supplemental materials. The large size of this mechanism compared to the one used by Lifshitz *et al.* is due to exploration of other secondary chemistry by RMG, which improves the range of applicability of this mechanism.

In the shock-tube experiment, 1-iodonaphthalene was used as a precursor to generate naphthyl radicals. Because RMG does not yet have the capability of modeling iodine compounds, a few reactions were manually added to model the decomposition routes for iodonaphthalene, shown in Table 2.2. The rate for iodonaphthalene bond scission to form naphthyl and iodine was taken to be the same as the analagous reaction for iodobenzene, measured by Tranter *et al.*[11] Rate estimates made by Lifshitz *et al.* were used for the other two reactions with a hydrogen atom.[10]

Table 2.2: Iodonaphthalene reactions added to RMG mechanism. Rate parameters are in units of kJ, mol, cm, and s.

Reaction	A	n	Ea	Reference
1-$C_{10}H_7I = $ 1-$C_{10}H_7^\bullet + I^\bullet$	1.052e19	-0.98	285.02	Tranter *et al.*[11]
1-$C_{10}H_7I + H^\bullet = $ 1-$C_{10}H_7^\bullet + HI$	6.00e13	0	26	Lifshitz *et al.*[10]
1-$C_{10}H_7I + H^\bullet = C_{10}H_8 + I^\bullet$	3.00e13	0	39	Lifshitz *et al.*[10]

Reactor simulations were done using Cantera 2.3.0 [12] across the experimental temperature range of 900-1200 K, pressure of 2 atm, and initial composition of 5% acetylene and 0.05% iodonaphthalene in argon. An adiabatic, constant volume reactor was used, with a reaction time of 2 microseconds. As seen in Figure 2-8, the iodonaphthalene reactions are able to fairly accurately capture its consumption. Following iodonaphthalene decomposition, the RMG mechanism is also able to accurately predict the formation of the main product, acenaphthalene, shown in Figure 2-9, although the formation of the side product naphthalene was underpredicted by about a factor of five.

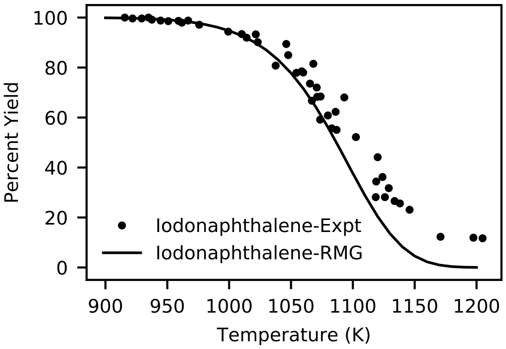


Figure 2-8: Remaining iodonaphthalene content as percent yield relative to initial mole fraction.

In addition to the main products, the RMG mechanism also predicted the formation of side products from acetylene self-reaction, most notably vinylacetylene, benzene, and ethene, shown in Figure 2-10. While none of these products were reported in the experiment, they are all feasible predictions. Experimental work by Rokstad *et al.* on pure acetylene pyrolysis has shown that vinylacetylene and benzene are the major products at low conversion.[13] Ethene formation is a result of the hydrogen atoms generated by acenaphthylene formation.

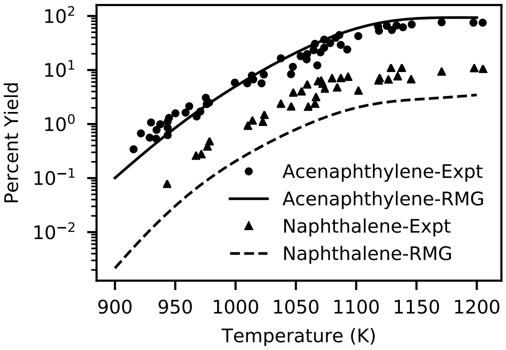


Figure 2-9: Percent molar yields of acenaphthylene and naphthalene relative to initial moles of iodonaphthalene.

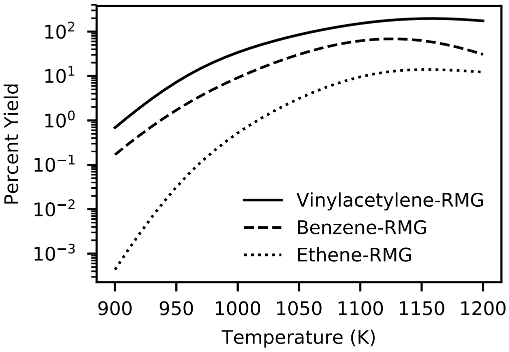


Figure 2-10: Predicted percent molar yields of vinylacetylene, benzene, and ethene relative to initial moles of iodonaphthalene. Although these are side products formed from acetylene, they are presented relative to iodonaphthalene to allow comparison to Figure 2-9.

Interestingly, RMG did not predict any ethynylnaphthalene formation, which was also not observed by Lifshitz *et al.* but has been observed at similar temperatures but lower pressure by Parker *et al.*[14]

It is important to note that no additional model refinement was done. These results demonstrate that RMG can now successfully build a mechanism for an aromatic system as a result of the improvements made to the algorithm itself. Local uncertainty analysis was performed at 1200 K using the uncertainty module in RMG, which does uncertainty propagation based on estimated uncertainties of the data sources. The total variance and most significant contributors are shown in Table 2.3, which reveals much greater uncertainty in the prediction for naphthalene compared to acenaphthalene. The total variance of naphthalene corresponds to an uncertainty of a factor of 2.36 in concentration, while the the total variance of acenaphthalene corresponds to an uncertainty of a factor of 1.15. For both species, kinetics has a greater contribution than thermochemistry, suggesting that future work should be done on refining these reaction rates.

Table 2.3: Reaction rates and species thermo with highest contributions to uncertainty in naphthalene and acenaphthalene concentration at 1200 K and 0.002 s.

	$C_{10}H_8$	$C_{12}H_8$
Total Variance $(d\ln(c))^2$	0.7375	.01925
Kinetics Contributions (%)		
1-$C_{10}H_7$ + C_2H_2 = 1-$C_2H_2C_{10}H_7^\bullet$	31	14
$C_2H_3^\bullet$ + $C_2H_3^\bullet$ = C_2H_4 + C_2H_2	13	0.95
1-$C_{10}H_7$ + $H^\bullet$ = $C_{10}H_8$	12	0.87
1-$C_{10}H_7I$ + $H^\bullet$ = $C_{10}H_8$ + $I^\bullet$	7.8	0.56
1-$C_{10}H_7$ = 2-$C_{10}H_7$	0.047	72
Thermochemistry Contributions (%)		
$C_{10}H_8$	9.5	–
C_2H_2	4.4	0.058
$H^\bullet$	2.5	0.12
1-$C_{10}H_7$	0.10	2.9
2-$C_{10}H_7$	0.044	1.6

2.5 Conclusions

Many changes have been made to RMG to accurately and robustly represent and react aromatic species. Most notable is the introduction of Clar structures to capture aromaticity and reactivity in PAHs in replacement of Kekulé structures. Clar structures have been used in many applications as a simple descriptor that matches well with experimental and computational observations regarding aromaticity. This work represents the first application of Clar structures in automatic mechanism generation.

A new kekulization algorithm was also implemented that enables RMG to seamlessly react benzene bonds and allows for more flexible and intuitive group definitions for rate rules. This leads to more accurate rate predictions by robustly differentiating rates for aromatics and non-aromatic alkenes. To further improve rate predictions, a chemically motivated method for calculating reaction degeneracy was implemented, eliminating overestimation in cases where duplicate reactions are found due to resonance.

These changes provide a more solid foundation for many of RMG's core functions and greatly improves their robustness. As a result, RMG was able to build a model for iodonaphthalene and acetylene co-pyrolysis with decent agreement with experiment. Uncertainty analysis provides guidance for future work to refine the rate coefficients and thermochemistry which control the chemistry in these PAH systems.

References

(1) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(2) RDKit: Open-source cheminformatics., 2018.

(3) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. *Int. J. Chem. Kinet.* **2018**, *50*, 294–303.

(4) Solà, M. *Front. Chem.* **2013**, *1*, 22.

(5) Portella, G.; Poater, J.; Bofill, J. M.; Alemany, P.; Solà, M. *J. Org. Chem.* **2005**, *70*, 2509–2521.

(6) Clar, E.; Zander, M. *J. Chem. Soc.* **1958**, 1861–1864.

(7) Hansen, P.; Zheng, M. *J. Math. Chem.* **1994**, *15*, 93–107.

(8) Berkelaar, M.; Eikland, K.; Notebaert, P. lpsolve: Open source (Mixed-Integer) Linear Programming system., 2016.

(9) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.

(10) Lifshitz, A.; Tamburu, C.; Dubnikova, F. *J. Phys. Chem. A* **2009**, *113*, 10446–10451.

(11) Tranter, R. S.; Klippenstein, S. J.; Harding, L. B.; Giri, B. R.; Yang, X.; Kiefer, J. H. *J. Phys. Chem. A* **2010**, *114*, 8240–8261.

(12) Goodwin, D. G.; Moffat, H. K.; Speth, R. L. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes., 2017.

(13) Rokstad, O. A.; Lindvaag, O. A.; Holmen, A. *Int. J. Chem. Kinet.* **2014**, *46*, 104–115.

(14) Parker, D. S. N.; Kaiser, R. I.; Bandyopadhyay, B.; Kostko, O.; Troy, T. P.; Ahmed, M. *Angew. Chemie Int. Ed.* **2015**, *54*, 5421–5424.

Chapter 3

Automated chemical resonance generation and structure filtration for kinetic modeling

3.1 Introduction

Predictive kinetic models are often relatively large, and may consist of thousands of inter-reacting chemical species.[1] While handcrafting such models is error-prone,[2] automated model generation, as implemented in various forms (*e.g.*, EXGAS,[3] Genesys,[4] RMG[5]), provides a systematic, reliable approach to this challenge.

The main open-source software project in this area is the Reaction Mechanism Generator (RMG) suite.[5] RMG automatically generates comprehensive kinetic models consisting of elementary as well as pressure-dependent well-skipping reactions for a given reacting mixture at pre-defined physical conditions. The software uses kinetics and thermodynamic libraries containing experimental and calculated data, and estimates the remaining necessary parameters using reaction templates[5] and the group additivity method,[6] respectively. The underlying algorithm was previously thoroughly described elsewhere.[5] Briefly, a flux-based algorithm determines which species to incorporate into the final model using a core–edge framework, while the user-set tolerance controls the final model size and therefore the model truncation error, which arises from the inability of a model to include all possible species and reactions.

Automating the process of model generation requires a convenient and efficient method for representing species along with all potentially reactive sites such as π-bonds, radical electrons, and lone electron pairs, as well as partial charge distribution. These representations must be consistent with the reaction templates so the computer can discover all the viable reactions. These computerized representations should describe the desired molecular properties as closely as possible to the physical reality. Lewis structures[7, 8] are attractive candidates for this end; in addition to describing reactive sites, they are intuitive for the human chemist and are relatively easily implemented in computer software. Nevertheless, they have two significant drawbacks: (1) being two-dimensional models of physical objects, they fail to represent three-dimensional molecular properties such as intra-molecular interactions (*e.g.*, hydrogen bonds) and steric effects; (2) Lewis structures describe localized electronic configurations only, hence if some atoms have partial valences,[9, 10] *i.e.*, as in chemical resonance, the species cannot be represented by a single Lewis structure due to electron delocalization. The present work describes efforts to overcome the second hurdle.

Species with delocalized electrons in organic chemistry are commonly represented using several Lewis structures (also commonly referred to as resonance, localized, canonical, or contributing structures). These structures differ from one another by the arrangement of electrons, while the relative nuclei positions remain fixed. A Lewis diagram can represent a single structure with an integer number of electrons ascribed to each atom. The physical species, referred to as the resonance hybrid, can be described by some combination of the respective contributing structures, though not necessarily with equal weights.

For instance, abstracting a hydrogen atom from the primary carbon in acetaldehyde (3.1) and abstracting the hydroxyl group hydrogen from ethenol (3.2) both form vinoxy radical (Eq. 3.1), an important species in combustion and atmospheric chemistry. However, each reaction ostensibly produces a different localized structure. The model generation software has to recognize this allyl radical shift resonance type (Eq. 3.1) to identify the two generated localized structures as belonging to the same species. The software also has to be able to generate all representative localized structures for any arbitrary species to further apply relevant reaction templates during model enlargement, *e.g.*, so that the computer would know that the reverse reactions of 3.1 and 3.2 both exist.

$$\mathrm{CH_3CHO + H^\bullet \rightleftharpoons {}^\bullet CH_2CHO + H_2} \tag{R3.1}$$

$$\mathrm{CH_2CHOH + H^\bullet \rightleftharpoons CH_2CHO^\bullet + H_2} \tag{R3.2}$$

$$[\mathrm{{}^\bullet CH_2CHO \longleftrightarrow CH_2CHO^\bullet}] \tag{3.1}$$

Chemical resonance is known to have a significant effect on thermochemical stability and kinetic rates.[11] An example of resonance stability is the thermochemical effect of the radical position in buteneyl. While but-2-ene-1-yl (a conjugated system) has a standard heat of formation of 31.40 $\mathrm{kcal\,mol^{-1}}$, but-3-ene-1yl (a non-conjugated similar system) has a significantly higher value of 48.57 $\mathrm{kcal\,mol^{-1}}$;[12] hence, the latter is less stable and more reactive. The kinetic effect is demonstrated here for the vinoxy radical case (Eq. 3.1): The localized structure with the radical on the carbon site has a larger relative contribution to the resonance hybrid than the structure with the radical on the oxygen site (a qualitative explanation for the unequal weights is the electronegativity ratio of oxygen and carbon atoms; a quantitative basis for this determination is given in Section 4). Consequently, the rate of hydrogen abstraction by the oxygen atom radical site in vinoxy is several orders of magnitude lower than the respective non-resonating saturated case, ethoxy radical, in which the radical is localized on the oxygen atom (Fig. 3-1).

In many cases, the number of localized structures describing the resonance hybrid may become relatively large due to the combinatorial effect of the different resonance pathway types. In some cases, tens or even hundreds of structures could be assigned to a single species. Fortunately, often only a few localized structures are significant for describing the species reactivity (*e.g.*, the software principally requires one structure to signify a radical site while others might not always bear new information), while most are non-representative of the resonance hybrid. To feasibly identify representative resonance structures during large-scale automated kinetic model generation and to avoid spending computational resources on

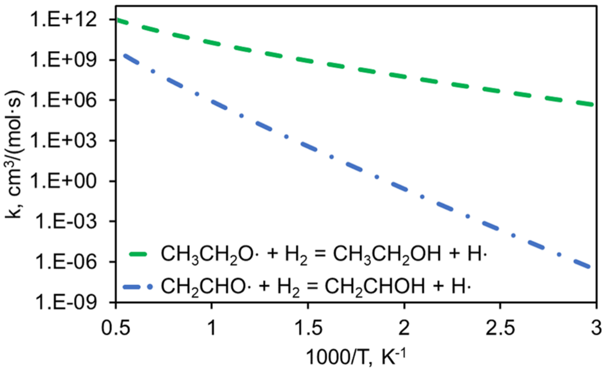


Figure 3-1: Rate coefficients for hydrogen abstraction from molecular hydrogen by ethoxy[13] and vinoxy[14] radicals. The rate coefficient for vinoxy was determined from the reverse reaction using appropriate thermodynamic data.[15, 16]

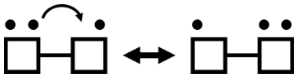
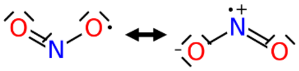
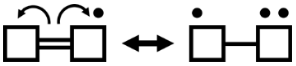
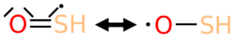
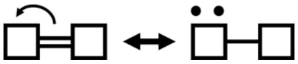
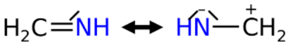
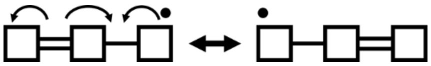
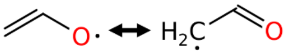
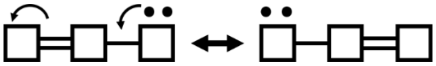
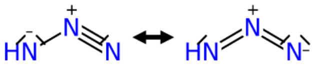
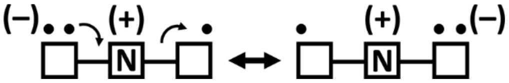
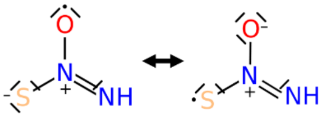
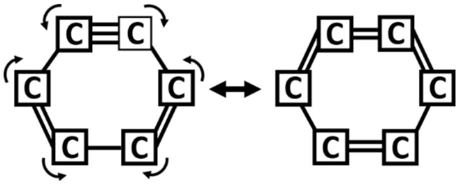
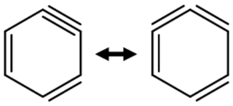
unimportant reactions, an efficient and accurate filtration method must be implemented.

The main objective of the present work is to describe and demonstrate an automated and efficient approach for localized structure reactive site discovery of arbitrary species, starting from a single localized structure. Knowledge of localized reactive sites is critical for automated reaction identification and classification using reaction templates. This work focuses on chemical systems consisting of H/C/N/O/S elements, for which RMG has been relatively well trained.[17–19] The resonance generation and filtration algorithms discussed herein were recently implemented in RMG, and are available starting from version 2.3.0.[20] These features are also freely accessible from the RMG website[21] without any required installation to support researchers and educators.

3.2 Resonance generation

Resonance pathways (Table 3.1) represent series of actions applied to localized electronic structure, yielding additional localized structures of the same species. These actions may include modifications to bond orders, lone electron pairs, and radicals, keeping the total number of electrons fixed. A set of resonance pathways is considered fundamental if no pathway within the set can be reconstructed using any combination of the others. When creating new species, RMG generates localized resonance structures by applying all relevant pathways, starting from a given localized structure. If additional localized structures are found, the resonance generation continues recursively by applying the pathways to all non-isomorphic structures generated during this process (the definition of isomorphism is adopted from graph theory[22]). Currently, the software implements a fundamental set of seven resonance pathway types, shown in Table 3.1, and three global representations specific to aromatic species, as described below.

Table 3.1: Fundamental delocalization types defined in RMG

	Pathway type	Template	Example
1	Adjacent lone pair/ radical shift		
2	Adjacent radical lone pair/multiple bond shift		
3	Adjacent lone pair/ multiple bond shift		
4	Allyl radical shift		
5	Lone pair/multiple bond shift		
6	Lone pair/radical shift mediated by N5		
7	Aryne bond shift		

Electron delocalization in a two-atom moiety is only possible if at least one participating atom is able to possess a lone electron pair (*i.e.*, a carbene or a heteroatom). If the group has a radical electron, there could principally be four resonance pathways, forming a non-fundamental set (Fig. 3-2A). Two of these pathways were implemented in RMG (Table 3.1, pathways 1, 2), which is suffice to describe resonance pathways within such group. The interplay of these two pathways alone may form a variety of localized structures (Fig. 3-2B). It is noted that the final structure list is independent of the starting structure since each resonance pathway is reversible. An additional two-atom group resonance pathway was implemented, describing a non-radical case where two delocalized electrons have characteristics of both a lone pair as well as a π-bond in the resonance hybrid (Table 3.1, pathway 3).

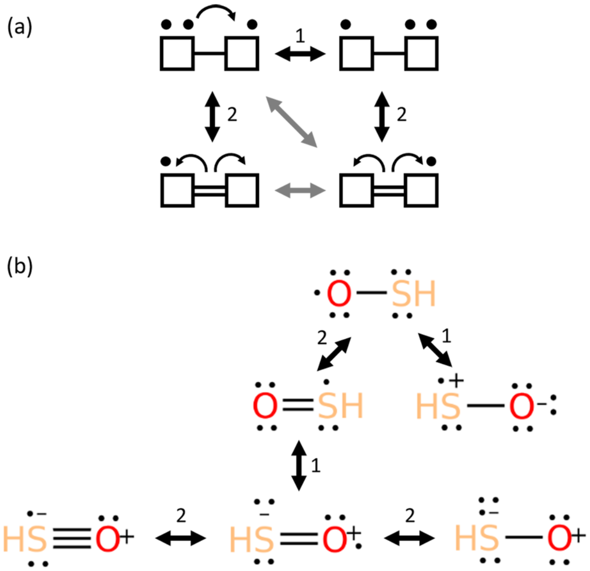


Figure 3-2: (A) Resonance pathway types of a radical two-atom system. Gray arrows represent redundant pathways which were not implemented in RMG. (B) All resonance structures found by RMG for the HSO radical. Numbers on arrows correspond to pathways in Table 3.1.

The ally radical shift resonance pathway (Table 3.1, pathway 4) is relatively ubiquitous and well-described in organic chemistry textbooks.[23] A sibling three-atom pathway involves a similar lone electron pair shift with a multiple bond (Table 3.1, pathway 5), and is important, for example, when describing pathways in some of the localized structures relevant for azide groups (RNNN), nitrous oxide (N_2O), or aniline ($C_6H_5NH_2$).

A unique three-atom pathway mediated by hyper-valence nitrogen is also considered (Table 3.1, pathway 6). This pathway is important for correctly describing localized structures in systems such as NO3 radical if some of the oxygen atoms are tracked isotopes. Each of the three-atom group resonance pathways is concerted, and cannot be described by consecutively applying any combination of several two-atom resonance types. For example, the lone electron pair / multiple bond shift resonance pathway (Table 3.1, pathway 5) resembles the pattern of the adjacent electron lone pair / multiple bond shift resonance pathway (Table 3.1, pathway 3). Nevertheless, implementing the two-atom pathway consecutively will result in an infeasible intermediate localized structure with an unphysical atom valance, *e.g.*, for the middle nitrogen in N_2O (Fig. 3-3). Allowing such infeasible electronic configurations would indeed alleviate the need to include some of the three-atom resonance pathways in the software, but it would also result in many undesired and unphysical localized structures.

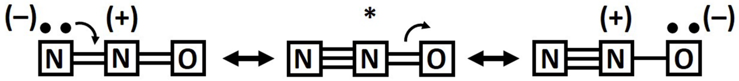


Figure 3-3: Illustrating the lone pair / multiple bond shift type (Table 3.1, pathway 3) applied successively in resonance structure generation of N_2O, passing through an infeasible structure (marked with an asterisk), which is not allowed in RMG.

Several adjacent resonance pathways may form a conjugated system of connected p orbitals, if allowed by stereo-effects. The CH_2CNO system is an example of a conjugated system, where several of the pathways in Table 3.1 apply. The spin density of CH_2CNO is indeed shared across all the atoms in the molecule (Fig. 3-4). A special case of conjugated p orbitals arises in aromatic molecules, leading to unusual thermodynamic stability and substantially different reactivity. As such, it is important that the representations which RMG uses for these molecules indicate their aromaticity in order to model them properly.

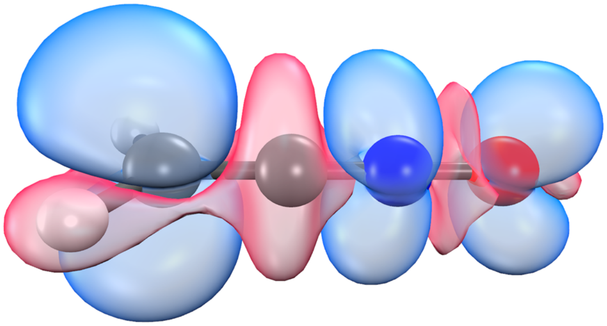


Figure 3-4: Spin density of CH_2CNO calculated using the NBO 6.0 population analysis software[24] implemented in Q-Chem 4.4[25] at uB3LYP/6-311G++(3df,3pf), and visualized using IQmol[26] using a 1% iso-value.

For aromatic species, RMG relies on global resonance pathways instead of the fundamental pathways shown in Table 3.1. While the fundamental pathways search for specific patterns of bonds, radicals, and lone pairs, global pathways use guidelines to completely rearrange bonds in the molecule. One reason for implementing this approach is that pattern searches may become intractable with the variety of bond patterns that can occur in relatively large aromatic species. Another reason is that delocalized structures which can indicate the aromaticity of the molecule are more useful than standard Lewis structures. This is typically demonstrated using benzene, which has two possible localized structures with alternating single and double bond assignments, referred to as Kekulé structures. Commonly, benzene is instead depicted with a single inscribed circle indicating that the π electrons are delocalized and shared by all six atoms in the ring. While this is no longer a Lewis structure, it becomes more useful to RMG by correctly indicating that the bonds in the ring are all equivalent and aromatic. Generation of this resonance form is done by the Aromatic resonance structure method (Fig. 3-5).

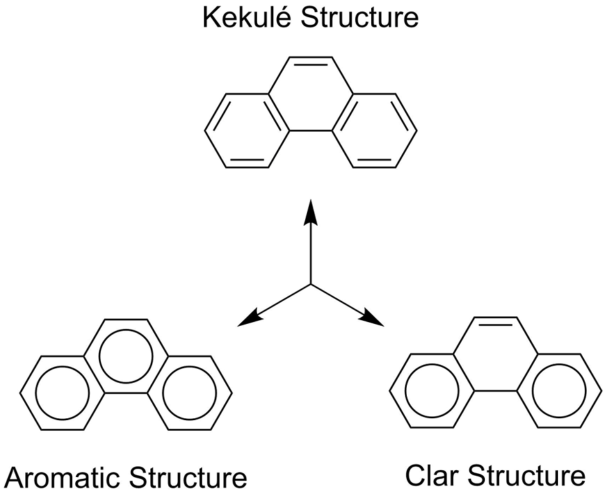


Figure 3-5: Global representations for aromatic species shown for phenanthrene. RMG can readily convert between the three types of structures.

For radical aromatic species, it is also necessary to capture the ability of the radical to delocalize into the aromatic ring. Here, the localized Kekulé structure is necessary in order to evaluate potential delocalization paths using successive allyl radical shifts. Therefore, RMG is able to generate Kekulé structures as well (Fig. 5). The current version of RMG uses an algorithm which generates a single Kekulé structure for each molecule. Of course, polycyclic aromatic hydrocarbons can have numerous different Kekulé structures; enumerating them is very computationally challenging[27] and unnecessary in RMG because of the use of Aromatic structures.

RMG is also capable of generating Clar structures (Fig. 3-5), which are a more detailed approach to depicting aromaticity.[28] Instead of generally indicating delocalization of π electrons, it assigns sets of 6 π electrons to certain rings in the molecule, subject to atom valence constraints, and the remaining π electrons are then assigned to double bonds. The structure with the most π sextets is proposed to be most representative of the molecule's true behavior. This was implemented in RMG as an improved way to indicate aromaticity for polycyclic aromatic hydrocarbons, which can have an effect on reactivity in different parts of the molecule.[29]

Finally, there is a specialized resonance pathway for considering arynes (Table 3.1, pathway 7), of which benzyne is the most well-known representative, but also includes polycyclic aromatic hydrocarbons. Such species have two resonance forms, either with three adjacent double bonds (cumulene form) or a single triple bond (aryne form).

Of course, aromaticity is not limited to benzenoid molecules and also includes heterocyclic aromatics such as furan or pyridine, to name just a few. Currently, RMG does not have special resonance treatment for heterocyclic aromatics, and instead it uses Lewis structures. Part of the challenge in this area is determining whether the aromatic bonds in such molecules can be accurately represented with a single bond type, or whether more detailed aromatic bond descriptors are necessary. While this area is still unexplored, it is of interest for future development.

Accounting for resonance pathways is also important for reaction degeneracy calculation, defined as the number of different elementary routes generating the same products from a particular reactant set via a similar transition state.[30] For example, Reactions 3.3 and 3.4 have degeneracies of one and two, respectively, since NO_2 has only one possible nitrogen radical site, yet two equivalent oxygen radical sites. RMG uses an atom labeling approach proposed by Bishop and Laider to accurately count the number of degenerate reactions which form the same product.[29, 31] Each atom is assigned a unique integer ID for tracking. The main idea is that for a given reactant labeling, isomorphic products with identical atom IDs will not contribute towards reaction degeneracy (Fig. 3-6A, B), while isomorphic products with different atom IDs will (Fig. 3-6A, B). Here, isomorphic refers to graph isomorphism, meaning that a one-to-one correspondence between atoms and bonds exists between the two molecules. If two structures can be made to coincide by a rigid bond rotation, they are isomorphic. If the atom IDs also match, then the two molecules are considered identical, using RMG's terminology.

$$^{\bullet}\mathrm{NO_2} + {}^{\bullet\bullet}\mathrm{NH} \rightleftharpoons {}^{\bullet}\mathrm{NHNO_2} \quad \text{degeneracy} = 1 \tag{R3.3}$$

$$^{\bullet}\mathrm{NO_2} + {}^{\bullet\bullet}\mathrm{NH} \rightleftharpoons {}^{\bullet}\mathrm{NHONO} \quad \text{degeneracy} = 2 \tag{R3.4}$$

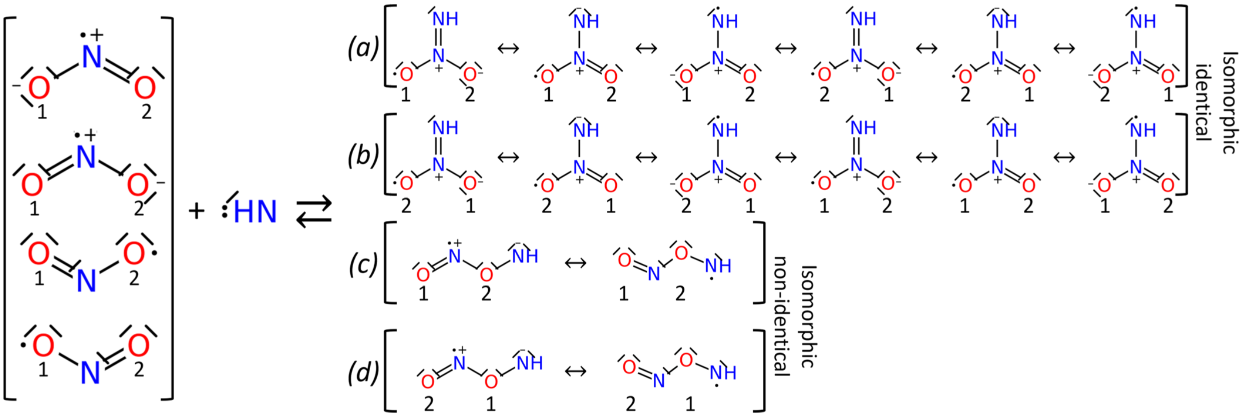


Figure 3-6: An example of the degeneracy determination algorithm based on resonance structures and atom ID labeling for NO_2 + NH $\rightleftharpoons$ $NHNO_2$ (reaction 3.3) and NO_2 + NH $\rightleftharpoons$ NHONO (reaction 3.4).

3.3 Attaining representative localized structures

In many cases, the combinatorial effect of applying resonance pathways (Table 3.1) could result in a relatively large number of localized structures. Many of these structures are not representative of the resonance hybrid (*i.e.*, have a relatively low contribution to the resonance hybrid). It is undesirable to use such non-representative, low-contributing structures in reactive species objects: while the extra memory required for storing the excess localized structures is arguably manageable, substantial time and computational resources are required for keeping track of all possible cross-reactions between all localized structures of any pair of species in the model core. Consequently, considering unrepresentative localized structures as reactive could lead to significant computational challenges even for medium-size systems, spending valuable computational resources on unimportant tasks. In some cases, and depending on hardware, if all localized structures were considered reactive RMG would not be able to terminate successfully.

Two approaches were adopted to identify the representative localized structures (*i.e.*, the important resonance forms). First, the localized electronic configuration search space for structure generation was constrained; additionally, a quantitative heuristic-based run-time filtration procedure was developed. The two approaches are thoroughly described below.

Constraining the localized electronic structure search space was achieved by carefully formulating atom types, which describe feasible configuration of bond orders, lone electron pairs, and formal charges of elements. Each reactive element in RMG (currently H/C/N/O/Si/S/-Cl/F/I) is represented by such atom types. For example, 'O2s' is an atom type describing an oxygen atom with two bonding (or radical) electrons and only single bonds, as in H_2O or $^{\bullet}OH$, while 'O4tc' describes an oxygen atom with four bonding electrons, a triple bond, and a formal charge, as in CO.[32] Atom types cumulatively define the broad configuration space in which new structures could be generated. A thorough description of all atom types is available online in RMG's documentation.[21] If any of the elements in a generated localized structure does not have a respective atom type representation, *i.e.*, has an invalid electronic configuration, the structure is ignored (*e.g.*, see Fig. 3-3 above and respective discussion). If such structures are generated while applying the various resonance pathways, they will not be considered for representing the respective species, and the software will not attempt to apply additional resonance pathways to them.

To further reduce the feasible structure list into a succinct list of representative structures, run-time localized structure filtration heuristics were implemented in RMG in the following order of importance: (a) minimization of deviation from an octet; (b) minimization of formal charges; (c) charge stabilization by electronegativity consistency; (d) charge stabilization by proximity considerations; (e) aromaticity considerations. These heuristics were previously mentioned elsewhere in different variations,[33–35] yet to the best of our knowledge were never before automated. The specific implementation in RMG, along with various exceptions, are thoroughly described below.

The purpose of the "octet rule" is to avoid assigning empty atomic orbitals to stable species. Minimal absolute deviation from an octet is given priority in the heuristics: all additional filtration procedures are applied to the already octet-conforming, or near-octet, list of structures. It is worth noting at this point in the discussion that there is extensive disagreement in the literature regarding the correct representative Lewis structures of third row elements (such as sulfur). A review of chemistry textbooks showed that Lewis structures, formal charges, and expanded octets are among the most inconsistently treated chemical properties.[36] While some advocate expanding the octet rule for these elements (where molecular orbital hybrids of atomic d orbitals are occupied, accounting for dectet and even duodectet) to reduce formal charges[33, 34, 36, 37] (Fig. 3-7A), others support strictly adhering to the octet rule on the expense of adding formal charges[38] (Fig. 3-7B). Papers by Weinhold *et al.* and by Purser[36, 38, 39] provide more details on this disagreement.

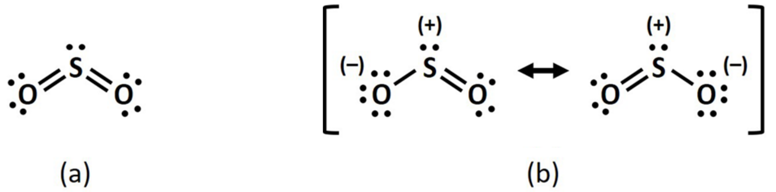


Figure 3-7: Two opposing approaches for representative Lewis structures of SO_2, a common textbook example species. (A) An expanded octet (dectet) structure. (B) Isomorphic octet-conforming structures with formal charge separation.

It is an unresolved question whether one could get by with not enforcing the octet rule in the context of kinetic model generation. RMG is currently set to allow expanded octet structures, such as in Fig. 3-7A; this choice is not set in stone, and was made primarily since the present reaction estimation templates as well as the group additivity thermodynamic estimation method in RMG were not designed for partially charged groups. However, since octet-conforming localized structures were found to describe the electronic structure better than expanded octet structures,[38] future versions of RMG should consider third row elements as adhering to the octet rule. New methods underway to improve thermodynamic data estimation using a neural network approach and to improve and automate kinetics family tree generation will allow transitioning to using octet-conforming structure representations. The octet-conforming structure representation approach would also be advantageous once RMG supports ionic interactions to better represent charge dispersion in species containing third row elements. Still, the current implementation in RMG is expected to perform reasonably well for sulfur species since both the training sets for reaction rate estimations and the rates to be estimated have consistent representations.

Additional heuristics are mainly concerned with partial charges, and result in increased localized structure stabilization. The first of these heuristics is minimization of formal charge separation. Although structures with the least formal charges are preferred, an additional charge separation is allowed if such structures introduce reactive sites (*i.e.*, radicals or π-bonds) not present in all octet-conforming structures with lower charge separation. A prominent example of such case is the partially charged localized structure of NO_2 (Table 3.1, pathway 1). It is noted that some species may only be represented by structures containing formal charges, *e.g.*, N_2O, HNO_3, HCNO, HON, azide, NSN, and singlet H_2NN.

The electronegativity heuristic is used to confirm that negative charges are attributed to more electronegative atoms where possible. For example, the $CH{\equiv}[N^+][O^-]$ structure is preferred over $[CH^-]{=}[N^+]{=}O$ as representative of the CHNO species. Since this heuristic is applied after the octet rule heuristic, some species are ultimately represented by structures violating the electronegativity heuristic; examples include CO ($[C^-]{\equiv}[O^+]$), HNC ($[C^-]{\equiv}[NH^+]$), and HON ($[OH^+]{=}[N^-]$). The stabilization by charge proximity heuristic is applied to give priority to structures in which same-sign charges are as far apart as possible, and vice versa.

Finally, for aromatic species, representations which capture the aromaticity are prioritized over those that do not, while balancing with the inclusion of potential radical sites. For stable monocyclic aromatic species, the Aromatic resonance structure is sufficient to characterize the reactivity. On the other hand, representations of radical aromatics such as benzyl radical need to consider delocalization of the radical into the aromatic ring, which disrupts the aromaticity of that ring in the resonance structure. For polycyclic aromatics, radical delocalization is limited to locations where at least one aromatic ring is maintained. For example, in the case of 1-methylphenanthrene radical (Fig. 3-8), RMG keeps the structures with the radical on the methylene group or carbons 2, 4, 9, and 10a (IUPAC numbering) (Fig. 3-8B), but discards the structures with the radical on carbons 4b, 6, and 8 because they disrupt the aromaticity of all three rings (see Fig. S1 for all non-representative structures). Note that all Kekulé structures are also considered non-representative, and Aromatic and Clar structures are kept instead. This simple heuristic matches well with the observed spin density in that molecule (Fig. 3-8A), which indicates that the ring farthest from the methylene group indeed does not have significant spin density.

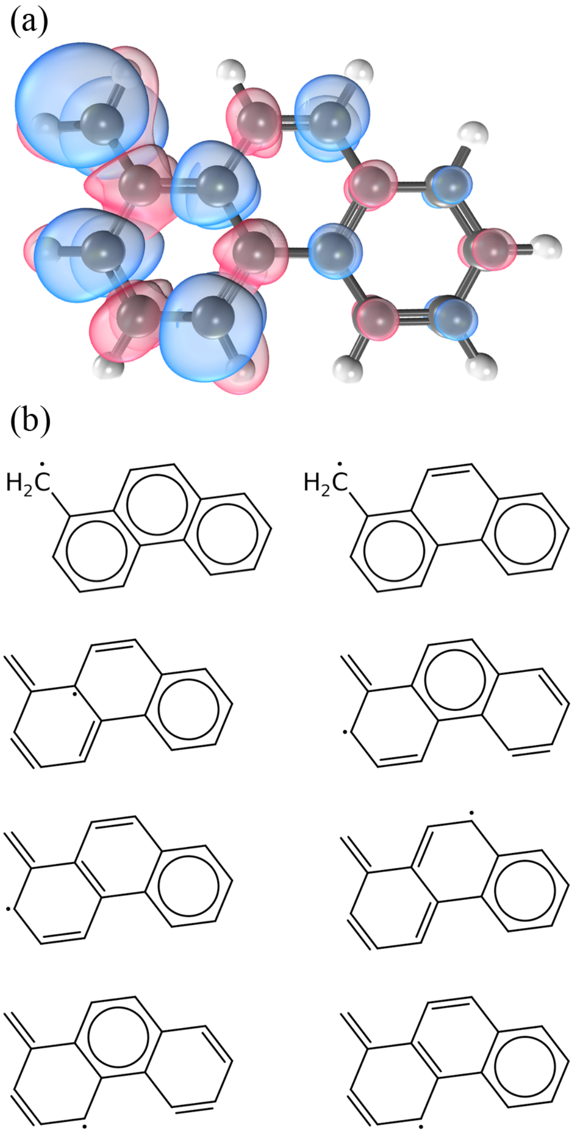


Figure 3-8: (A) Spin-density of 1-methylphenanthrene calculated using the NBO 6.0 population analysis software[24] implemented in Q-Chem 4.4[25] at the uB3LYP/6-311G(2d,d,p) level of theory, and visualized using IQmol[26] using a 1% iso-value. (B) Representative localized structures for 1-methylphenanthrene radical as determined by RMG.

Two species were selected to demonstrate in detail the above filtration heuristics (Fig. 3-9). Formaldiminoxy radical, CH_2NO, is an important intermediate generated by methylene reburning of nitric oxide in combustion processes,[40] and it is also important in nitromethane decomposition. Within the localized electronic structure search space defined by atom types and using the resonance pathways in Table 3.1, RMG discovered twelve resonance structures of this species (Fig. 3-9A). Applying the octet rule heuristic resulted in five structures with minimal octet deviations, as noted in Fig. 3-9A. The formal charge minimization heuristic rules-out structure (4) in the figure; note that both structures (3) and (5) introduce a radical site on the nitrogen atom, hence they are not filtered at this stage. However, since structure (3) violates the electronegativity charge stabilization heuristic, the final localized structures representative of CH_2NO are (1), (2), and (5) only.

Hydroxysulfonyl radical, HSO_3, is an important intermediate in one of the atmospheric H2SO4 formation pathways via $SO_2 + OH \rightleftharpoons HSO_3$ and subsequently $HSO_3 + OH \rightleftharpoons H_2SO_4$,[41] as well as in combustion of fuel-air-H_2S systems.[42] RMG discovered 25 resonance structures for this species. There are only three octet-conforming structures, one of which has an additional charge separation and is filtered out. Consequently, the two representative octet-conforming structures are (1) and (2) in Fig. 3-9B. If expanded octet structures are considered, structures (7) and (8), with no charge separation, are preferred as representatives of the resonance hybrid. Generating and filtering the resonance structures in Fig. 3-9 on a conventional computer (Intel Core i7-4790 CPU) took ~ 5 ms ($\sigma = 1.5$ ms, $n = 1000$) and ~ 20 ms ($\sigma = 4.7$ ms, $n = 1000$) for CH_2NO and HSO_3, respectively. This approach is therefore fast enough for implementation in large-scale automated model generation.

To expedite structure filtration, an on-the-fly filtration procedure was implemented. During localized structure exploration using the relevant pathways to each species (Table 3.1), the algorithm identifies structures which significantly deviate from an octet or have a relatively large charge separation (both with respect to the already generated pool of structures). Although these identified structures are kept, they are not explored further. Using on-the-fly filtration can result in a significant reduction of resonance generation time. For example, 224 localized structures were generated for $CH_2CC(O)OO(T)$ without implementing on-the-fly filtration vs. 147 structures when this procedure is implemented. This $\sim 35\%$ decrease in number of structures translates to a 77% reduction in execution time relative to the values reported above when running on the same machine: ~ 400 ms ($\sigma = 22$ ms, $n = 1000$) vs. only ~ 90 ms ($\sigma = 12$ ms, $n = 1000$).

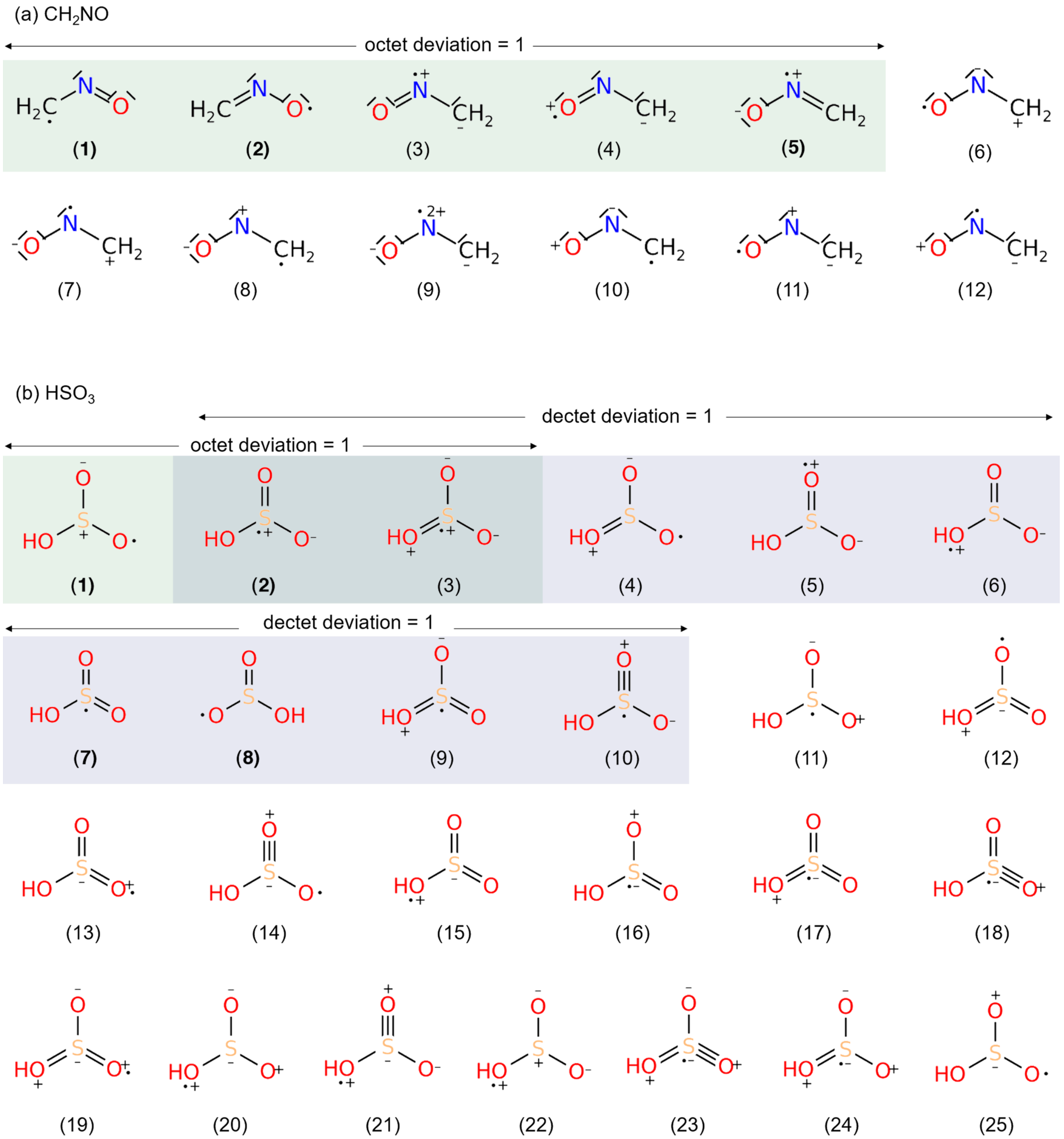


Figure 3-9: Unfiltered non-isomorphic localized structures generated for (A) CH_2NO and (B) HSO_3 radicals. Structures with the lowest octet deviation (or lowest dectet deviation if expanded octet is considered) are highlighted.

Automated localized structure filtration still has challenges to overcome. For example, species with several resonating moieties, particularly if those are conjugated, result in more than a handful of representative localized structures.

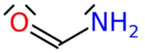
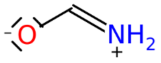
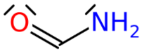
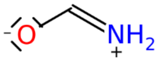
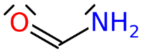
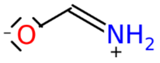
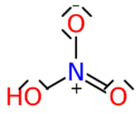
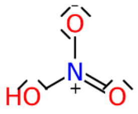
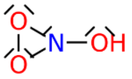
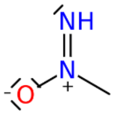
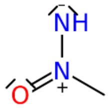
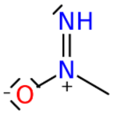
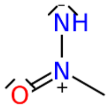
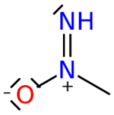
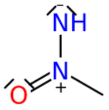
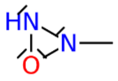
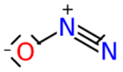
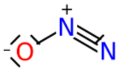
3.4 Validation

Two methods were used to validate the representative resonance structures proposed by RMG, both yielding relative structure contributions to the overall resonance hybrid wavefunction. The Hückel-Lewis configuration interaction (HL-CI) method[35] is based on the Hückel theory, one of the simplest quantum methodologies, using a very simple basis set of carbon atom p_Z orbitals with corrections for heteroatom. For a hydrocarbon, this method approximates all off-diagonal terms of the secular determinant $|H - ES|$ to be equal (H, E, S are the Hamiltonian matrix, the total energy, and the atomic orbital overlap matrix, respectively). The HL-CI algorithm, elaborated with heteroatom correction parameters,[43] was recently implemented as a standalone code as part of this work.[44] HL-CI can be used to identify which resonance structures contribute significantly to the lowest energy wavefunction for a given spin multiplicity. This method has to be initialized with the desired localized structures to be studied, and our current implementation utilizes RMG for this end. Therefore it only serves in this work to determine structure weights and particularly to identify cases where RMG's proposed representative structures actually have relatively low contributions. It is a simple, fast, and convenient method, but cannot be used to identify structures not proposed by the user (or, in our implementation, by RMG).

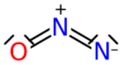
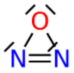
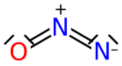
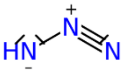
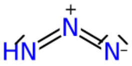
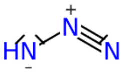
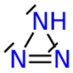
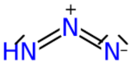
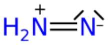
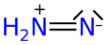
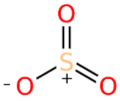
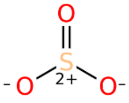
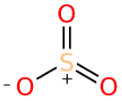
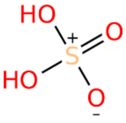
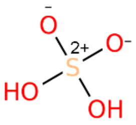
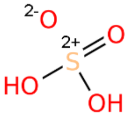
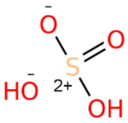
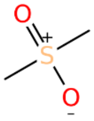
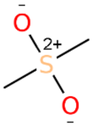
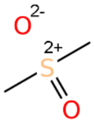
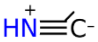
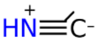
Subsequent to a Natural Bond Orbital (NBO)[45] calculation, the Natural Resonance Theory (NRT)[46–48] analysis method expresses Schrödinger's wave equation solutions in the chemically intuitive language of Lewis-like bonding patterns and associated resonance-type interactions. The electronic structure calculation has to be performed using a host software. Here, NBO 6.0[24] implemented in Q-Chem 4.4[25] was used. For the present purpose, we employed routine DFT with expanded polarization functions, B3LYP/6-311++G(3df,3pd); doublets and triplets were treated with an unrestricted method. The NRT delocalization threshold was set to 0.1 $\mathrm{kcal\,mol^{-}1}$. For singlets, we report the weighting of resonance structures given by NRT. The NRT approach has known practical difficulties with open-shell species, since the α and β electrons experience different exchange forces and may hence lead to different spatial distributions and localization patterns. Therefore, for radical and triplet resonance hybrids we instead report NBO bond orders (BOs).

The HL-CI and NBO/NRT methods were used to test the reliability of RMG's heuristic methods for deciding which resonance structures are representative of the true bonding. The species in Table 3.2 were selected to demonstrate the resonance pathways discussed in Table 3.1 (except for pathway 6 which is mainly relevant for degeneracy determination) and in Fig. 3-5, and include the various examples discussed above in the context of filtration heuristics.

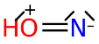
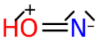
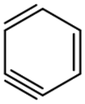
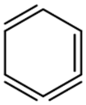
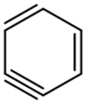
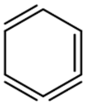
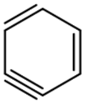
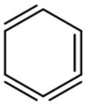
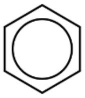
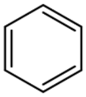
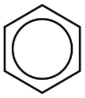
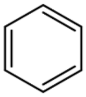
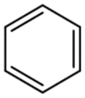
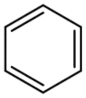
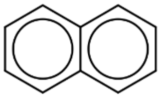
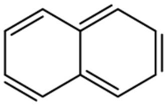
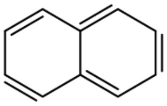
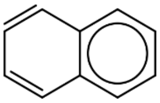
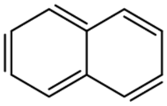
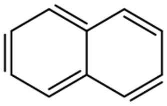
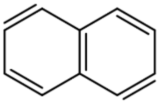
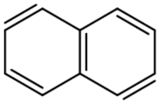
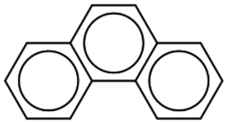
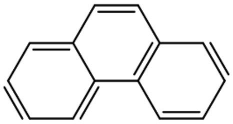
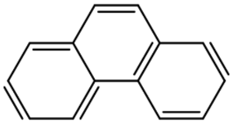
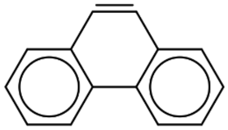
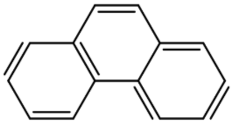
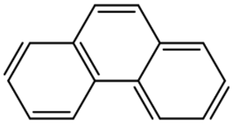
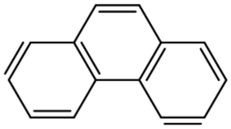
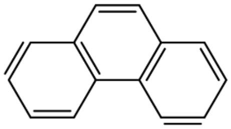
Table 3.2: Comparison of RMG's representative localized structures to weight contributions calculated by the HL-CI method and to representative structures determined by the NBO/NRT method.

Species	RMG	HL-CI	NBO/NRT
NH_2CHO	 	66.2%  33.8% 	62.6%  29.5% 
HNO_3		N/A	68.5%  10.4% 
$CH_3N(O)NH$	 	59.4%  40.6% 	50.5%  50.5%  19.4% 
N_2O		N/A	57.41% 

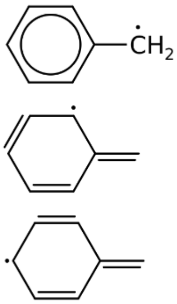
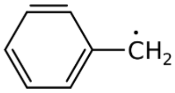
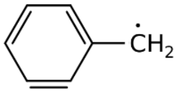
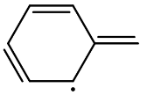
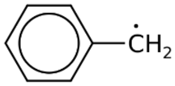
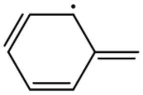
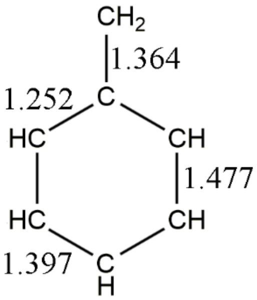
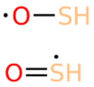
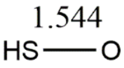
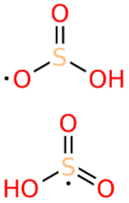
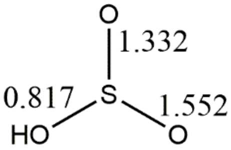
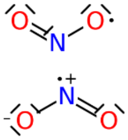
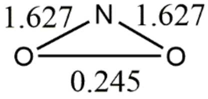
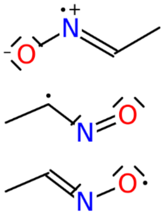
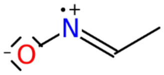
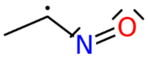
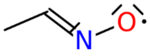
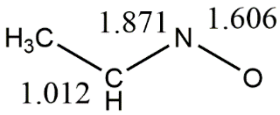
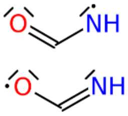
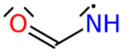
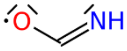
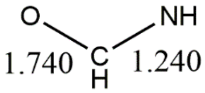
(Continued)

Species	RMG	HL-CI	NBO/NRT
			30.63%  5.10% 
NH_3	 	N/A	50.12%  20.69%  18.54% 
H_2NN		N/A	89.12% 
SO_3		N/A	66.9%  9.9% 
H_2SO_4		N/A	40.7%  21.7%  15.0% 
$DMSO_2$		N/A	48.2%  26.5% 
HNC		N/A	98.9% 

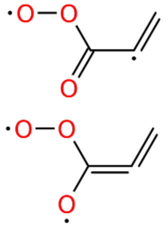
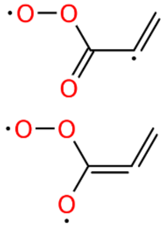
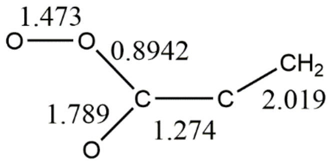
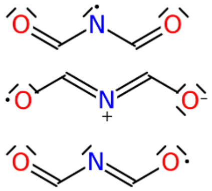
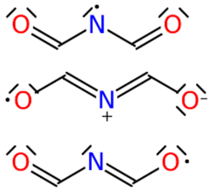
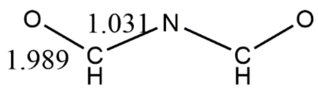
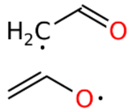
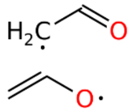
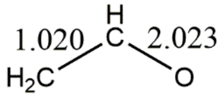
(Continued)

Species	RMG	HL-CI	NBO/NRT
HON		N/A	96.2% 
Benzyne		56.7% 	40.5% 
		43.3% 	32.5% 
Benzene		50.0% 	37.1% 
		50.0% 	37.1% 
Naphthalene		33.3% 	23.0% 
		33.3% 	23.0% 
		33.3% 	19.4% 
Phenanthrene		33.3% 	14.1% 
		33.3% 	14.1% 
		33.3% 	11.0% 

(Continued)

Species	RMG	HL-CI	NBO/NRT
Benzyl radical		28.1%  28.1%  15.3%  15.3%  13.1% 	
HSO		N/A	
HSO_3		N/A	
NO_2		N/A	
CH_3CHNO		74.3%  19.8%  5.9% 	
NHCHO		58.6%  41.4% 	

(Continued)

Species	RMG	HL-CI		NBO/NRT
$CH_2CC(O)OO(T)$		66.4% 33.6%		
OCNCO		37.5% 33.2% 29.4%		
CH_2CHO		67.8% 32.2%		

A benchmark species, NH_2CHO, for which literature data is available,[49] was selected for an initial assessment of all methods (Table 3.2). The two representative structures proposed by RMG were also the major structures according to NBO/NRT, and the weights determined by HL-CI and NBO/NRT were similar. The values reported by NBO/NRT do not sum to unity since other minor structures are also present (structures contributing less than 5% are not shown). Normalizing the NBO/NRT values of major structures resulted in an even closer agreement between HL-CI and NBO/NRT (within few percent), which is encouraging considering that the former method is not based on a modern electronic structure software. The NBO/NRT calculated contributions were found to be in satisfying agreement with the previously published data.[49]

The structures proposed by RMG for the species in Table 3.2 were also those identified by NBO/NRT as the major contributors, with three exceptions: (a) Cyclic resonance structures were suggested by NBO/NRT to be important for HNO_3, $CH_3N(O)NH$, HN_3, N_2O, and NO_2. These cyclic resonance structures are a way of showing there is some interaction between electrons on non-bonded atoms. To our knowledge, this weak interaction does not materially affect which reactions are possible, so RMG ignores these structures. (b) Since RMG currently allows expanded octet structures, hypervalance sulfur species have different representative structures than those determined by NBO/NRT. (c) For aromatic species, RMG's use of global structures to represent delocalization differs from the localized structures determined by NBO/NRT. While the HL-CI method often failed to converge when solving the non-linear equation system, it was nevertheless valuable in assessing relative structure contributions of radical species, for which only total NRT BOs are available.

The resonance hybrid BOs calculated by NBO/NRT are in reasonable agreement with RMG's predictions. Both HSO and HSO_3, which demonstrate the adjacent radical lone pair / multiple bond shift pathway (Table 3.1), have bond orders of about 1.5 between the relevant S and O atoms. This supports RMG's proposed structures with an alternating single/double S–O bond. Further support comes from visualizing the spin density of HSO, showing radical characteristic on both S and O atoms (Fig. 3-10A). Likewise, the CH_3CHNO analysis is in agreement with a BO of ~ 1 for the C–C bond and intermediate BOs for C–N and N–O. Both NHCHO and $CH_2CC(O)OO(T)$ have higher BOs next to the more electronegative O rather than C–N or C–C bonds, in agreement with the higher HL-CI weight ascribed to the structures with the C=O bond. However, NBO/NRT ascribed a BO of 1.473 for the O–O bond in $CH_2CC(O)OO(T)$, which is not reflected in the other methods. Moreover, the OCNCO resonance hybrid shows a BO close to 1 for the C–N bond, in contrast to the relatively high HL-CI weight ascribed to the respective structure with the localized C=N bond (33.2%). Last, the vinoxy radical was previously reported to have BOs of 1.74 and 1.31 C–C and C–O, respectively, using NBO/NRT calculated at B3LYP/6-311++G(d,p).[50] The present analysis, using an updated NRT algorithm implemented in NBO 6.0 and calculated at B3LYP/6-311++G(3df,3pd) level, determined these BOs to be quite different: 1.02 and 2.02, respectively. This discrepancy stems from an inaccurate analysis of an older NBO version used in Ref [50] implementing a preliminary NRT algorithm. The soon-to-be-released NBO 7.0 software[51] determined these values to be 1.18 and 1.97, respectively, using B3LYP/6-311++G(d,p).[52] The NBO 7.0 results support our NBO 6.0 analysis, both suggesting that the localized C=O bond has a higher contribution to the resonance hybrid than the localized C=C bond, in agreement with the HL-CI weights and experimental measurements.[53, 54] The updated NBO 7.0 data also suggests a higher degree of delocalization.

The analysis in Table 3.2 supports the filtration heuristics discussed above. The NH_2CHO,

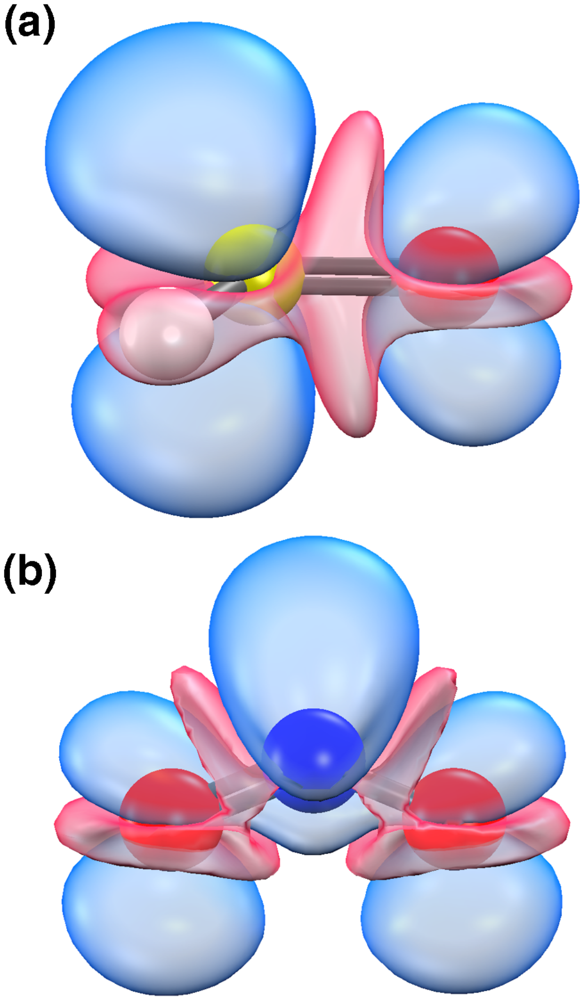


Figure 3-10: Spin densities of (A) HSO and (B) NO_2 calculated using the NBO 6.0 population analysis software[24] implemented in Q-Chem 4.4[25] at uB3LYP/6-311G++(3df,3pf), and visualized using IQmol[26] using a 1% iso-value. Representative structures for these molecules must account for the fact that the unpaired electron resonates over several atoms.

NO_2 (Fig. 3-10B), OCNCO, and CH_3CHNO species demonstrate the importance of including structures with a higher charge separation degree. The importance of accounting for an additional radical site by allowing a structure with larger charge separation is demonstrated by visualizing the spin density of NO_2 (Fig. 3-10B), where the radical density is clearly seen on the N atom in addition to the O atoms. The electronegativity heuristic is also reflected in the selection of the representative CH_3CHNO structure with formal charges. The octet rule is very important in selecting representative structures (including expanded octet for third-row elements), as can be seen, for example, in the singlet H_2NN case; octet-violating localized structures such as $NH_2{-}N$ are much less important than the double bonded structure with formal charges, $[NH_2^+]{=}[N^-]$.

The HL-CI method ascribed a low weight ($\sim 5\%$) to the C=NO moiety of the resonance structure with the radical localized on the O atom, which is predicted by RMG to be present in contributing structures of CH_3CHNO (Table 3.2) and CH_2NO (not shown), for example. In this sense, RMG might be considered conservative.

For aromatic species, we see a decent level of agreement, taking into account RMG's use of global structures. Benzene is a relatively trivial example, since the global structure is simply the average of two potential assignments of double bonds. For naphthalene and phenanthrene, while the exact localized structures predicted by NBO/NRT are not generated by RMG, the Clar structures are essentially an average of those structures. For benzyl radical, averaging the bond orders across all of the RMG-predicted structures gives BOs which are reasonably close to those predicted by NBO/NRT. Accounting for the fact that the resonance structures are not all equal contributors, it is possible to get even closer to the NBO/NRT BOs. Finally, for benzyne, RMG includes both of the structures identified by NBO/NRT.

RMG is unaware of resonance structure weights. Ideally, knowledge of the relative structure weights is unnecessary during automated model generation, since the database is trained (if enough relevant data is available) with experimental or calculated thermodynamic data and reaction rates that already consider the physical electron delocalization. However, discriminating between "representative" structures and ignorable "non-representative" structures is imperative for automatic model generation and is completely independent of being able to estimate relative structure weights.

If a species has no thermodynamic data from respective libraries, RMG will using the group additivity approach[6] to estimate the thermodynamic properties. This procedure is applied for all representative resonance structures of a species, and the dataset belonging to the structure with the lowest ΔH_f^o will be used for the species, as discussed elsewhere.[5] Reaction rate estimates depend on matching templates of radicals, lone electron pairs, and bonding patterns in the reacting structures. Since such localized electronic configurations are affected by delocalization, it is crucial that RMG considers all major resonance structures representing the resonance hybrid to capture important reactions. The algorithms presented and validated herein successfully identified the representative resonance structures for a variety of chemical systems of interest to gas phase kinetics, and therefore have an important role in automated reaction mechanism generation.

3.5 Conclusions

This work presents an efficient and automated method for discovery of representative resonance structures of arbitrary chemical species in the C/H/O/N/S system, recently implemented in the Reaction Mechanism Generator (RMG) software.

Localized structures are generated using a fundamental set of resonance pathway types, accounting for two- and three-atom systems as well as global approaches for aromatic species. These pathways are applied to discover new localized chemical structures within a pre-defined electronic search space based on atom types ascribed to elements. Since the number of generated resonance structures is often relatively large, leading to an impractical algorithm and the inclusion of unrealistic reactions, the structure list is filtered using the octet rule (allowing dectet for third-row elements) along with formal-charge-related heuristics.

The representative structures identified by RMG were validated against two methods, the Hückel-Lewis configuration interaction (HL-CI) and the Natural Resonance Theory implemented in the Natural Bond Orbital software (NBO/NRT). These existing methods are not convenient for automated mechanism generation for large kinetic models. The discussed algorithm and both literature methods were found to be in satisfying agreement. With only three exceptions (out of two dozen), the structures proposed by RMG were also the major contributors in the NBO/NRT analysis. For some non-cyclic species, a cyclic structure (*i.e.*, a structure with a significant bond-order between nominally non-bonded atoms) was identified by NBO/NRT as having a significant contribution. Furthermore, currently RMG allows expanded octet for third-row elements, in contrast to the suggested NBO/NRT structures. The HL-CI method suggested that a the moiety C=NO with the radical site on the oxygen atom, predicted as representative by RMG, has a relatively low ($\sim 5\%$) contribution to the resonance hybrid. Finally, the global structures used for aromatics contrasts with the basic design of NBO/NRT. The above discrepancies are all acceptable, and the heuristic-based predictions proved to be efficient and accurate enough for large scale automated model generation.

The approach presented herein and implemented in RMG is essential for identification of reactive sites and consequently reaction templates. Automatically discovering potential localized structures along with filtration to identify the representative structures was shown to be robust and relatively fast. This algorithm has also been made accessible online from the RMG website (https://rmg.mit.edu/)[21] without any required installation to support researchers and educators.

References

(1) Lu, T.; Law, C. K. *Prog. Energy Combust. Sci.* **2009**, *35*, 192–215.

(2) Chen, D.; Wang, K.; Wang, H. *Combust. Flame* **2017**, *186*, 208–210.

(3) Warth, V.; Battin-Leclerc, F.; Fournet, R.; Glaude, P.; Côme, G.; Scacchi, G. *Comput. Chem.* **2000**, *24*, 541–560.

(4) Vandewiele, N. M.; Van Geem, K. M.; Reyniers, M.-F.; Marin, G. B. *Chem. Eng. J.* **2012**, *207-208*, 526–538.

(5) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(6) Benson, S. W.; Cruickshank, F. R.; Golden, D. M.; Haugen, G. R.; O'Neal, H. E.; Rodgers, A. S.; Shaw, R.; Walsh, R. *Chem. Rev.* **1969**, *69*, 279–324.

(7) Lewis, G. N. *J. Am. Chem. Soc.* **1916**, *38*, 762–785.

(8) IUPAC, *Lewis formula (electron dot or Lewis structure)*, 2nd ed.; McNaught, A. D., Wilkinson, A., Eds.; Blackwell Scientific Publications: Oxford, 1997.

(9) Thiele, J. *Justus Liebig's Ann. der Chemie* **1899**, *306*, 87–142.

(10) Pauling, L. *J. Am. Chem. Soc.* **1931**, *53*, 3225–3237.

(11) Benson, S. W. *J. Am. Chem. Soc.* **1965**, *87*, 972–979.

(12) Wang, H. et al. A high-temperature chemical kinetic model of n-alkane (up to n-dodecane), cyclohexane, and methyl-, ethyl-, n-propyl and n-butyl-cyclohexane oxidation at high temperatures, JetSurF version 2.0., 2010.

(13) Sivaramakrishnan, R.; Su, M.-C.; Michael, J. V.; Klippenstein, S. J.; Harding, L. B.; Ruscic, B. *J. Phys. Chem. A* **2010**, *114*, 9425–9439.

(14) Rao, H.-B.; Zeng, X.-Y.; He, H.; Li, Z.-R. *J. Phys. Chem. A* **2011**, *115*, 1602–1608.

(15) Burke, M. P.; Chaos, M.; Ju, Y.; Dryer, F. L.; Klippenstein, S. J. *Int. J. Chem. Kinet.* **2012**, *44*, 444–474.

(16) Goos, E.; Burcat, A.; Ruscic, B. Extended Third Millennium Ideal Gas and Condensed Phase Thermochemical Database for Combustion with Updates from Active Thermochemical Tables., 2010.

(17) Merchant, S. S.; Zanoelo, E. F.; Speth, R. L.; Harper, M. R.; Van Geem, K. M.; Green, W. H. *Combust. Flame* **2013**, *160*, 1907–1929.

(18) Class, C. A.; Aguilera-Iparraguirre, J.; Green, W. H. *Phys. Chem. Chem. Phys.* **2015**, *17*, 13625–13639.

(19) Grinberg Dana, A.; Buesser, B. A.; Merchant, S. S.; Green, W. H. *Int. J. Chem. Kinet.* **2018**, *50*, 243–258.

(20) RMG-Py version 2.3.0., 2018.

(21) The Reaction Mechanism Generator Website.

(22) Deo, N., *Graph theory with applications to engineering & computer science*, Dover ed.; Dover Publications, Inc.: Mineola, New York, 2016.

(23) Anslyn, E. V.; Dougherty, D. A., *Modern Physical Organic Chemistry*; University Science: Sausalito, CA, 2006.

(24) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. **2013**.

(25) Shao, Y. et al. *Mol. Phys.* **2015**, *113*, 184–215.

(26) Gilbert, A. T. B. IQmol., 2018.

(27) Džonova-Jerman-Blažič, B.; Trinajstić, N. *Comput. Chem.* **1982**, *6*, 121–132.

(28) Clar, E.; Zander, M. *J. Chem. Soc.* **1958**, 1861–1864.

(29) Liu, M.; Green, W. H. *Proc. Combust. Inst.* **2019**, *37*, 575–581.

(30) Schlag, E. W. *J. Chem. Phys.* **1963**, *38*, 2480–2482.

(31) Bishop, D. M.; Laidler, K. J. *J. Chem. Phys.* **1965**, *42*, 1688–1691.

(32) Frenking, G.; Loschen, C.; Krapp, A.; Fau, S.; Strauss, S. H. *J. Comput. Chem.* **2007**, *28*, 117–126.

(33) Ahmad, W.-Y.; Omar, S. *J. Chem. Educ.* **1992**, *69*, 791.

(34) Carroll, J. A. *J. Chem. Educ.* **1986**, *63*, 28.

(35) Hagebaum-Reignier, D.; Girardi, R.; Carissan, Y.; Humbel, S. *J. Mol. Struct. THEOCHEM* **2007**, *817*, 99–109.

(36) Purser, G. H. *J. Chem. Educ.* **1999**, *76*, 1013.

(37) Lever, A. B. P. *J. Chem. Educ.* **1972**, *49*, 819.

(38) Suidan, L.; Badenhoop, J. K.; Glendening, E. D.; Weinhold, F. *J. Chem. Educ.* **1995**, *72*, 583.

(39) Weinhold, F. *J. Chem. Educ.* **2005**, *82*, 527.

(40) Shapley, W. A.; Bacskay, G. B. *Theor. Chem. Accounts Theory, Comput. Model. (Theoretica Chim. Acta)* **1998**, *100*, 212–221.

(41) Wayne, R. P., *Chemistry of atmospheres : an introduction to the chemistry of the atmospheres of earth, the planets, and their satellites.* Oxford [England] ; New York : Oxford University Press, 2000.: 2000.

(42) Savel'ev, A. M.; Starik, A. M.; Titova, N. S. *Combust. Explos. Shock Waves* **2002**, *38*, 609–621.

(43) Van-Catledge, F. A. *J. Org. Chem.* **1980**, *45*, 4801–4802.

(44) Grinberg Dana, A. An iPython notebook for localized structure weight calculations using the HLCI method., 2018.

(45) Glendening, E. D.; Landis, C. R.; Weinhold, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 1–42.

(46) Glendening, E. D.; Weinhold, F. *J. Comput. Chem.* **1998**, *19*, 593–609.

(47) Glendening, E. D.; Weinhold, F. *J. Comput. Chem.* **1998**, *19*, 610–627.

(48) Glendening, E. D.; Badenhoop, J. K.; Weinhold, F. *J. Comput. Chem.* **1998**, *19*, 628–646.

(49) Glendening, E. D.; Hrabal II, J. A. *J. Am. Chem. Soc.* **1997**, *119*, 12940–12946.

(50) Weinhold, F. In *Comput. Methods Photochem. Vol. 13 Mol. Supramol. Photochem.* Kutateladze, A. G., Ed.; CRC Press, Taylor & Francis Group: Boca Raton, FL, 2005.

(51) Glendening, E. D.; Badenhoop, J. K.; Reed, A. E.; Carpenter, J. E.; Bohmann, J. A.; Morales, C. M.; Weinhold, F. **2018**.

(52) Weinhold, F. Private Communication., 2018.

(53) Foster, S. C.; Miller, T. A. *J. Phys. Chem.* **1989**, *93*, 5986–5999.

(54) Endo, Y.; Saito, S.; Hirota, E. *J. Chem. Phys.* **1985**, *83*, 2026–2034.

Chapter 4

Improving parameter estimation for polycylic species in RMG

4.1 Introduction

In modeling the chemistry of polycyclic aromatic hydrocarbons, there are two main aspects which can significantly affect parameter estimation: aromaticity and ring strain. Correctly capturing these effects is essential to accurately simulating these systems in automatic mechanism generation. Representation and handling of aromaticity in the Reaction Mechanism Generator (RMG) software has been discussed previously in the context of resonance structures.[1, 2] However, ring strain is even more challenging to model given the limitations of the 2D molecule representations which are common in cheminformatics and used almost exclusively in RMG. As a result, there is no concept of 3D geometry or ring strain effects aside from what might be implicitly encoded in thermochemistry or kinetics groups and data.

This lack of ring strain knowledge is often made readily apparent by the types of molecules which may be generated by RMG. A sample of some infeasible polycyclic structures which have been generated in the past is shown in Figure 4-1. The creation of these species and their inclusion in the model can be attributed to failures in both thermochemistry and kinetics estimation; the algorithm thinks that the reaction is sufficiently fast and the product is sufficiently stable.

It's not particularly surprising that this is an issue, given the enormous number of possible

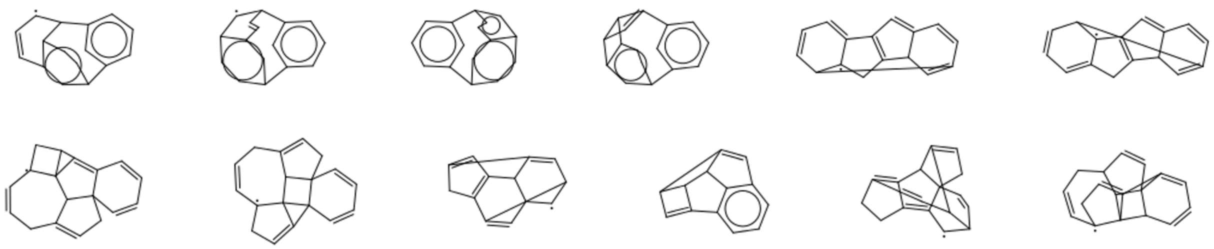


Figure 4-1: A selection of some infeasible polycyclic species which RMG generates.

cyclic structures and the generally greater complexity associated with them compared to acyclic structures. Figure 4-2 compares the number of possible trees (*i.e.*, acyclic graphs)[3] to the number of possible connected graphs[4] as a function of the number of nodes. This provides a useful illustration to give an order of magnitude approximation of the number of linear vs. cyclic molecules. From this we can see the number of possible cyclic structures quickly surpasses the number of acyclic ones.

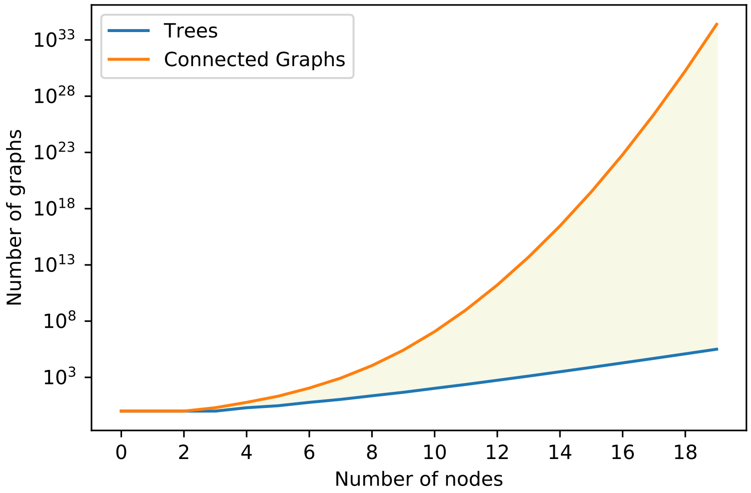


Figure 4-2: Total number of trees (*i.e.*, acyclic graphs) compared to all connected graphs as a function of the number of nodes. The difference between the two represents the number of graphs with at least one cycle.

However, only a small fraction of all possible cyclic molecules will be chemically significant. If a naïve computer program were used to generate molecular structures, limited only by basic understanding of bonding and valence limits (*e.g.*,, RMG without good data), generation of unstable structures like those shown in Figure 4-1 would quickly overwhelm computational capacity. In order to properly identify only important species, RMG must have accurate thermochemistry and kinetics estimation. Thus, improvements in both areas are needed to reduce the number of infeasible polycyclics predicted by RMG.

4.2 Ring perception

Before actually reaching the parameter estimation stage, one of the first considerations when working with cyclic molecules is how to identify cycles or rings. This ring perception task is a well explored problem in graph theory. Although the definition of a ring is very clear (a set of connected vertices/atoms which form a closed loop), there are many ways to define a list of rings in a given graph. Previously, the only ring perception algorithm implemented in RMG was a method for finding the smallest set of smallest rings (SSSR), also called the minimal cycle basis set. The idea of this particular ring set is to find the minimal set (*i.e.*, involving the fewest atoms/bonds) of linearly independent rings which can span the entire graph. In other words, all possible rings in the graph can be constructed by some linear combination of rings in the SSSR.

The SSSR is very commonly used in cheminformatics for identifying rings in molecules and is also implemented in RDKit[5], OpenBabel[6], and the CDK[7] in addition to RMG. However, one major problem with the SSSR is that it can be non-deterministic in cases where there are multiple candidate rings which are the same size. A common example here is cubane, shown in Figure 4-3. The correct SSSR for cubane includes only five of the six potential rings, since the last ring can be obtained from a linear combination of the other five. The SSSR implmentation in RMG was actually an incomplete implementation of an algorithm proposed by Fan *et al.*[8] As a result, RMG would only return four of the six rings.

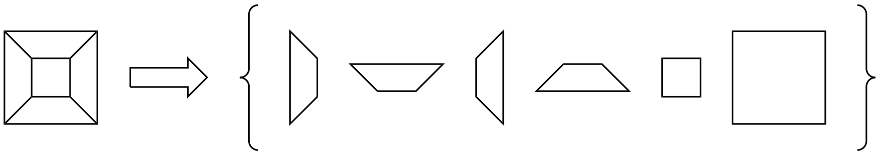


Figure 4-3: Six smallest rings of cubane. The SSSR is comprised of five out of the six. The SSSR algorithm in RMG only returns four of the expected five.

The main issue with a non-deterministic ring set like the SSSR is that the seemingly simple question of “How many rings is this atom a part of?” can have different answers depending on the exact set of rings which are returned. Here, there were two goals: first to fix or replace the existing SSSR algorithm in RMG to return the correct result for species like cubane, second to find a different cycle set which would be deterministic and give a more representative result. For the second goal, the set of relevant cycles (RC), also known as the set of $\mathcal{K}$-cycles was identified as a better option.[9] The set of relevant cycles is comprised of the union of all possible SSSR, meaning that it eliminates the non-determinism of the SSSR. In the case of cubane, the RC includes all six possible rings. It is also implemented as the largest set of smallest rings (LSSR) in OpenBabel.

A newer algorithm for enumerating the SSSR was proposed by Figueras,[10] shown to be faster than the older algorithm by Fan *et al.*[8] by using a breadth-first search algorithm. An initial implementation of the new algorithm was completed, before discovering that the algorithm itself would return incorrect results in certain cases, as noted by Berger *et al.*[11] Berger also reviewed a number of other algorithms for generating a various types of cycle sets in the context of chemical structures, including known cases for which each algorithm would produce incorrect results. For generating the set of relevant cycles, an algorithm by Vismara [12] has been well-received for its speed and accuracy. An extension of Vismara's algorithm had already been implemented as part of the RingDecomposerLib (RDL) package.[13]

Specifically, RDL is based on a slightly different concept called the Unique Ring Families (URFs) proposed by Kolodzik *et al.*[14] The basic idea is to identify at a higher level "families" of rings which align more closely with chemically intuitive understanding of the rings present in a molecule. Each URF is comprised of RCs which have the same size, share some number of bonds, and can be inter-converted via XOR operations with smaller RCs. From these URFs, both the SSSR and set of RC can be readily identified. As such, the RDL package provided the exact functionality which was needed by RMG. The code is written in C with a Python wrapper, which made integration into RMG very straightforward. Currently, RMG only uses RDL for calculating SSSR and RC; however, future integration of more URF features would be straightforward to implement and potentially valuable.

In addition to improved accuracy and reliability over the old SSSR implementation, RDL also provides improved performance and scaling. Figures 4-4–4-7 compare the performance of the old SSSR method and the new SSSR and RC methods which call RDL. In general, RDL is much faster at computing these cycle sets, and scales almost linearly with the number of rings in the molecule, while the original SSSR implementation scales roughly exponentially. However, Figure 4-7 shows how performance changes for a set of cases where the RC behaves poorly. Figure 4-8 shows a sample molecule in the series used for the test. While the SSSR for this molecule includes all eight cyclohexane rings and one of the 24-membered rings going around the entire molecule, the RC includes all 256 potential paths for the 24-membered ring. This results in the much worse scaling for enumerating the RC compared to the SSSR.

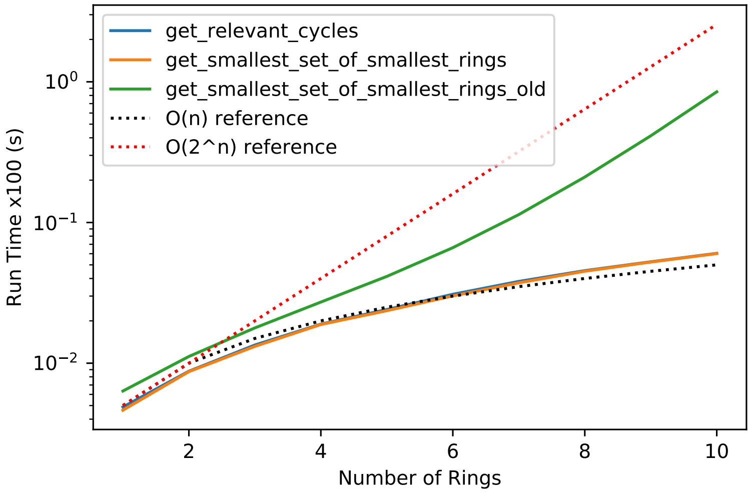


Figure 4-4: Performance comparison of ring enumeration methods for linearly fused cyclohexane rings (*e.g.*, cyclohexane, decalin, etc.). Vertical axis shows total run time for 100 consecutive calls. Each data point is also averaged over 10 trials. The previous SSSR algorithm is denoted get_smallest_set_of_smallest_rings_old, while the other two use RDL.

4.3 Thermochemistry estimation

For thermochemistry, others have worked on improving estimation of ring strain for group additivity estimates as well as completely new approaches using convolutional neural networks.

In standard group additivity (GAV), ring strain corrections can be applied by matching the full molecule to a single group. The most specific correction matching the molecule will be applied, but if there are no matching groups, zero correction is applied. As an extension of this base GAV method, RMG now uses heuristic methods shown in Figure 4-9(a) to estimate

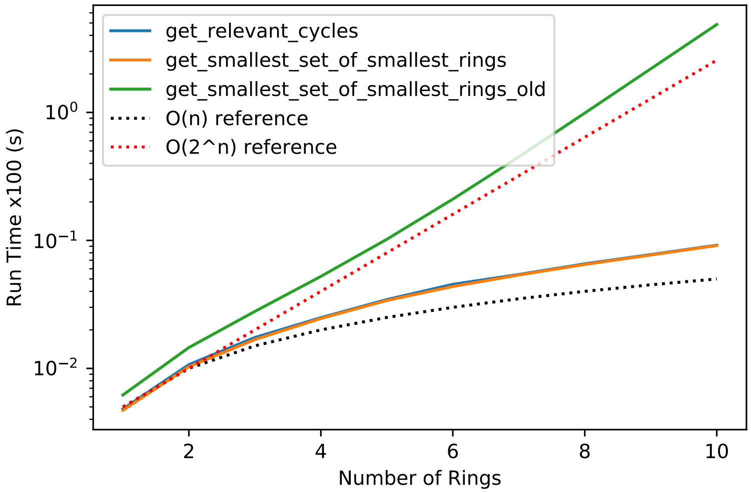


Figure 4-5: Performance comparison of ring enumeration methods for connected cyclohexane rings (*e.g.*, cyclohexane, bicyclohexyl, etc.). Vertical axis shows total run time for 10 consecutive calls. Each data point is also averaged over 10 trials. The previous SSSR algorithm is denoted get_smallest_set_of_smallest_rings_old, while the other two use RDL.

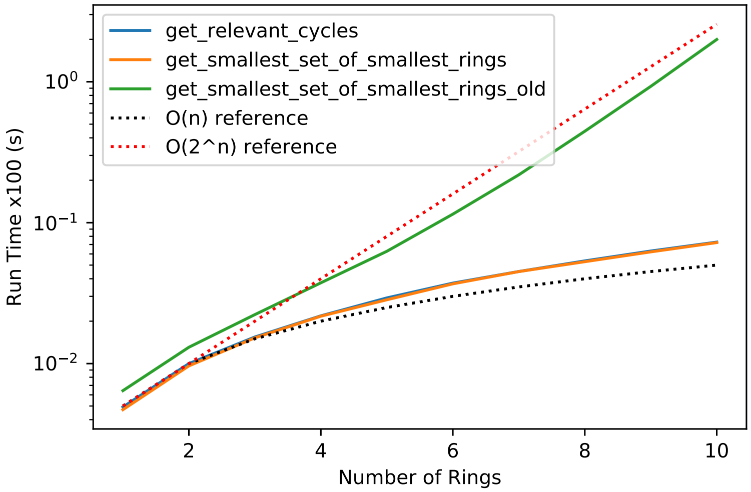


Figure 4-6: Performance comparison of ring enumeration methods for spiro fused cyclohexane rings (*e.g.*, cyclohexane, spiro[5.5]undecane, etc.). Vertical axis shows total run time for 10 consecutive calls. Each data point is also averaged over 10 trials. The previous SSSR algorithm is denoted get_smallest_set_of_smallest_rings_old, while the other two use RDL.

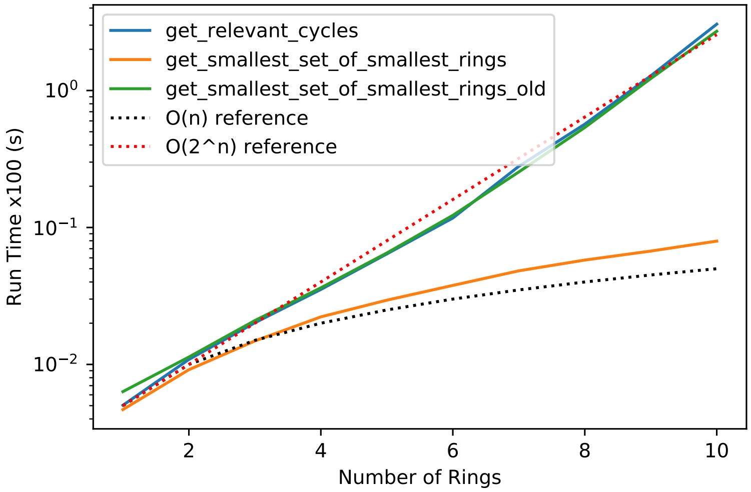


Figure 4-7: Performance comparison of ring enumeration methods for spiro fused cyclohexane rings where the rings form a larger ring (*e.g.*, Figure 4-8). Vertical axis shows total run time for 10 consecutive calls. Each data point is also averaged over 10 trials. The previous SSSR algorithm is denoted get_smallest_set_of_smallest_rings_old, while the other two use RDL.

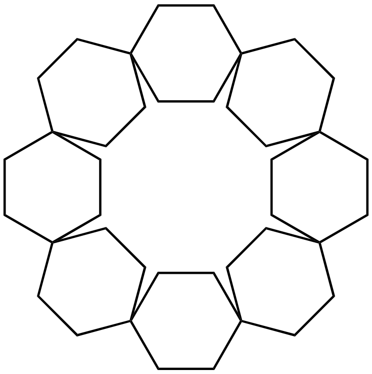


Figure 4-8: Example of a worst case molecule for enumerating relevant cycles. Adapted from Kolodzik *et al.*[14]

ring strain in polycyclic molecules if there is not a group which matches the full molecule.[15] The method uses a database of calculated ring strain values for selected bicyclics. In cases where there is no exact match for a bicyclic structure, RMG uses a separate heuristic to estimate it, shown in Figure 4-9(b). The key benefit of this approach is that a polycyclic molecule is guaranteed to get a ring strain correction, although its accuracy may vary.

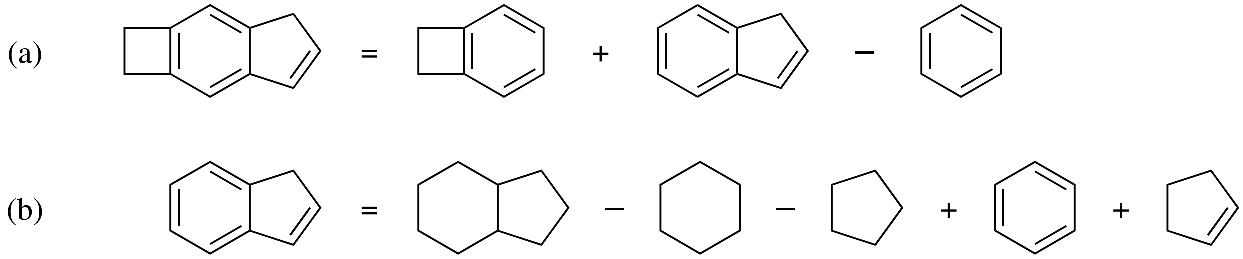


Figure 4-9: (a) Bicyclic decomposition method to predict ring strain of large polycyclics. (b) Heuristic to estimate bicyclic ring strain for unknown bicyclics.

This heuristic approach assumes that ring strain contributions from bicyclic substructures are independent, which is a decent approximation in most cases of linear polycyclics like the one in Figure 4-9. However, the majority of polycyclic species are fused polycyclics which are poorly captured by bicyclic decomposition because of the extra strain experienced by the central atom which is part of all three rings. Additionally, modeling polycyclic aromatic hydrocarbon formation generally leads to the formation of unsaturated polycyclics, which have unique ring strain which is often poorly captured by the second heuristic.

A completely different approach using machine learning has also been implemented in RMG.[16] The machine learning estimator (MLE) uses graph convolution to convert molecular graphs into a fingerprint, which is then used in a standard feed-forward neural network to estimate thermochemistry. The neural network was first trained on a database of thermochemistry calculated at DFT levels, and transfer learning was then used to learn from a smaller dataset calculated using CCSD(T)-F12.

One benefit of group additivity and the polycyclic heuristic is that improving the estimate for a specific polycyclic structure is very straightfoward. Throughout the course of building various models, many polycyclic structures were discovered to be consistently generated by RMG despite seeming unstable. In almost all cases, the primary reason was poor thermo estimates resulting in excessively low enthalpies of formation, which let RMG to believe that the molecules were more stable than in reality. The solution was to calculate thermochemistry for these species and create new polycyclic corrections. These species are summarized in Table 4.1 along with a comparison of $\Delta H_{f,298}$ values estimated by GAV and the machine learning estimator and calculated using CBS-QB3.

GAV estimates generally underestimate the enthalpy of these species as expected, although it does overestimate the enthalpy by a substantial amount for species 7. MLE performs similarly and tends to underestimate enthalpy, but its performance relative to GAV is unpredictable. For some species, it does better, and for others, it does worse. This illustrates one downside of the MLE, namely that the estimation method is very non-transparent, and it is unclear exactly what contributes to a particular estimate. Additionally, the poor estimates for these particular species can likely be attributed to the absence of similar species from the training set. Compared to GAV, it seems that MLE could be worse at extrapolating to species outside of the training set. However, the MLE can be continuously improved by adding new data to the training set (such as these calculations) and re-training the model, which will improve future estimates.

The new polycyclic group additivity corrections fitted from these calculations are provided in Table 4.2. The labels are simply identifiers used when adding these entries to the RMG-database. Roughly speaking, the 's' numbers indicate number of shared atoms between rings, the plain numbers indicate sizes of those rings, and the remaining descriptors identify unsaturated bonds. With the addition of these corrections to the database, GAV estimates for the molecules shown in Table 4.1 are now identical to the calculated values. Additionally, species with the exact same cyclic core structure (*e.g.*, molecules with radicals or side chains) also benefit from the corrections.

However, while this approach of adding new ring strain corrections works well on a case by case basis, it does not significantly improve the prediction accuracy for polycyclics in general. It would be theoretically possible to implement a tricyclic decomposition method, but complexity of such a method would increase substantially. Determining how to decompose a complex polycyclic structure into tricyclic components would be non-trivial, and would have the complicating factor that two tricyclic components could overlap by either one or two rings. Enumerating and calculating data for potential tricyclic substructures would also

Table 4.1: Comparison of GAV, MLE, and CBS-QB3 values for $\Delta H_{f,298}$ of selected strained polycyclic species.

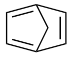
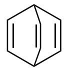
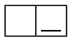
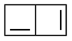
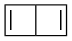
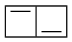
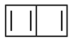
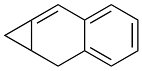
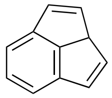
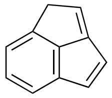
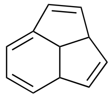
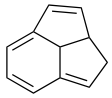
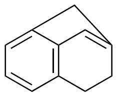
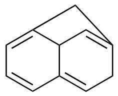
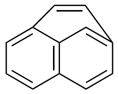
		$\Delta H_{f,298}$ (kcal/mol)		
	Species	GAV	MLE	CBS-QB3
1		61.9	69.9	143.8
2		60.6	71.8	71.4
3		64.6	61.3	83.0
4		92.2	106.1	123.5
5		97.4	103.3	97.3
6		91.6	102.1	132.4
7		210.3	151.8	168.8
8		66.2	57.1	80.3
9		65.5	69.7	87.8
10		66.7	66.6	79.4
11		48.0	59.6	74.9
12		61.0	52.4	74.8
13		40.6	41.0	161.9
14		69.8	61.1	176.2
15		95.0	64.4	177.9

Table 4.2: Fitted group additivity ring strain corrections for selected strained polycyclic species.

		$\Delta H_{f,298}$	ΔS_{298}	C_p (cal/mol K)						
	Label	(kcal/mol)	(cal/mol K)	500 K	600 K	700 K	800 K	900 K	1000 K	1500
1	s3_5_5_triene	103.170	67.392	-7.643	-8.911	-8.948	-8.325	-6.710	-5.427	-4.153
2	s4_6_6_barrelene	19.066	49.625	-6.525	-4.619	-3.437	-2.596	-1.814	-1.708	-1.904
3	s2_4_4_ene_2	80.045	60.555	-5.463	-5.577	-5.448	-4.796	-4.053	-3.693	-2.850
4	s2_4_4_diene_1_3	96.567	67.309	-6.387	-6.461	-6.218	-5.581	-4.592	-4.193	-3.443
5	s2_4_4_diene_1_4	65.133	65.027	-6.490	-5.314	-4.653	-4.048	-3.323	-3.248	-2.934
6	s2_4_4_diene_2_5	106.085	65.391	-4.942	-6.178	-6.329	-5.779	-4.891	-4.528	-3.455
7	s2_4_4_triene_1_4_m	118.453	76.077	-1.380	-3.176	-4.085	-4.386	-4.685	-4.712	-5.913
8	s2_s2_6_6_3_ben_ene	49.858	57.209	-6.078	-6.636	-6.774	-6.170	-4.869	-3.677	-2.520
9	s2_s2_s2_6_5_5_ben_diene1	31.613	54.952	-7.094	-6.439	-6.029	-5.410	-4.401	-3.268	-1.804
10	s2_s2_s2_6_5_5_ben_diene2	27.606	67.124	-7.726	-7.742	-7.568	-7.171	-6.048	-4.988	-3.887
11	s2_s2_s2_6_5_5_diene_ene_ene1	15.048	90.457	-13.880	-13.484	-12.673	-11.404	-8.898	-6.838	-4.917
12	s2_s2_s2_6_5_5_diene_ene_ene2	20.095	90.893	-14.117	-14.900	-14.467	-13.113	-10.292	-7.892	-5.481
13	s2_s2_s3_6_6_5_ben_ene	132.105	58.708	-6.450	-5.745	-5.424	-5.174	-4.423	-3.607	-2.536
14	s2_s2_s3_6_6_5_diene_diene	117.332	83.403	-11.166	-12.168	-11.642	-10.223	-7.567	-5.820	-4.412
15	s2_s2_s3_6_6_6_ben_triene	105.475	58.725	-7.245	-7.052	-6.442	-5.661	-4.296	-3.190	-3.003

be much more difficult than it was for bicyclics simply due to the larger number of possible structures.

4.4 Kinetics estimation

4.4.1 Ring membership

Kinetics estimation in RMG uses rate rules which are associated with group definitions, typically focused on describing the reacting site in the molecule. As such, structural information contained in kinetics groups is often limited to descriptions of the local environment, unlike the polycyclic group additivity corrections which are matched to the global structure of the molecule. Therefore, it can be challenging to distinguish rate rules for linear and cyclic molecules without writing elaborate groups which effectively define the entire molecule. For example, the two reactions in Figure 4-10 share the same substructure relevant to the intramolecular cyclization reaction. Although they would be expected to have very different reaction rates in reality, they would most likely match the same group in the kinetics tree and be assigned the same rate.

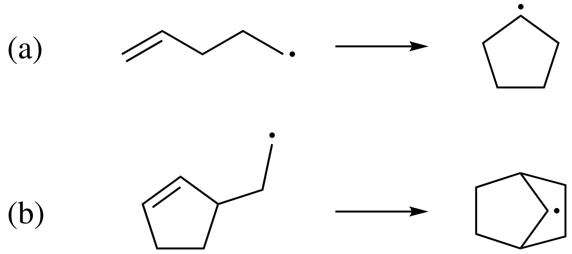


Figure 4-10: Two reactions with identical sub-structures but very different rates.

Therefore, a method to introducing information about global structure to the local group definitions has the potential to substantially improve rate estimates in such cases. There are numerous directions in which this could be taken, with varying levels of complexity in how much information is stored and how it is used. As an initial step, a simple attribute indicating whether each atom is part of a ring was implemented to provide more global context at an atom level. In the example above, such an attribute would enable straightforward distinction between the substructures of the two reactants.

In terms of implementation, the ring membership attribute was attached to the `Atom` and `GroupAtom` classes, such that each atom in a molecule would be associated with a value for whether or not it was part of any ring. The second necessary step was adding support for reading and writing the attribute in group adjacency lists, which enables defining groups with specific ring membership requirements. This was accomplished by extending the existing adjacency list syntax with a new optional parameter, prefaced by the ‘r’ key, as shown in Figure 4-11. This increases the amount of information which can be encoded in a group definition, thereby improving their flexibility and specificity.

```
# No ring specification, equivalent to wildcard
1 C u0 p0 c0 {2,S} {3,S} {4,S}
# Carbon not in any rings
1 C u0 p0 c0 r0 {2,S} {3,S} {4,S}
# Carbon in at least 1 ring
1 C u0 p0 c0 r1 {2,S} {3,S} {4,S}
```

Figure 4-11: Examples of the new ring attribute in the RMG’s adjacency list format.

The current implementation was specifically designed to be easily extensible in the future. The potential next step may be to convert the attribute from a boolean value to an integer indicating the number of rings which an atom is a member of. Additionally, size information about the ring(s) which the atom is in could be encoded as well.

4.4.2 Intramolecular addition

With the implementation of the new ring perception algorithms and the ring attribute based on membership in the set of RC, the next step was to apply it to reaction families. In this case, the reactions of interest were the intramolecular cycloaddition families, `Intra_R_Add_Endocyclic` and `Intra_R_Add_Exocyclic`, shown in Figure 4-12. These two families differ by which side of the double bond the radical attaches to, which is very important for linear compounds. In cases where the double bond is on a ring, the template can match in either direction around the ring, leading to duplicate reactions between the two families.

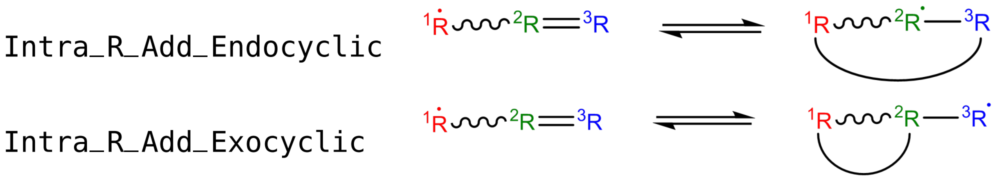


Figure 4-12: Reaction templates for `Intra_R_Add_Endocyclic` (top) and `Intra_R_Add_Exocyclic` (bottom).

Initially, the primary focus was to reduce the number of duplicate reactions generated for cyclic species by manually restructuring the group trees using the ring attribute. To do so, more groups were designed specifically for differentiating linear and cyclic structures. A portion of the original backbone tree for `Intra_R_Add_Endocyclic` is shown in Figure 4-13. This tree describes the main templates which are used to generate the reactions. The first level of the tree defines the distance between the radical site and the multiple bond, and deeper levels define the types of bond along that chain. For example, `R4_S_D` is a template where the radical atom is two single bonds away from the double bond which it will attach to.

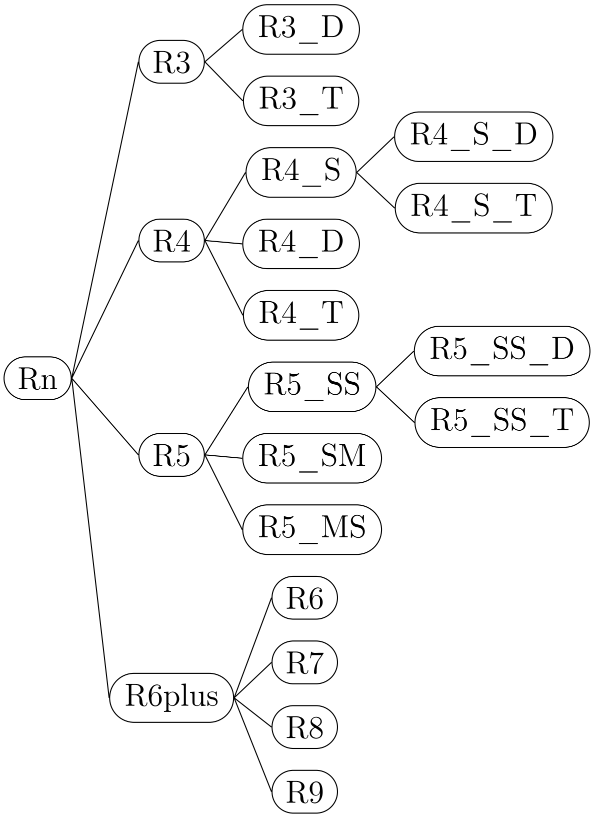


Figure 4-13: Selected portion of the original group tree for `Intra_R_Add_Endocyclic`.

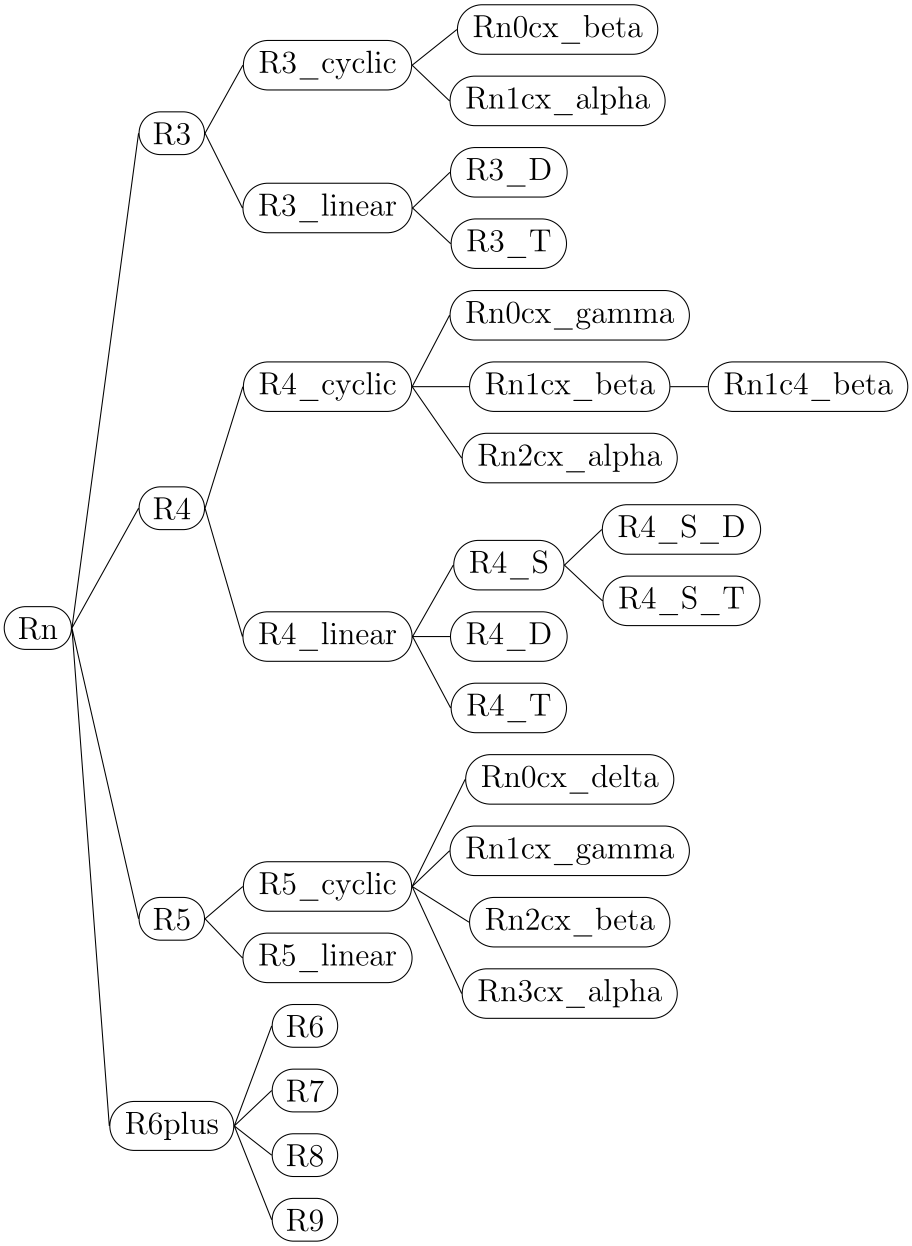


Figure 4-14: Selected portion of the new group tree for `Intra_R_Add_Endocyclic`.

First, to ensure that `Intra_R_Add_Endocyclic` and `Intra_R_Add_Exocyclic` would be mutually exclusive when generating reactions for cyclic species, forbidden structures were added to `Intra_R_Add_Exocyclic` in order to force all species with the multiple bond on a ring to react in `Intra_R_Add_Endocyclic`. The ring attribute greatly simplified this task, requiring only three forbidden structures (as a result of different labeling options) to cover all possible cyclics. An early attempt without the ring attribute required manual enumeration of all possible ring sizes and labeling possibilities.

In `Intra_R_Add_Endocyclic`, more groups were added for cyclic structures o improve differentiation between rates for linear and cyclic species. For each backbone length, the previous linear groups were placed under a new node called `Rn_linear`, where n corresponds to the backbone length. A sibling node was then created called `Rn_cyclic` containing templates for cyclic cases. One of the challenges in creating these templates was that none of the cyclic templates could be subgraph isomorphic to each other since they involved different ring sizes or labeling. Thus, the `Rn_cyclic` group itself had to be a pseudo-template consisting of the union of all of its children (in RMG syntax, an `OR` group). To briefly explain the naming syntax for one of these cyclic groups, `Rn2c4_beta` is has a four membered ring (c4) with the radical site two atoms away (n2) attacking the second atom in the ring (beta, counting from the ipso site). Groups were created for ring sizes from 3 to 8, with side chain lengths from 0 (radical on the ring) to 5.

Leave-one-out cross-validation was performed on the `Intra_R_Add_Endocyclic` family to compare performance before and after restructuring. In short, an estimate for each training rate was obtained by removing the training point from the tree, re-averaging the tree, and estimating the rate for the training reaction. Figure 4-15 shows the performance before and after this restructuring, with the same set of training data. We see that prediction accuracy was actually slightly reduced by the restructuring, with the mean absolute error (MAE) in $\log_{10}(k)$ increasing from 0.992 to 1.056. A possible explanation for this is that almost all of the training reactions at this point were for linear species. Therefore, the original tree, which was designed specifically for linear species, performed slightly better.

Following the restructuring, efforts were put into adding more training data to the `Intra_R_Add_Endocyclic` family, with a focus on reactions relevant to polycyclic aromatic

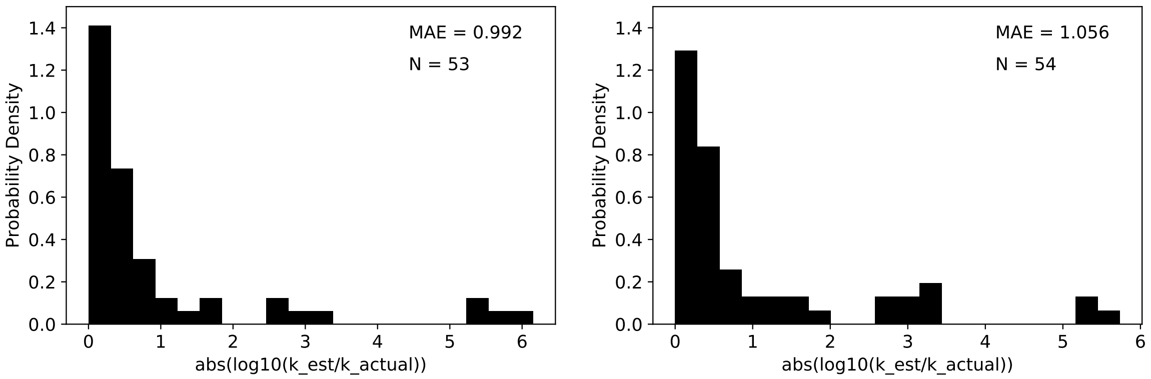


Figure 4-15: Leave-one-out performance for `Intra_R_Add_Endocyclic` before (left) and after (right) tree restructuring using ring attribute.

hydrocarbon formation. The left plot in Figure 4-16 shows the prediction accuracy for the newly added training reactions, *i.e.*, how well the old tree estimated rates of the new reactions. The right plot shows the leave-one-out performance after adding the new reactions into the training set. With the additional training, the MAE is reduced from 2.596 to 2.090. There is a clear improvement in the prediction accuracy when the new cyclic reactions are used to train the tree.

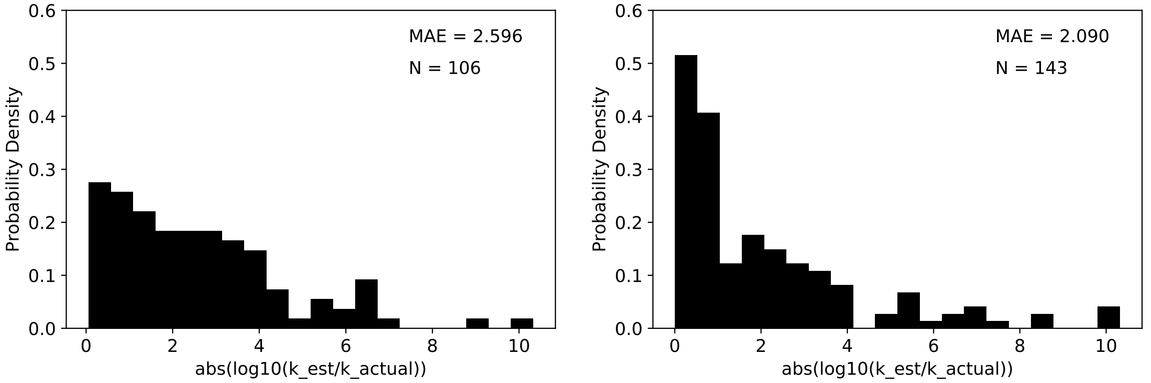


Figure 4-16: Prediction accuracy for new cyclic training reactions before training (left) and leave-one-out performance after training (right) in `Intra_R_Add_Endocyclic`.

However, the overall error in predictions is substantially worse when compared to the performance for linear reactions as shown in Figure 4-15. This suggests that the *a priori* design of the tree based on molecular structure considerations was not well-suited for the actual data for cyclic species. Another contributing factor is that the spread in the distribution of rates was greatly increased with the addition of new training reactions, which covered a wider chemical space. The distributions of the old and new training reactions are shown in Figure 4-17, where we can see that the new training reactions span many more orders of magnitude. Thus, training an estimator to cover the expanded data set is inherently more challenging.

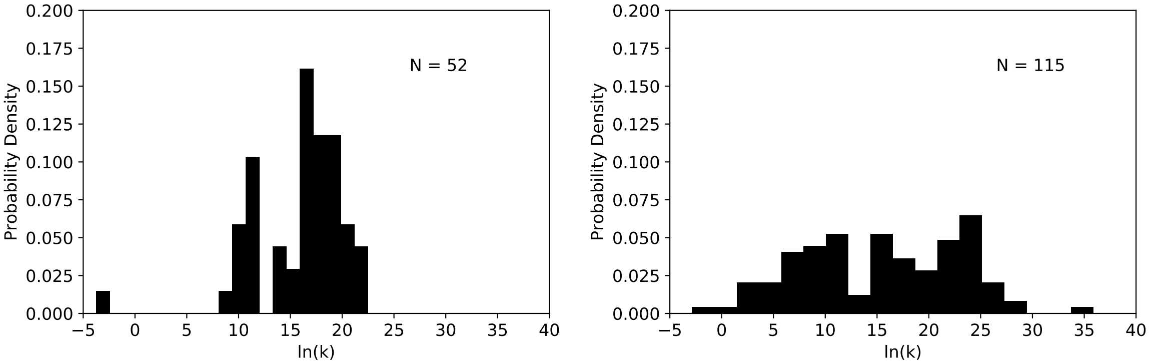


Figure 4-17: Distribution of reaction rates for original set of training reactions (left) and newly added training reactions (right).

One comparison which is not shown here is the performance of the original tree in estimating the rates of all of the new training reactions. The primary reason is that the old tree was not able to estimate a large portion of the new training reactions due to the group definitions being too specific and not including the species involved in the new templates. Thus, without modifications, the reactions which could be estimated by the old tree would be more chemically similar to the original training set. On the other hand, if the tree were to be modified to enable estimation of the new training reactions, the required modifications would be substantial enough for the comparison to be a poor representation of the original tree.

A new feature which has been added to RMG is the capability of automatically generating kinetics trees. The basic idea to create a decision tree based on a pool of training data in the form of reactions and their associated rates. Starting with a generic reaction template, the algorithm can gradually increase the specificity of the template by creating extensions such as adding an atom or a bond. It then chooses which extension to use at each level of the tree based on how well it splits the set of training data. The algorithm then fits a Blowers-Masel interpolant to the reactions at each node in the tree, which incorporates the enthalpy of reaction as another parameter for estimating the rate coefficient.

In some ways, this approach can be seen as being in between manually generated kinetics trees and non-transparent machine learning approaches like neural networks. The types of allowable extensions are predefined and based on chemical structure changes which are very intuitive to understand, but the actual process of building the tree relies on unbiased, numerical algorithms which can optimally design the tree for the available training data.

Using this feature, the `Intra_R_Add_Endocyclic` tree was completely regenerated, using all of the training reactions. Figure 4-18 shows the leave-one-out cross-validation performance of the automatically generated tree, with and without use of the ring attribute. An important note here is that this validation method is slightly different from that used for the old-style trees. With the old-style trees, the rate rule generated from a given training reaction is first removed, the tree is re-averaged, and then an estimate is generated for the reaction. With auto-generated trees, the Blowers-Masel fitting replaces tree averaging. Thus, for the leave-one-out test, only the Blowers-Masel function at the matched node is re-fitted after removing the rule corresponding to the training reaction.

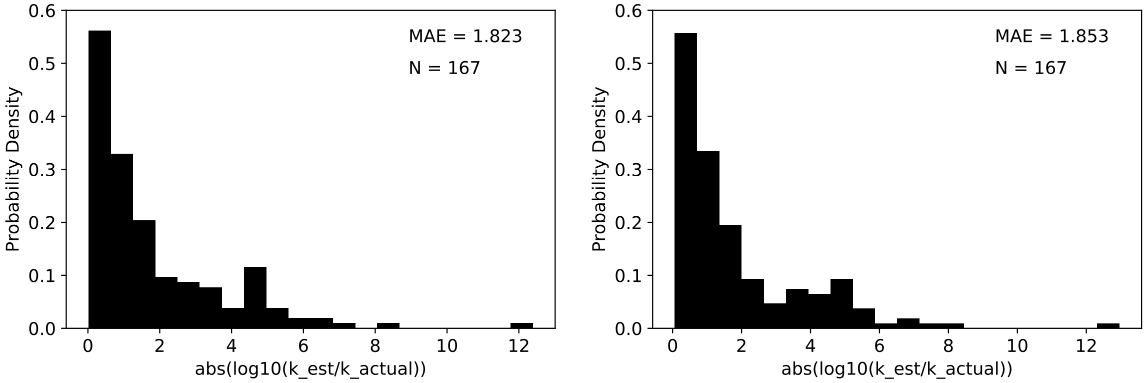


Figure 4-18: Leave-one-out performance for `Intra_R_Add_Endocyclic` with new cyclic training reactions and new automatically generated tree with ring attribute (left) and without ring attribute (right).

We see that the MAE with the ring attribute is 1.823, and the MAE without is only slightly higher at 1.853. Compared to Figure 4-16, this demonstrates that the automatic tree generation algorithm is more effective than manual tree generation. However, the MAE is still much larger than the base case with fewer training reactions in Figure 4-15. When using the ring attribute, the first split which the algorithm identified was specifying the ring attribute. This implies that after evaluating all of the potential ways to extend the base reaction template, the ring attribute performed the best in reducing the variance in the resulting child nodes. This confirms the intuition that the presence of an existing ring has a substantial effect on the rate of ring closure. When the algorithm was not allowed to use the ring attribute, the first split was to add a new atom to the template.

4.5 Conclusions

This work highlights some challenges and solutions to modeling polycyclic species in RMG. The primary contributor to these challenges in both thermochemistry and kinetics estimation is the lack of knowledge about ring strain and 3D geometry in general. One fundamental task necessary to working with cyclic species is being able to properly identify the rings in the molecule. While there are many possibilities for defining cycle sets, two of the most common are the SSSR and RC. New algorithms were implemented for both by interfacing RMG with the RingDecomposerLib, which enumerate both cycle sets faster and with better scaling than the previous SSSR method in RMG.

For thermochemistry estimation, both group additivity with the polycyclic heuristic and new machine learning estimator were shown to perform poorly for highly strained polycyclic species which are substantially different from molecules used in the training set. In general, both methods tend to under-predict enthalpy, which results in the over-representation of such species in RMG molecules. As a solution, new polycyclic group additivity values were fitted to calculated thermochemistry data for these species, at the CBS-QB3 level of theory. These calculations can also be incorporated into the training set for the machine learning estimator to improve future predictions.

For kinetics estimation, a new ring membership attribute was implemented in order to encode global structural information at an atom level. Using the new attribute allowed the `Intra_R_Add_Endocyclic` family tree to be redesigned to reduce generation of duplicate reactions and improve differentiation of linear and cyclic groups. However, leave-one-out cross-validation showed that the new tree performed slightly worse than the previous tree for the original set of mostly linear training reactions, indicating the limitations of manually designing a tree based on chemical intuition. Addition of more training reactions for cyclic molecules further reduced the performance of the family, most likely due to the wider chemical space being considered. Automatic tree generation out-performed the manually designed tree, providing the best accuracy for the full set of training reactions. Additionally, the first split identified by the algorithm was defining the ring membership attribute, confirming its importance in determining the rate.

References

(1) Liu, M.; Green, W. H. *Proc. Combust. Inst.* **2019**, *37*, 575–581.

(2) Grinberg Dana, A.; Liu, M.; Green, W. H. *Int. J. Chem. Kinet.* **2019**, *51*, 760–776.

(3) Sloane, N. J. A. Number of trees with n unlabeled nodes., `https://oeis.org/A000055`.

(4) Sloane, N. J. A. Number of connected graphs with n nodes., `https://oeis.org/A001349`.

(5) RDKit: Open-source cheminformatics., 2018.

(6) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. *J. Cheminform.* **2011**, *3*, 33.

(7) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. *J. Cheminform.* **2017**, *9*, 1–19.

(8) Fan, B. T.; Panaye, A.; Doucet, J. P.; Barbu, A. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 657–662.

(9) Plotkin, M. *J. Chem. Doc.* **1971**, *11*, 60–63.

(10) Figueras, J. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 986–991.

(11) Berger, F.; Flamm, C.; Gleiss, P. M.; Leydold, J.; Stadler, P. F. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 323–331.

(12) Vismara, P. *Electron. J. Comb.* **1997**, *4*, 1–15.

(13) Flachsenberg, F.; Andresen, N.; Rarey, M. *J. Chem. Inf. Model.* **2017**, *57*, 122–126.

(14) Kolodzik, A.; Urbaczek, S.; Rarey, M. *J. Chem. Inf. Model.* **2012**, *52*, 2013–2021.

(15) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. *Int. J. Chem. Kinet.* **2018**, *50*, 294–303.

(16) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

Chapter 5

Reaction Mechanism Generator v3.0: Advances in automatic mechanism generation

5.1 Introduction

Detailed chemical kinetic modeling is continuing to gain interest as an approach to study reactive chemical systems, ranging in application from combustion and pyrolysis of fuels to degradation of active pharmaceutical ingredients. This growth can be attributed to a combination of demand for studying increasingly complex chemistries and supply of computational power and quantum chemistry capabilities. By taking advantage of these computational resources, automatic mechanism generation tools are able to systematically enumerate and evaluate potential chemical pathways, reducing the chance of human error or bias. This is largely a data-driven task, requiring good estimation algorithms for thermochemical and rate parameters, which in turn rely on accurate training data from experiments or quantum chemistry calculations.

The Reaction Mechanism Generator (RMG) software has been in development for over a decade, with the current Python version (RMG-Py) having begun development in 2008. RMG-Py v1.0 was previous described in 2016 [1], and development has continued rapidly since.

Here, we are excited to present RMG-Py v3.0, which brings many new features including Python 3 compatibility, heterogeneous catalysis modeling, and new parameter estimation algorithms. With these and other improvements, the codebase has doubled to over 120,000 lines of Python code. Many new developments have been focused on improving heteroatom chemistry, including nitrogen and sulfur, and aromatic chemistry, to study formation of soot and coke. RMG has recently been used successfully to model ethylamine pyrolysis,[2] di-tert-butyl sulfide pyrolysis,[3] hexylbenzene pyrolysis,[4] effect of substituted phenols on ignition delay,[5] and PAH formation in methane oxidation.[6]

The structure and concept behind RMG has been described previously,[1] so only a brief overview will be given here. RMG is a tool for automatically constructing detailed chemical mechanisms which is largely comprised of three components:

1. a cheminformatics framework for representing molecules, reactions, and various data classes for thermochemistry and kinetics

2. a database and parameter estimation framework for predicting thermochemistry and kinetics parameters

3. a mechanism construction framework, primarily using a flux-based species selection algorithm, including functionality for automatic construction of pressure-dependent networks.

RMG uses a core/edge reaction model, where the core contains species and reactions which have already been identified as being important and the edge contains species and reactions which are under consideration. In each iteration, RMG will identify one or more species to move from the edge to the core based on a homogeneous batch reactor simulation, and then generate new reactions between the new species and other species in the core. The model is considered converged when no edge species exceed the tolerance for selection. The latest release of RMG includes updates across all three components to expand modeling capabilities and improve accuracy, robustness, and performance.

5.2 New features

5.2.1 Python 3 compatibility

Until this release, RMG-Py has always been only Python 2 compatible. Python 3 was first released at the end of 2008, around the same time as RMG-Py development started. Despite that, Python 2 remained more widely-used until the last few years.[7] The official end-of-life for Python 2 is January 1, 2020, which strongly motivated the transition to Python 3. In addition, many software packages have ended Python 2 support in their latest releases, including RMG dependencies such as RDKit[8] and Cantera[9]. As such, we chose to also transition to Python 3 with this release of RMG-Py.

The transition for RMG-Py included many steps. The first step was ensuring that Python 3 versions of all of our dependencies were available. This was straightforward for widely-used packages since all of them already supported Python 3. However, some packages developed by our group also had to be updated with Python 3 support, namely PyDAS and PyDQED.[10, 11] The second step was making the necessary changes in RMG-Py to enable Python 3 compatibility. This was relatively straightforward with the aid of packages to automate this transition, like python-future. In the final step, we took this opportunity to standardize function names throughout our API to comply with PEP-8 recommendations, effectively the official Python style guide. In total, transition tasks took approximately 500 developer hours to complete.

With the v3.0 release, RMG-Py is now fully compatible with Python 3.7. The Python 2 version of RMG-Py will no longer be actively supported, although a legacy version will be made available for users. Of course, installation of the newest Python 3 version is recommended.

5.2.2 Heterogeneous catalysis

RMG v3.0 also introduces support for generating heterogeneous catalysis models, which was previously developed independently as the RMG-cat project.[12] This feature involved additions to all aspects of the model generation process.

Molecule representations have been extended to include catalyst sites, which are represented as a generic "X" element. New bond types have been implemented to represent the metal-adsorbate bond, including Van der Waals bonds (internally represented with a bond order of 0) and quadruple bonds (*e.g.*, for adsorption of a carbon atom). These extend the existing single, double, triple, and benzene bond orders.

Thermochemistry estimation has been expanded to estimate parameters for surface species by applying adsorption corrections. For a given surface species, the metal first removed to obtain an estimate for the gas-phase species using existing methods (*e.g.*, group additivity or libraries), then an adsorption correction is determined from a group additivity tree and added to the gas-phase value. Thermo libraries are also supported for surface species. By default, RMG uses binding energies for Pt(111), but energies for an arbitrary catalyst can be specified in the input file (Figure 5-1). Adsorption corrections are then scaled appropriately based on the specified binding energies.

```
catalystProperties(
    bindingEnergies={
        'H': (-2.479, 'eV/molecule'),
        'O': (-3.586, 'eV/molecule'),
        'C': (-6.750, 'eV/molecule'),
        'N': (-4.352, 'eV/molecule'),
    },
    surfaceSiteDensity=(2.72e-9, 'mol/cm^2'),
)
```

Figure 5-1: Example input file block for specifying catalyst properties.

Kinetics estimation has been expanded with new families (detailed in Section 5.4.2) for estimating various types of surface reactions, such as adsorption and dissociation. To support these surface reactions, new data classes have also been added for surface rate constants (`SurfaceArrhenius` and `SurfaceArrheniusBEP` for Bronsted-Evans-Polanyi relationships) and sticking coefficients (`StickingCoefficient` and `StickingCoefficientBEP`).

Surface simulations require use of the new `SurfaceReactor` which has been added. This module performs the reactor simulations necessary for the flux-based algorithm for model growth. It is modeled as a zero-dimensional, isothermal, isochoric batch reactor which tracks surface coverage in addition to gas-phase mole fractions. User specification of surface area to volume ratio and surface site density are required.

For surface jobs, RMG will output separate gas- and surface-phase Chemkin mechanism files along with a single Cantera mechanism file.

5.2.3 Uncertainty analysis

Beyond generating chemical mechanisms, RMG also provides some features for model analysis. Previously, local first-order sensitivity analysis was available to calculate sensitivities of species concentrations to thermochemistry and rate constants. New methods for both local and global uncertainty analysis have been implemented in RMG.[13] Local uncertainty analysis builds on those first-order sensitivity by incorporating estimated uncertainties for thermochemical and rate parameters to obtain uncertainties for species concentrations. Global uncertainty analysis uses the MIT Uncertainty Quantification Library (MUQ 2)[14] to construct polynomial chaos expansions (PCEs) based on reactor simulations at random points within the uncertainty space of the input thermochemical and rate parameters. Reactor simulations are performed using Cantera.[9] A key feature of the RMG uncertainty module is the ability to track correlated uncertainties in model input parameters, such as correlations arising from group additivity estimates for thermochemistry and rate rule estimates for rate coefficients. This can have significant effects on uncertainty propagation and the resulting uncertainties on output parameters.

Uncertainty analysis can be requested via the RMG input file, which will lead to it being performed upon completion of model generation. Using uncertainty analysis does require that sensitivity analysis settings also be provided, since sensitivity analysis is required part of local uncertainty analysis. Local uncertainty analysis is also used to determine the parameters to vary for global uncertainty analysis, in order to minimize computational cost. For global analysis, PCE fitting can be controlled by specifying either a maximum runtime, error tolerance, or maximum number of model evaluations. Jupyter notebooks are also available for interactive use of these features. These new tools can provide insights beyond first-order sensitivity analysis to aid in the model development process.

```
uncertainty(
    localAnalysis=True,
    globalAnalysis=True,
    uncorrelated=True,
    correlated=True,
    localNumber=10,
    globalNumber=5,
    pceRunTime=1800,
    pceErrorTol=None,
    pceMaxEvals=None,
)
```

Figure 5-2: Example input file block for requesting uncertainty analysis.

5.2.4 Ranged reactors

In RMG, the reaction conditions of interest (*i.e.*, temperature (T), pressure (P), composition (X)) are provided by defining reactors in the input file. RMG supports three reactor types for mechanism generation, distinguished by the phases involved: `SimpleReactor` for gas phase, `LiquidReactor` for liquid phase, and the new `SurfaceReactor`.

Ranged reactors are a new feature in RMG v3.0 to simplify the task of specifying a range of initial conditions. Because RMG uses a flux-based algorithm for identifying important species and reactions, the reactor conditions used to generate a model directly affect the conditions at which the model is applicable. Previously, the recommended approach for building a model applicable at a range of conditions was to define multiple reactors spanning the space of conditions of interest. For example, if the goal was to develop a model valid for temperatures from 1000 K to 200K and pressures from 1 bar to 10 bar, the user may need to define 10 reactors with all combinations of $T = \{1000, 1200, 1400, 1600, 1800, 2000\}$ K and $P = \{1, 10\}$ bar. This can be annoying to the user, and risks missing important chemistry which may occur in between the chosen points.

With the new feature, ranges for TPX can be directly specified for a single reactor block. Internally, RMG will automatically select points within the space of conditions for each iteration, using a weighted stochastic grid sampling algorithm. On each iteration, a coarse grid with 20 points in each dimension is constructed, and the desirability of each point is evaluated based on the number of iterations since it was last chosen. The desirability values are normalized to form probabilities, and a random point is chosen using those probabilities. The algorithm then takes a random step from the chosen point, with a maximum distance of $\sqrt{2}/2$ times the distance between grid points. A simplified example of the algorithm for two dimensions is shown in Figure 5-3.

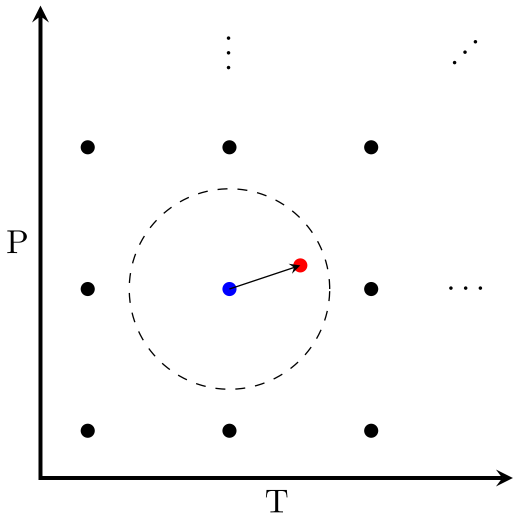


Figure 5-3: Schematic representation of how RMG selects conditions for a given simulation when using ranged reactors. The blue point indicates the initial point chosen from the coarse grid. A random step is then taken within the bounds of the dotted line, which results in the final set of conditions represented by the red point.

5.2.5 Isotopic mechanisms

RMG can now generate isotopically labeled reaction mechanisms via a post-processing algorithm.[15] After a normal RMG job is completed, the isotopes module can generate all combinations of isotopically labeled species and reactions. Importantly, RMG modifies species' entropy based on changes to molecular symmetry and modifies kinetic Arrhenius factors based on reaction path degeneracy and basic kinetic isotope effects (KIE). Given the information available to RMG and the requirement for automatic calculation, only classical, mass-dependent KIE has been implemented.

One challenge with this approach is that the mechanism size increases substantially with the inclusion of all isotopologues. However, it is still very useful for generating detailed isotopic

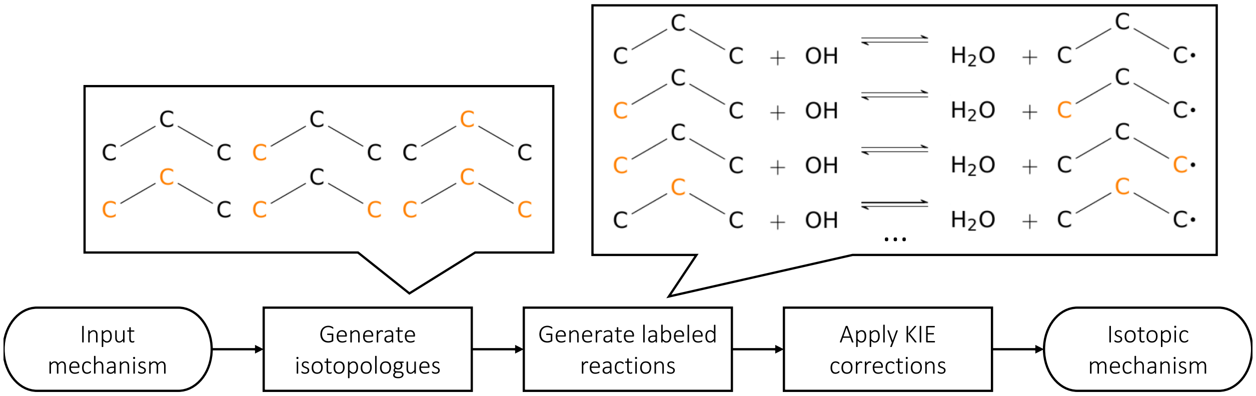


Figure 5-4: Algorithm for constructing isotopic reaction mechanisms.

mechanisms, and has been shown to provide good agreement and insight into position-specific isotope analysis experiments.[15] Currently, the algorithm is limited to generation of isotopic mechanisms for ^{13}C, though the framework is easily extensible to other isotopes.

5.3 Molecular representation

5.3.1 Atom types

RMG uses atom types as a way to describe the local environment around an atom when defining group structures. They enable definition of highly specific groups which can accelerate graph isomorphism or more generic groups which improve flexibility. The set of available atom types has been revised and expanded to improve representation of heteroatoms. Particular focus has been placed on expanding atom type descriptors for the various bonding configurations of nitrogen and sulfur, for which the full list of updated atom types has been recently reported.[16] New carbon and oxygen atom types for representing formal charges and varying numbers of lone pairs have been added, along with additional halogen atom types. For surface chemistry, atom types representing generic surface sites have been added, along with quadruple bonds for carbon and silicon. A list of these new atom types is shown in Table 5.1.

Table 5.1: New atom types available in RMG v3.0.

Atom Type	Description
New carbon atom types	
Ca	Carbon atom with two lone pairs
Csc	Carbon with all single bonds and formal charge of +1
Cdc	Carbon with one double bond and formal charge of +1
Cq	Carbon with quadruple bond (for surface adsorption)
C2s	Carbon with one lone pair and single bonds
C2sc	Carbon with one lone pair, single bonds, and formal charge of -1
C2d	Carbon with one lone pair and one double bond
C2dc	Carbon with one lone pair, one double bond, and formal charge of -1
C2tc	Carbon with one lone pair, one triple bond, and formal charge of -1
New oxygen atom types	
O0sc	Oxygen with three lone pairs, single bonds, and formal charge of -1
O2s	Oxygen with two lone pairs and single bonds
O2sc	Oxygen with two lone pairs, single bonds, and formal charge of +1
O2d	Oxygen with two lone pairs and one double bond
O4sc	Oxygen with one lone pair, single bonds, and formal charge of +1
O4dc	Oxygen with one lone pair, one double bond, and formal charge of +1
O4tc	Oxygen with one lone pair, one triple bond, and formal charge of +1
O4b	Oxygen with one lone pair and two benzene bonds
New halogen atom types	
F	Fluorine with any local bonding structure
F1s	Fluorine with three lone pairs and one single bond
Cl	Chlorine with any local bonding structure
Cl1s	Chlorine with three lone pairs and one single bond
I	Iodine with any local bonding structure
I1s	Iodine with three lone pairs and one single bond
New silicon atom types	
Siq	Silicon with quadruple bond (for surface adsorption)
New surface site types	
X	Generic surface site
Xv	Vacant surface site
Xo	Occupied surface site

5.3.2 Resonance structures

Resonance structures are an important aspect of molecule representation in RMG. Given that RMG uses localized representations of molecules (*i.e.*, Lewis structures), it is important that the algorithm can generate and identify the structures which are most representative of the true behavior of a molecule. Thus, significant improvements have been made to resonance structure generation algorithms, in particular for aromatic species and heteroatoms.[16, 17] For aromatic species, RMG can now generate Clar structures[18, 19] in replacement of Kekulé structures which are now considered unrepresentative. For heteroatom molecules with lone pairs, more delocalization pathways are now recognized by RMG. To address the increase in computational requirements for handling additional resonance pathways and structures, a heuristic-based filtration algorithm will identify representative resonance structures on-the-fly.

5.4 Parameter estimation

Parameter estimation is possibly the most important step in mechanism generation, especially for flux-based algorithms like the one used in RMG. Because the criteria for selecting species to add into the model depends on the calculated reaction flux to those species, thermochemistry and rate constant predictions must not only be accurate for important species, but they must be reasonably correct for unimportant species, so that they can be properly neglected.

5.4.1 Thermochemistry

For thermochemistry estimation, RMG relies primarily on group additivity, where the thermochemistry for a molecule is derived from the sum of contributions from each heavy atom.[20, 21] However, a major limitation of base group additivity is that only local features are captured and longer range effects such as steric interactions and ring strain must be treated separately.

For polycyclic species, ring strain can substantially affect the thermochemistry of a species. A previous limitation of the group additivity algorithm was that ring strain corrections would only be applied if there was an exact match to the molecule. To address this, a new estimation algorithm was developed to provide an estimate for the ring strain of any molecule based on a heuristic algorithm which decomposes the molecule into mono- and bicyclic substructures.[22]

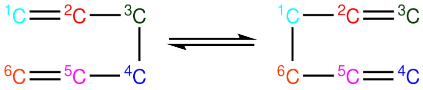
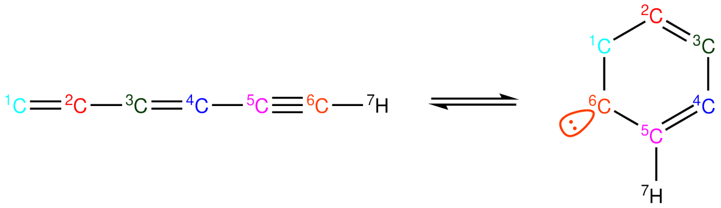
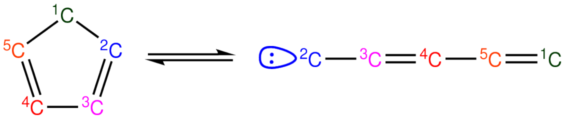
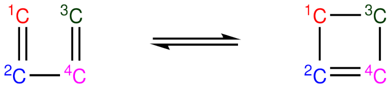
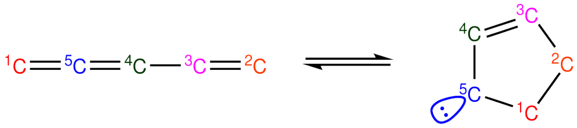
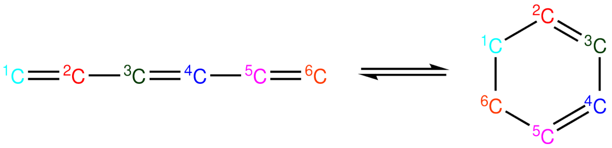
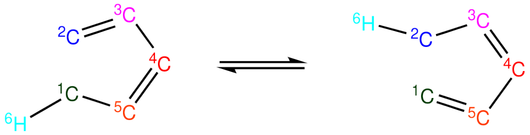
RMG v3.0 also includes an updated neural network based thermochemistry estimator, developed using the *chemprop* package for molecular property prediction.[23] This is a new version of the previously reported thermochemistry estimator which was first introduced in RMG v2.3.[24] Many molecular property prediction models are based on DFT data, including the previous version of the RMG thermochemistry estimator,[24] because they are readily available in large databases or can be calculated with low computational cost. However, RMG strongly benefits from more accurate predictions. Therefore, the new thermochemistry estimator was designed using a transfer learning approach that is able to learn accurate models from small high-quality data sets composed of experimental and coupled cluster calculations. [25] As described in the *chemprop* publication, the deep learning models use a message passing neural network (MPNN) to encode molecular graphs into fixed-length feature vectors which are passed through additional fully-connected neural network layers to make the thermochemistry predictions. Instead of using the featurization for atoms and bonds implemented by *chemprop*, we removed features that depend on resonance structure and added ring membership features, which we have shown to be beneficial. [24, 25] Two separate models were trained, one to predict enthalpies of formation and one to predict entropy and heat capacities simultaneously.

5.4.2 Kinetics

Kinetics families

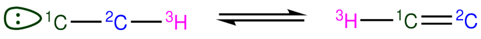
New kinetics families have been implemented in RMG to allow automatic enumeration of new reaction pathways. All of the new families which have been added since RMG v1.0 are shown in Table 5.2. These new kinetics families include reactions involved in the propargyl recombination pathway to benzene formation,[6] peroxide reactions relevant in liquid phase oxidation chemistry, surface reaction types for heterogeneous catalysis simulations,[12] and a few other reactions types which have been found to be important for various systems.

Table 5.2: New kinetics families available in RMG v3.0.

Propargyl recombination reaction families
`6_membered_central_C-C_shift`

`Concerted_Intra_Diels_alder_monocyclic_1,2_shiftH`

`Cyclopentadiene_scission`

`Intra_2+2_cycloaddition_Cd`

`Intra_5_membered_conjugated_C=C_C=C_addition`

`Intra_Diels_alder_monocyclic`

`Intra_ene_reaction` (Previously `H_shift_cyclopentadiene`)


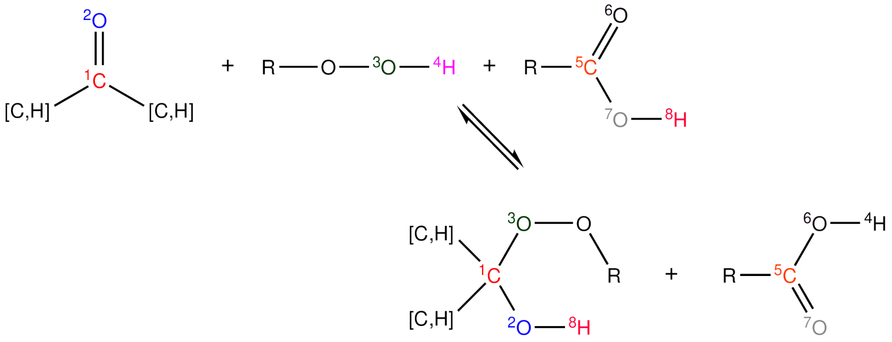
(Continued)

Singlet_Carbene_Intra_Disproportionation

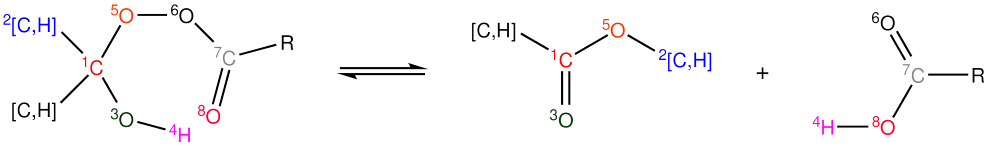


Liquid phase peroxide oxidation reaction families

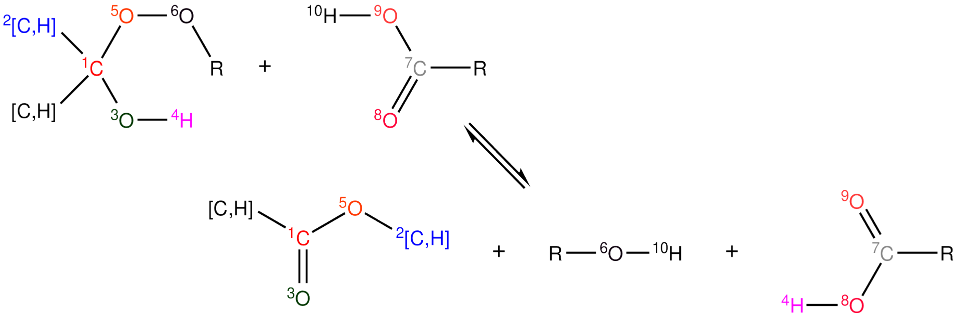
Baeyer-Villiger_step1_cat



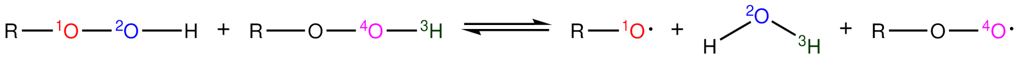
Baeyer-Villiger_step2



Baeyer-Villiger_step2_cat

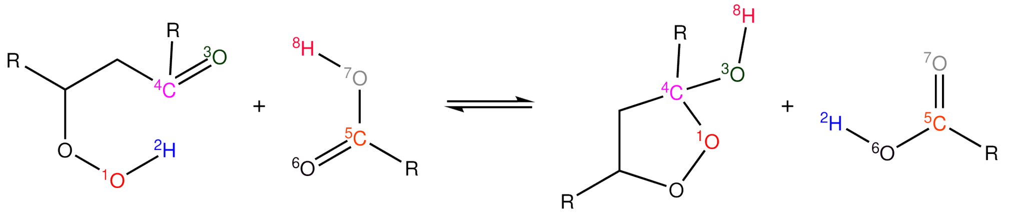


Bimolec_Hydroperoxide_Decomposition

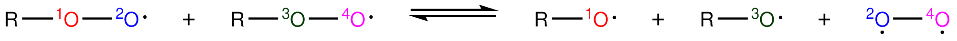


(Continued)

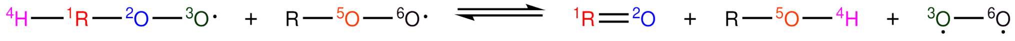
Korcek_step1_cat



Peroxyl_Disproportionation

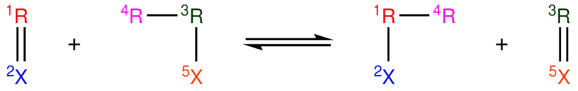


Peroxyl_Termination

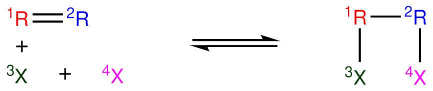


Surface reaction families

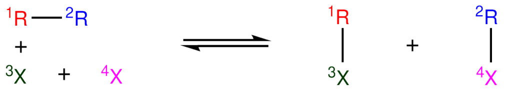
Surface_Abstraction



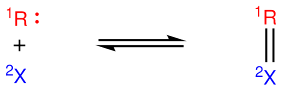
Surface_Adsorption_Bidentate



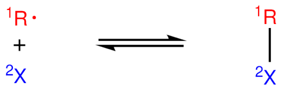
Surface_Adsorption_Dissociative



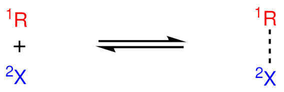
Surface_Adsorption_Double



Surface_Adsorption_Single

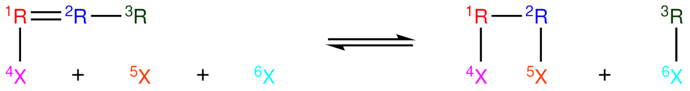


Surface_Adsorption_vdW

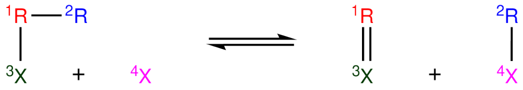


(Continued)

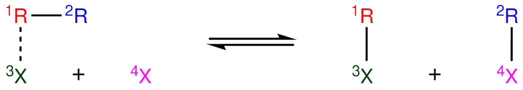
Surface_Bidentate_Dissociation



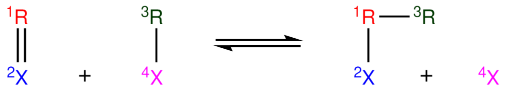
Surface_Dissociation



Surface_Dissociation_vdW

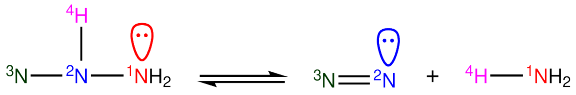


Surface_Recombination

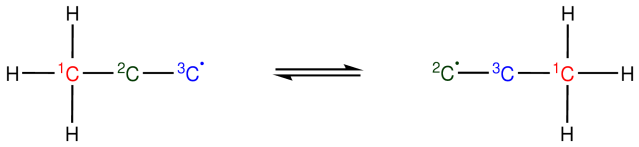


Other new reaction families

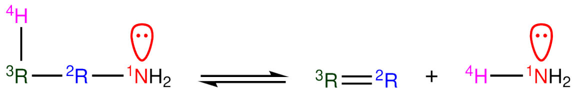
1,2_NH3_elimination



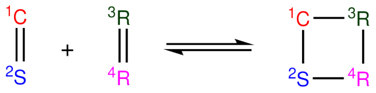
1,2_shiftC



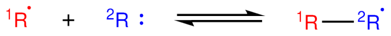
1,3_NH3_elimination



2+2_cycloaddition_CS

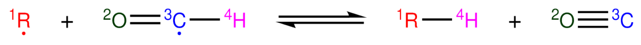


Birad_R_Recombination (Previously Oa_R_Recombination)

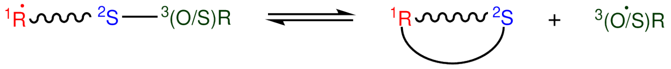


(Continued)

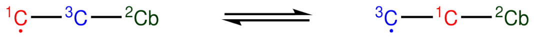
CO_Disproportionation



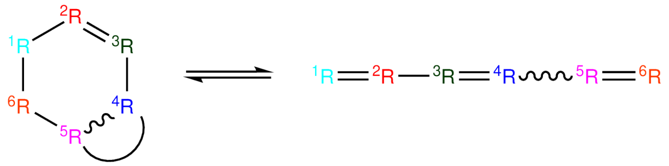
Cyclic_Thioether_Formation



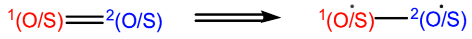
Intra_R_Add_Exo_scission



Intra_Retro_Diels_alder_bicyclic (Previously Intra_Diels_alder)



Singlet_Val6_to_triplet



Automated tree generation

One major challenge with the original kinetics family format was the need to manually maintain and design the tree structure for each family. When adding new training reactions, it is often necessary to extend the tree with new group structures in order to optimize the utilization of training reactions in generating new rate rules.

The solution which has been implemented in RMG v3.0 is the capability of automatically generating the tree. The new method uses machine learning approaches to automatically generate a decision tree based on the available training reactions. Starting with a generic reaction template, new groups are generated based on pre-defined types of extensions, *e.g.*, adding an atom, adding a bond, specifying an element, etc. An optimal extension is chosen at each level of the tree by determining information gain based on the reduction in reaction rate variance. More details of the algorithm will be described in a separate publication.

In the v3.0 release, the R_Recombination family has been updated with an automatically generated tree. Updates to other reaction families can be expected in upcoming releases.

5.5 Performance improvement

5.5.1 Parallelization

One general approach which can be used to improve the performance of a computer program is parallelization, which allows running simultaneously on multiple processes. In the case of RMG, parallelization is challenging to implement overall since the core algorithm is very serial in nature, *i.e.*, tasks must be performed in order because they rely on the results or prior tasks. However, there are certain portions of the algorithm which are more amenable to parallelization. RMG v2.0 implemented an initial approach for parallelizing reaction generation and thermochemistry estimation, but encountered numerous challenges which made it difficult to use. In RMG v3.0, parallelization has been completely revamped using the built-in `multiprocessing` module in Python, providing parallel processing support for reaction generation and quantum calculations for the QMTP (Quantum Mechanics for Thermochemical Properties).[26]

5.5.2 Molecule comparison

One task which often requires substantial computing time in RMG is molecule comparison, which is typically done to identify whether two molecules in RMG are or are not the same chemical species. Part of the challenge is because RMG uses localized resonance structures to represent molecules, so simply comparing two structures may not be sufficient to determine whether or not they are the same. Instead, all of the resonance structures must be compared. Therefore, the original approach to comparing molecules was to generate all resonance structures for the two species and comparing them to each other using graph isomorphism. To confirm that two molecules are the same, the comparison can quit as soon as a matching pair of resonance structures is found. However, to confirm that two molecules are different, all combinations of resonance structures must be checked.

The previous approach was very time consuming, especially when considering resonance structure generation. A timing comparison of various methods for comparing molecules is shown in Figure 5-5. Two test cases are shown, Case A is for two molecules which are the same, while Case B is for two molecules which are slightly different. The first (blue) bar indicates the time required for generating resonance structures and then performing isomorphism comparison, which takes a substantial amount of time. The second (orange) bar shows the time for isomorphism if resonance structure generation is excluded from the timing. In Case B, we see that isomorphism itself takes longer because all pairs of resonance structures must be compared to confirm that the two molecules are different.

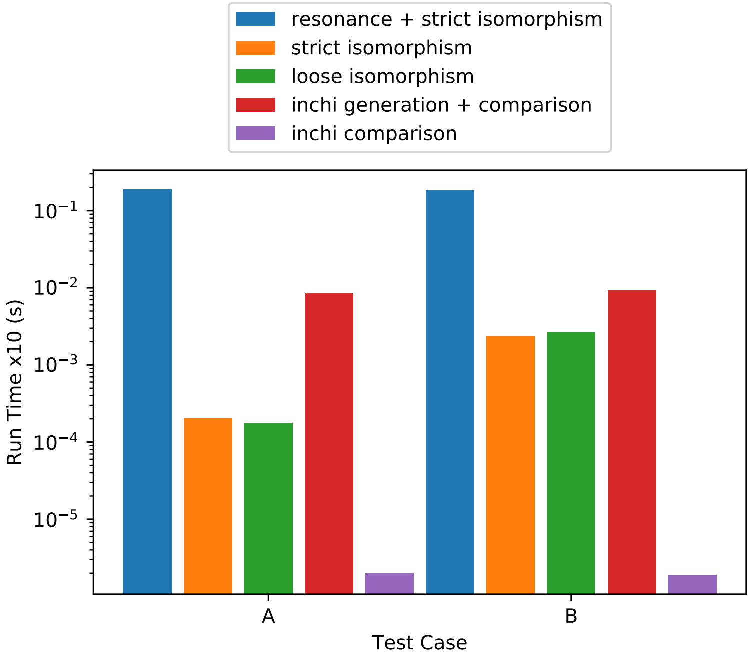


Figure 5-5: Performance comparison for various methods for comparing molecules.

A newly implemented approach to isomorphism, referred to here as “loose isomorphism,” relies on ignoring electron related features, such as radicals, lone pairs, and bond orders. The purpose here is to have an isomorphism approach which is independent of resonance structures and only focuses on the actual molecular structure. This eliminates the need to generate resonance structures. The timing for this method is shown by the third (green) bar in Figure 5-5. We see that this method is comparable in performance to normal isomorphism. However, there is a guaranteed performance improvement because resonance structure generation is avoided.

A third option which also has significant potential is InChI (International Chemical Identifier) comparison. An InChI is a string identifier for a molecule which is also designed to be independent of resonance structures and would therefore give the same result as the loose isomorphism method. Additionally, string comparison is extremely fast. Unfortunately, InChI generation time is non-trivial. The fourth (red) bar in Figure 5-5 shows the time required for generating InChI strings for two molecules and comparing them, and the fifth (purple) bar shows the time for the string comparison only. Though the string comparison is fast as expected, InChI generation makes the overall process take longer than graph isomorphism.

It is important to note that these timings are not completely representative of actual operation. Importantly, resonance structures and InChI strings can be stored, such that they only need to be generated once. Then subsequent comparisons would require much less time. However, a large portion of comparisons in RMG are with newly generated molecules, where the data would always need to be generated. As a result, the true cost of these comparisons would be in between the total time and just the comparison time.

In RMG v3.0, most molecule comparisons have been changed to use loose isomorphism because it can be considered an absolute improvement over resonance structure generation plus isomorphism. However, InChI comparison could be considered in the future if InChI generation speed is improved.

5.6 Development practices

With continued growth of the RMG development team and user-base, good software development practices have become increasingly important. In recent years, additional emphasis has been placed on implementing best-practices for open-source software development. All RMG source code is publicly available on GitHub.[27] Code review and continuous integration testing are emphasized as part of the development workflow, which has been formalized via official contributor guidelines.[28] Elements of git-flow [29] and semantic versioning [30] have also been implemented into the development workflow to improve version release planning.

5.7 Conclusions

RMG v3.0 is now available, and we recommend existing users to update their installations to take advantage of new features. Compared to RMG v1.0, there are many new features and substantial improvements across all aspects of the software. Python 3 support ensures that RMG is up to date with the latest scientific packages and will be for the foreseeable future. New chemistry features like surface mechanism generation and isotopic mechanism generation enable application of RMG to more systems than ever before. Uncertainty analysis provides new ways to analyze models to quantify the overall uncertainty in a model and identify the parameters which contribute most to that uncertainty. Fundamental improvements to molecular representation in the form of new atom types and resonance transformations work together to improve the the accuracy of the localized molecular representations. Parameter estimation, as the key to generating good models, has been improved via expansion of the database as well as addition of new algorithms like the neural network thermochemistry estimator. Finally, performance improvement is always an ongoing focus, and the recent implementation of parallelization and improved molecule isomorphism comparison are steps towards faster model generation. Of course, there are countless other developments which have not been mentioned here, but detailed release notes for all RMG releases can be found in the documentation.[31]

References

(1) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(2) Grinberg Dana, A.; Buesser, B. A.; Merchant, S. S.; Green, W. H. *Int. J. Chem. Kinet.* **2018**, *50*, 243–258.

(3) Class, C. A.; Liu, M.; Vandeputte, A. G.; Green, W. H. *Phys. Chem. Chem. Phys.* **2016**, *18*, 21651–21658.

(4) Lai, L.; Gudiyella, S.; Liu, M.; Green, W. H. *Energy & Fuels* **2018**, *32*, 5489–5500.

(5) Zhang, P.; Yee, N. W.; Filip, S. V.; Hetrick, C. E.; Yang, B.; Green, W. H. *Phys. Chem. Chem. Phys.* **2018**, *20*, 10637–10649.

(6) Chu, T.-C.; Buras, Z. J.; Oßwald, P.; Liu, M.; Goldman, M. J.; Green, W. H. *Phys. Chem. Chem. Phys.* **2019**, *21*, 813–832.

(7) JetBrains Python Developers Survey 2018., https://www.jetbrains.com/research/python-developers-survey-2018/ (accessed 10/24/2019).

(8) RDKit: Open-source cheminformatics., 2018.

(9) Goodwin, D. G.; Moffat, H. K.; Speth, R. L. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes., 2017.

(10) Allen, J. W.; Gao, C. W. PyDAS: A Python wrapper for the DASSL, DASPK, and DASKR differential algebraic system solvers., https://github.com/ReactionMechanismGenerato PyDAS.

(11) Allen, J. W. PyDQED: A Python wrapper for the DQED constrained nonlinear optimization code., https://github.com/ReactionMechanismGenerator/PyDQED.

(12) Goldsmith, C. F.; West, R. H. *J. Phys. Chem. C* **2017**, *121*, 9970–9981.

(13) Gao, C. W.; Liu, M.; Green, W. H. *Submitted.*

(14) Conrad, P. R.; Parno, M. D.; Davis, A. D.; Marzouk, Y. M. MIT Uncertainty Quantification Library (MUQ 2)., http://muq.mit.edu (accessed 10/28/2019).

(15) Goldman, M. J.; Vandewiele, N. M.; Ono, S.; Green, W. H. *Chem. Geol.* **2019**, *514*, 1–9.

(16) Grinberg Dana, A.; Liu, M.; Green, W. H. *Int. J. Chem. Kinet.* **2019**, *51*, 760–776.

(17) Liu, M.; Green, W. H. *Proc. Combust. Inst.* **2019**, *37*, 575–581.

(18) Clar, E.; Zander, M. *J. Chem. Soc.* **1958**, 1861–1864.

(19) Solà, M. *Front. Chem.* **2013**, *1*, 22.

(20) Benson, S. W.; Buss, J. H. *J. Chem. Phys.* **1958**, *29*, 546–572.

(21) Benson, S. W., *Thermochemical kinetics : methods for the estimation of thermochemical data and rate parameters.* New York : Wiley, c1976.: 1976.

(22) Han, K.; Jamal, A.; Grambow, C. A.; Buras, Z. J.; Green, W. H. *Int. J. Chem. Kinet.* **2018**, *50*, 294–303.

(23) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(24) Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. *J. Phys. Chem. A* **2019**, *123*, 2142–2152.

(25) Grambow, C. A.; Li, Y.-P.; Green, W. H. *J. Phys. Chem. A* **2019**, *123*, 5826–5835.

(26) Jocher, A.; Vandewiele, N. M.; Han, K.; Liu, M.; Gao, C. W.; Gillis, R. J.; Green, W. H. *Comput. Chem. Eng.* **2019**, 106578.

(27) RMG-Py GitHub., https://github.com/ReactionMechanismGenerator/RMG-Py (accessed 10/28/2019).

(28) Liu, M.; Han, K.; Goldman, M. J.; Payne, M. A. RMG Contributor Guidelines., https://github.com/ReactionMechanismGenerator/RMG-Py/wiki/RMG-Contributor-Guidelines (accessed 10/28/2019).

(29) Driessen, V. A successful Git branching model., `https://nvie.com/posts/a-successful-git-branching-model/` (accessed 10/28/2019).

(30) Preston-Werner, T. Semantic Versioning., `https://semver.org/` (accessed 10/28/2019).

(31) RMG Documentation: Release Notes., `http://reactionmechanismgenerator.github.io/RMG-Py/users/rmg/releaseNotes.html` (accessed 11/07/2019).

Chapter 6

Predicting polycyclic aromatic hydrocarbon formation with an automatically generated mechanism for acetylene pyrolysis

6.1 Introduction

Polycyclic aromatic hydrocarbon (PAH) formation occurs in a wide range of systems, from combustion and pyrolysis of conventional fuels[1] to carbon-rich circumstellar envelopes.[2, 3] PAHs are known to be harmful to human health [4] and are also precursors to soot, which lead to additional health and environmental hazards.[1, 5] There has been ongoing interest in understanding the chemical mechanisms involved in PAH formation for many decades, yet there is still much which is unclear.

Numerous PAH formation pathways have been proposed and studied to date. Broadly speaking, possible routes to forming the first aromatic ring (*i.e.*, benzene) include $C_2 + C_4$ pathways,[6–8], $C_3 + C_3$ pathways,[9, 10] and $C_5 + C$ pathways.[11–13] Even more pathways have been studied for formation of the second aromatic ring (*i.e.*, indene and naphthalene), most of which were summarized thoroughly by Mebel *et al.*[14] Many pathways to forming the second ring are also extensible to formation of larger PAHs. Most notable is the well-known hydrogen abstraction, C_2H_2 addition (HACA) pathway, which was proposed by Frenklach *et al.*,[6] and a closely related route proposed by Bittner and Howard.[15] HACA pathways have been studied starting from phenyl,[16] naphthyl,[17–19] biphenylyl,[20] and phenanthryl.[21] Vinylacetylene addition also provides a route to adding another aromatic ring, which has been calculated for phenyl [22] and naphthyl.[23] Finally, addition of phenyl [24, 25] and benzyne [25, 26] have also been studied as a direct pathway to an additional aromatic ring.

The initiation mechanism for acetylene pyrolysis is also interesting. There has been much discussion regarding the relative importance of radical initiation pathways (e.g. addition and H-loss to C_4H_3 or disproportionation to C_2H_3 and C_2H) or isomerization to vinylidene.[27–29] It is also possible that acetone impurities which are typically present in acetylene may contribute to initiation,[30] although there has also been work suggesting that there is negligible effect.[31] More recently, there have been a few computational studies of acetylene initiation steps,[32–34] and in particular, Zador *et al.* concluded that the vinylidene formation and addition to acetylene is the primary initiation pathway.

There have been a number of previously published mechanisms specifically for acetylene pyrolysis including PAH formation, including early work by Frenklach *et al.*,[6] and more recent ones by Norinaga *et al.*,[35] Slavinskaya *et al.*,[36] and Tao *et al.*[37] Norinaga *et al.* compiled elementary reactions reported in literature to obtain a mechanism predicting up to the formation of coronene. Slavinskaya *et al.* focused on optimizing model parameters using numerous experimental works from literature, focusing on C_1-C_4 chemistry, with the final mechanism including species up to benzo[a]pyrene. Tao *et al.* combined previously published acetylene mechanisms to construct an improved mechanism for predicting the formation of 7 EPA targeted PAHs up to coronene.

Developing detailed kinetic models for PAH formation is challenging because the variety of products and inclination to form soot make experimental investigations difficult, the size of PAHs make accurate quantum calculations difficult, and the number of species and reactions can make the mechanism generation process itself difficult. Automatic mechanism generation provides a useful tool to aid the latter by keeping track of all of the species and reactions, and automatically identifying ones which are relevant to the system of interest. Additionally, parameter estimation methods (*e.g.*, thermochemical group additivity and kinetics rate rules) enable straightforward prediction of unknown parameters. One such software is the Reaction Mechanism Generator (RMG), an open-source mechanism generation package written in Python.[38]

Previously, we have described efforts towards expanding the RMG-database with the necessary pathways and data to predict formation of one- and two-ring aromatic species.[39] Those efforts included the addition of new kinetics families for propargyl recombination and rate calculations for pathways forming naphthalene and acenaphthylene, which were combined to generate a pressure-dependent mechanism for methane oxidation using RMG. In this work, we focus on further improving the RMG database with key thermochemical and rate parameters for modeling PAH formation in acetylene pyrolysis. While the focus of these additions was on PAH formation pathways, initiation reactions for acetylene pyrolysis were also considered.

6.2 Methods

6.2.1 Kinetics libraries

High-pressure-limit rate coefficients have been calculated for a number of PAH formation pathways. In some cases, electronic structure results were obtained from literature, and in others, electronic structure calculations were performed using Gaussian 16 [40] (for CBS-QB3 and B3LYP calculations) and Molpro [41] (for CCSD(T) calculations). Rate coefficients were calculated using Arkane, a TST and master equation solver packaged with RMG-Py.[42] Table 6.1 summarizes the reaction pathways for which rate coefficients were calculated.

The phenyl + benzene and benzyne + benzene surfaces have been previously reported by Comandini *et al.* at the uCCSD(T)/cc-pVDZ//uB3LYP/6-311+G(d,p) level of theory.[26] Here, we recalculated the addition and H-loss reactions using CBS-QB3 to obtain more accurate energies.

The naphthyl + acetylene surfaces have been previously reported by Kislov *et al.*[17] and Frenklach *et al.*[18] For the $C_{14}H_{11}$ surface, 1D hindered-rotor scans were performed, and for

Table 6.1: Potential energy surfaces for which high-pressure limit rate constants were calculated using Arkane. Reference column refers to source of electronic structure data.

PES	Reactants	Level of Theory	Reference
$C_{12}H_{11}$	phenyl + benzene	CBS-QB3	this work
$C_{12}H_{10}$	benzyne + benzene	CBS-QB3	this work
$C_{12}H_9$	naphthyl + acetylene	G3(MP2,CC)//B3LYP/6-311G(d,p)	this work
$C_{14}H_{11}$	vinylnaphthyl + acetylene	G3(MP2,CC)//B3LYP/6-311G(d,p)	this work
$C_{14}H_{11}$	naphthyl + vinylacetylene	G3(MP2,CC)//B3LYP/6-311G(d,p)	[23]
$C_{14}H_{11}$	biphenylyl + acetylene	G3(MP2,CC)//B3LYP/6-311G(d,p)	[20]
$C_{16}H_{11}$	phenanthryl + acetylene	G3(MP2,CC)//B3LYP/6-311G(d,p)	[21]

the $C_{12}H_9$ surface, hindered-rotor scans were obtained from Frenklach *et al.*

A kinetics library was also created for reactions on the $C_{14}H_9$ surface (*i.e.*, ethenylnaphthyl + acetylene) with high-pressure limit rate coefficients reported by Liu *et al.*[19] They performed calculations at the B3LYP//6-311+G(d,p) level of theory, and applied a correction calculated from the difference between CBS-QB3 and B3LYP//6-311+G(d,p) energies for the analogous benzene system.

To accurately capture acetylene initiation reactions, a kinetics library was created with reactions on the C_4H_4 potential energy surface using pressure-dependent rate coefficients calculated by Zádor *et al.*[34] Their calculations were performed at the CCSD(T)-F12b/cc-pVQZ-F12//M06-2X/MG3S level of theory, and master equation calculations were done using MESS.[43]

6.2.2 Thermochemistry libraries

Separate from the rate calculations, thermochemistry calculations were performed for all species from the PAH formation pathways which have been added to the RMG database. Electronic structure calculations using CBS-QB3 were performed using Gaussian 09 [44] and Gaussian 16, including 1D hindered-rotor calculations using B3LYP/CBSB7. Hindered rotor scans were performed automatically using ARC (Automatic Rate Calculator), a new Python package developed in our group for automating quantum chemistry calculations. Thermochemistry calculations were also performed automatically by ARC using Arkane, with bond-additivity corrections from Petersson *et al.*[45]

6.2.3 Model generation

RMG v2.4.1 was used to generate the acetylene pyrolysis model. An initial composition of 98% acetylene, 1.8% acetone, and 0.2% methane was used, based on the composition reported by Norinaga *et al.*[35] Reactor conditions were set with a temperature range of 1000-1500 K and a pressure of 0.2 atm. The pressure dependence feature of RMG was enabled, so that pressure dependent networks would be automatically constructed for species with up to 16 atoms. Species constraints were set, limiting the maximum number of carbon atoms in any molecule to 20, and the maximum number of radicals to 1.

The kinetics library for the C_4H_4 surface was included as a seed mechanism. The other kinetics libraries containing high-pressure limit reaction rates for PAH formation pathways were included and appended to the final mechanism. In post-processing, a few highly-strained polycyclic species which were identified as being unreasonable were manually removed from the model. In general, these species were identified by RMG as being important due to inaccurate thermochemistry estimation. In the end, the final model contains 1594 species and 8924 reactions.

6.2.4 Model simulation and analysis

In this work, we chose to validate the model predictions using data from Norinaga *et al.* for acetylene pyrolysis in a flow reactor.[46] The primary reason for this choice was that they reported mole fractions of both small molecules and PAH species up to coronene as well as measured temperature profiles for their reactor.

To simplify the simulation, we chose to simulate a single fluid element flowing through the reactor as a homogeneous batch reactor. Using the provided reactor geometry and temperature profiles as a function of position, we calculated temperature profiles as a function of residence time, assuming ideal plug flow. The resulting temperature profiles are shown in Figure 6-1. Reactor simulations were performed using Cantera.[47] Rate-of-production (ROP) analyses were performed with Chemkin-Pro,[48] using the same reactor idealization and temperature profiles for the 1073 K and 1373 K set points.

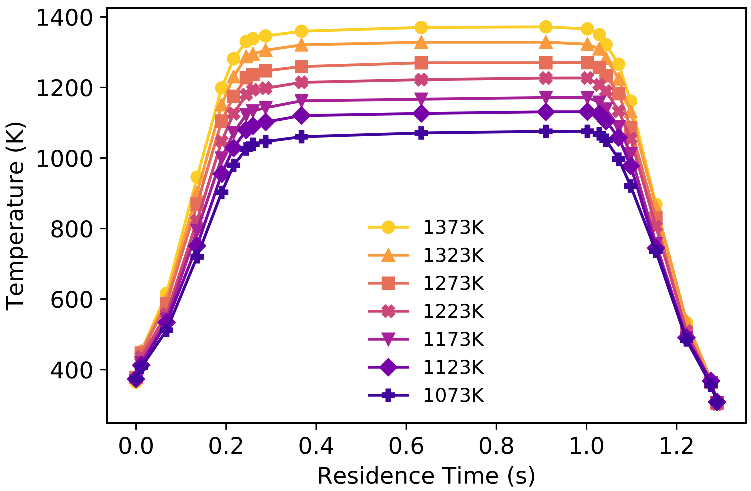


Figure 6-1: Temperature profiles for each set point temperature as a function of residence time.

6.3 Results and discussion

The RMG model predictions are shown in Figures 6-2–6-4, in comparison with the experimental measurements of Norinaga *et al.*[46] and predictions from published mechanisms by Norinaga *et al.*,[35] Slavinskaya *et al.*,[36] and Tao *et al.*[37] It is important to note that while the RMG and Norinaga mechanisms include acetone and related reactions, neither the Slavinskaya mechanism nor the Tao mechanism include acetone. As a result, the initial composition described by Norinaga *et al.*[35] including methane and acetone was used to simulate the RMG and Norinaga mechanisms, while pure acetylene was used as the initial composition to simulate the Slavinskaya and Tao mechanisms. The RMG mechanism was also simulated using pure acetylene for comparison. By comparing the two RMG simulations, we see that inclusion of acetone affects both acetylene conversion and product distributions. These differences will be discussed in more detail shortly.

6.3.1 Acetylene consumption

In Figure 6-2, we see that at higher temperatures, the RMG model predicts larger final mole fractions of acetylene than was observed in the experiment. A major contributing factor is the lack of large PAHs in the RMG model. In the experiment, PAHs with three or more

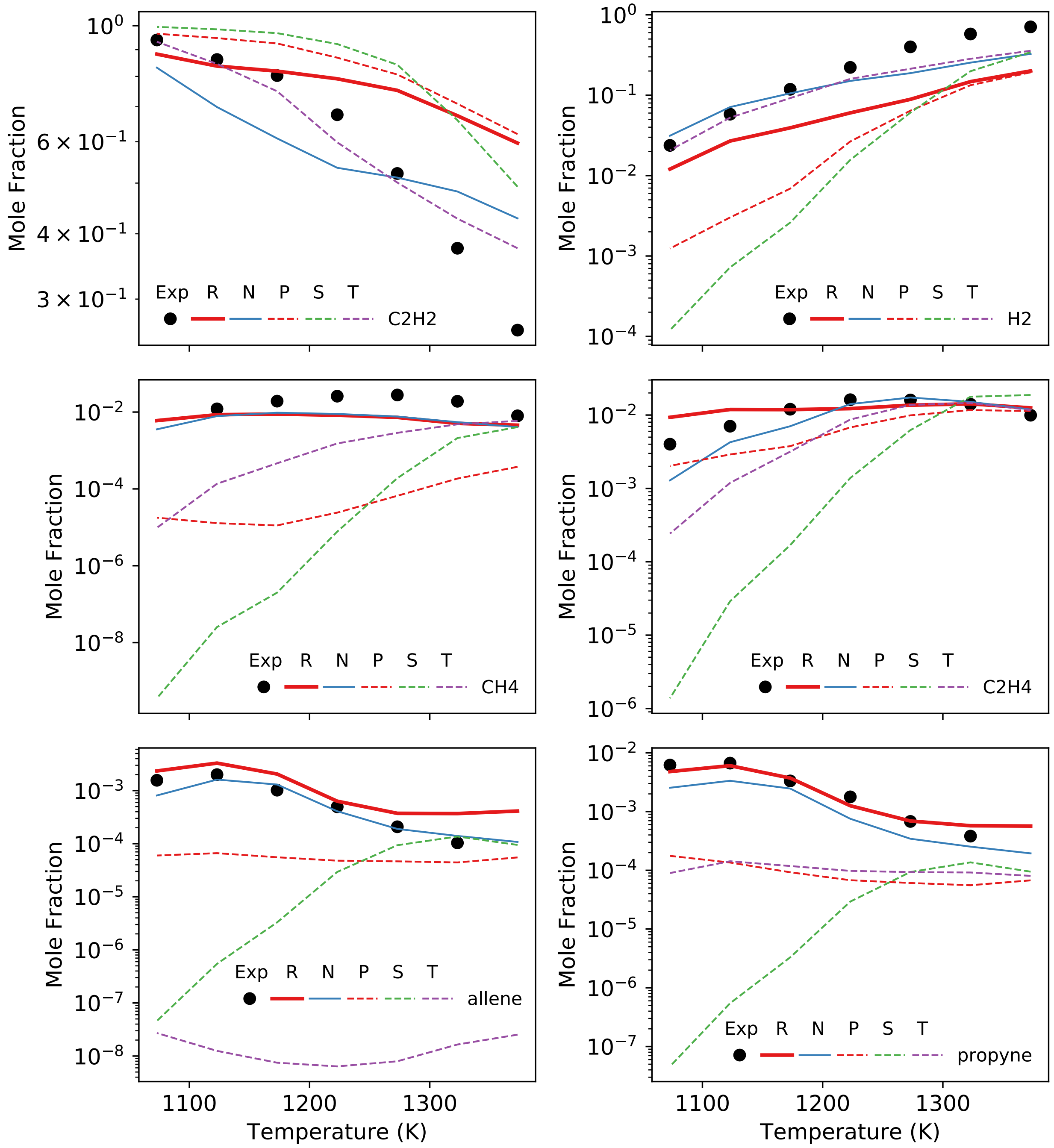


Figure 6-2: Final mole fractions as a function of reactor temperature. Points are the experimental data from ref. [46], lines are the RMG model (R), Norinaga model (N), RMG model with pure acetylene (P), Slavinskaya model (S), and Tao model (T). Dashed lines are model simulations where the initial composition is pure acetylene.

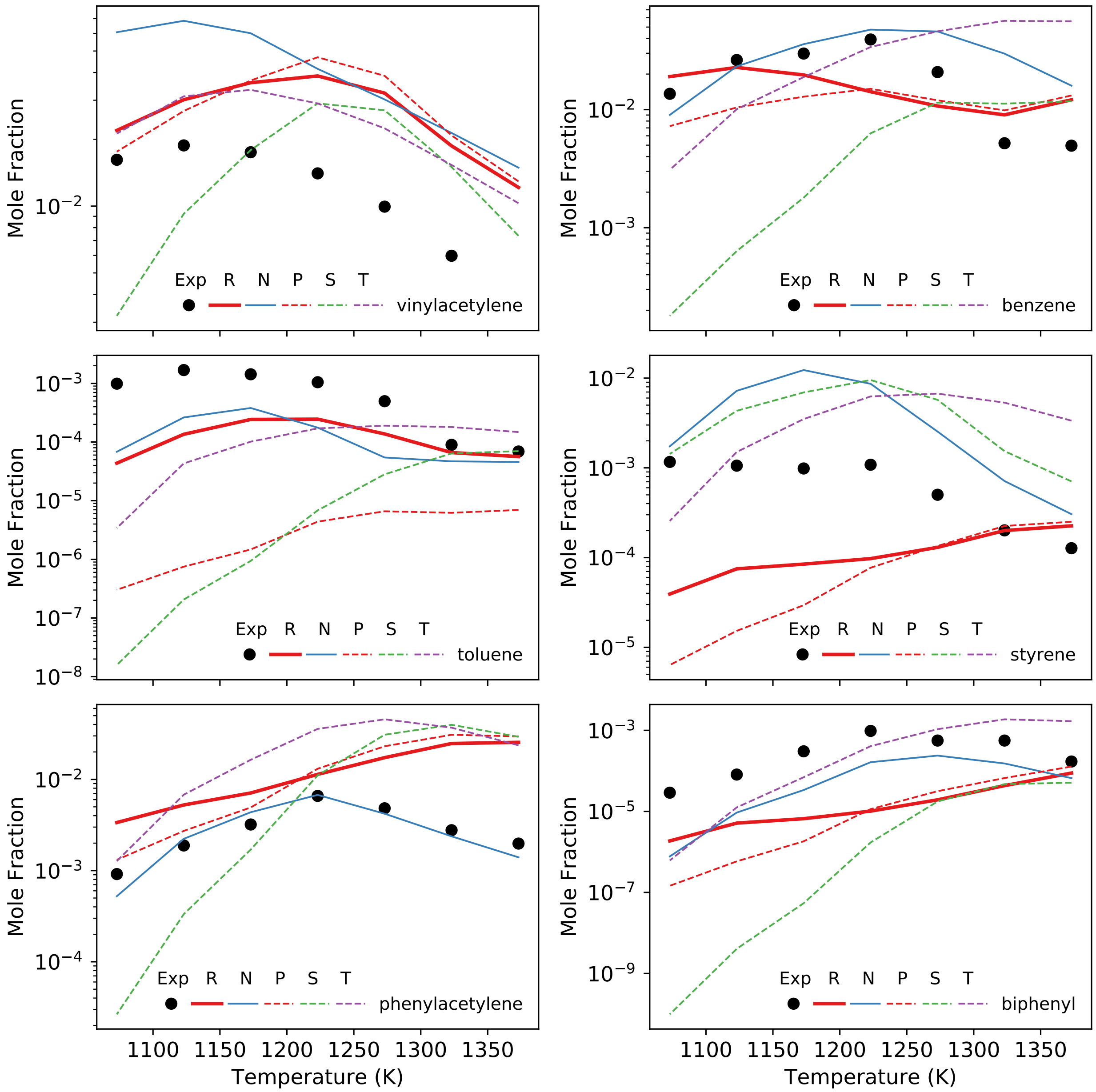


Figure 6-3: Final mole fractions as a function of reactor temperature. Points are the experimental data from ref. [46], lines are the RMG model (R), Norinaga model (N), RMG model with pure acetylene (P), Slavinskaya model (S), and Tao model (T). Dashed lines are model simulations where the initial composition is pure acetylene.

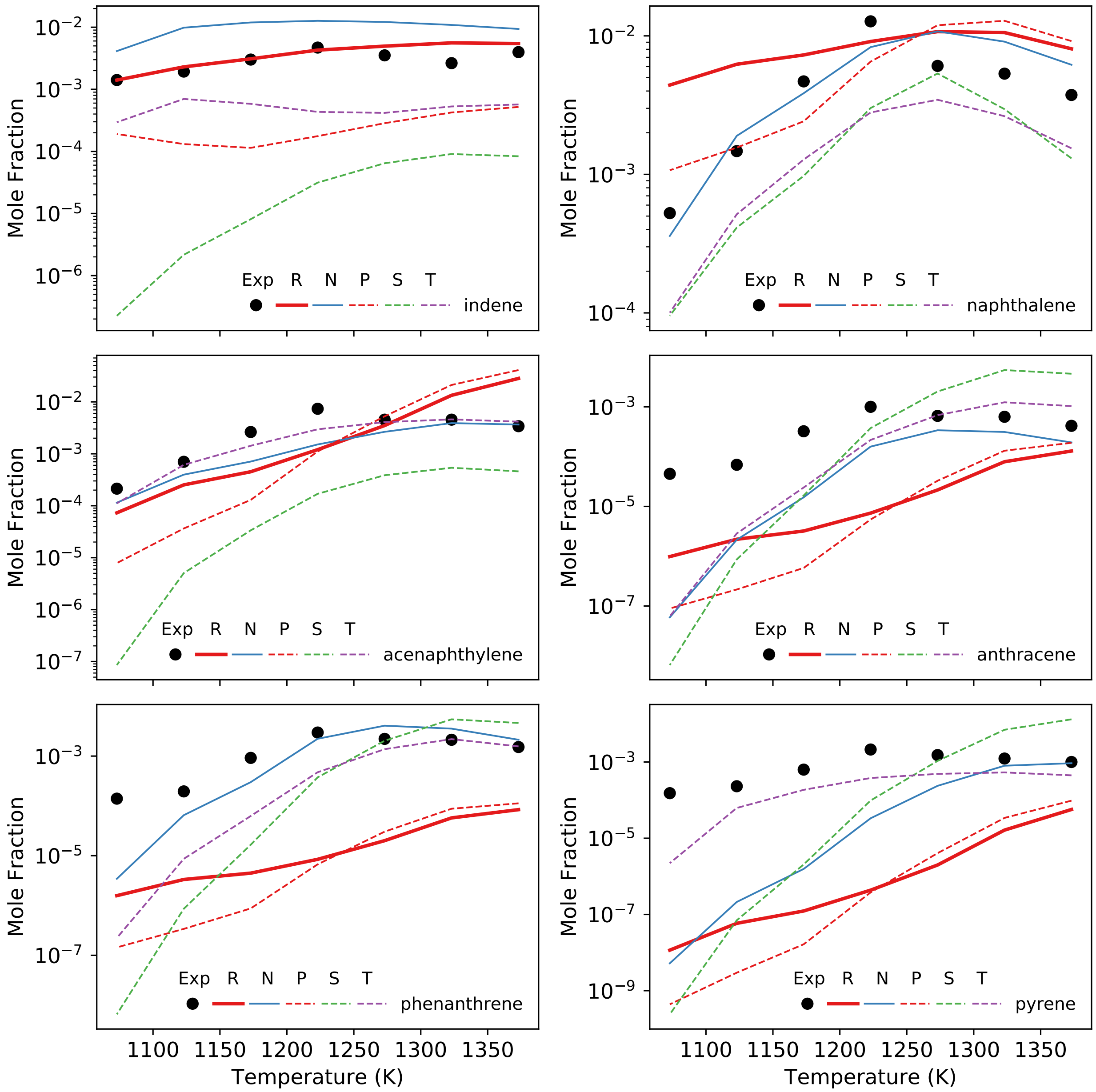


Figure 6-4: Final mole fractions as a function of reactor temperature. Points are the experimental data from ref. [46], lines are the RMG model (R), Norinaga model (N), RMG model with pure acetylene (P), Slavinskaya model (S), and Tao model (T). Dashed lines are model simulations where the initial composition is pure acetylene.

rings accounted for 28% of product carbons at 1373 K, with 12% in PAHs not included in the RMG model. Additionally, the C/H ratio of the experimental product composition at 1373 K is approximately 1:2.78, which suggests that some carbon-rich products (*i.e.*, soot or coke) were formed which were not quantified. At 1073 K, large PAHs only make up 2.7% of the product carbons, and the experimental C/H ratio is 1:1.03, suggesting minimal soot formation. Since the RMG model stops at pyrene, there are no further consumption channels for acetylene, leading to the over-prediction.

Figures 6-5 and 6-6 show the top 10 pathways for acetylene consumption at 1373 K and 1073 K, respectively. An overall observation is that acetylene consumption is not highly dominated by a single channel, but is instead relatively spread out. At 1373 K, the top 10 pathways only account for 69% of total acetylene consumption, and the top 58 pathways have to be considered before reaching 99%. We see that almost all of the top pathways involve radical addition to acetylene, the exceptions being isomerization to vinylidene and vinylidene addition. Additionally, most of these pathways are well-skipping pathways in pressure dependent networks, indicating the importance of considering pressure-dependence for this system.

Notably, acetylene dimerization to either vinylacetylene (R6.1) or diacetylene (R6.2) do not appear in the top acetylene consumption pathways according to the RMG model. This is in stark contrast to the other models which were evaluated.

$$\mathrm{C{\equiv}C + C{\equiv}C \rightleftharpoons C{\equiv}C{-}C{=}C} \tag{R6.1}$$

$$\mathrm{C{\equiv}C + C{\equiv}C \rightleftharpoons C{\equiv}C{-}C{\equiv}C + H_2} \tag{R6.2}$$

The vinylacetylene formation reaction is present in all four models, with vast differences in the rate constant, as shown in Figure 6-7. The RMG model uses the pressure dependent rate calculated by Zádor *et al.* for the C_4H_4 potential energy surface.[34] The Slavinskaya model cites Melius *et al.*,[49] although there is approximately a factor of 2 difference in the activation energy. The Norinaga and Tao models have very similar rates, for which Norinaga cited Dúran *et al.*[50] and Tao cited Saggase *et al.*[51]

The diacetylene formation reaction was intentionally removed from the Slavinskaya model,

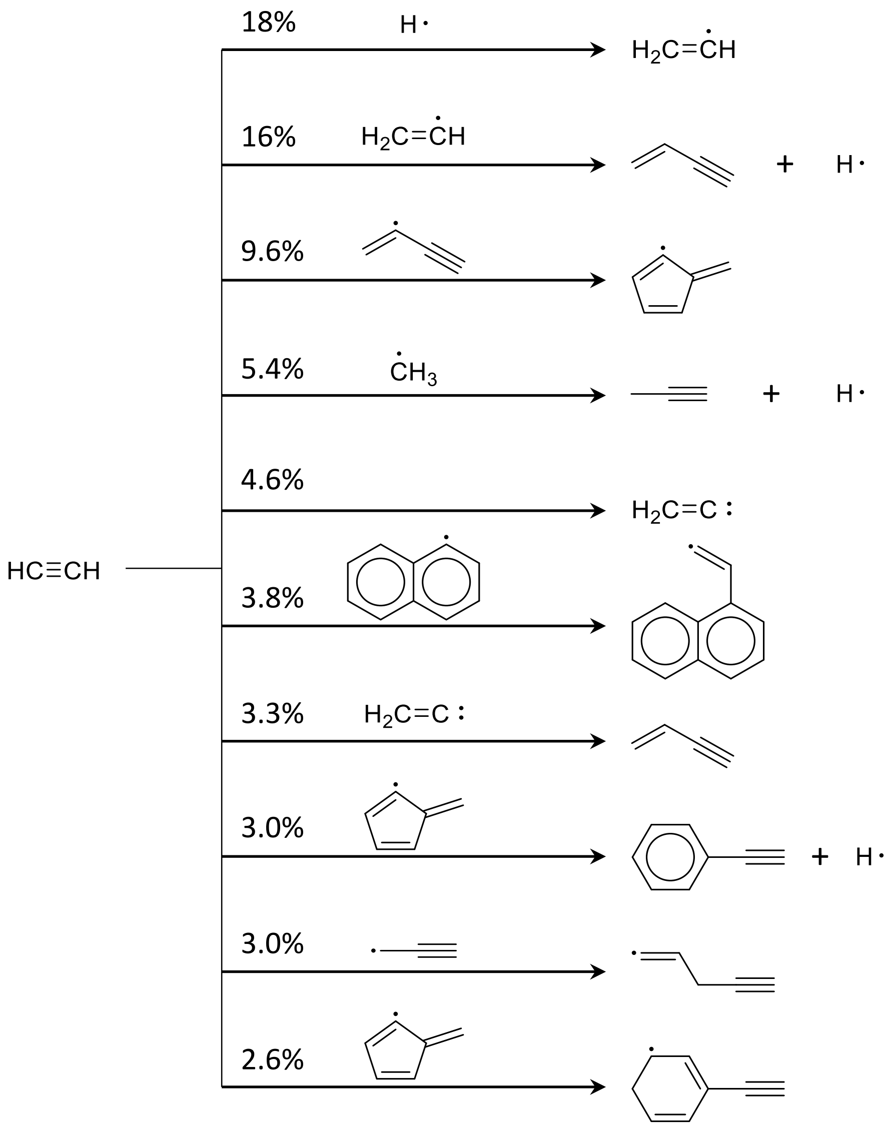


Figure 6-5: Top 10 consumption pathways of acetylene at conditions of ref. [46], 1373 K set point. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux.

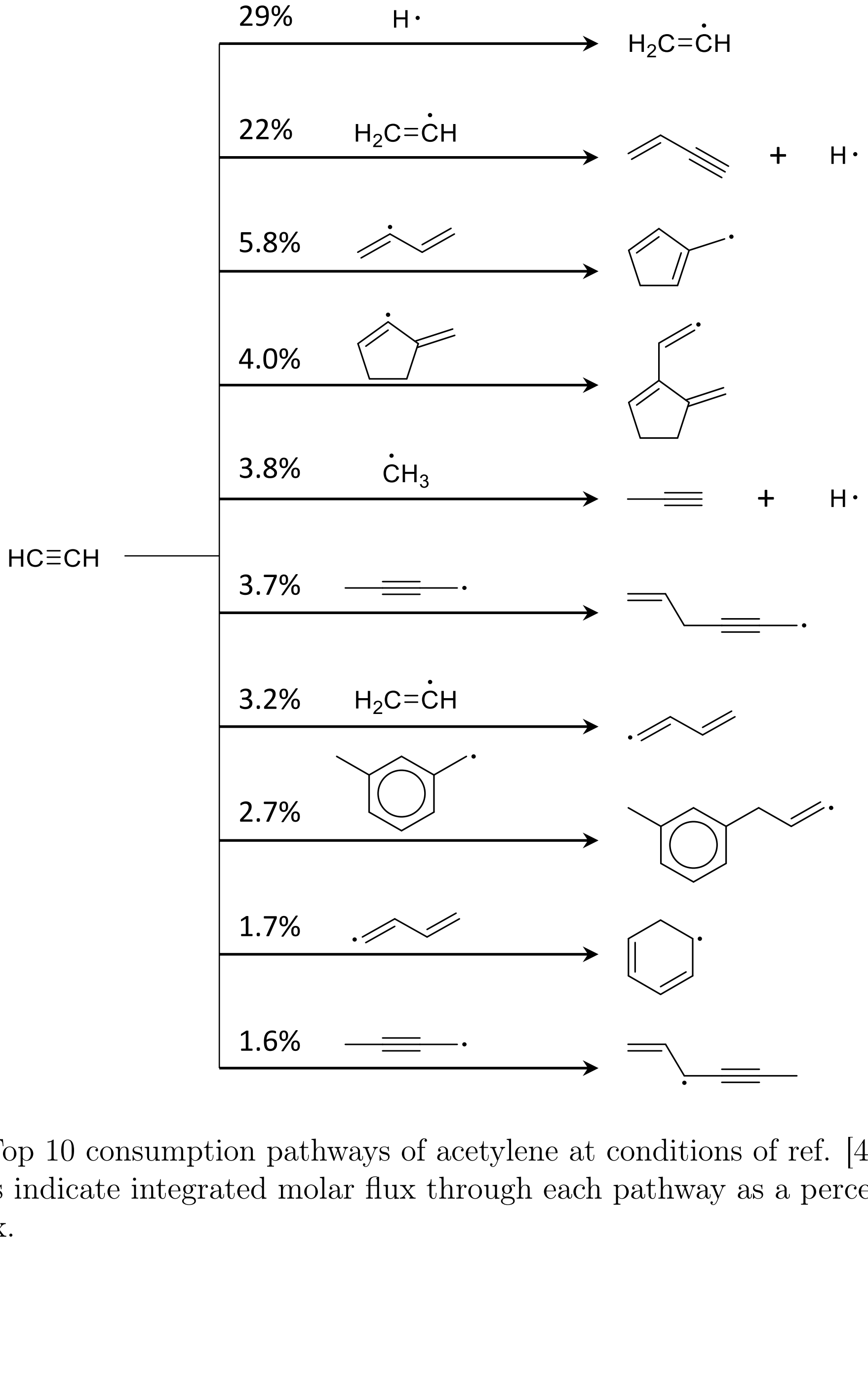


Figure 6-6: Top 10 consumption pathways of acetylene at conditions of ref. [46], 1073 K set point. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux.

because the authors noted that it was the sum of two reactions proceeding via $H_2CCCCH + H$. The reaction is included in the other models, with the RMG model again using the pressure-dependent rate computed by Zádor *et al.* The Norinaga and Tao models have identical rates for this reaction, taken from Fournet *et al.*[52] There is also a notable difference between the two rates for this reaction.

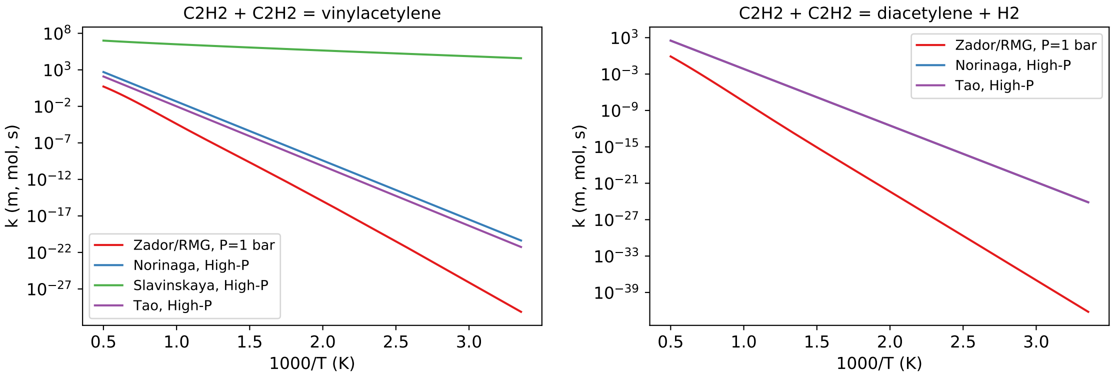


Figure 6-7: Comparison of reaction rates for formation of vinylacetylene (left) and diacetylene (right) via acetylene dimerization.

With the updated rates by Zádor *et al.*, it seems that direct formation of vinylacetylene or diacetylene (realistically through a chemically-activated reaction) is not a particularly significant pathway at these conditions.

6.3.2 Role of acetone

As mentioned previously, the inclusion of acetone in the model affected both acetylene conversion and product distributions. Acetone primarily decomposes into ketene or carbon monoxide, releasing up to two methyl radicals in the process, as shown in Figure 6-8. Ketene is favored at lower temperatures, while carbon monoxide is favored at higher temperatures, leading to increased formation of methyl radicals. The methyl radicals then contribute significantly to the formation of odd-carbon species. Methyl addition to acetylene can proceed through a well-skipping pathway to form propyne or allene. These C_3 species can then form propargyl radicals, which greatly simplify the process of reaching C_5 (*e.g.*, cyclopentadiene), C_7 (*e.g.*, toluene), and C_9 (*e.g.*, indene). Simulating the RMG model using pure acetylene as the initial composition results in substantial under-prediction of all of these odd-carbon species.

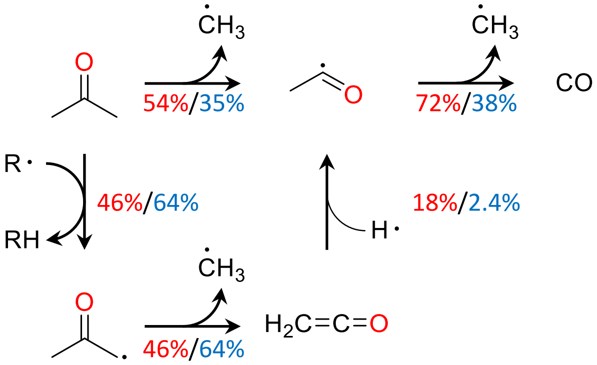


Figure 6-8: Acetone decomposition pathways. Values indicate integrated molar flux through each pathway as a percentage of total acetone flux. First value in red indicates flux at 1373 K and second value in blue indicates flux at 1073 K.

6.3.3 Small molecule products

Low molecular weight products which were reported in the experiment include H_2, CH_4, C_2H_4, allene, and propyne. Overall, the RMG model predictions match the experimental data very well for these species. The most significant deviation is for hydrogen, which is under-predicted by about a factor of two at low temperatures and a factor of three at higher temperatures. This discrepancy is likely also related to the under-prediction of large PAH formation in the RMG model, since conversion of acetylene to carbon-rich PAHs results in elimination of hydrogen.

Hydrogen, methane, and ethylene are all formed via hydrogen abstraction reactions, with the top pathways being formation of resonance-stabilized radicals (RSRs) like propargyl, i-C_4H_3, cyclopentadienyl, indenyl, and 1-methylvinoxy. These reactions are key chain-propagation steps to formation of larger compounds, as the RSRs will then add to acetylene or vinylacetylene to continue molecular weight growth, with the exception of 1-methylvinoxy which primarily breaks apart into ketene and a methyl radical as mentioned above.

Propyne and allene, on the other hand, are primarily formed via chemically-activated pathways, as shown in Figure 6-9. The most significant pathway is direct formation of propyne via methyl addition to acetylene, which can then isomerize to form allene.

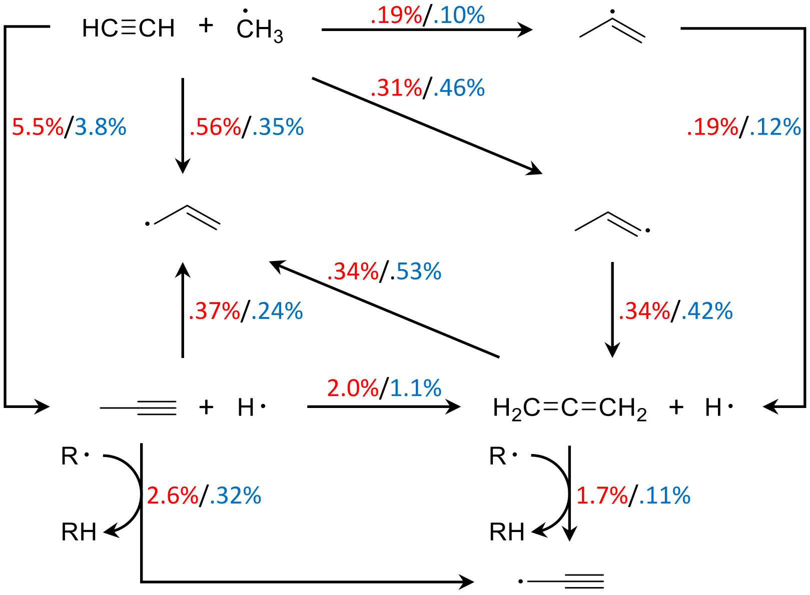


Figure 6-9: Main pathways to C_3 species. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux. First value in red indicates flux at 1373 K and second value in blue indicates flux at 1073 K set point conditions of ref. [46]. Pathways with values less than 0.1% are omitted.

Vinylacetylene is primarily formed via chemically activated pathways, the primary pathway being vinyl addition to acetylene, and the secondary pathway being vinylidene addition to acetylene, as shown in Figure 6-10. The vinylidene pathway becomes more significant at higher temperatures, due to higher flux of acetylene isomerization to vinylidene. The vinyl addition pathway can also proceed via C_4H_5 radical intermediates, which play important roles in aromatic ring formation. At lower temperatures, formation of both C_4H_5 radicals increases, and the net flux of H-elimination from i-C_4H_5 goes in the reverse direction, consuming vinylacetylene instead. Diacetylene is formed via H-elimination or disproportionation from C_4H_3 radicals, which are mostly formed via hydrogen abstraction from vinylacetylene, with a small contribution from C_2H addition to acetylene. As mentioned previously, dimerization of acetylene was not found to contribute significantly to either vinylacetylene or diacetylene at these conditions.

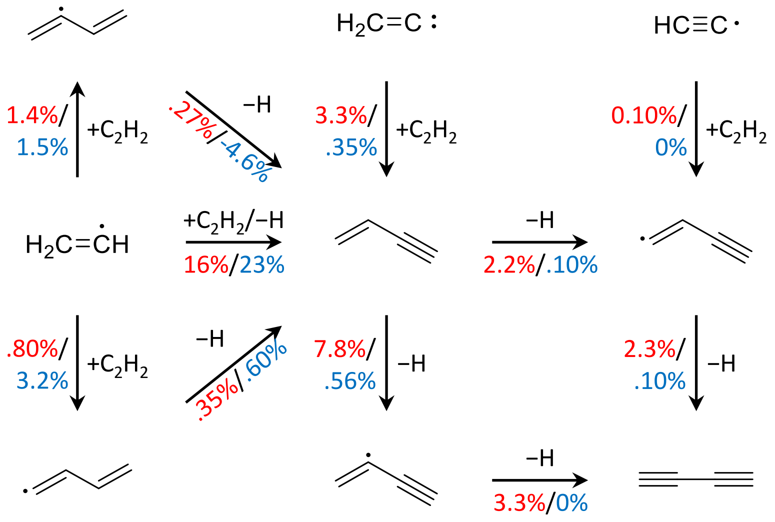


Figure 6-10: Main pathways to C_4 species. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux. First value in red indicates flux at 1373 K and second value in blue indicates flux at 1073 K set point conditions of ref. [46]. Pathways with values less than 0.1% are omitted. −H can include H-elimination, H-abstraction, or disproportionation.

6.3.4 One- and two-ring aromatic species

Looking at single ring aromatics, benzene and toluene are reasonably predicted, but styrene diverges from experiment at low temperatures and phenylacetylene diverges at high temperatures. From the reaction path analysis for benzene, shown in Figure 6-11, we see that the primary pathway for benzene formation is via methylcyclopentadienyl, which is generated through multiple $C_4 + C_2H_2$ pathways, all involving many isomerization and H-shift steps. A secondary pathway is via cyclohexadienyl radical, which is formed by 1-butadienyl addition to acetylene. From benzene, phenyl addition is the only pathway to biphenyl which was captured by the RMG model.

The RMG model predicts styrene well at high temperatures, but begins to under-predict its formation below 1273 K. Interestingly, this is in contrast with the other literature models which tend to over-predict styrene formation. For phenylacetylene, the RMG model over-predicts formation at high temperatures, which is similar in behavior to both the Tao and Slavinskaya models. In the RMG model, the fulvenyl and phenylvinyl radical are the primary precursors to both phenylacetylene and styrene. From phenylvinyl, H-elimination to phenylacetylene is much more favored than H-abstraction to form styrene. An alternative pathway to styrene is via vinyl addition to benzene, which has fairly low flux since it breaks the aromaticity of the benzene ring.

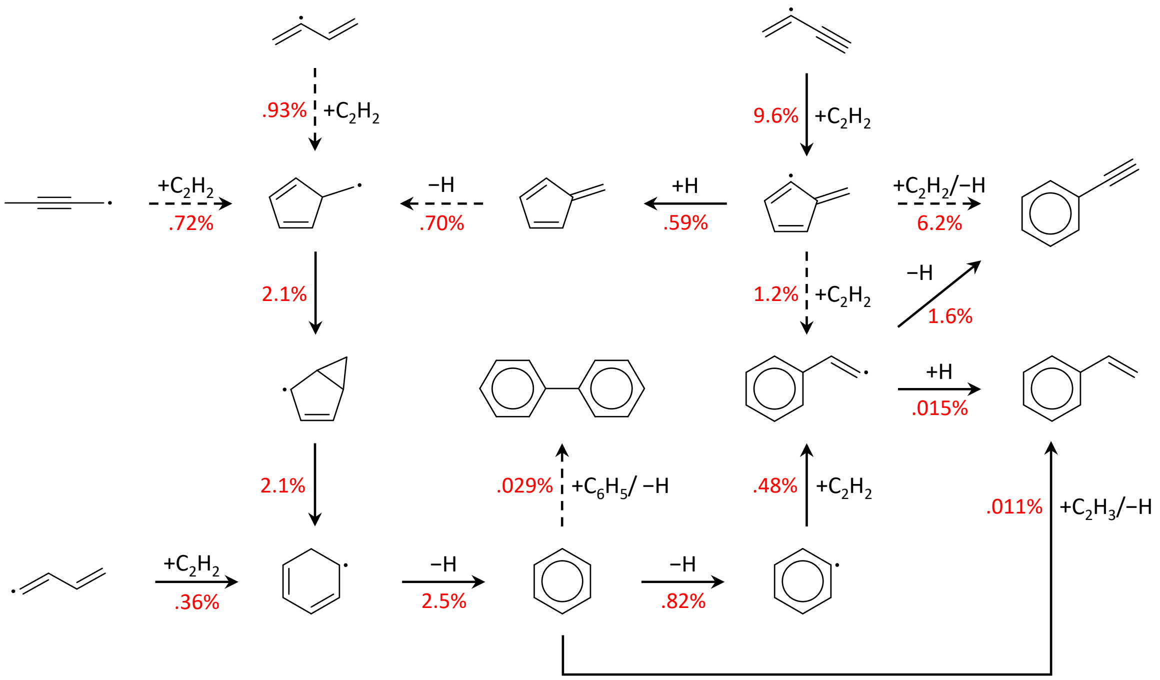


Figure 6-11: Main pathways to mono-aromatic species and biphenyl. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux at 1373 K conditions of ref. [46]. Dashed lines represent a combination of multiple reactions. ±H can include H-elimination, H-abstraction, disproportionation, or recombination.

Both toluene and indene are well predicted by the RMG model. Since these are both odd-carbon species, propargyl plays a major role in their formation. The primary route to toluene is via propargyl addition to vinylacetylene, which proceeds via a well-skipping route to directly form benzyl radical, as shown in Figure 6-12. Alternatively, benzyl can also be formed by cyclopentadienyl (CPDyl) radical addition to acetylene followed by a few isomerization steps. However, toluene is actually a very small consumption pathway of benzyl, which at the conditions of ref. [46] instead prefers to add another acetylene and ring-close to form indene. The importance of propargyl in these pathways to toluene and indene offers a clear explanation for the low prediction of these species when acetone is not included in the initial composition.

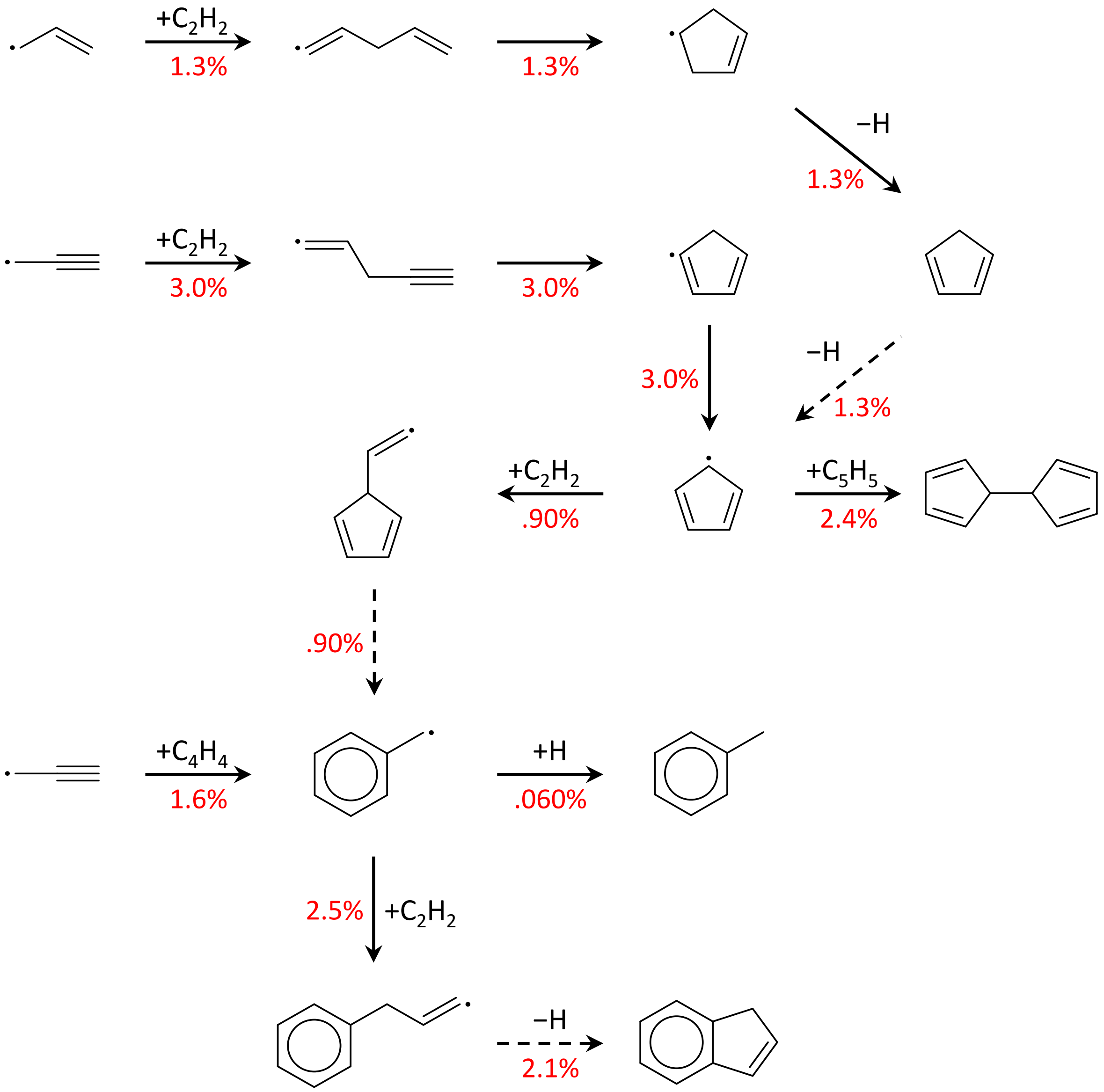


Figure 6-12: Main pathways to toluene and indene. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux at 1373 K conditions of ref. [46]. Dashed lines represent a combination of multiple reactions. ±H can include H-elimination, H-abstraction, disproportionation, or recombination.

Naphthalene is also predicted reasonably well by the RMG model, although it is over-predicted by about an order of magnitude at 1073 K. Naphthalene can be formed by many different paths, which are shown in Figure 6-13. The most significant pathway is via cyclopentadienyl recombination, which involves multiple isomerizations and two H-loss steps. HACA growth also contributes to naphthalene formation, starting from three different styrene radicals. The most significant route starts from the 1-phenylvinyl radical, which can ring-close to form methylindenyl radical, which can then isomerize and undergo H-loss to naphthalene. The other two routes follow the Bittner-Howard and Modified-Frenklach pathways. The final pathway is via vinyl addition to phenylacetylene, which proceeds through many of the same intermediates as the other two pathways before the final H-elimination step to form naphthalene.

6.3.5 Three- and four-ring aromatic species

While the acenaphthylene prediction by the RMG model matches reasonably with experiment, once we get to three ring aromatics, we see significant under-prediction by the model. In short, this is due to the limited number of PAH formation pathways which are included in this RMG model. As previously mentioned, a number of pathways for which accurate quantum chemistry data were available were included as kinetics libraries for this model. The RMG simulation was not able to fully explore additional reaction pathways beyond those due to computational limitations of the RMG algorithm in handling such complex chemistry.

At 1373 K, the amount of these three- and four-ring PAHs predicted by the RMG model comes fairly close to matching the experimental measurements, within about a factor of five or better. All of the pathways to these large PAHs fall under the HACA scheme (Figure 6-14). The main precursor to three-ring aromatics is the naphthalen-1-yl radical, formed either via H-abstraction from naphthalene or the Frenklach pathway for HACA growth from phenylacetylene, which directly forms naphthalen-1-yl without passing through naphthalene. Following addition of acetylene, the most favorable route to acenaphthylene involves an H-shift from the 8-position on the ring to the vinyl group. The ring-closing pathway following the H-shift goes through a stable benzylic intermediate and accounts for about 80% of acenaphthylene formation. The remainder is mostly from the slower, direct ring-closing route

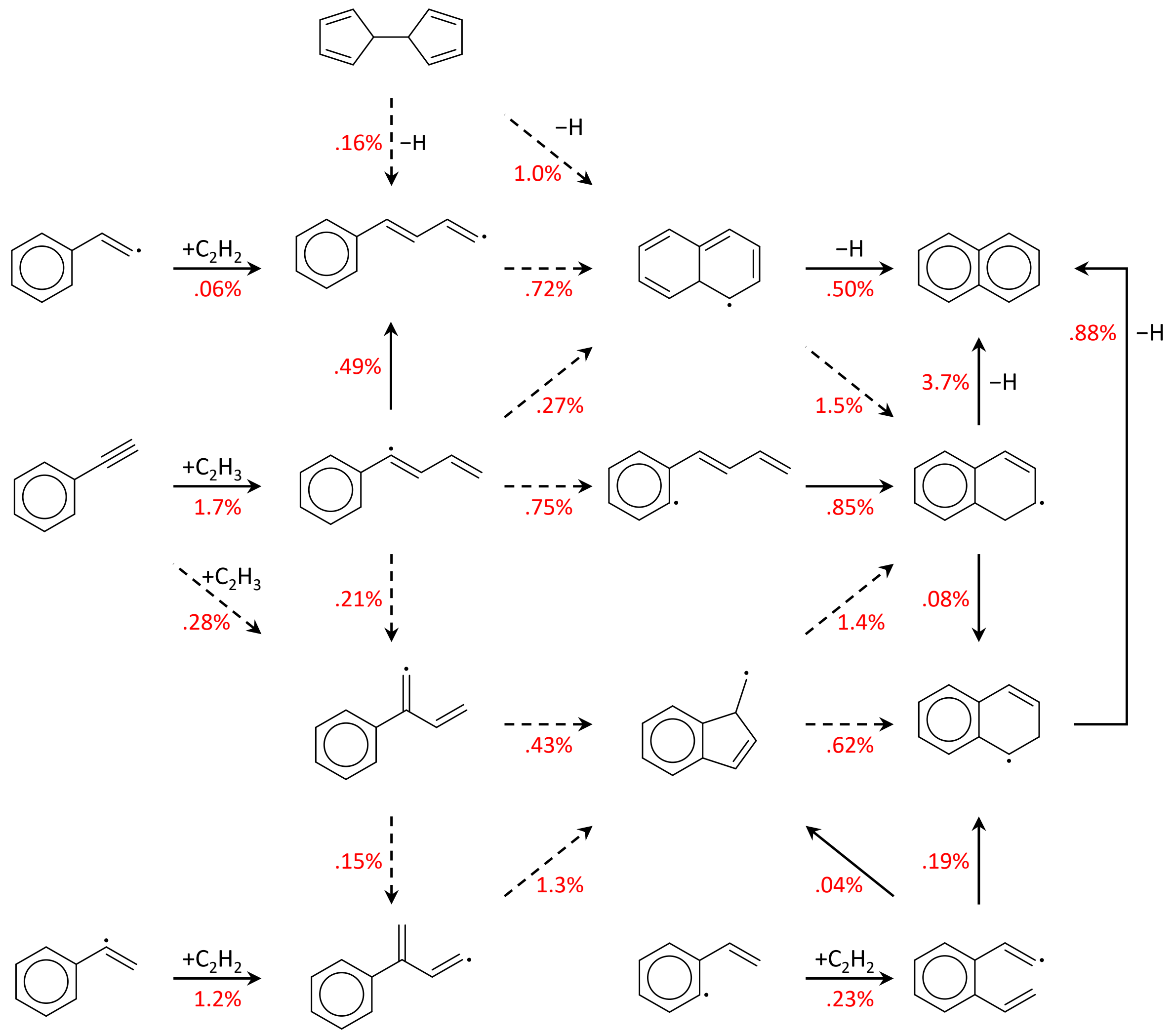


Figure 6-13: Main pathways to naphthalene. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux at 1373 K conditions of ref. [46]. Dashed lines represent a combination of multiple reactions. ±H can include H-elimination, H-abstraction, disproportionation, or recombination.

which involves disrupting the aromaticity of the ring.

The other major precursor to three-ring aromatics is the naphthalen-2-yl radical, which exclusively forms via H-abstraction. A small fraction undergoes H-shift to naphthalen-1-yl, while the majority adds to acetylene. The 2-naphthalen-2-ylvinyl radical can then go through one of four pathways. The highest-flux pathway involves an H-shift from the 1-position on the ring to the vinyl side-chain forming 2-vinylnaphthalen-1-yl, followed interestingly by isomerization to the 2-naphthalen-1-ylvinyl radical via a three-member ring intermediate, which proceeds to form acenaphthylene. A small fraction of the 2-vinylnaphthalen-1-yl radical continues along the Modified-Frenklach pathway via a second acetylene addition to form phenanthrene. The second most significant pathway involves H-elimination to form 2-ethynylnaphthalene, following the Frenklach pathway to form anthracene or phenanthrene in about a 2:1 ratio. The third pathway follows the Bittner-Howard HACA scheme with a second acetylene addition, followed by ring-closing and H-elimination to phenanthrene or anthracene in about a 3:1 ratio. The final pathway is H-shift to the 3-position on the ring, following the Modified-Frenklach pathway to form anthracene. Although reactions for the naphthalenyl + vinylacetylene pathway to phenanthrene and anthracene were included in the model, almost no flux passed through the pathway despite the relative abundance of vinylacetylene.

Finally, biphenyl provides a minor route to phenanthrene, via a single HACA step which closes one of the bay sites in biphenyl. An analogous pathway closes the bay site in phenanthrene to form pyrene. Even though this is the only reaction path to pyrene included in the RMG model, it does account for a substantial amount of pyrene formation. If phenanthrene formation were to increase (*e.g.*, by adding more pathways or improving precursor concentrations), it is likely that this pathway would be sufficient to explain pyrene formation at the high end of the temperature range.

6.4 Conclusions

In this work, we have demonstrated the ability of RMG to automatically generate a mechanism to predict PAH formation in acetylene pyrolysis. Kinetics and thermochemistry data for

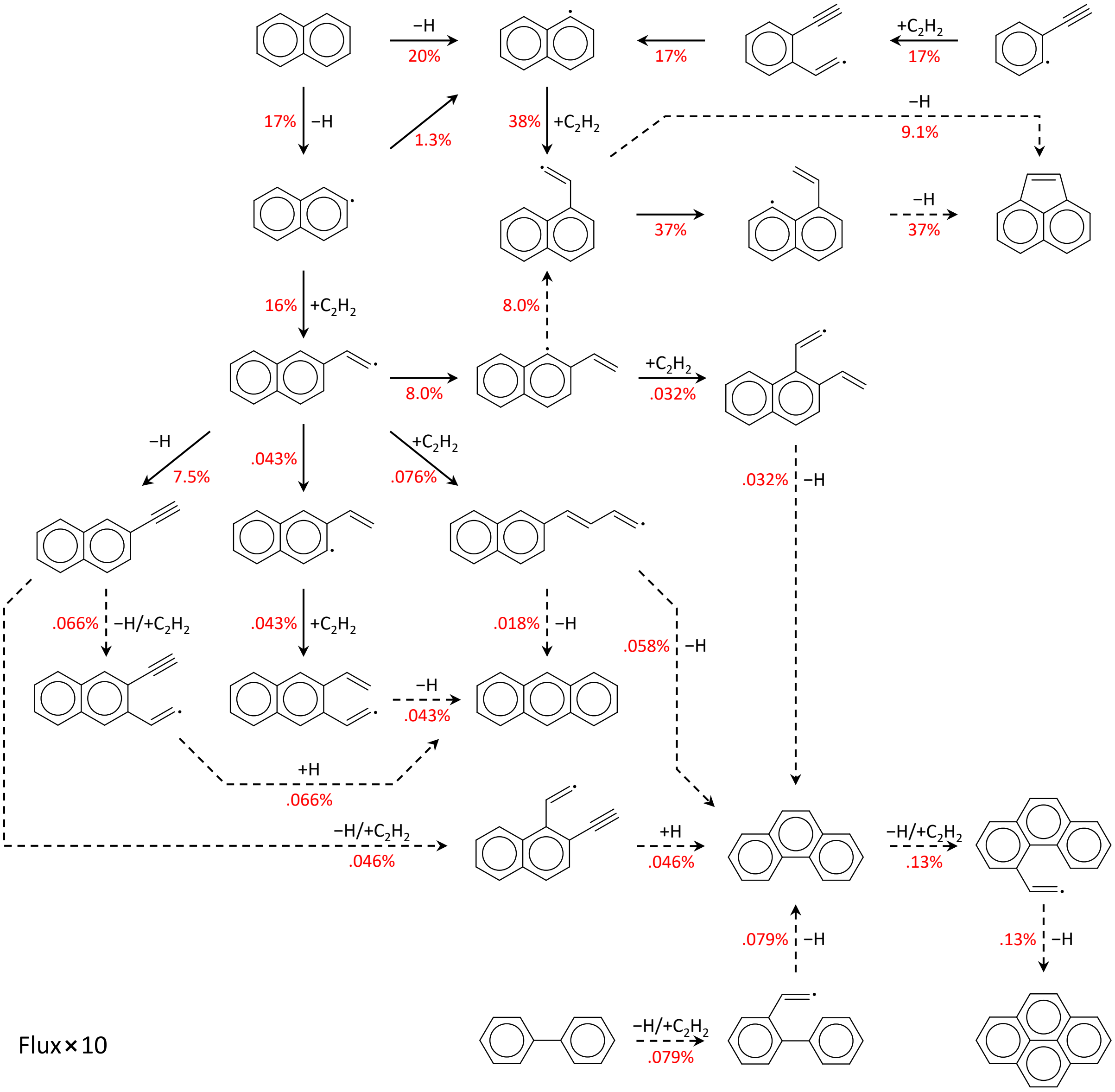


Figure 6-14: Main pathways to three and four ring aromatic species. Values indicate integrated molar flux through each pathway as a percentage of total acetylene flux at 1373 K conditions of ref. [46], multiplied by 10. Dashed lines represent a combination of multiple reactions. ±H can include H-elimination, H-abstraction, disproportionation, or recombination.

key PAH formation pathways and acetylene initiation pathways have been calculated using Arkane and added to the RMG database. Using the newly added data, a detailed chemical mechanism with 1594 species and 8924 reactions was generated using RMG. Importantly, no manual adjustment or optimization of thermochemical and kinetic parameters was performed following model generation. The model predictions agree well with experimental flow reactor data, particularly for small molecule products and one- to two-ring aromatics. Acetylene consumption is under-predicted, which is most likely linked to the under-prediction of three- and four-ring aromatic species.

Acetylene consumption was found to be relatively spread out across many different pathways, in contrast to dominance of direct dimerization to vinylacetylene or diacetylene which was characteristic of other literature models. The impurity acetone present in the experiments was found to play an important role in the formation of odd-carbon species, from methane up to indene, despite its low initial concentration. Because the Slavinskaya and Tao models did not include acetone, they tended to under-predict the formation of these species.

For PAHs, the majority of formation pathways captured by the RMG model can be classified as part of the HACA mechanism. The main exception is cyclopentadienyl recombination, which contributed significantly to the formation of naphthalene. The three main types of HACA pathways (Bittner-Howard, Modified-Frenklach, and Frenklach) were all observed to contribute to formation of naphthalene from benzene and phenanthrene/anthracene from naphthalene. The under-prediction of these large PAHs suggests that the included pathways are not sufficient to explain their formation and that other pathways may be missing.

This work represents an important milestone for the RMG software in automatically generating a mechanism predicting up to pyrene. However, there are still improvements which can be made to both the software and the chemical mechanism. Notably, generation of PAH formation models is still computationally challenging due to the sheer number of potential species and reactions which RMG evaluates. Additionally, highly-strained species tend to be over-represented in the RMG model due to poor thermochemistry estimates. On the chemistry side, accurate thermochemical and kinetic data for more pathways is important for improving the accuracy of these *ab initio* models.

References

(1) Richter, H.; Howard, J. B. *Prog. Energy Combust. Sci.* **2000**, *26*, 565–608.

(2) Léger, A.; D'Hendecourt, L.; Boccara, N., *Polycyclic Aromatic Hydrocarbons and Astrophysics*; NATO ASI Series, Series C: Mathematical and Physical Sciences, 1389-2185: 191; Dordrecht : Springer Netherlands, 1986.: 1986.

(3) Frenklach, M.; Feigelson, E. D. *Astrophys. J.* **1989**, *341*, 372.

(4) Kim, K.-H.; Jahan, S. A.; Kabir, E.; Brown, R. J. *Environ. Int.* **2013**, *60*, 71–80.

(5) Frenklach, M. *Phys. Chem. Chem. Phys.* **2002**, *4*, 2028–2037.

(6) Frenklach, M.; Clary, D. W.; Gardiner, W. C.; Stein, S. E. *Symp. Combust.* **1985**, *20*, 887–901.

(7) Westmoreland, P. R.; Dean, A. M.; Howard, J. B.; Longwell, J. P. *J. Phys. Chem.* **1989**, *93*, 8171–8180.

(8) Wang, H.; Frenklach, M. *J. Phys. Chem.* **1994**, *98*, 11465–11489.

(9) Miller, J. A.; Klippenstein, S. J. *J. Phys. Chem. A* **2003**, *107*, 7783–7799.

(10) Marinov, N. M.; Castaldi, M. J.; Melius, C. F.; Tsang, W. *Combust. Sci. Technol.* **1997**, *128*, 295–342.

(11) Melius, C. F.; Colvin, M. E.; Marinov, N. M.; Pitz, W. J.; Senkan, S. M. *Symp. Combust.* **1996**, *26*, 685–692.

(12) Moskaleva, L. V.; Mebel, A. M.; Lin, M. C. *Symp. Combust.* **1996**, *26*, 521–526.

(13) Sharma, S.; Green, W. H. *J. Phys. Chem. A* **2009**, *113*, 8871–82.

(14) Mebel, A. M.; Landera, A.; Kaiser, R. I. *J. Phys. Chem. A* **2017**, *121*, 901–926.

(15) Bittner, J. D.; Howard, J. B. *Symp. Combust.* **1981**, *18*, 1105–1116.

(16) Mebel, A. M.; Georgievskii, Y.; Jasper, A. W.; Klippenstein, S. J. *Proc. Combust. Inst.* **2017**, *36*, 919–926.

(17) Kislov, V. V.; Sadovnikov, A. I.; Mebel, A. M. *J. Phys. Chem. A* **2013**, *117*, 4794–816.

(18) Frenklach, M.; Singh, R. I.; Mebel, A. M. *Proc. Combust. Inst.* **2019**, *37*, 969–976.

(19) Liu, P.; Li, Z.; Bennett, A.; Lin, H.; Sarathy, S. M.; Roberts, W. L. *Combust. Flame* **2019**, *199*, 54–68.

(20) Yang, T.; Kaiser, R. I.; Troy, T. P.; Xu, B.; Kostko, O.; Ahmed, M.; Mebel, A. M.; Zagidullin, M. V.; Azyazov, V. N. *Angew. Chemie Int. Ed.* **2017**, *56*, 4515–4519.

(21) Zhao, L.; Kaiser, R. I.; Xu, B.; Ablikim, U.; Ahmed, M.; Joshi, D.; Veber, G.; Fischer, F. R.; Mebel, A. M. *Nat. Astron.* **2018**, *2*, 413–419.

(22) Zhao, L.; Kaiser, R. I.; Xu, B.; Ablikim, U.; Ahmed, M.; Zagidullin, M. V.; Azyazov, V. N.; Howlader, A. H.; Wnuk, S. F.; Mebel, A. M. *J. Phys. Chem. Lett.* **2018**, *9*, 2620–2626.

(23) Zhao, L.; Kaiser, R. I.; Xu, B.; Ablikim, U.; Ahmed, M.; Evseev, M. M.; Bashkirov, E. K.; Azyazov, V. N.; Mebel, A. M. *Nat. Astron.* **2018**, *2*, 973–979.

(24) Park, J.; Burova, S.; Rodgers, A. S.; Lin, M. C. *J. Phys. Chem. A* **1999**, *103*, 9036–9041.

(25) Comandini, A.; Brezinsky, K. *J. Phys. Chem. A* **2011**, *115*, 5547–5559.

(26) Comandini, A.; Abid, S.; Chaumeix, N. *J. Phys. Chem. A* **2017**, *121*, 5921–5931.

(27) Durán, R. P.; Amorebieta, V. T.; Colussi, A. J. *J. Phys. Chem.* **1988**, *92*, 636–640.

(28) Kiefer, J. H.; Von Drasek, W. A.; Von Drasek, W. A. *Int. J. Chem. Kinet.* **1990**, *22*, 747–786.

(29) Benson, S. W. *Int. J. Chem. Kinet.* **1992**, *24*, 217–237.

(30) Colket, M. B.; Seery, D. J.; Palmer, H. B. *Combust. Flame* **1989**, *75*, 343–366.

(31) Kern, R.; Xie, K.; Chen, H.; Kiefer, J. *Symp. Combust.* **1991**, *23*, 69–75.

(32) Cremer, D.; Kraka, E.; Joo, H.; Stearns, J. A.; Zwier, T. S. *Phys. Chem. Chem. Phys.* **2006**, *8*, 5304.

(33) Mebel, A. M.; Kislov, V. V.; Kaiser, R. I. *J. Chem. Phys.* **2006**, *125*, 133113.

(34) Zádor, J.; Fellows, M. D.; Miller, J. A. *J. Phys. Chem. A* **2017**, *121*, 4203–4217.

(35) Norinaga, K.; Deutschmann, O. *Ind. Eng. Chem. Res.* **2007**, *46*, 3547–3557.

(36) Slavinskaya, N. A.; Mirzayeva, A.; Whitside, R.; Starke, J.; Abbasi, M.; Auyelkhankyzy, M.; Chernov, V. *Combust. Flame* **2019**, *210*, 25–42.

(37) Tao, H.; Wang, H.-Y.; Ren, W.; Lin, K. C. *Fuel* **2019**, *255*, 115796.

(38) Gao, C. W.; Allen, J. W.; Green, W. H.; West, R. H. *Comput. Phys. Commun.* **2016**, *203*, 212–225.

(39) Chu, T.-C.; Buras, Z. J.; Oßwald, P.; Liu, M.; Goldman, M. J.; Green, W. H. *Phys. Chem. Chem. Phys.* **2019**, *21*, 813–832.

(40) Frisch, M. J. et al. Gaussian 16, Revision C.01., 2016.

(41) Werner, H.-J. et al. MOLPRO, version 2015.1, a package of ab initio programs., Cardiff, UK, 2015.

(42) Grinberg Dana, A.; Liu, M.; Green, W. H. *Int. J. Chem. Kinet.* **2019**, *51*, 760–776.

(43) Georgievskii, Y.; Miller, J. A.; Burke, M. P.; Klippenstein, S. J. *J. Phys. Chem. A* **2013**, *117*, 12146–12154.

(44) Frisch, M. J. et al. Gaussian 09, Revision A.02., Wallingford, CT, 2009.

(45) Petersson, G. A.; Malick, D. K.; Wilson, W. G.; Ochterski, J. W.; Jr., J. A. M.; Frisch, M. J. *J. Chem. Phys.* **1998**, *109*, 10570.

(46) Norinaga, K.; Janardhanan, V. M.; Deutschmann, O. *Int. J. Chem. Kinet.* **2008**, *40*, 199–208.

(47) Goodwin, D. G.; Speth, R. L.; Moffat, H. K.; Weber, B. W. Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes., \url{https://www.cantera.org}, 2018.

(48) CHEMKIN-PRO 15151., 2015.

(49) Melius, C. F.; Miller, J. A.; Evleth, E. M. *Symp. Combust.* **1992**, *24*, 621–628.

(50) Durán, R. P.; Amorebieta, V. T.; Colussi, A. J. *Int. J. Chem. Kinet.* **1989**, *21*, 947–958.

(51) Saggese, C.; Sánchez, N. E.; Frassoldati, A.; Cuoci, A.; Faravelli, T.; Alzueta, M. U.; Ranzi, E. *Energy & Fuels* **2014**, *28*, 1489–1501.

(52) Fournet, R.; Bauge, J. C.; Battin-Leclerc, F. *Int. J. Chem. Kinet.* **1999**, *31*, 361–379.

Chapter 7

Recommendations for Future Work

In this thesis, I have focused on improving the fundamental infrastructure for modeling polycyclic aromatic hydrocarbons in RMG. In particular, significant effort has been placed on redesigning the entire resonance structure generation algorithm to provide more accurate representations of molecules. Contributions have also been made towards improving perception and handling of polycyclic structures where 3D geometry is important to consider, although there are still much to be explored in this area. Additionally, thermochemical and kinetic parameters for important PAH formation pathways have been added to the RMG database to allow RMG to find these pathways and estimate reasonable parameters. All of this work has culminated in the generation of an acetylene pyrolysis model which can predict formation of three- and four-ring PAHs such as phenanthrene and pyrene. Throughout this work, many challenges have been identified, related to automatic mechanisms generation in general as well as understanding PAH formation. While some of these challenges have been addressed, there still remains many which require attention. This chapter discusses several possibilities for further improving RMG and PAH formation models.

7.1 3D geometry considerations in RMG

One of the challenges which I encountered during this work was the lack of 3D geometry perception in RMG. In Chapter 4, I described work to encode information about ring membership into atom-level attributes. This provided a basic way to inform RMG about global structural properties at an atomic level. However, there is still much which can be done to improve and expand on this representation. With the ring attribute itself, the current implementation is easily extensible, so it would be interesting to explore whether more encoding more specific information such as the number and size of the rings could further aid in improving kinetics estimation.

However, I think that more substantial changes are also needed to address this challenge. One limitation of the ring attribute is that it still depends heavily on the existence of suitable training data. Even if RMG knows that a ring is involved in a particular reaction, it doesn't help if there isn't a data point for a similar reaction with a ring. Instead, there needs to be a heuristic method to evaluate the physical feasibility of an intramolecular reaction. With reactions like intramolecular addition or hydrogen migration, the transition state must bring the two reacting sites close to each other. If the two sites are sterically hindered and unable to reach each other, then the reaction should not be generated. One approach could be to use force-fields to estimate the energetic barrier to bringing the reacting sites close to each other. However, while force-field calculations are relatively cheap, they are still not on the desired order-of-magnitude for on-the-fly calculations during mechanism generation.

An alternative could be to implement a method to calculate geometric distances using VSEPR (valence shell electron pair repulsion) theory. Since RMG already has detailed atom and types, it could be possible to associate bond angles with each atom type and bond lengths with each bond type. This would enable a rough calculation of the minimum and maximum possible distance between any two atoms in the molecule. Heuristics could then be implemented to prevent reaction generation between atoms which are too far from each other. Depending on the implementation details, such an algorithm has the potential to be very fast. It could then be used to filter reaction generation for intramolecular reactions and potentially speed-up reaction generation by limiting it to only reasonable reactions.

Another 3D consideration which is currently completely neglected is stereochemistry, including cis-trans isomerism, conformational isomerism, and enantiomerism. In pyrolysis and combustion systems, cis-trans isomerism and conformational isomerism can significantly affect whether a reaction is possible. For example, some cyclization reactions in PAH formation pathways are only possible from the cis conformer, and hydrogen migration reactions in cyclic molecules may be impossible if the sites are in opposite axial positions. Enantiomers are generally disregarded in combustion chemistry, but are very important in pharmaceuticals, which are an active area of work in RMG. Implementation of these considerations in RMG would require infrastructure to represent them and perceive them for molecules, and methods for identifying when stereoisomerism is gained or lost as a result of a reaction.

7.2 Limiting reactivity by region within PAHs

When generating mechanisms for PAH formation, or mechanisms for large molecules in general, reaction generation often becomes the dominating factor in terms of computational cost.[1] This can be explained easily by the increase in number of possible reaction sites and resonance structures, given a rough scaling for the number of reaction generation calls as

$$n_{\text{reaction generation calls}} \sim \left(n_{\text{species}} \cdot \bar{n}_{\text{resonance structures}} \cdot \bar{n}_{\text{reaction sites}}\right)^2 \tag{7.1}$$

where $\bar{n}$ represents average values per species. The dependence on the number of resonance structures was the primary motivation for the implementation of resonance structure filtering, which was discussed in Chapter 3. The number of possible reaction sites is directly related to size of the reacting molecules, and to a slightly lesser extent, number of reactive features such as multiple bonds, radicals, or lone pairs.

For PAHs, the number of resonance structures has already been reduced with the implementation of Clar structures and resonance structure filtering, but the number of reaction sites is still very large. Thus, to improve performance when modeling PAH formation, one avenue may be to reduce the number of reaction sites being considered. PAHs are generally very stable, and are most reactive along their edges. This fact is used in work by Kraft *et al.* where they use a Monte Carlo approach to simulate PAH growth, focusing on reactions of various types of edge sites.[2] Doing so limits the number of possible reactions and improves performance, enabling simulation up to very large molecules.

In the case of RMG, this particular approach of limiting reactivity to edges has not been particularly applicable until now, given that it has not be able to simulate the formation of sufficiently large PAHs. However, in order to further extend RMG's capabilities to even larger PAHs, such considerations will definitely be necessary to maintain feasible computational costs. Implementing such an approach would be challenging, but could perhaps be done by adding an atom-level "reactive" flag to allow marking certain atoms as non-reactive, causing them to be ignored when matching reaction templates.

7.3 DLPNO-CCSD(T) calculations for PAHs

One major challenge of studying PAH chemistry is the general lack of accurate thermochemical and kinetic data. The most common way to obtain these parameters is to use quantum chemistry methods to obtain the electronic structure of a molecule, from which its thermochemistry can be calculated. If transition state geometries can be found, then rate coefficients can also be calculated. However, the accuracy of these calculations depends heavily on the method and basis sets used, and the large size of PAH molecules limits the methods which are computationally feasible. One of the best methods for obtaining accurate energies is CCSD(T)-F12, which is only feasible for very small molecules, up to around 8 heavy atoms, due to poor $O(N^7)$ scaling with molecule size. CBS-QB3, a composite method which is very popular due to its good accuracy/cost ratio, is feasible for up to around 16 heavy atoms. This means that calculations could be performed for species up to pyrene, though the calculations will take a significant amount of time which would be undesireable for exploring a large potential energy surface.

A promising solution to these challenges is the set of DLPNO (domain-based local pair natural orbital) methods,[3, 4] which are implemented in the ORCA software for quantum calculations.[5] The basic idea behind DLPNO methods is to reduce the total number of electron orbitals under consideration by localizing them to parts of the molecule. Available methods include DLPNO-CCSD(T) and DLPNO-CCSD(T)-F12, which can achieve similar results to conventional CCSD(T) while providing near-linear scaling.[4] Paulechka et al. evaluated DLPNO-CCSD(T), showing that DLPNO-CCSD(T)/def2-QZVP//B3LYP-D3(BJ)/def2-TZVP has an uncertainty of about 3 kJ/mol for their test set of molecules up to biphenyl.[6]

Up to now, there have been many detailed studies of potential energy surfaces up to C_{16} using methods like CBS-QB3 or G3(MP2,CC). For molecules larger than that, most available studies have been performed using B3LYP or comparable DFT methods, which are computationally reasonable, but are not sufficiently accurate to provide good thermochemistry and rate constants. DLPNO-CCSD(T) has the potential to be a powerful and accurate tool for investigating this next range of PAH chemistry, going from three- and four-ring molecules to six- and seven-ring molecules.

7.4 Code organization and optimization

RMG has experienced very rapid growth in the last few years. RMG-Py development began in 2008, and in 2016, RMG v1.0 was released with 60,000 lines of code. In 2019, RMG v3.0 was released with 120,000 lines of code. As the code base and user base continues to grow, it becomes more and more important that the code is clean and maintainable. This is necessary to streamline future code development and can also have performance implications.

Broadly, there are three kinds of general code organization issues which exist within RMG. One issue is code which is completely unused, left over from features which were updated or refactored. A second issue is unnecessary code which is executed (thereby contributing to computation time) but not for useful purpose. This can easily happen when changes or additions are made without cleanly removing old code. A third issue is convoluted code which can result from ad hoc addition of code or be left over from refactoring. The challenge is that all these issues can be difficult to detect, and while the first two issues can be easily resolved by removing code, the third issue requires additional refactoring. For identifying the first issue, code coverage reports can be useful, as they will indicate exactly which lines of code are executed or not. Of course, the coverage report will depend on the particular task being performed. Identifying the second and third issues will generally require reading the code and understanding the broader context of what is being done.

The second and third issues are also related to code optimization. One of the early weaknesses of RMG-Py was that it was much slower than RMG-Java. Since then, RMG-Py has become even slower, as a trade-off with improved accuracy. Historically, performance has not been a key concern with each individual change. Instead, minor performance losses build up until they become very noticeable, at which point effort is placed in searching for ways to improve performance. An alternative workflow would be to increase the importance of performance evaluation when implementing changes. With any change that has potentially has performance effects, performance tests should be run on a relevant test case before approving the change. While this should become the standard workflow for future development, it does not account for existing inefficiencies in the code. First, profiling should be used to identify parts of the code which contribute significantly to runtime. Generally, reaction generation and reactor simulation tend to be the two most expensive tasks. One potential avenue for optimizing these tasks is using Cython. While many parts of RMG are already cythonized, the speed-up factor which can be attained by converting Python to Cython ranges from 2x to over 100x depending on how C-like the code is. Unfortunately this means that very pythonic idioms (such as list comprehensions) actually hurt Cython performance. The majority of cythonized code in RMG is essentially pure Python, which generally results in 2x speed-up. Therefore, there is substantial room for improvement by improving the efficiency of cythonized code.

7.5 RMG input/output improvements

Input and output files are an important aspect of using RMG, but there are a number of deficiencies with the current system. Currently, input files and database files are all Python syntax files which are loaded by executing the file using the Python interpreter. This was an attractive option because Python syntax is very human-readable and this provided an easy way to read the files. However, executing Python files can pose significant security risks, especially in the case of the RMG-website, which was previously subject to code-injection attacks. Additionally, this approach directly links input file syntax with code structure, which can cause issues with backward compatibility of input files when updating code. Reaction mechanisms are primarily output as Chemkin files, with metadata stored as comments. This has lead to many issues with recovering the metadata by parsing the comments, which also has many backward compatibility issues.

With that said, there is an opportunity to implement a new file format which can unify the existing files for RMG input files, database files, and output mechanisms. Such a file could use YAML, which is a markup language useful for storing data in a human-readable way. YAML files can be easily loaded into RMG without the security concerns of executing Python files. It has already been implemented in RMG for storing data in Arkane jobs. The next step would be to implement it as a format for mechanisms, which could easily be adapted to store database data, and input files. Cantera has already developed a new YAML format for chemical mechanisms, so it would be ideal if RMG could adopt the same syntax. Making this transition to YAML would provide a number of notable benefits: 1) a consistent, universal format across databases and mechanisms, 2) more robust metadata storage, 3) code-independent file formats, and 4) improved security.

References

(1) Jocher, A.; Vandewiele, N. M.; Han, K.; Liu, M.; Gao, C. W.; Gillis, R. J.; Green, W. H. *Comput. Chem. Eng.* **2019**, 106578.

(2) Celnik, M.; Raj, A.; West, R.; Patterson, R.; Kraft, M. *Combustion and Flame* **2008**, *155*, 161–180.

(3) Riplinger, C.; Neese, F. *J. Chem. Phys.* **2013**, *138*, 034106.

(4) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. *J. Chem. Phys.* **2013**, *139*, 134101.

(5) Neese, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2018**, *8*, e1327.

(6) Paulechka, E.; Kazakov, A. *J. Phys. Chem. A* **2017**, *121*, 4379–4387.