REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
30-06-2022	Final Report	1-Jul-2020 - 31-Oct-2021

4. TITLE AND SUBTITLE
Final Report: Radical Artificial Intelligence for Multiphase Environmental Systems

5a. CONTRACT NUMBER
W911NF-20-1-0172

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
611102

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of California - Irvine
141 Innovation Drive, Suite 250

Irvine, CA 92697 -7600

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
75368-CH-H.2

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Ann Marie Carlton
UU	UU	UU	UU		19b. TELEPHONE NUMBER 949-824-5651

RPPR Final Report
as of 05-Jul-2022

Agency Code: 21XD

Proposal Number: 75368CHH **Agreement Number: W911NF-20-1-0172**
INVESTIGATOR(S):

Name: Ann Marie Carlton
Email: agcarlto@uci.edu
Phone Number: 9498245651
Principal: Y

Name: David Van Vranken
Email: dlvanvra@uci.edu
Phone Number: 9498245455
Principal: N

Name: PIERRE BALDI
Email: pfbaldi@uci.edu
Phone Number: 9498245809
Principal: N

Organization: **University of California - Irvine**
Address: 141 Innovation Drive, Suite 250, Irvine, CA 926977600
Country: USA

DUNS Number: 046705849 EIN: 952226406
Report Date: 31-Jan-2022 Date Received: 30-Jun-2022
Final Report for Period Beginning 01-Jul-2020 and Ending 31-Oct-2021
Title: Radical Artificial Intelligence for Multiphase Environmental Systems
Begin Performance Period: 01-Jul-2020 **End Performance Period:** 31-Oct-2021
Report Term: 0-Other

Submitted By: Ann Marie Carlton Email: agcarlto@uci.edu
Phone: (949) 824-5651

STEM Degrees: 0 **STEM Participants:** 6

Major Goals: The major goals of this project were to
1.) instruct a deep learning system with radical training reactions
2.) combine radical and polar rule-based chemistry for single step reactions in a deep learning system
3.) build on the single step approach for combined radical and polar chemistry in multi-generation oxidation
4.) test and evaluate the deep learning system at all stages of model development.

Accomplishments: We have constructed an initial data set of radical reactions, we call "RMechDB". We prioritized and are in the process of publishing the dataset because when presenting at annual ARO meeting, other grantees expressed interest in the dataset itself to help provide structural information for their spectra.

RMechDB consists of more than 5,500 pedagogically chosen elementary radical mechanistic steps based on published transformations. The majority of the published mechanistic steps had to be further decomposed into elementary reaction steps with individual transition states. Over 880 steps were taken from eight introductory organic chemistry textbooks, advanced organic chemistry books and an atmospheric chemistry textbook. Over 800 reactions were taken from the primary research literature including mechanisms for common synthetic transformations (atom transfer, tin chemistry, radical cyclizations), autoxidation, atmospheric reactions, and explosives. The literature mechanisms also included steps leading to 14 common industrial polymers: ethylene, propylene, butadiene, chloroprene, isoprene, acrylamide, acrylic acid, methyl acrylate, ethyl acrylate, butyl acrylate, methyl methacrylate, acrylonitrile, styrene, pmethylstyrene,
vinyl chloride, vinyl fluoride, tetrafluoroethylene, chlorotrifluoroethylene, vinylidene
fluoride, vinyl acetate, N-vinylpyrrolidinone. The conditions for polymerization, often including more than one type of initiator, were taken from the research literature and are not necessarily the proprietary initiators and conditions used for industrial synthesis. The data from textbooks and research literature are considered the core of the

RMechDB database. The core data set has been augmented with a large number of mechanistic steps related to atmospheric oxidation of organic molecules. A large number (847) of steps were taken from a comprehensive review of atmospheric isoprene oxidation that traced the fate of each individual carbon atom detailing the highly branched pathways from reaction with HO•, O2, NO, Cl• and other species (Wennberg et al., Chemical Reviews) For simplicity, we focus on daytime atmospheric chemistry of isoprene at atmospherically relevant conditions (average atmospheric T=278 K), neglecting elementary steps involving NO3, which is a dominant nighttime oxidant. Most of the elementary steps were inferred from composite transformations. About 3,000 mechanistic steps were coded from the first two stages of the major oxidation pathways in the Master Chemical Mechanism (MCM).13 The MCM contains mechanisms for atmospheric oxidation of 143 volatile organic compounds initiated by both HO• and NO3, including reactions of isoprene. Steps more than ten times slower than the fastest process (with the same reactants) were also excluded. Steps second-order in reactive intermediates were excluded on the assumption that they would not slow under typical conditions. For both the Wennberg and MCM steps, transformations initiated by pericyclic [3+2] cycloaddition of O3 with alkenes were excluded from this initial data set, but depicting the cycloaddition as a diradical process could be an expedient.34 Photolysis steps were also excluded. Any steps left out of this initial data set can be introduced in the future. Theindividual mechanistic steps fall into one of seven different categories: homolysis, recombination, abstraction, addition to pi bonds, retro-addition to pi bonds, pi (e.g., allylic) and alpha lone pair resonance (e.g., ketyls). In RMechDB, resonance is represented as a mechanistic step, even though there is no transition state. Homolysis and recombination are the mechanistic reverse of each other, like addition and retro-addition. Alpha resonance is represented with a single curved half-arrow, but it is acknowledged that the half-arrow falsely implies formation of a partial double bond.

Training Opportunities: Five graduate students and one undergraduate student worked on developing reactions for RMechDB. We met every week (online during the pandemic) and students took turns 'sharing their screen' and leading discussions.

Results Dissemination: We are in the process of submitting an article to J. of Chemical Information and Modeling. The draft has been uploaded. We have nearly finished public website to share the data freely. (There is currently an internal UCI site).
Yinting Chiu will present on this work later in the year at the ACS meeting. I will share her slides with ARO when they are completed.

Honors and Awards: Ann Marie Carlton was promoted to Full Professor during the time period of the grant and elected to be Vice Chair of the Chemistry Department.

Protocol Activity Status:

Technology Transfer: Nothing to Report

PARTICIPANTS:

Participant Type: PD/PI
Participant: Ann Marie Grover Carlton
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: David VanVraken
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Co PD/PI
Participant: Pierre Baldi

Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yinting Chiu
Person Months Worked: 6.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Yuzo Kanomato
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Alexnder Shmakov
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

Participant Type: Graduate Student (research assistant)
Participant: Amin Tavolkoli
Person Months Worked: 1.00 **Funding Support:**
Project Contribution:
National Academy Member: N

ARTICLES:

RPPR Final Report
as of 05-Jul-2022

Publication Type: Journal Article Peer Reviewed: Y **Publication Status:** 5-Submitted
Journal: Journal of Chemical Information and Modeling
Publication Identifier Type: Publication Identifier:
Volume: Issue: First Page #:
Date Submitted: 6/27/22 12:00AM Date Published: 12/31/22 5:00AM
Publication Location:
Article Title: RMechDB: A Public Database of Elementary Radical Reaction Steps
Authors: Mohammadamin Tavakoli, Yin Ting T. Chiu, Pierre Baldi, Ann Marie Carlton, David Van Vranken
Keywords: deep learning, chemical prediction, radicals,
Abstract: We introduce RMechDB, a public database and web-server of elementary radical reaction steps, each with a single transition state at or around room temperature. The initial data set of RMechDB consists of over 5,500 manually curated plausible arrow-pushing steps for organic radical reactions. The steps were taken from a variety of sources. About 2,000 mechanistic steps were taken from textbooks and research publications. Another 3,000 were taken from gasphase atmospheric reactions of isoprene and other organic molecules in the Master Chemical Mechanism v3.3.1. Reactions are encoded in SMIRKS format with accurate atom mapping and annotations for arrow-pushing mechanisms. The RMechDB web-server includes an interactive search interface and a form for submitting a download request. It also offers an interactive interface for submitting new reactions to expand the data set through community contributions. Although there are several applications for RMechDB, it is primarily designed as a cent
Distribution Statement: 2-Distribution Limited to U.S. Government agencies only; report contains proprietary info
Acknowledged Federal Support: **Y**

Partners

,

I certify that the information in the report is complete and accurate:
Signature: Ann Marie Carlton
Signature Date: 6/30/22 11:30AM

RMechDB: A Public Database of Elementary Radical Reaction Steps

Mohammadamin Tavakoli,[†,¶] Yin Ting T. Chiu,[‡,¶] Pierre Baldi,[*,†] Ann Marie Carlton,[‡] and David Van Vranken[*,‡]

†*Department of Computer Science, University of California, Irvine, Irvine, California 92697, United States*

‡*Department of Chemistry, University of California, Irvine, Irvine, California 92697, United States*

¶*Equal Contribution*

E-mail: pfbaldi@uci.edu; david.vv@uci.edu

Abstract

We introduce RMechDB, a public database and web-server of elementary radical reaction steps, each with a single transition state at or around room temperature. The initial data set of RMechDB consists of over 5,500 manually curated plausible arrow-pushing steps for organic radical reactions. The steps were taken from a variety of sources. About 2,000 mechanistic steps were taken from textbooks and research publications. Another 3,000 were taken from gas-phase atmospheric reactions of isoprene and other organic molecules in the Master Chemical Mechanism v3.3.1. Reactions are encoded in SMIRKS format with accurate atom mapping and annotations for arrow-pushing mechanisms. The RMechDB web-server includes an interactive search interface and a form for submitting a download request. It also offers an interactive interface for submitting new reactions to expand the data set through community contributions. Although there are several applications for RMechDB, it is primarily designed as a central platform for aggregating, curating, and distributing reliable data about elementary radical reaction steps for both machine learning- and rule-based reaction models.

Introduction

A free radical is a chemical compound (e.g. atom, molecule) with at least one half occupied orbital. The presence of the half occupied orbitals make a radical compound highly reactive. Because of this high reactivity, free radicals have the potential to both serve as powerful chemical tools and be extremely harmful contaminants. Chemical reactions involving a free radical are radical reactions which are an essential part of synthetic, biochemical, atmospheric, and plasma chemistry. For instance, the climate crisis has dramatically altered fire activity worldwide. Wild land fires are increasing in frequency, duration, intensity, and size. The chemistry of flames is dominated by radical reactions and the chemical composition of fire smoke changes during atmospheric transport. This so-called "aging" of smoke is poorly understood, but known to be largely driven by free radical processes. As another example from the pharmaceutical industry, the composition of drug formulations changes gradually upon storage. As a result, all drug companies are required to study those changes through forced degradation studies under several conditions, including photochemical and oxidative conditions, which mostly involve radical reactions. Thus, it is of great importance to study the chemistry of radical reactions and their outcomes.

During the past few years, data driven methods such as deep learning have provided new powerful tools for addressing chemoinformatics problems.[?] Due to important applications ranging from automated drug discovery to computer-aided synthetic chemistry, there has been an increasing interest in developing deep learning models to predict the outcome of chemical reactions. While the deep learning models have been evolving in sophistication and complexity, a major stumbling block has remained the lack of comprehensive, public, reaction data.[1] The majority of current models is being trained using the data set of chemical transformations from the US Patent office, as well as a few other smaller data sets. However, these data sets come with significant limitations in terms of overall size, chemistry coverage and balance, and lack of meta-data, atom-mapping, reactant or product balance, and elementary reaction step information. For instance, the USPTO data set of chemical reactions restrictively represent chemical reactions in the form of overall transformations leading to one single major product. It contains little information about underlying mechanisms, and

about key intermediates and side products. Furthermore, radical reactions are hard to extract and appear to be underrepresented. In contrast, radical reactions often proceed through a complex series of chemical steps and highly branched mechanistic pathways. Developing an accurate machine learning model to predict the outcome of radical reactions requires a training data set of purely radical reactions with information about the mechanistic pathways and intermediate products. Here we introduce RMechDB as a central platform for aggregating, curating, and distributing elementary step radical reaction. RMechDB is available in the form of an online database and interactive web server where users can search, filter, download, and insert elementary step radical reactions. The initial version of the RMechDB data set consists of over 5500 manually curated radical reaction and is accessible through the DeepRXN platform at `https://deeprxn.ics.uci.edu/rmechdb`.

Mechanistic Pathways VS Overall Transformations

The term reaction can be ambiguous and is most commonly used to describe either: 1) a chemical transformation with reactants, products, chemical conditions, and yields; or 2) a single step in an arrow-pushing mechanistic pathway. Thus in this work, instead of using the term reaction which is rather vague, we use the more specific terms of transformation and elementary step. Every mechanistic pathway can be decomposed into a series of discrete elementary steps, each with a single transition state. In several aspects, it is advantageous to show every step in a mechanistic pathway. When all the steps in a pathway are elementary, there is no chance of missing key intermediates that give rise to competing pathways during chemical transformation. For example, when the transformation of ISOPAO to C524O2 is depicted as a one-step process, it misses the potential for the allyl radical intermediate to form an isomeric peroxy radical and downstream products (Figure 1). Another advantage to mechanistic pathways based on elementary reaction steps, is that they can be described using curved half-arrows that correspond to the interaction of a singly occupied molecular orbitals with a HOMO and/or LUMO[2]. The curly arrows, also known as electron flow specifications or arrow-pushing mechanisms, are depicting the interaction

between molecular orbitals. This representation of elementary steps is highly informative and, when elementary steps are chained together, an interpretation of the corresponding transformation can readily be derived. This is important for deep learning approaches to reaction prediction for at least three reasons. First, prediction of mechanistic pathways leads to prediction that are interpretable. Interpretability is a an important consideration in machine learning, especially for so-called "black-box" approaches such as deep learning. Second, when machine learning models operate at the level of elementary steps, balance between reactants and products is always preserved together with the underlying atom mapping. And third, all intermediary and final products can be accounted for, which is an important consideration in synthetic chemistry applications.

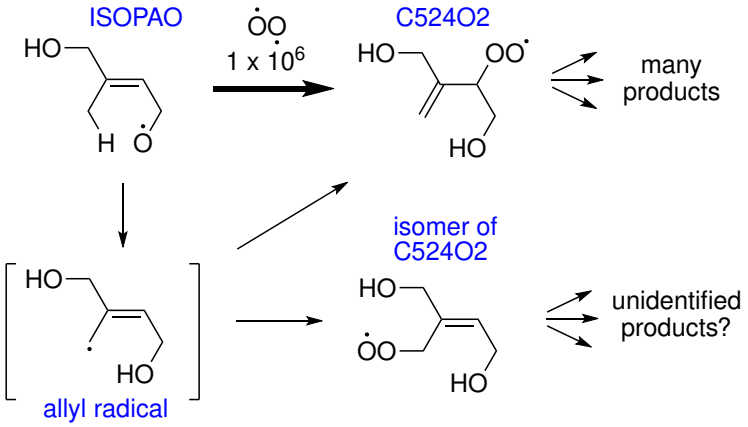

Figure 1: Missing steps and intermediates prevent identification of products. The formation of an allyl radical was not depicted for the transformation of ISOPAO to C524O2 in the MCM. It isn't clear why the missing allyl radical intermediate would not also generate an isomer of C524O2 and account for more downstream products.

Data Driven Models to Predict the Outcome of Chemical Reactions

An open-source, publicly available database of pedagogical elementary reaction steps will facilitate training and development of tools for prediction of reaction mechanisms. There are two common approaches to the prediction of stepwise mechanisms of organic transformations using databases of

elementary reaction steps. The quantitative approach uses a database of kinetic and thermodynamic parameters to accurately predict the products of the reactions and the pathways by which they form. This approach [RMG, MCM, atmospheric chemistry] is not restricted to elementary reaction mechanisms, but it does require kinetic parameters. The approach is best applied to cases where the product structures are known but the abundances are not known. The qualitative approach [Reaction Predictor] is to use a database of diverse plausible (fast at or below 100 °C) mechanistic steps, to match chemical structures (and mechanistic pathways) to [mysterious/unknown/not structurally characterized] analytes in readily available spectra or chromatograms. This approach is best applied when the abundance is known, but the chemical structure is unknown. Chemical structure can provide powerful insight into biological effects, phase partitioning, and reactivity under changing reaction conditions. Public databases of mechanistic steps will empower the use of machine learning to create tools that assign chemical structures and mechanisms to products of environmental, synthetic, and environmental transformations of organic compounds.

Existing Data Sets of Elementary Reaction Steps

Readers may question the need for databases of mechanistic steps since there are several large databases of organic transformations such as REAXYS, SciFinder, and the Open Reaction Database (ORD). Those databases are composed of recipes that describe reactants, conditions, yields, and a list of products that rarely sums to 100%. The proprietary REAXYS database currently has over 57 million transformations. The SciFinder Scholar database has over 126 million transformations, which includes sequential reactions. Organic transformations were mined from US Patents from 1976-2016 and are publicly available. The growing public database ORD already gathers about 2 million chemical transformations from other available sources.[3] These powerful databases of chemical transformations allow synthetic organic chemists, or systems trained with machine learning,[4] to plan out synthetic routes composed of sequential laboratory experiments, but the data don't reveal the underlying mechanisms of any individual transformations. Databases of transformations are not

new; and neither is the application of AI to planning of synthetic routes. Why is there no database of elementary arrow-pushing reaction steps? Sadly, when curved arrows were first introduced in 1922,[5,6] the connection between curved arrows, frontier orbitals and transition states, was not recognized, so there was no incentive to apply them solely to elementary mechanistic steps. As a result, curved arrow mechanisms and half-arrow radical mechanisms have been used inconsistently, throughout the organic chemistry literature and are rendered in graphical forms that are not easily recoverable through data mining. Reaction Mechanism Generator (RMG) supports the only existing database of elementary mechanistic reaction steps. RMG predicts mechanistic pathways through a quantitative approach, using thermochemical and kinetic parameters to model species concentrations and rates for each step.[7] RMG is supported by a searchable database, consisting of 98 families of reaction types.[8] Almost half (40/98) of the reaction families in the current RMG database involve radicals. About a fourth of the reaction families supported by RMG do not correspond to elementary reaction steps at or around room temperature (e.g., unimolecular keto-enol tautomerization). The current RMG data set currently contains over 30,000 elementary mechanistic steps in a searchable database. Most of the mechanistic steps and kinetic data were developed to support high temperature processes up to 2000 K. The estimated kinetic parameters are expected to be less accurate at lower temperature.[9] Many of the steps would be implausibly slow at room temperature. For example, the kinetic parameters for homolysis of a CH_3 group from isoprene would proceed with a half-life of over 10^{42} years. Many of the steps that proceed through a single transition state at high temperatures (e.g., over 1500 K) would involve more than one mechanistic step at room temperature.[8] For example, at room temperature, the addition of HO• to the double bond of alpha-pinene should not be concerted with ring opening. The requirement for accurate thermochemical and kinetic creates a major hurdle for applications involving complex organic structures. RMG development has so far been focused on processes involving simple reactants with just a single organic functional group and up to one heteroatom: CH_4, CH_3CH_3, $CH_3CH_2CH_3$, exo-tetrahydrodicyclopentadiene, $C_{10}H_{16}$, CH_3OCH_3, $CH_3(CH_2)_3OH$, $CH_3(CH_2)_5CH_3$, $((CH_3)_2CH)_2CO$, $CH{=}CHCH{=}CHCH_2CH_3$, $HCC(CH_2)_4CCH$, $C_6H_5(CH_2)_5CH_3$, $(CH_3)_2CHCH_2OH$, $CH_3(CH_2)_4CH_3$, $H_2NCH_2CH_3$, and $((CH_3)_3C)_2S$, C_6H_5OH.

The NIST Chemical Kinetics Database is another rich source of mechanistic information for radical reactions.[10] For example, a search of reactions of HO• yields over 3,390 entries (11,000 records), but over half (1,811) were for transformations with unspecified products. Other entries (e.g., CH_2O + $\cdot OH \rightarrow$ HCOOH + H$\cdot$)[11] clearly involve multi-step transformations. The Chemical Kinetics Database is not currently searchable by chemical structure or substructure. M-CSA (Mechanism and Catalytic Site Atlas) is a searchable database of plausible enzymatic reaction mechanisms.[12] Almost half (401 out of 944) of the entries have at least one plausible multi-step mechanism, although some of the mechanistic steps are not elementary. M-CSA is a rich source of information but the mechanistic steps within an enzyme active site are not necessarily applicable to reactions of freely diffusible species. The Master Chemical Mechanism,[13 14 15 16 17 18 19] contains an extensive set of mechanisms and rate parameters for the reaction of a wide range of volatile organic reactants with species such as HO, O_2, NO, O_3, NO_3 and light. Most of the reactions are composite processes and do not depict the underlying elementary mechanistic steps. Similarly, Wennberg and coworkers have published a nearly exhaustive set of mechanistic pathways for the atmospheric oxidation of isoprene.[20] Moreover, many of the composite processes in those isoprene pathways are not decomposed into elementary reaction steps.

RMechDB: Underlying Data Set

[TODO: change the database to data set wherever needed]

A Data Set of Radical Elementary Steps

Organic transformations in databases such as REAXYS, SciFinder and ORD are easily validated because published products are rigorously characterized using convenient spectroscopic techniques such as mass spectrometry, NMR and IR. In contrast, mechanistic steps with one transition state are not easily validated. Experimental proof of a mechanistic step usually requires electronic structure calculations and/or laborious experimental tools such as chemical kinetics, isotopic

labeling, crossover experiments, etc. It is often quoted that one can never prove a mechanism, but only dis-prove the plausible alternatives.[21] We set out to construct a database of plausible elementary reaction steps, which are useful to chemists in constructing mechanistic pathways and predicting byproducts of organic reactions. Plausibility is subjective. For RMechDB, we define an elementary mechanistic step as plausible if a half-life of a day or less is expected at room temperature under the conditions cited. If more than one pathway has been postulated in the literature, it is expedient to include steps from both potential pathways in the data set until the discrepancy is resolved. That way, any pathway proposed using the data will reflect the ambiguity in the body of literature. In theory, the plausibility of any elementary reaction step can ultimately be validated using electronic structure calculations. Resonance interconversions are depicted as mechanistic steps.

Structure of the Data

In previous work, a limited dataset of about 100 radical mechanistic steps was created and used to train Reaction Predictor with machine learning to rank plausible steps from new query reactants.[22] Entries consisted of elementary reaction steps in the SMIRKS format including atom mapping and electron flow specification (corresponding to curved half-arrows) introduced for training Reaction Predictor. In this work we have created the first large publicly available database of plausible elementary radical reaction steps in the Reaction Predictor SMIRKS format. Data fields are included for temperature (blank if 298 K; "heat" if unspecified). A field is included for reactions involving light. An additional data field is included for the scholarly source of information. Additional fields can be added later for important parameters such as phase, solvent, wavelength, enthalpy change, etc.

Composition of the Current Data Set

The initial data set in RMechDB consists of over 5,500 pedagogically chosen elementary radical mechanistic steps based on published transformations. The majority of the published mechanistic steps had to be further decomposed into elementary reaction steps with individual transition states.

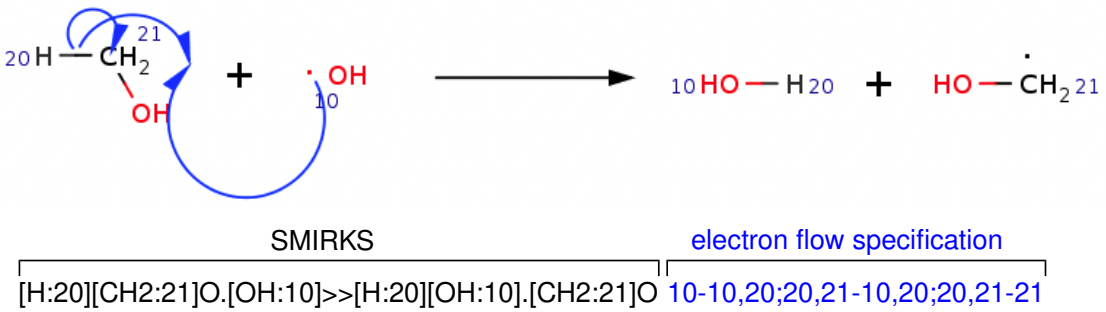

Figure 2: RMechDB format for depicting reactions and arrow pushing mechanisms.

Over 880 steps were taken from eight introductory[23 24 25 26 27 28 29 30] organic chemistry textbooks, advanced organic chemistry books[31 32], and an atmospheric chemistry textbook.[33] Over 800 reactions were taken from the primary research literature including mechanisms for common synthetic transformations (atom transfer, tin chemistry, radical cyclizations), autoxidation, atmospheric reactions, and explosives. The literature mechanisms also included steps leading to 14 common industrial polymers: ethylene, propylene, butadiene, chloroprene, isoprene, acrylamide, acrylic acid, methyl acrylate, ethyl acrylate, butyl acrylate, methyl methacrylate, acrylonitrile, styrene, p-methylstyrene, vinyl chloride, vinyl fluoride, tetrafluoroethylene, chlorotrifluoroethylene, vinylidene fluoride, vinyl acetate, N-vinylpyrrolidinone. The conditions for polymerization, often including more than one type of initiator, were taken from the research literature and are not necessarily the proprietary initiators and conditions used for industrial synthesis. The data from textbooks and research literature are considered the core of the RMechDB database. The core data set has been augmented with a large number of mechanistic steps related to atmospheric oxidation of organic molecules. A large number (847) of steps were taken from a comprehensive review of atmospheric isoprene oxidation that traced the fate of each individual carbon atom detailing the highly branched pathways from reaction with HO•, O_2, NO, Cl• and other species.[20] For simplicity, we focus on daytime atmospheric chemistry of isoprene at atmospherically relevant conditions (average atmospheric T=278 K), neglecting elementary steps involving NO_3, which is a dominant nighttime oxidant. Most of the elementary steps were inferred from composite transformations. About 3,000 mechanistic steps were coded from the first two stages of the major oxidation pathways in the

Master Chemical Mechanism (MCM).[13] The MCM contains mechanisms for atmospheric oxidation of 143 volatile organic compounds initiated by both HO• and NO_3, including reactions of isoprene. Steps more than ten times slower than the fastest process (with the same reactants) were also excluded. Steps second-order in reactive intermediates were excluded on the assumption that they would not slow under typical conditions. For both the Wennberg and MCM steps, transformations initiated by pericyclic [3+2] cycloaddition of O_3 with alkenes were excluded from this initial data set, but depicting the cycloaddition as a diradical process could be an expedient.[34] Photolysis steps were also excluded. Any steps left out of this initial data set can be introduced in the future. The individual mechanistic steps fall into one of seven different categories: homolysis, recombination, abstraction, addition to pi bonds, retro-addition to pi bonds, pi (e.g., allylic) and alpha lone pair resonance (e.g., ketyls). In RMechDB, resonance is represented as a mechanistic step, even though there is no transition state. Homolysis and recombination are the mechanistic reverse of each other, like addition and retro-addition. Alpha resonance is represented with a single curved half-arrow, but it is acknowledged that the half-arrow falsely implies formation of a partial double bond. The steps in radical chain mechanisms are often classified as initiation, propagation, or termination steps, but many transformations involving radicals do not involve chain mechanisms. Homolysis is a typical chain initiation step. Atom abstraction, addition, retro-addition, and resonance are typical chain propagation steps. Recombination is a typical chain termination step.

Additional Information Useful to Users Using the Data Set for Training

A test set that consists of 500 reaction steps was created to use as a comparison point for our system. 450 mechanistic steps were taken out of the initial training set and put into the test set. An additional 50 mechanisms that show the first two steps of the oxidation of myrcene (7-methyl-3-methyleneocta-1,6-diene), a compound of interest, were also put into the test set.

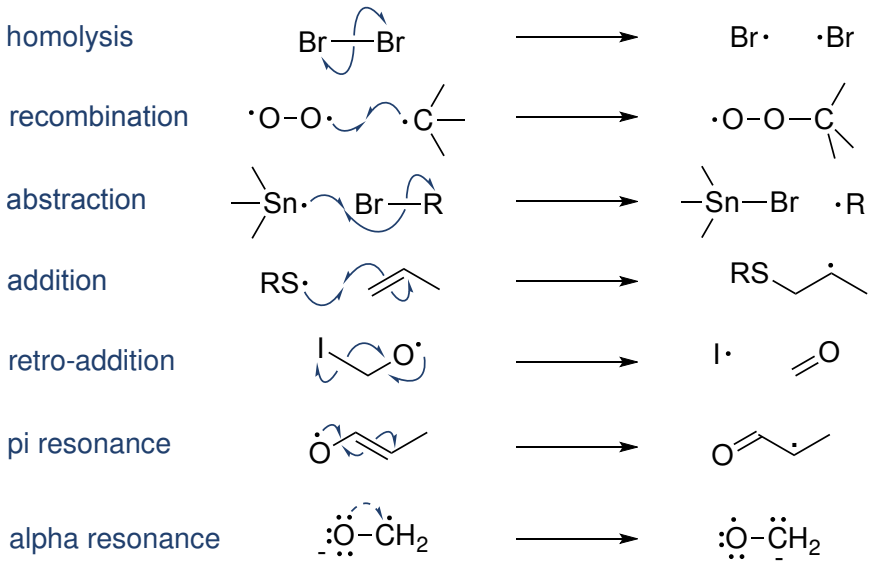

Figure 3: Seven different categories of mechanistic steps involving radicals.

Biasing a Dataset Against Unproductive Processes Using Spectator Species

Systems such as Reaction Predictor are not designed to take species concentrations into account, even though elementary reaction rates depend on both rate constants and species concentrations. In many cases, users may want to artificially bias the training data to avoid unproductive reactions. For example, individual reactions such as HO• + H–OO•; •OO• + H–OO•, and •OO• + •N=O maybe be highly plausible when the two reactants are present at equimolar concentration; but for the atmospheric oxidation of organic molecules, the reactions between these species would be unproductive (Fig. 4A). Furthermore, the chain terminating recombination of two radicals such as R• and •OH, would be slow due to their low concentrations. The training set can be biased to minimize these kinds of unproductive steps by creating additional copies of the training steps that include these species as spectators (Fig. 4B). To empower this approach, a duplicate set of the atmospheric training steps was designed to include spectator species: HO•, HOO•, O_2, •NO, and •NO_2 even when not involved in the bond-forming process. These species are common in the atmosphere, recorded in the literature and included the MCM database were considered sufficiently atmospherically relevant.[13,20] A system trained with this additional spectator dataset would be better able to avoid unproductive pathways involving reactions between these species.

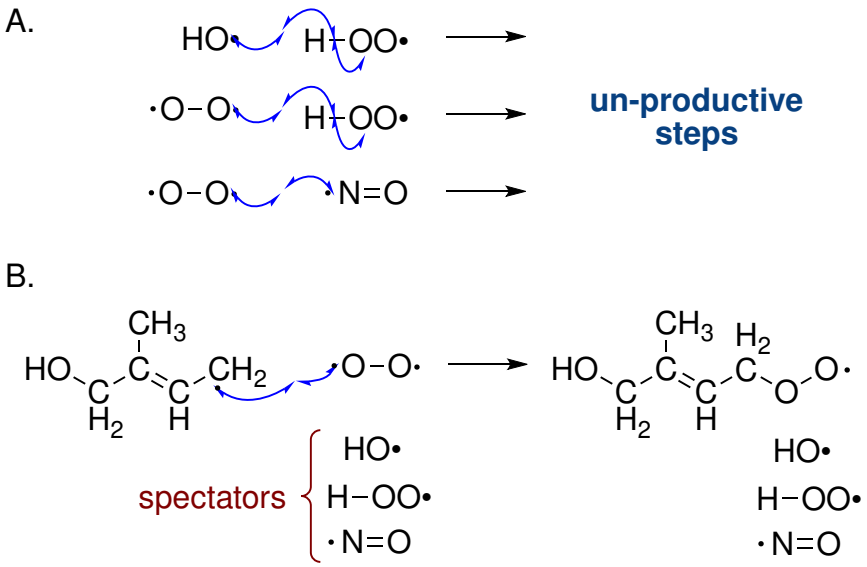

Figure 4: A. Many unproductive processes would be plausible if the species were present at high equimolar concentrations. B. Inclusion of spectator species throughout a data set will bias training to avoid unproductive bond-forming events.

RMechDB: Online Database

RMechDB is available online and accessible as a part of the DeepRXN platform. It contains the first publicly available data set of radical reactions in the form of elementary steps designed for computer-aided studies of radical chemistry and in particular the training of interpretable predictive models using machine learning methods.

The initial version of RMechDB contains the 5500+ radical elementary step reactions described above, categorized into two major classes of: core reactions (textbook and literature), and specific reactions (atmospheric). The scheme of the database in RMechDB consists of three fundamental models: (1) `Reaction`, (2) `Molecule`, and (3) `Atom`. The integration of these three models allows for fast and efficient search and reaction retrieval. Several properties are associated with each instance of these models, and in particular with the instances of the `reaction` model. For the `reaction` model, these properties are as follows:

1. **Reaction ID:** Each reaction is associated with a unique and searchable ID number.

2. **Canonicalized atom mapped SMILES of the reactants:** The SMILES string of the reactants

molecules, with integer labels for atoms that are participating the reaction. We use a labeling convention where the labels of the participating atoms on the nucleophile part starts from 10 and increments by one per atom and the labels of the participating atoms on the electrophile part starts from 20 and increments by one per atom.

3. **Canonicalized atom mapped SMILES of the products:** The SMILES string of the reactants molecules, with integer labels for atoms that are participating the reaction. The labeling convention is the same as reactant labeling.

4. **Canonicalized SMILES of the products:** The unique SMILES representation of the product molecules generated from the reactive reactants.

5. **Canonicalized arrow codes:** The codes for arrow pushing mechanisms containing the integer labels of the participating atoms on the reactant sides. The arrow codes start from the arrow on the nucleophilic group.

6. **Spectator molecules:** The unique SMILES representation of the molecules that are present in the reaction but not react. These might be reagents or spectators molecules that are implying a certain environment.

7. **Reactive molecules:** The molecules on both sides of the reaction that are participating in the reaction.

8. **Reactive atom:** The two atoms corresponding to the reactive molecular orbitals that are describing the electron transfer in the mechanism.

9. **Type of the reaction:** Core or specific reaction.

10. **Initial heat or energy:** A binary value for specifying the need for the exposure to an external source of energy or heat.

11. **Reaction category I:** The labels according to the 3-class categorization scheme shown in Figure 6.

12. **Reaction category II:** The labels according to the 7-class categorization scheme shown in Figure 6.

An instance of the `Reaction` model in RMechDB can be uniquely retrieved form the database using either the **Reaction ID** or from the combined properties 2-5.

The other two essential models `Molecule` and `Atom` are separately integrated with the `Reaction` model to facilitate the most efficient search (Figure 5). Each entry of the `Molecule` model is associated with properties such as a unique ID, canonicalized SMILES string, the pair of (`Reaction ID`, `Molecule ID`), role in the reaction, and all the accessible properties of the OEMolBase class in OEChem. Similarly, the entries of the `Atom` model are associated with properties such as a unique ID, canonicalized atom mapped SMILES string of the parent molecule, the pair of (`Reaction ID`, `Atom ID`), role in the reaction, and all the accessible properties of the OEAtomBase class in OEChem.

RMechDB: Web Server

The web server of the RMechDB includes three interfaces for: (1) Searching the data; (2) Downloading the data; and (3) Uploading new data.

Searching the Data

RMechDB provides an interactive search interface where users can search through the database using a variety of methods available through `http://deeprxn.ics.uci.edu/rmechdb/search`. At the highest level, the interface allows searching by reaction or compound.

Reaction Search

1. **Exact search:** Using the exact search method, the user inputs the query in the form of the SMIRKS of an elementary step containing reactants and products (no arrow code needed).

Then the user can opt to search for this precise elementary step, or to search for this precise step but with additional molecules involved as reagents or spectators.

2. **Similarity search:** Using the similarity search method, the user again inputs the query in the form of the SMIRKS of an elementary step containing reactants and products (no arrow code needed). Then the user specifies a similarity metric and the number of similar reactions (N) to be retrieved under this query. Upon hitting the search button, N elementary steps sorted from the most similar to the least similar to the input query are displayed.

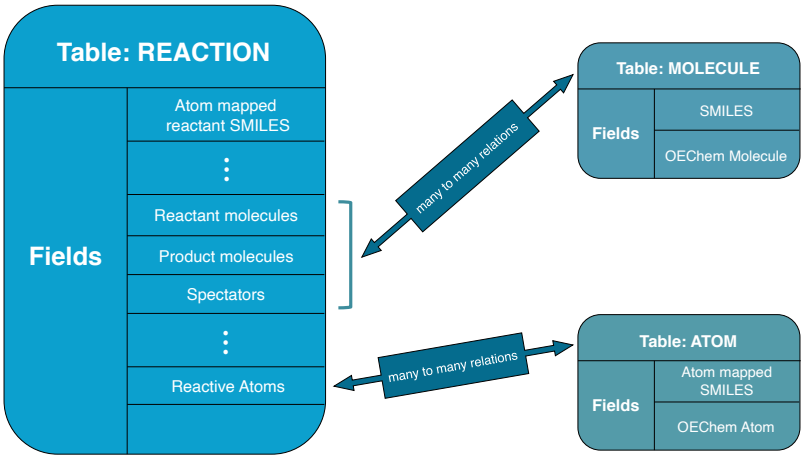

Figure 5: Depiction of the main tables and relations within the RMechDB database.

The current version of RMechDB is equipped with the following similarity metrics computed on various representations of the elementary steps:

1. The Tanimoto, dice, and cosine distance between the binary Extended Connectivity Fingerprints (ECFP) of the elementary steps.

2. The Euclidean distance between the embedding of the elementary steps derived using a pretrained transformer architecture, trained on the SMIRKS of the USPTO data set.[??]

3. The Euclidean distance between the embedding of the elementary steps derived using using the pretrained RxnHypergraph method.[?]

Compound Search

In addition to search capabilities based on elementary steps, RMechDB provides search capabilities based on smaller chemical entities as follows:

1. **Molecule search:** In this search, the user inputs the SMILES string of a desired molecule. After testing the validity of the input SMILES, RMechDB displays those elementary steps in the database that contain the desired molecule in the reactant or product side of the elementary step.

2. **Reactive atom (molecular orbital) search:** In this search, the user inputs the atom mapped SMILES string of the molecule where the reactive atom is labeled using an integer between 1 and 9, while the other atoms are not labeled. After testing the validity of the input SMILES with the labeled atom, RMechDB displays all the elementary steps in the database where the labeled atom is acting as one of the two main reactive atoms in the elementary step.

3. **Substructure search:** In this search, the user inputs the SMARTS of a chemically valid substructure. RMechDB displays all the elementary steps in the database with molecule(s) containing the input substructure. The molecule that contains the input substructure can be in the reactant or product side of the elementary step.

In addition, the results of each search can also be filtered using the following properties: (1) the type of the elementary steps (core or atmospheric); and (2) the category of the elementary step based on either of the two categorization schemes described in Section .

Downloading the Data

The data set of the chemcial reactions in RMechDB is available for download at `http://deeprxn.ics.uci.edu/rmechdb/download`. The data set is licensed under the *Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND)* license, which limits its free public usage to non-commercial purposes. Under this license the users are not allowed to modify

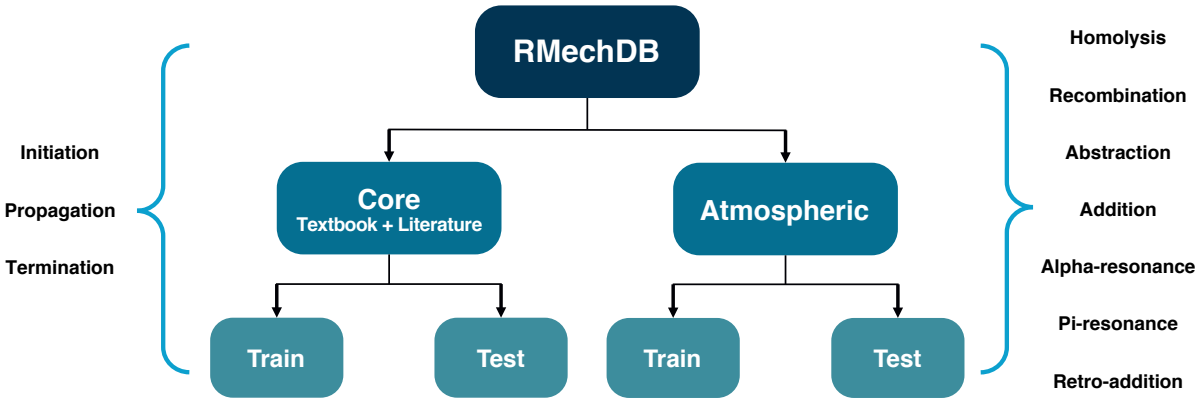

Figure 6: The general scheme of the radical MechDB.

and distribute the data set or to distribute the original data set without referencing the original source. After submitting basic information (name, email, and institution) and accepting the license terms, users receive an email with a comma separated value (CSV) file containing all the data and metadata.

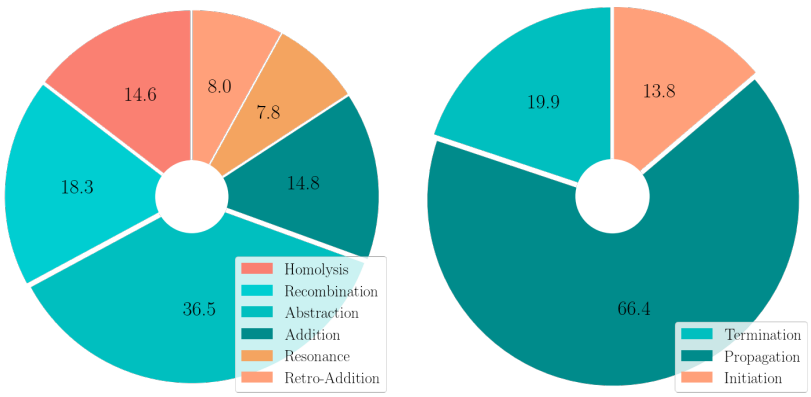

Figure 7: The distribution of the different classes of reaction in the current version of the RMechDB data set.

Uploading New Data

While we continue to insert new data in RMechDB, we invite the community to contribute new radical elementary steps. Uploading new data can be done at: `http://deeprxn.ics.uci.edu/rmechdb/expand`.

Contributing users must fill out two fields: (1) the SMIRKS of the elementary step; and (2) the corresponding electron flow specification (codes for arrow pushing) as shown in Figure 2. There is also a third optional field where the user can provide information about the source of the elementary step (e.g. the title of a textbook, a publication). After uploading the elementary step, it will be checked for validity, duplication, and plausibility.

Validity Check

A submitted elementary step is considered to be valid if it satisfies the following three criteria:

1. The SMILES string of all the molecules on both sides of the submitted elementary step must be correct and convertible to graphs representing valid molecules. An example of a valid and invalid elementary step is shown in Figure 8 (a).

2. The annotations for the arrow pushing mechanisms must be correct. This implies that the reacting atoms on the reactant side of the elementary step must be labeled with distinct integers. These integers form the basis for the arrow pushing mechanisms associated with electron transfers. The arrow codes must be consistent with the integers used to label the reacting atoms. An example of a valid and invalid arrow code is shown in Figure 8 (b).

3. The entered SMIRKS and arrow codes are then used to extract the interacting orbitals. We used our offline elementary step model[35–37] to find the interacting molecular orbitals and their corresponding atoms. If this model fails, the SMIRKS is considered invalid. An example of a SMIRKS that can be correctly modeled, and of a SMIRKS that cannot be correctly modeled, is shown in Figure 8 (c).

[TODO: describe the elementary step model as a part of the RMechDB and reword accordingly.]

Duplication Check

In this step, we check that the valid uploaded elementary step is not equivalent to any elementary step already included in the RMechDB data set. We consider two steps to be equivalent if they have the same:

1. Canonicalized SMILES string of the reacting molecules.
2. Canonicalized SMILES string of the product molecules.
3. Canonicalized SMILES string of the spectator molecules.
4. Conventional representation of the codes for arrow pushing mechanism. The labels of the participating atoms on the nucleophilic component start from 10 with increments of one per atom, and the the labels of the participating atoms on the electrophilic component start from 20 with increments of one per atom. It is important to mention that the user can use any integers to label the participating atoms. The conventional arrows codes will be automatically generated by RMechDB.

Once an elementary step is uploaded, RMechDB performs the validity and duplication tests automatically. In case of failure of either test, an informative error message is displayed with details about the corresponding errors.

Plausibility Check

Once the submitted elementary step passes both tests, it is further reviewed by the RMechDB committee members for plausibility verification. If the input SMIRKS is deemed to be plausible, it is then added to RMechDB.

Acknowledgement

This work was in part supported by NSF grant 195811 to DVV and PB, and ARO grant #W911NF2010172 to AC, DVV and PB. We are deeply grateful to OpenEye Scientific Software for their free academic license of the OEChem toolkit.

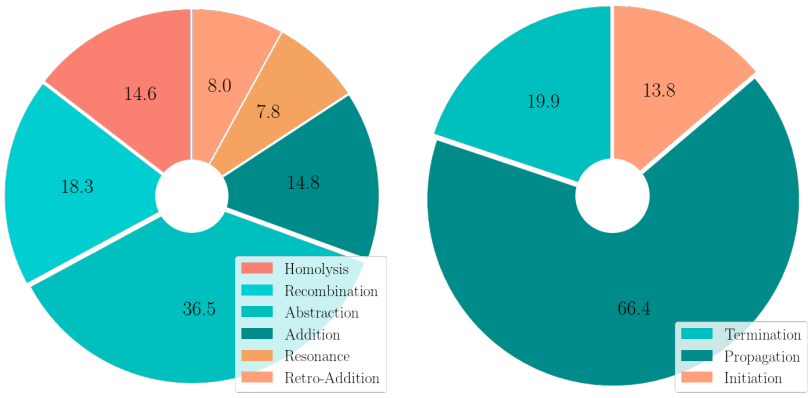

Figure 8: The distribution of the different classes of elementary steps in the current version of RMechDB.

References

(1) Baldi, P. Call for a Public Open Database of All Chemical Reactions. *Journal of Chemical Information and Modeling*

(2) Fleming, I. Curly arrows, when used with a molecular description of bonding, work as well as they do simply because they illustrate the electron distribution in the frontier orbital, and for reaction kinetics, it is the frontier orbital that is most important. *Frontier Orbitals and Organic Chemical Reactions* **1977**,

(3) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *Journal of the American Chemical Society* **2021**, *143*, 18820–18826, PMID: 34727496.

(4) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Central Science* **2017**, *3*, 434–443, PMID: 28573205.

(5) O'Hagan,; Lloyd, The Iconic Curly Arrow. `https://www.chemistryworld.com/features/the-iconic-curly-arrow/3004840.article`.

(6) Kermack, W. O.; Robinson, R. LI.—An explanation of the property of induced polarity of atoms and an interpretation of the theory of partial valencies on an electronic basis. *J. Chem. Soc., Trans.* **1922**, *121*, 427–440.

(7) Liu, M.; Grinberg Dana, A.; Johnson, M. S.; Goldman, M. J.; Jocher, A.; Payne, A. M.; Grambow, C. A.; Han, K.; Yee, N. W.; Mazeau, E. J., et al. Reaction mechanism generator v3. 0: Advances in automatic mechanism generation. *Journal of Chemical Information and Modeling* **2021**, *61*, 2686–2696.

(8) `https://rmg.mit.edu/database/`.

(9) RMG-Introduction to RMG. `https://www.youtube.com/watch?v=NSgghuMwAtw&list=PLZUqt5RldKbQWR1iJalegC-VjcxItpaXF&index=1`, it is noted that "parameter error [(k(T) and K(T))] much worse at low T (errors in Boltzmann factors exponentially amplified at low T)".

(10) Manion, J. A. et al. `https://kinetics.nist.gov/`.

(11) Zhao, Y.; Wang, B.; Li, H.; Wang, L. Theoretical studies on the reactions of formaldehyde with OH and OH. *Journal of Molecular Structure: THEOCHEM* **2007**, *818*, 155–161.

(12) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. *Nucleic Acids Research* **2018**, *46*, D618 – D623.

(13) `http://mcm.york.ac.uk/`.

(14) Jenkin, M. E.; Saunders, S. M.; Pilling, M. J. The tropospheric degradation of volatile organic compounds: a protocol for mechanism development. *Atmospheric Environment* **1997**, *31*, 81–104.

(15) Saunders, S. M.; Jenkin, M. E.; Derwent, R. G.; Pilling, M. J. Protocol for the development of

the Master Chemical Mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmospheric Chemistry and Physics* **2003**, *3*, 161–180.

(16) Jenkin, M. E.; Saunders, S. M.; Wagner, V.; Pilling, M. J. Protocol for the development of the Master Chemical Mechanism, MCM v3 (Part B): tropospheric degradation of aromatic volatile organic compounds. *Atmospheric Chemistry and Physics* **2003**, *3*, 181–193.

(17) Bloss, C.; Wagner, V.; Jenkin, M. E.; Volkamer, R.; Bloss, W. J.; Lee, J. D.; Heard, D. E.; Wirtz, K.; Martin-Reviejo, M.; Rea, G.; Wenger, J. C.; Pilling, M. J. Development of a detailed chemical mechanism (MCMv3.1) for the atmospheric oxidation of aromatic hydrocarbons. *Atmospheric Chemistry and Physics* **2005**, *5*, 641–664.

(18) Jenkin, M. E.; Wyche, K. P.; Evans, C. J.; Carr, T.; Monks, P. S.; Alfarra, M. R.; Barley, M. H.; McFiggans, G. B.; Young, J. C.; Rickard, A. R. Development and chamber evaluation of the MCM v3.2 degradation scheme for -caryophyllene. *Atmospheric Chemistry and Physics* **2012**, *12*, 5275–5308.

(19) Jenkin, M. E.; Young, J. C.; Rickard, A. R. The MCM v3.3.1 degradation scheme for isoprene. *Atmospheric Chemistry and Physics* **2015**, *15*, 11433–11459.

(20) Wennberg, P. O.; Bates, K. H.; Crounse, J. D.; Dodson, L. G.; McVay, R. C.; Mertens, L. A.; Nguyen, T. B.; Praske, E.; Schwantes, R. H.; Smarte, M. D.; St Clair, J. M.; Teng, A. P.; Zhang, X.; Seinfeld, J. H. Gas-Phase Reactions of Isoprene and Its Major Oxidation Products. *Chemical Reviews* **2018**, *118*, 3337–3390, PMID: 29522327.

(21) Buskirk, A.; Baradaran, H. Can Reaction Mechanisms Be Proven? *Journal of Chemical Education* **2009**, *86*, 551.

(22) Chen, J. H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *Journal of Chemical Information and Modeling* **2009**, *49*, 2034–2043, PMID: 19719121.

(23) Brown,; Foote,; Iverson,; Anslyn, *Organic Chemistry, 5th Ed.*; Brooks-Cole, 2008.

(24) Ege, S.; W., K. R.; Zitek, P. *Organic Chemistry, Structure and Reactivity*; Cengage Learning, Mifflin Company, 2004.

(25) Loudon, M.; Parise, J. *Organic Chemistry 6th Ed.*; W. H. Freeman, 2015.

(26) McMurry, J. E. *Organic Chemistry with Biological Applications*; Cengage Learning, 2014.

(27) Smith, J. *Organic Chemistry 5th Ed.*; McGraw Hill, 2016.

(28) Solomons, T. W.; Fryhle, C. B. *Organic Chemistry 11th Ed.*; Wiley, 2013.

(29) Vollhardt, P. *Organic Chemistry Structure and Function*; W. H. Freeman, 2005.

(30) Solomons, T. W.; Fryhle, C. B. *Organic Chemistry 8th Ed.*; Pearson, 2012.

(31) Carey, F. A.; Sundberg, R. J. *Advanced Organic Chemistry Part B: Reactions and Synthesis, 5th Ed*; Springer, 2010.

(32) Bruckner, R. *Organic Mechanisms: Reactions, Stereochemistry and Synthesis*; Springer, 2010.

(33) Seinfeld, J.; Pandis, S. N. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, 2nd Ed.*; Wiley, 2016.

(34) Chan, W.-T.; Hamilton, I. Mechanisms for the ozonolysis of ethene and propene: Reliability of quantum chemical predictions. *The Journal of Chemical Physics* **2003**, *118*, 1688–1701.

(35) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level Using Machine Learning. *Journal of Chemical Information and Modeling* **2012**, *52*, 2526–2540, PMID: 22978639.

(36) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *Journal of Chemical Information and Modeling* **2011**, *51*, 2209–2222, PMID: 21819139.

(37) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **2018**,