Mechanistic Reaction Data for Interpretable Deep Learning in Chemistry

Ryan J. Miller
Department of Computer Science
University of California, Irvine
rjmille3@uci.edu

Alexander E. Dashuta
Department of Chemistry
University of California, Irvine
adashuta@uci.edu

Pierre Baldi
Department of Computer Science
University of California, Irvine
pfbaldi@uci.edu

David Van Vranken
Department of Chemistry
University of California, Irvine
david.vv@uci.edu

Ann Marie Carlton
Department of Chemistry
University of California, Irvine
agcarlto@uci.edu

Abstract

The lack of openly accessible, well-curated reaction databases remains a major obstacle to data-driven research in chemistry. Many existing chemical datasets are proprietary and/or limited to unbalanced overall transformations that map reactants directly to products without revealing underlying mechanisms, intermediates, or byproducts. As a result, machine learning models trained on such data often act as "black boxes," predicting products without explaining how or why they form. To address this gap, we present the largest and most comprehensive publicly available dataset of manually curated elementary reaction steps, integrated into a platform that supports continuous curation, search functionality, and community contribution at scale. Our datasets cover polar and radical elementary steps, complete mechanistic pathways, and combinatorially generated mechanisms, with each reaction represented as a balanced, canonicalized SMIRKS string with reactive atom mapping and mechanistic annotations. By making mechanistic reaction data widely available, we aim to enable the development of interpretable and more accurate machine learning models for reaction and pathway prediction. We make the platform publicly available at https://deeprxn.ics.uci.edu/.

1 AI Task Definition

Chemists reason about reactions through detailed stepwise mechanisms that outline how atoms and electrons rearrange, and what drives each step. In contrast, many ML models map reactants directly to products without intermediates or mechanistic context, limiting interpretability and generalization. Our dataset enables a new class of tasks: predicting and generating individual elementary steps and complete pathways with explicit mechanistic detail. Models trained on such data can explain how and why products form, identify byproducts, reveal alternative branches and intermediates, and support interpretable discovery of new reactions in a way that mirrors the reasoning chemists naturally use.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

2 Dataset Rationale

While some mechanistic datasets exist, they remain limited in scope. Gas-phase resources such as the NIST Chemical Kinetics Database [12], the Master Chemical Mechanism [15], and the Reaction Mechanism Generator (RMG) database [6] primarily cover radical reactions and omit multi-phase or solution-phase processes critical for pharmaceutical, biochemical, and environmental chemistry. The USPTO dataset [11] does include many polar reactions, and recent work by Coley et al. [2] and Jung et al. [9] has algorithmically converted USPTO reactions into mechanistic steps using templates. However, these USPTO-derived datasets are restricted to the reaction types represented in patents, omit key context such as solvents and temperatures, and may suffer from incomplete reactions or inaccurate mappings due to automated extraction.

To overcome these limitations, we propose an adaptable and continuously expanding dataset of manually curated elementary step mechanisms. Each step is curated and verified by expert organic chemists for plausibility and includes: (1) reactant SMIRKS strings, (2) product SMIRKS strings, (3) precise reactive atom mapping and arrow-pushing annotations, and, where available, (4) temperature, (5) solvent, (6) numerical rate constants, and (7) scholarly references. Manual verification ensures greater accuracy and reliability than automated approaches. Crucially, our dataset is designed to be flexible: we actively expand its coverage and can incorporate missing or underrepresented reaction types by identifying and adding relevant mechanisms from the literature. This adaptability ensures a diverse and balanced representation of mechanistic classes that can evolve to meet the needs of the community.

3 Data Creation Pathway

We have assembled a diverse collection of elementary mechanistic steps spanning multiple reaction classes. For polar ($2e^-$) processes, approximately 13K steps were curated from literature and textbooks. To supplement these polar reactions, combinatorial datasets were generated by pairing atom mapped structures together, creating roughly 100K nucleophilic steps with Mayr-derived rate constants [13], and 51M proton-transfer steps derived from the Eigen relationship [4, 5]. For pathway testing, a set of 1K textbook [3] reaction pathways was curated, and a comprehensive polar set exceeding 350K elementary steps is currently undergoing verification.

For radical ($1e^-$) processes, aprroximately 2K literature steps and 3K atmospheric reactions (from MCM [15]) have been collected, with an additional 8K plausible room-temperature steps being incorporated from RMG [6].

We are also incorporating pericyclic processes, including 3K [4+2] cycloadditions which are currently undergoing validation.

The ultimate objective is a master mechanistic dataset encompassing polar, radical, and pericyclic steps, with balanced representation, rate data where available, and broad mechanistic diversity.

4 Cost & Scalability

The main cost is expert labor; unlike automated pipelines, our approach prioritizes accuracy and interpretability via manual expert verification. This is a relatively inexpensive operation which requires a small, dedicated team rather than major investments in hardware or experimental facilities:

Chemistry Graduate Students (1–2 people): Responsible for curating, labeling, and balancing elementary reaction steps across polar, radical, and pericyclic classes. They manually verify plausibility and ensure consistency across datasets.

Computer Science Graduate Student (1 person): Maintains the platform, validates reactions using chemoinformatics libraries, performs combinatorial reaction generation, and coordinates model training, validation, and benchmarking.

5 Acceleration Potential

Historically, proprietary databases like Reaxys and CAS dominated reaction "recipes," describing reactants and products without mechanistic detail. The release of the open access USPTO dataset enabled searchable resources such as ORD [10] and spurred ML advances using these recipes. Likewise, a balanced, high-quality public dataset of elementary steps could drive models that not only predict products but also uncover new mechanistic pathways.

Rapid identification of products and their underlying pathways would be transformative. Gaps in mechanistic knowledge hinder pharmaceutical synthesis, compromise drug safety and shelf life, obscure harmful species in secondary organic aerosols, and affect the fate of pharmaceuticals once they enter the environment. A high-quality mechanistic dataset can address these challenges by enabling machine learning models to explain the steps leading to a product. This mechanistic grounding gives models a deeper understanding of chemical reactivity, better equipping them to handle such challenges while providing interpretable predictions that chemists can readily validate.

Acknowledgements

This work was supported by the National Science Foundation under grant 1955811, awarded to Professors Pierre Baldi and David Van Vranken. We thank OpenEye Scientific Software for providing access to the OEChem toolkit, which was used to parse and manipulate molecular graphs. We also thank Chemaxon for providing free academic licenses; Chemaxon Marvin was used to draw and visualize chemical structures.

References

[1] Claude F Bernasconi, Douglas E Fairchild, Robert L Montañez, Perdram Aleshi, Huaiben Zheng, and Edward Lorance. Kinetics of proton transfer from cationic carbon acids in water and aqueous dmso. effect of activating groups and solvent on intrinsic rate constants. *The Journal of Organic Chemistry*, 70(19):7721–7730, 2005.

[2] Shuan Chen, Ramil Babazade, Taewan Kim, Sunkyu Han, and Yousung Jung. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Scientific Data*, 11(1):863, 2024.

[3] Jonathan Clayden, Nick Greeves, and Stuart Warren. *Organic chemistry*. Oxford university press, 2012.

[4] Manfred Eigen. Proton transfer, acid-base catalysis, and enzymatic hydrolysis. part i: elementary processes. *Angewandte Chemie International Edition in English*, 3(1):1–19, 1964.

[5] Mo Eigen. Fast elementary steps in chemical reaction mechanisms. *Pure and Applied Chemistry Seventh*, 6(1):97–116, 1963.

[6] Connie W Gao, Joshua W Allen, William H Green, and Richard H West. Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms. *Computer Physics Communications*, 203:212–225, 2016.

[7] J Peter Guthrie. Hydration of thioesters. evaluation of the free-energy changes for the addition of water to some thioesters, rate-equilibrium correlations over very wide ranges in equilibrium constants, and a new mechanistic criterion. *Journal of the American Chemical Society*, 100(18):5892–5904, 1978.

[8] J Peter Guthrie, Jonathan Barker, Patricia A Cullimore, Jinqiao Lu, and David C Pike. The tetrahedral intermediate from the hydration of n-methylformanilide. *Canadian journal of chemistry*, 71(12):2109–2122, 1993.

[9] Joonyoung F Joung, Mun Hong Fong, Jihye Roh, Zhengkai Tu, John Bradshaw, and Connor W Coley. Reproducing reaction mechanisms with machine-learning models trained on a large-scale mechanistic dataset. *Angewandte Chemie International Edition*, 63(43):e202411296, 2024.

[10] Steven M Kearnes, Michael R Maser, Michael Wleklinski, Anton Kast, Abigail G Doyle, Spencer D Dreher, Joel M Hawkins, Klavs F Jensen, and Connor W Coley. The open reaction database. *Journal of the American Chemical Society*, 143(45):18820–18826, 2021.

[11] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, 2012.

[12] William G Mallard, F Westley, JT Herron, Robert F Hampson, and DH Frizzell. *NIST chemical kinetics database*, volume 126. National Institute of Standards and Technology Washington, DC, USA, 1992.

[13] Herbert Mayr and Armin R Ofial. Do general nucleophilicity scales exist? *J. Phys. Org. Chem.*, 21(7-8):584–595, 2008.

[14] Hans J. Reich. pk_a values in water. `https://organicchemistrydata.org/hansreich/resources/pka/#ka-water`, 2025. Bordwell pKa Table. ACS Division of Organic Chemistry. Accessed 17 Jan 2025.

[15] Sandra M Saunders, Michael E Jenkin, Richard G Derwent, and Mike J Pilling. 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*, 3(1):161–180, 2003.

A Appendix / supplemental material

In this appendix, we provide additional details about the data curation and formatting.

A.1 Data Format

We provide an example of an atom mapped SMIRKS with arrow pushing annotation:

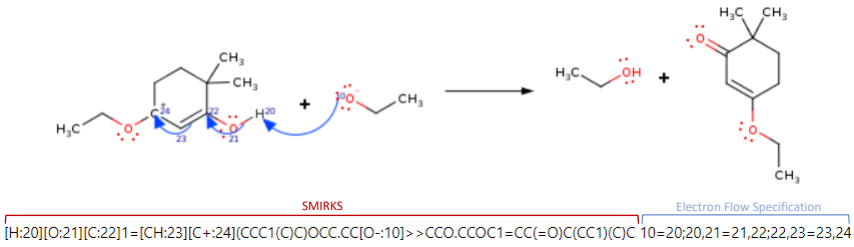

Figure 1: The reaction is stored as a SMIRKS string which contains atom mappings for the reactive atoms. The nucelophilic source atom is mapped as atom 10, and the electrophilic sink atom is mapped as atom 20. The reactions come with arrow-codes, which specify the flow of electrons. For example, 10=20 indicates the electrons from atom 10 move to atom 20, then 20,21=21,22 indicates the electrons move from the bond between atoms 20,21 to the bond between atoms 21,22 etc.

A.2 Pathway Dataset

To build a pedagogically diverse set of organic reactions, we curated 1,187 single-step transformations from the intermediate-level textbook by Clayden, Greeves, and Warren [3]. This text includes many modern reactions not typically covered in introductory courses, such as allylsilane additions, enol silyl ethers, boron-mediated aldol reactions, organosulfur and organophosphorus chemistry, electron-transfer reductions, and heterocycle syntheses. Each textbook example was translated into a dataset entry that includes reactant(s), reaction temperature, and product(s) for use in product-prediction tasks. Many of these transformations are accompanied by explicit arrow-pushing mechanisms for the key steps. Implied multi-step workups were excluded to ensure that each entry represents a

single-step transformation. All entries include scholarly literature references, and when textbook schemes used generic substituents (R groups), representative examples from the research literature were selected to match the transformation. Each entry also records an estimate of the minimum number of elementary mechanistic steps needed to reach the product.

A.3 Mayr Combinatorial Reaction Generation

Additional reactions were generated from Mayr–Ofial nucleophilicity (N, s_N) and electrophilicity (E) parameters [13]. We selected 853 of 1,254 nucleophiles and 313 of 352 electrophiles, excluding sterically hindered species, protic-solvent data, and certain hard anions (e.g., acetate, benzoate, methyl carbonate). This yielded 257,712 nucleophile–electrophile pairs, of which 96,558 with $s_N(N + E) \geq 3$ were converted to mapped SMIRKS with explicit electron flow. One thousand of the least reactive pairs were manually checked for plausibility. While S_N2 reactions are key in organic chemistry, classic sp^3 electrophiles like alkyl halides are missing from the Mayr dataset.

A.4 Proton Transfer Combinatorial Reaction Generation

A comprehensive set of acid–base proton transfer steps was generated combinatorially using aqueous pK_a values. Starting from 7,613 heteroatom acids and their conjugate bases (primarily from the DataWarrior set, supplemented with work from Reich [14] and Guthrie [7, 8]), all pairwise combinations were enumerated. For arrow-pushing specification, the acidic proton was mapped as atom 20, the attached heteroatom as atom 21, and the basic site as atom 10.

Using a simplified Eigen relationship [4, 5], second-order rate constants at 25 °C were estimated from ΔpK_a values using the following equation:

$$\log k_1 = \Delta pK_a + \log k_{-1} \approx \Delta pK_a + 9$$

Only steps with $\log k_1 \geq 3$ ($\approx 10^3$ M^{-1} s^{-1}) were retained as kinetically plausible. This conservative cutoff yielded 51,505,065 heteroatom–heteroatom proton transfer steps encoded as SMIRKS with explicit electron flow.

Carbon acids were treated separately using intrinsic rate constants and Brønsted β values in the Eigen–Bernasconi equation [1]. From 65 carbon acids and 7 classes of heteroatom bases, 4,902 plausible proton transfers with $\log k_1 \geq 3$ were included. An additional 49 literature-reported proton transfers from heteroatom acids to carbon bases (with experimental rates) were also added.

The resulting dataset spans acids with pK_a values from -15 to $+37$, emphasizing structural diversity over reaction conditions. All entries are mechanistically annotated and suitable for machine learning applications.