Multistep Synthesis Tool for Organic Chemistry Education

Raelyn Brooks

October 15, 2025

1 Introduction

The automation and visualization of organic synthesis pathways have become critical in both educational and research contexts within computational chemistry. Traditional reaction depiction methods, while effective for static textbook examples, often fail to convey the complexity of multistep synthetic logic and reagent interactions. This project addresses that gap by implementing a visualization pipeline using SMIRKS-based reaction representations and the RDKit cheminformatics library. The core objective is to generate interpretable and data-driven visual representations of chemical transformations, making synthetic pathways more accessible for analysis, teaching, and algorithmic reasoning.

Organic chemistry specifically has long been considered one of the most difficult undergraduate courses, with national failure rates often exceeding 50 percent. Among the challenges faced by students, is multistep synthesis complexity standing out as a major stumbling block. Success in synthesis problems requires balancing starting materials, reagents, intermediates, conditions, and stereochemistry, all while reasoning backwards (retrosynthesis) and forwards (reaction progression). Students often describe synthesis as a "black box": they know reactions individually, but cannot connect them into coherent strategies. This project aims to design and implement an interactive software system that teaches multistep synthesis through decision trees and constraint satisfaction techniques. By breaking synthesis into a sequence of guided yes/no decisions, the system will reinforce fundamentals and provide educational feedback at each step.

The literature reflects a growing interest in translating chemical reaction data into computationally interpretable formats. Previous studies have leveraged symbolic representations like SMILES (Simplified Molecular Input Line Entry System) and its generalized form, SMIRKS, to encode reactions and reaction rules. Within this context, RDKit serves as a key enabling tool, offering molecule parsing, coordinate computation, and visualization capabilities that underpin many modern cheminformatics pipelines. The purpose of this review is to situate this project's methodology—an RDKit-based multistep pathway visualization tool—within the broader landscape of chemical data representation, machine-assisted reaction modeling, and visualization frameworks.

2 Related Works

2.1 Reaction Representation and SMIRKS Formalism

The foundational step in computational reaction modeling is the translation of chemical transformations into a symbolic language interpretable by algorithms. The SMIRKS format, derived from SMILES, provides a flexible mechanism to encode both specific reactions and generalized transformation rules. Literature on chemical data representation emphasizes that these formats enable reaction pattern recognition, rule-based synthesis prediction, and large-scale database integration. Works in this domain highlight that robust SMIRKS parsing supports not only mechanistic modeling but also serves as a data layer for visualization systems and generative synthesis algorithms [4].

Works in this domain highlight that robust SMIRKS parsing supports not only mechanistic modeling but also serves as a data layer for visualization systems and generative synthesis algorithms. The annotated literature identifies several efforts to standardize chemical notation and link structural encoding with reaction prediction models, a crucial link to the project's educational synthesis visualization goals.

2.2 RDKit and Cheminformatics Frameworks

Among open-source cheminformatics platforms, RDKit is one of the most widely adopted due to its robust handling of molecular representations and transformations. Numerous studies in cheminformatics emphasize RDKit's versatility for tasks including substructure searching, molecular fingerprinting, and reaction enumeration. Researchers have particularly noted its extensibility for educational and visualization purposes, allowing developers to generate 2D molecular depictions, reaction schemes, and datasets for computational learning systems.

In this project, RDKit's molecule parsing (MolFromSmiles) and coordinate generation (Compute2DCoords) functions serve as the computational backbone for converting textual reaction definitions into visual chemical diagrams. By leveraging RDKit's drawing modules (Draw.MolToImage), the project extends the toolkit into a visual pedagogy tool—making chemical synthesis pathways interpretable even for non-specialist learners. This positions the work within a subfield of cheminformatics focused on interpretable visual analytics, bridging chemical informatics and visual communication.

2.3 Computational Visualization and Educational Tools

The use of visualization in computational chemistry extends beyond aesthetics; it plays a vital cognitive role in supporting chemical reasoning. Prior work on reaction visualization tools—ranging from commercial platforms like ChemDraw to research-oriented frameworks like Indigo and Open Babel—demonstrates how symbolic chemistry can be rendered graphically to assist comprehension and verification. However, many of these systems rely on manual input or lack automated support for multi-step reaction sequences.

By contrast, the system developed in this project automates pathway rendering from large reaction datasets. Each reaction, formatted as Reactants → Reagents → Products, is parsed and visually represented as a network of chemical structures connected by arrows annotated with reagent labels. This approach aligns with modern pedagogical chemistry tools that emphasize interactivity and visual logic to improve understanding of synthesis design. The use of PIL (Python Imaging Library) to dynamically compose RDKit-rendered molecules onto a unified canvas represents a practical integration of informatics and visual communication principles highlighted in the literature.

2.4 Data-Driven Synthesis Prediction and Multistep Modeling

Recent studies in reaction prediction, retrosynthesis, and synthesis planning have leveraged large datasets encoded in SMILES/SMIRKS to train machine learning models capable of generating plausible synthetic routes [2]. Although this project does not directly implement predictive modeling, it shares methodological continuity with such research by structuring reactions in a format compatible with machine learning frameworks [3]. Literature examining data-driven synthesis design emphasizes that visualization and explainability remain key challenges; thus, tools that render reaction logic interpretable, like this project's visualization pipeline, contribute to advancing both human understanding and computational transparency in chemistry.

Moreover, by including thousands of curated textbook and large-scale reactions, the dataset architecture supports potential future expansion into reaction pathway decision trees or knowledge graph representations, which several annotated studies identify as emerging directions in cheminformatics research [5].

3 Methodology

Dijkstra's algorithm was chosen for single-source shortest path computation on graphs with non-negative edge weights, following the implementation presented in CLRS [1].

4 Results

5 Conclusion

References

[1] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms, Third Edition*, 3rd ed. The MIT Press, 2009.

[2] Dai, H., Li, C., Coley, C. W., Dai, B., and Song, L. *Retrosynthesis prediction with conditional graph logic network.* Curran Associates Inc., Red Hook, NY, USA, 2019.

[3] Hormazabal, R., Park, C., Lee, S., Han, S., Jo, Y., Lee, J., Jo, A., Kim, S., Choo, J., Lee, M., and Lee, H. Cede: a collection of expert-curated datasets with atom-level entity annotations for optical chemical structure recognition. In *Proceedings*

of the 36th International Conference on Neural Information Processing Systems (Red Hook, NY, USA, 2022), NIPS '22, Curran Associates Inc.

[4] Schilter, O. T., Laino, T., and Schwaller, P. Cmd+v for chemistry: Image to chemical structure conversion directly done in the clipboard. *Applied AI Letters 5*, 1 (Jan. 2024).

[5] Souza, V. F., Cicalese, F., Laber, E. S., and Molinaro, M. Decision trees with short explainable rules. *Theor. Comput. Sci. 1047*, C (Aug. 2025).