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 Background

In organic chemistry, multistep synthesis refers to the process of transforming simple molecular starting materials into complex target molecules through a series of discrete, logically ordered chemical reactions. Each reaction step modifies the structure of a molecule, typically by altering functional groups—specific arrangements of atoms that dictate chemical behavior. From a computer science perspective, these functional groups can be viewed as data types or states: just as an integer might be cast into a string or an array into a list, a hydroxyl group (–OH) might be converted into a halide (–Cl), or an alkene might be transformed into an alcohol. Each synthetic step is thus a state transition in a well-defined, though highly constrained, chemical state machine.

Designing a synthetic pathway is not unlike algorithm design, where chemists aim to find an efficient and feasible sequence of operations to reach a target state. Here, the "output" is a desired molecule, and the "input" is a set of commercially available or easily accessible starting materials. The multistep aspect arises because direct, one-step conversions are rarely possible; instead, chemists must plan a sequence of reactions, each altering the molecular structure in a controlled manner. This is analogous to chaining multiple functions in a program to transform an initial data structure into a final desired format.

A central algorithmic strategy in synthesis planning is retrosynthesis, introduced by E.J. Corey, organic chemist and 1990 Chemistry Nobel Prize Winner. Retrosynthesis involves working backward from the target molecule to break it down recursively into simpler precursors. This is strikingly similar to goal decomposition in AI planning or backtracking algorithms in search problems. Starting from the target, chemists iteratively identify strategic bonds to "disconnect," yielding simpler molecules that could plausibly be converted into the target in one step. This continues until reaching molecules that are known or easily obtainable starting materials. In computational terms, retrosynthesis resembles a reverse search through a tree or graph, where each node represents a molecular state and each edge represents a known chemical transformation.

During synthesis planning, selectivity and functional group compatibility act as constraints, analogous to conditional logic in code or constraint satisfaction problems (CSPs) in AI. For example, a reaction designed to modify an alcohol group might also inadvertently react with a nearby amine group, producing unwanted side products. Chemists must therefore choose reaction sequences that respect these constraints, ensuring that each transformation occurs at the correct site and does not disrupt other parts of the molecule. This is akin to managing function side effects in programming—ensuring that an operation modifies only what it is intended to, without corrupting unrelated data.

An additional layer of complexity involves stereochemistry (the 3D arrangement of atoms) and protecting groups (temporary modifications used to "mask" reactive sites). These elements function like temporary state variables or conditional flags that preserve critical information during intermediate steps. For example, protecting groups are akin to placing certain variables in a "read-only" or "inactive" state to prevent unwanted changes until the program reaches the correct point in execution to "unprotect" them. Stereochemical relationships act like metadata that must be preserved across transformations, ensuring that the final molecule has the correct 3D orientation—just as a compiler must preserve type information across optimizations.

To manage these complex transformations systematically, chemists often model reactions in a linear data flow format:

$$\text{Reactants} + \text{Reagents} \rightarrow \text{Products}$$

This format mirrors data flow in functional programming or pipeline architectures in software engineering. Each step is a transformation function that takes molecular "inputs" and produces "outputs," which can then feed into the next step. From a computational perspective, this linear structure offers several advantages:

1. **Traceability:** Each transformation can be logged and indexed, enabling the reconstruction of synthetic routes and error checking, similar to how logs are used in debugging.

2. **Database Integration:** Standardized formats like SMILES and SMIRKS allow reactions to be stored as structured data, facilitating search, retrieval, and machine learning. This mirrors how normalized database schemas support efficient querying.

3. **Modularity:** Each reaction step is a modular operation that can be reused across different pathways, analogous to reusable functions or classes in code.

4. **Visualization:** Linear reaction flows can be easily represented as directed acyclic graphs (DAGs), where nodes are molecular states and edges are transformations—making them ideal for cheminformatics pipelines and algorithmic reasoning.

In short, multistep synthesis can be understood as a sophisticated algorithmic problem: molecules represent structured data, functional groups represent types and states, reactions are transformation functions, and retrosynthesis is a backward search strategy under complex constraints. For computer scientists, this framing highlights why cheminformatics libraries like RDKit, symbolic languages like SMIRKS, and visualization pipelines are powerful—they translate the chemical logic into computational structures that can be analyzed, optimized, and taught using algorithmic principles.

3 Related Works

3.1 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.

3.2 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.

3.3 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 $\rightarrow$ Reagents $\rightarrow$ 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.

3.4 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].

4 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].

5 Results

6 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).