Multistep Synthesis Tool for Organic Chemistry Education

Raelyn Brooks

September 10, 2025

Organic chemistry 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, multistep synthesis stands 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. The purpose of this project is 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. While professional chemistry tools exist—such as ChemDraw or Reaxys–these platforms are not designed for novice learners. This project fills the gap by prioritizing transparency, interpretability, and pedagogy over raw prediction power.

A number of research directions inform this project. From a computer science perspective, foundational graph algorithms are necessary for representing molecules and reactions as nodes and edges [1]. This graph-based structure underpins most cheminformatics systems. On the theoretical side, Gale, Lobski, and Zanasi [2] propose a categorical model of organic chemistry, treating molecules as objects and reactions as morphisms. Their framework offers a formal way of reasoning about transformations, which can guide the development of rule-based synthesis logic. Decision trees provide a natural mechanism for structuring synthesis pathways.O'Sullivan, Ferguson, and Freuder show how decision trees can boost the efficiency of constraint satisfaction problems (CSPs) by reducing the search space and pruning infeasible solutions. [3] Similarly, Shati, Cohen, and McIlraith demonstrate how decision trees can be combined with constraints to produce interpretable and accurate models. This is essential for an educational setting, where students must understand why a particular path works or fails. [5] *Decision trees with short explainable rules* further emphasize explainability by introducing methods for building decision trees with short, clear rules, making them particularly well suited for teaching environments. [6]

Finally, the software must rest on an optimized data backbone. Reniers present schema recommender systems for document databases, which can be adapted to the chemical context. Their work highlights how workload-driven design improves query efficiency, an important consideration when storing and retrieving large numbers of molecules and reactions. [4]

Appendix

Planned features for the Multistep Synthesis Tool for Organic Chemistry include:

- Load dataset of organic molecules and reactions. (Existing dataset with names, structures, reactions, and reaction conditions.)

- Input molecule by name or drawing. (Name of a starting material or target molecule and convert it into a structured chemical format.)

- Confirm and display input molecules. (Show the user the structure of the input molecule to confirm correctness.)

- Search function for molecules. (Add a search feature that lets users find molecules, reagents, or intermediates within the dataset for use in pathway.)

- Generate single-step transformations. (Build logic to propose valid one-step transformations from the starting material toward the target product.)

- Generate multistep synthesis pathways. (Use decision trees and constraint satisfaction to create multistep synthesis routes from starting material to target product.)

- Step-by-step walkthrough mode with explanations. (Create an interactive mode where the user can walk through each reaction step with explanations of the reagents, conditions, and reasoning.)

- Constraint enforcement on pathways. (Implement rule checks that block or flag invalid or chemically impossible steps, with clear error explanations.)

- Store all molecules and reactions in a database. (MongoDB or similar NoSQL database to efficiently store and query chemical data.)

- User interface for interaction. (Simple web or desktop UI to input molecules, view structures, and navigate synthesis pathways.)

- Test on actual synthesis problems. (Use real-world synthesis problems from textbooks or literature to validate the tool's effectiveness.)

  Stretch Goals:

- Instructor problem mode (create custom synthesis problems for students to solve)

- Integrate cheminformatics library (e.g., RDKit) for advanced molecule handling and visualization

- User accounts and progress tracking (save synthesis attempts and scores over time)

- Machine learning suggestion layer (use ML models to suggest likely reagents or pathways based on historical data)

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] GALE, E., LOBSKI, L., AND ZANASI, F. A categorical model for organic chemistry. *Theor. Comput. Sci. 1032*, C (Apr. 2025).

[3] O'SULLIVAN, B., FERGUSON, A., AND FREUDER, E. C. Boosting constraint satisfaction using decision trees. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence* (USA, 2004), ICTAI '04, IEEE Computer Society, p. 646–651.

[4] RENIERS, V., VAN LANDUYT, D., RAFIQUE, A., AND JOOSEN, W. A workload-driven document database schema recommender (dbsr). In *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings* (Berlin, Heidelberg, 2020), Springer-Verlag, p. 471–484.

[5] SHATI, P., COHEN, E., AND MCILRAITH, S. Optimal decision trees for interpretable clustering with constraints. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (2023), IJCAI '23.

[6] SOUZA, V. F., CICALESE, F., LABER, E. S., AND MOLINARO, M. Decision trees with short explainable rules. *Theor. Comput. Sci. 1047*, C (Aug. 2025).