Annotated Bibliography

Raelyn Brooks

October 23, 2025

References

[1] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

This is a popular algorithms textbook which is well-cited. In particular, Part VI on graph algorithms will be of interest. Chapter 26 discusses flow networks and introduces commonly used notation. It formally describes the problem of obtaining a maximum flow and its equivalence to obtaining a minimum cut. The classical method of Ford and Fulkerson's algorithm for finding a maximum flow is described, and it includes several examples. Additional methods for obtaining a maximum flow, including the push-relabel method, are also described. The chapter notes include additional references to specific articles which may be helpful, such as those of historical interest (the article in which an algorithm was originally proposed) as well as state-of-the-art improvements (more recent articles to improve the approach).

[2] Hanjun Dai, Chengtao Li, Connor W. Coley, Bo Dai, and Le Song. *Retrosynthesis prediction with conditional graph logic network*. Curran Associates Inc., Red Hook, NY, USA, 2019.

This article presents a novel approach to retrosynthesis prediction using a Conditional Graph Logic Network (CGLN). Retrosynthesis is a fundamental problem in organic chemistry that involves identifying reactants capable of synthesizing a specified product molecule. Traditional methods often rely on template-based models with hard decision rules, which may not accurately capture the complexity of chemical reactions. The proposed CGLN model leverages graph neural networks to learn when to apply reaction templates, considering both chemical feasibility and strategic factors. The authors also introduce an efficient hierarchical sampling method to reduce computational costs. Experimental results demonstrate a significant improvement of 8.2Additionally, the model provides interpretability for its predictions, making it a valuable tool for chemists. This work is particularly relevant for those interested in the intersection of machine learning and chemistry, as it offers a sophisticated approach to solving complex chemical synthesis problems.

[3] Ella Gale, Leo Lobski, and Fabio Zanasi. A categorical model for organic chemistry. *Theor. Comput. Sci.*, 1032(C), April 2025.

This article presents a novel approach to modeling organic chemistry using category theory. The authors introduce a categorical framework that captures the structure and behavior of organic molecules and their reactions. The model allows for the representation of chemical reactions as morphisms between objects, providing a new perspective on retrosynthesis and disconnection rules. The paper includes several examples to illustrate the application of the categorical model to various organic chemistry problems. This work has potential implications for computational chemistry and the development of new algorithms for chemical synthesis planning. With a goal of this Junior IS project being to explore the intersection of computer science and chemistry, this article provides a foundational understanding of how advanced mathematical concepts can be applied to chemical problems.

[4] Rodrigo Hormazabal, Changyoung Park, Soonyoung Lee, Sehui Han, Yeonsik Jo, Jaewan Lee, Ahra Jo, Seunghwan Kim, Jaegul Choo, Moontae Lee, and Honglak Lee. 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*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

This article introduces CEDe, a comprehensive collection of expert-curated datasets designed to advance the field of Optical Chemical Structure Recognition (OCSR). OCSR focuses on translating chemical images into molecular structures, a critical task in scientific documentation. Traditional rule-based methods rely on detecting chemical entities such as atoms and bonds, followed by reconstructing the compound structure. However, recent neural network approaches, similar to image captioning, have shown data inefficiency, requiring millions of examples to achieve comparable performance. CEDe addresses this challenge by providing over 700,000 manually annotated chemical entity bounding boxes, facilitating structure reconstruction. Additionally, the authors release a large synthetic dataset with one million molecular images to support transfer-learning techniques for improved performance in low-data scenarios. Benchmark results demonstrate that detection-reconstruction models can match or exceed the performance of image captioning models while using significantly fewer training examples. This work is particularly relevant for researchers and practitioners in cheminformatics and machine learning, offering valuable resources to enhance OCSR methodologies.

[5] Suazette R. Mooring, Chloe E. Mitchell, and Nikita L. Burrows. Evaluation of a flipped, large-enrollment organic chemistry course on student attitude and achievement. *Journal of Chemical Education*, 93(12):1972–1983, 2016. Published: December 13, 2016.

This article evaluates the effectiveness of a flipped classroom model in a large-enrollment organic chemistry course. The authors investigate the impact of

this teaching approach on student attitudes and academic achievement. The study involves a comparison between traditional lecture-based instruction and the flipped model, where students engage with course material before class and participate in active learning activities during class time. The results indicate that the flipped classroom model leads to improved student attitudes towards the subject matter and higher achievement levels, as measured by exam scores and overall course performance. The authors discuss the implications of these findings for educators seeking to enhance student engagement and learning outcomes in large-enrollment science courses. This article is particularly relevant for those interested in innovative teaching methods in chemistry education, providing insights into how flipping the classroom can positively influence student experiences and success.

[6] Barry O'Sullivan, Alex Ferguson, and Eugene C. Freuder. Boosting constraint satisfaction using decision trees. In *Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence*, ICTAI '04, page 646–651, USA, 2004. IEEE Computer Society.

This article explores the use of decision trees to enhance the efficiency of constraint satisfaction problems (CSPs). The authors propose a hybrid approach that combines decision tree learning with traditional backtrack search methods. By leveraging knowledge from previously solved instances of CSPs, the method aims to reduce the search space and improve solution times. The paper presents experimental results demonstrating significant performance improvements, often achieving nearly an order-of-magnitude reduction in search effort. This approach is particularly relevant for applications such as product configuration and interactive constraint solving, where problems are solved repeatedly over time. The article provides valuable insights into how machine learning techniques can be integrated with constraint satisfaction to optimize problem-solving processes. More within current proposals for Junior IS project that is exploring building an organic chemistry tool using decision trees there is a need to understand constraint satisfaction, making this article a useful resource.

[7] RDKit Documentation Team. Rdkit overview, 2025. Accessed: October 22, 2025.

This document provides an overview of RDKit, an open-source cheminformatics software library widely used in computational chemistry. The RDKit Documentation Team introduces RDKit's main features, molecular representations (SMILES, SMARTS), and its integration with other scientific computing tools. The overview emphasizes RDKit's extensibility, enabling custom workflows for research in drug discovery, materials science, and related areas. This resource serves as a valuable guide for researchers and developers seeking to leverage RDKit for cheminformatics projects.

[8] Vincent Reniers, Dimitri Van Landuyt, Ansar Rafique, and Wouter Joosen. A workload-driven document database schema recommender (dbsr). In *Conceptual Modeling: 39th*

International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings, page 471–484, Berlin, Heidelberg, 2020. Springer-Verlag.

This article introduces a workload-driven document database schema recommender (DBSR) designed to optimize schema design for NoSQL document-oriented databases. The authors highlight the challenges of traditional normalization schemes used in relational databases, which minimize data redundancy, versus the redundancy-favoring approach of NoSQL databases that prioritize horizontal scalability and performance. The DBSR employs a systematic, search-based method to explore the complex schema design space, taking into account the application's data model and read workload. The recommender provides suggested document schemas, query plan recommendations, and a document utility matrix that evaluates costs and relative utility. Experimental evaluations using MongoDB and YCSB demonstrate significant improvements in read query performance. This work is particularly relevant for database designers seeking to optimize NoSQL schema design while balancing factors such as workload and data model. The article offers valuable insights into the systematic design of document databases, making it a useful resource for those involved in database management and optimization.

[9] Oliver Tobias Schilter, Teodoro Laino, and Philippe Schwaller. Cmd+v for chemistry: Image to chemical structure conversion directly done in the clipboard. *Applied AI Letters*, 5(1), January 2024.

This article introduces Clipboard-to-SMILES Converter (C2SC), a macOS application designed to facilitate the conversion of molecular structures directly from the clipboard. The app enables users to seamlessly convert screenshots of molecules into various molecular representations, including SMILES, SELFIES, InChI's, IUPAC names, RDKit Mol's, and CAS numbers. C2SC supports effortless conversion between these formats within the clipboard, enhancing workflow efficiency for chemists and researchers. The application automatically saves converted molecules to a local history file, displaying the last 10 entries for quick access. Additionally, it offers several SMILES operations, such as canonicalization and augmentation, along with price-searching capabilities for chemical vendors to find cost-effective purchasing options. C2SC also features continuous clipboard monitoring, automatically converting any supported representations or images detected into SMILES. The user-friendly interface, accessible directly from the status bar, makes C2SC suitable for a broad audience without requiring programming expertise. Most conversions are performed locally, ensuring privacy and efficiency, with internet access only needed for specific tasks like price lookups. Overall, C2SC provides a convenient and efficient solution for converting molecular structures from the clipboard, making it a valuable tool for the chemistry community. This will be great for understanding how to convert images of chemical structures into machine-readable formats, which is relevant for the Junior IS project focused

on building an organic chemistry tool where users will add the starting and end materials.

[10] Pouya Shati, Eldan Cohen, and Sheila McIlraith. Optimal decision trees for interpretable clustering with constraints. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23, 2023.

This article presents a novel approach to constrained clustering using decision trees, aiming to enhance interpretability while adhering to user-defined constraints. The authors introduce a SAT-based framework that allows for the incorporation of domain-specific knowledge through constraints, improving clustering accuracy. The proposed method addresses the limitations of previous approaches by providing strong theoretical guarantees on solution quality and supporting clustering constraints. Experimental results demonstrate the effectiveness of the framework in producing high-quality and interpretable clustering solutions across various datasets. This work is particularly relevant for applications where interpretability is crucial, such as in healthcare and finance, and contributes to the broader field of explainable artificial intelligence (XAI). Given the focus on decision trees in the current Junior IS project exploring organic chemistry tools, this article offers important insights into creating models that are both effective and understandable. Moreover, the emphasis on constraints aligns with the need to incorporate chemical knowledge into the decision-making process.

[11] Victor F.C. Souza, Ferdinando Cicalese, Eduardo Sany Laber, and Marco Molinaro. Decision trees with short explainable rules. *Theor. Comput. Sci.*, 1047(C), August 2025.

This article investigates the construction of decision trees that produce short and explainable rules for classification tasks. The authors propose a novel approach to building decision trees that prioritize interpretability without significantly compromising predictive accuracy. The paper introduces new metrics for evaluating the explainability of decision tree models and presents algorithms designed to optimize these metrics. Experimental results demonstrate that the proposed methods can generate decision trees with concise rules while maintaining competitive performance compared to traditional decision tree algorithms. This work is particularly relevant in contexts where model transparency is crucial, such as in healthcare and finance. The article provides valuable insights into balancing the trade-off between model complexity and interpretability in machine learning. Given the focus on decision trees in the current Junior IS project exploring organic chemistry tools, this article offers important perspectives on creating models that are both effective and understandable.

[12] Rik van der Lingen. Reaction smiles crd 1.37m dataset, January 2025. Dataset.

This dataset, titled "Reaction SMILES CRD 1.37M Dataset," is a comprehensive collection of chemical reaction data represented in the SMILES (Simplified Molecular Input Line Entry System) format. Compiled by Rik van der Lingen, it contains approximately 1.37 million reaction entries, making it a valuable resource for cheminformatics, computational chemistry, and machine learning research. The SMILES format allows efficient encoding of molecular structures and reactions as linear strings, enabling large-scale computational analysis. This dataset supports applications such as reaction prediction, retrosynthesis planning, and the training of chemical synthesis models.