Communications in Computational Chemistry
Commun. Comput. Chem., Vol. 7, No. 4, 289–310
doi: 10.4208/cicc.2025.153.01
Regular Article

Application of Modern Intelligent Algorithms in Retrosynthesis Prediction

Jianhan Liao[1], Xiaoxin Shi[2], Ya Gao[3,*], Xingyu Wang[4,*] and Tong Zhu[1,4,5,6,*]

[1]*Shanghai Engineering Research Center of Molecular Therapeutics and New Drug Development, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, 200062, China;*

[2]*School of Chemistry and Chemical Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Minhang, Shanghai, 200240, China;*

[3]*School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China;*

[4]*NYU-ECNU Center for Computational Chemistry, NYU Shanghai, Shanghai, 200062, China;*

[5]*Shanghai Innovation Institute, Shanghai, 200003 China;*

[6]*AI for Science Institute, Beijing, 100080, P.R. China.*

[*] Corresponding authors: gaoya@sues.edu.cn; xywang@nyu.edu; tongzhu.work@gmail.com

Received on 2 June 2025; Accepted on 23 June 2025

Abstract: In recent years, the rapid advancements in computer science have spurred the development of various cutting-edge intelligent algorithms. Among these, the transformer, which is built upon a multi-head attention mechanism, is one of the most prominent AI models. The advent of such algorithms has significantly advanced retrosynthesis prediction, though challenges remain in chemical interpretability and real-world deployment. Unlike traditional models, AI-based retrosynthesis prediction systems can automatically extract chemical knowledge from vast datasets to forecast retrosynthesis pathways. This review provides a comprehensive overview of modern intelligent algorithms applied to retrosynthesis prediction, with a particular focus on artificial intelligence techniques. We begin by discussing key deep learning models, then explore available chemical reaction datasets and molecular representations. The discussion extends to the latest state-of-the art in AI-assisted retrosynthesis models, including template-based, template-free, and semi-template-based approaches. Finally, we compare these models across various classifications, highlighting several challenges and limitations of current methods, and suggesting promising directions for future research.

Key words: artificial intelligence, retrosynthesis prediction, machine learning, deep learning.

1. Introduction

Organic synthesis, often regarded as an art due to its reliance on creativity, inspiration and aesthetic judgment [1,2], is a crucial branch of chemistry with extensive applications in drug design and synthetic biology [3–6]. Retrosynthetic analysis, a common method in organic synthesis design [7], involves deducing a synthetic route by working backward from the target compound. The process systematically identifies potential starting materials that can be reassembled to yield the desired compound. However, the growing complexity and diversity of target molecules have made the design of organic synthesis pathways increasingly challenging. To address these obstacles and enhance reproducibility and efficiency, there has been a burgeoning interest in automating organic retrosynthesis [8–10]. Computer-aided synthesis planning (CASP) emerged from this need, tracing its origins to Corey's groundbreaking rule-based synthesis prediction systems, the Logic and Heuristics for Automated Synthesis Analysis (LHASA) [11]. Despite its innovative approach, early rule-based approaches were limited by computational power and data availability.

Recent advancements in computer science have facilitated the development of intelligent algorithms capable of addressing a wide range of tasks [12], such as beam search algorithms, Monte Carlo tree search algorithms, genetic algorithms and neural network algorithms. The influx of big data has propelled the development of numerous artificial intelligence models [13,14], reinvigorating interest in AI applications across chemistry and drug discovery [15,16].

For chemists, CASP poses significant challenges, especially in retrosynthesis prediction, where limited input data can lead to numerous possible outputs. Recent models for retrosynthesis prediction have been designed to automate the identification of candidate reactants from a given product. When these reactants are not commercially available, a recursive expansion strategy is employed, iteratively breaking down the reactants into simpler precursors until all components are accessible or a predefined step limit is reached. Once reliable single-step retrosynthesis is achieved, multi-step retrosynthesis focuses on optimizing the reaction sequence to minimize synthesis steps, costs and waste production.

These models can be broadly categorized into three classes:

1. Template-based models utilize domain knowledge and formal rules derived from prior chemical experiences. Utilizing predefined reaction templates, which specify the transformation of reactants into products, they offer high interpretability and accuracy. However, their applicability is often constrained to scenarios within the predefined knowledge scope, limiting their generalizability.

2. Template-free models, typically devoid of explicit chemical knowledge, such as deep neural networks are considered black-box approaches. While these models have lower interpretability and are prone to violating chemical principles, they have shown promise in discovering new reaction pathways.

3. Semi-template-based models operate in two phases: identifying reaction centers to transform the product into synthons, and then completing these synthons into reactants.

This review provides a comprehensive exploration of contemporary retrosynthetic strategies, focusing on recent advancements in retrosynthesis prediction models. It examines the integration of artificial intelligence (AI) into retrosynthesis prediction, with comparative analysis of commonly utilized data sources and molecular representations. Furthermore, the review evaluates the applications of modern intelligent algorithms across template-based, template-free and semi-template-based models. Finally, it highlights the challenges and outlines potential future directions for this rapidly evolving field.

2. Deep learning algorithms

Artificial intelligence (AI) algorithms are designed to emulate human intelligence, extracting potential rules from datasets and leveraging these rules to make predictions with new data. Deep learning (DL), a rapidly advancing branch of AI, has demonstrated exceptional performance across a variety of tasks, fueled by enhanced computational power and modern algorithms. Generally, DL models can be divided into three categories: supervised learning, unsupervised learning, and reinforcement learning (RL).

In supervised learning, models are trained using datasets containing labeled samples, where the models learn to map input features to output labels. There are two primary types of supervised learning: classification and regression. Classification models predict discrete output labels, while regression models are designed to forecast continuous values. Unsupervised learning, on the other hand, involves training models with datasets of unlabeled samples. These models autonomously identify patterns and relationships in the data without explicit guidance. Reinforcement learning allows an agent to learn through trial and error within a specific environment. The agent receives rewards for actions that produce desirable outcomes and penalties for undesirable ones, aiming to develop a strategy that maximizes expected rewards over time. Most retrosynthesis prediction models employ the supervised learning strategy, as illustrated in **Figure 1**.

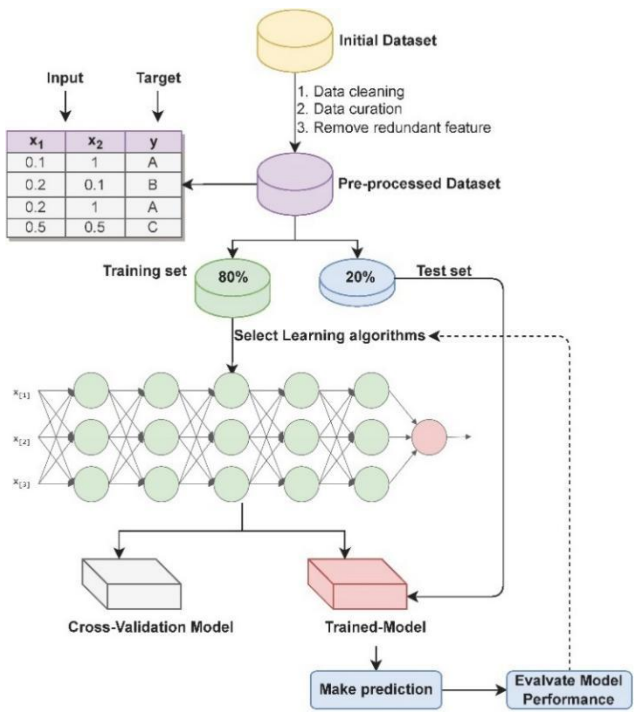

Figure 1. The process of supervised learning method.

Common DL algorithms used in retrosynthesis prediction include sequence to sequence (Seq2Seq) models, graph neural networks (GNNs), reinforcement learning (RL), and various search