REVIEW Artificial Intelligence

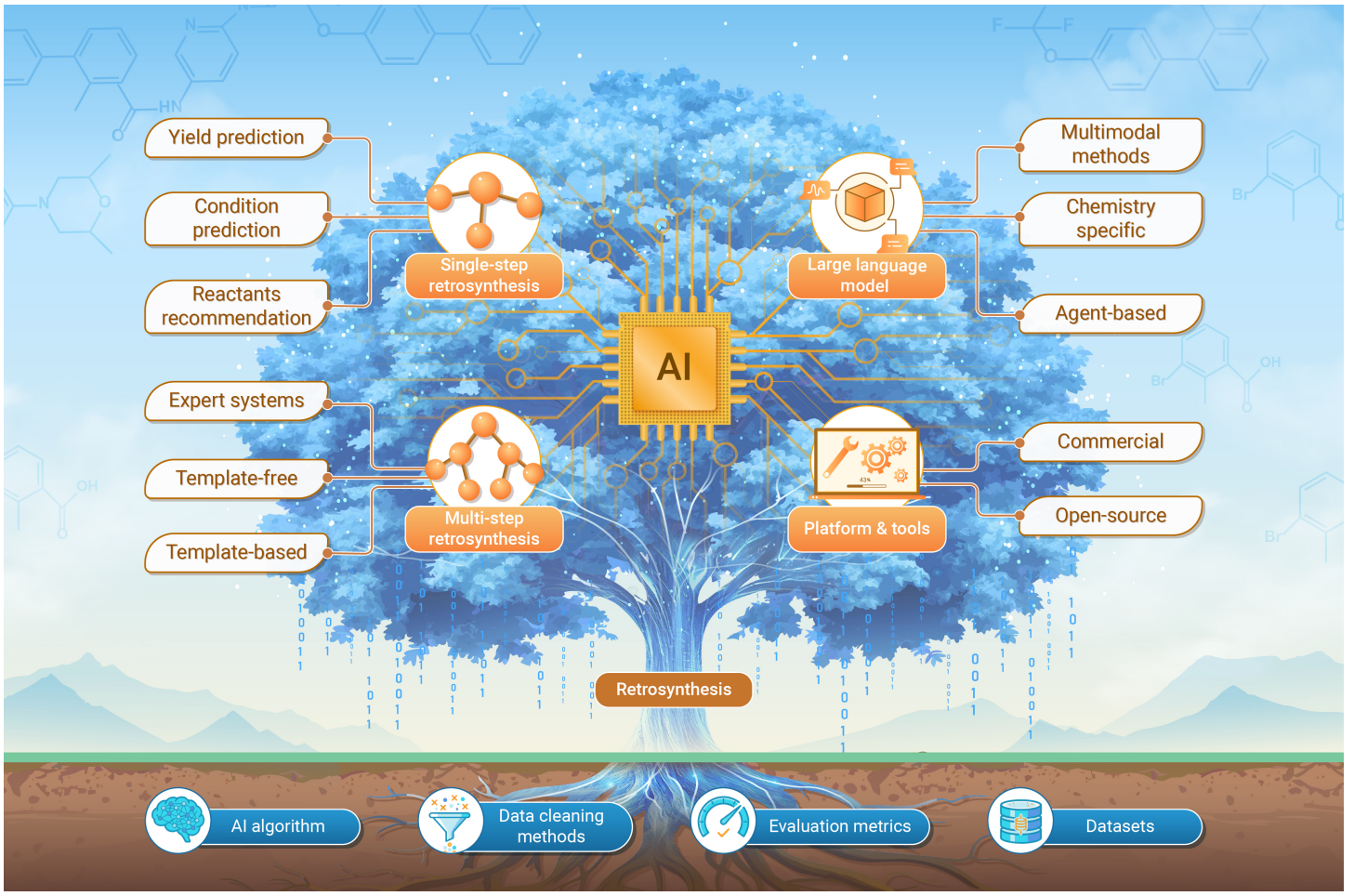
A comprehensive survey of AI-based retrosynthesis planning: Datasets, models, and tools

Peng-Cheng Zhao,[1,4] Hong-Ze Zhou,[1,4] Qiong Wang,[1] Zhen-Yi Wu,[1] Hui Yu,[2,*] and Jian-Yu Shi[1,3,*]

*Correspondence: huiyu@nwpu.edu.cn (H.Y.); jianyushi@nwpu.edu.cn (J.S.)

Received: September 19, 2025; Accepted: November 9, 2025; Published Online: December 9, 2025; https://doi.org/10.59717/j.xinn-inform.2026.100026

GRAPHICAL ABSTRACT



PUBLIC SUMMARY

- Retrosynthetic planning breaks complex molecules into simpler, purchasable starting materials.
- Advances in AI, especially LLMs, are rapidly transforming retrosynthesis.
- This review summarizes single-step tasks, including reactant, condition, and yield prediction.
- It also outlines multi-step planning methods, along with the relevant datasets and platforms.
- We highlight the growing role of LLMs in retrosynthesis and discuss key challenges and future directions.

OPEN ACCESS

Artificial Intelligence REVIEW

The Innovation Informatics

A comprehensive survey of AI-based retrosynthesis planning: Datasets, models, and tools

Peng-Cheng Zhao,[1,4] Hong-Ze Zhou,[1,4] Qiong Wang,[1] Zhen-Yi Wu,[1] Hui Yu,[2,*] and Jian-Yu Shi[1,3,*]

[1]School of Life Sciences, Northwestern Polytechnical University, Xi'an 710072, China

[2]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

[3]Beijing Fengming Technology Co., Ltd., Beijing 100000, China

[4]These authors contributed equally

*Correspondence: huiyu@nwpu.edu.cn (H.Y.); jianyushi@nwpu.edu.cn (J.S.)

Received: September 19, 2025; Accepted: November 9, 2025; Published Online: December 9, 2025; https://doi.org/10.59717/j.xinn-inform.2026.100026

Citation: Zhao P.-C., Zhou H.-Z., Wang Q., et al. (2026). A comprehensive survey of AI-based retrosynthesis planning: Datasets, models, and tools. *The Innovation Informatics* **2**:100026.

Retrosynthetic planning has long been recognized as a central challenge in organic synthesis, and in recent years it has increasingly become a key domain for the integration of artificial intelligence (AI) methodologies. This review provides a systematic overview of recent advances across retrosynthesis planning, including models, datasets, evaluation metrics, and practical platforms and tools that have shaped the field. Beyond the classical single-step retrosynthesis prediction (reactants recommendation), the work particularly emphasizes on two emerging aspects: (i) the prediction of reaction conditions and yields, which is critical for bridging theoretical retrosynthesis with experimental feasibility, and (ii) the application of large language models (LLMs) to retrosynthetic prediction, a rapidly growing direction which emphasizes interactive problem solving and the breaking down of disciplinary boundaries. Moreover, the survey highlights both the current achievements and the persistent challenges of intelligent retrosynthesis. Finally, it outlines future opportunities that may drive the next wave of innovation in AI-assisted synthesis planning.

INTRODUCTION

The prediction of reactants and the design of retrosynthetic routes constitute a central challenge within the domain of chemical synthesis.[1] Retrosynthetic analysis aims to develop methods that, given a target molecule, identify suitable starting materials and feasible synthetic pathways.[2] Such approaches provide chemists with systematic guidance in synthesis design, making the construction of complex molecules more tractable. Traditionally, experienced chemists relied on their knowledge of organic reactions to devise synthetic routes. However, due to the vastness of the possible reaction space and the incomplete understanding of reaction mechanisms, retrosynthetic planning remains an extremely challenging task.[1] To solve these problem, as early as 1969, Corey and Wipke pioneered the use of computer-assisted organic synthesis planning with the OCSS (Organic Chemical Simulation of Synthesis) system, introducing a new paradigm for retrosynthesis design.[3] In recent years, with the parallel development of computer-assisted synthesis planning (CASP) and artificial intelligence (AI) technologies, the design of synthetic routes has been significantly accelerated.[4]

Interestingly, researchers have observed that retrosynthesis shares structural similarities with many AI tasks, making it particularly amenable to AI-based solutions.[5] Retrosynthetic analysis can be broadly divided into two levels: single-step retrosynthesis and multi-step retrosynthesis.[6]

The single-step retrosynthesis task aims to design a synthetic route that can produce the target molecule in one reaction, whereas multi-step methods recursively apply single-step predictions to trace back from the target to purchasable starting materials.[7] This task formulation has inspired analogies with natural language processing (NLP) and graph learning.[8] For instance, drawing on the analogy to language translation, Transformer models have been developed to "translate" a product molecule's SMILES string into SMILES strings of potential reactants.[9] Similarly, molecules can be represented as graphs (with atoms as nodes and bonds as edges), enabling graph neural networks (GNNs) to predict bond disconnections required for retrosynthesis.[10] In multi-step retrosynthesis, the need to navigate through large solution spaces aligns naturally with search algorithms in AI. Representative methods include the A* search-based Retro* algorithm and Monte Carlo Tree Search (MCTS)-based strategies such as MEEA.[11]

The recent emergence of large language models (LLMs) has injected fresh momentum into retrosynthetic research. Their interactive capabilities allow researchers to pose retrosynthetic queries in natural language and receive synthesis suggestions in a conversational manner, thereby lowering the barrier for non-specialists. Moreover, LLMs learn extensive chemical semantics and patterns during pretraining, enabling them to capture latent relationships between molecular structures and reaction pathways even in the absence of explicit rules.[12] This human-like reasoning ability allows LLMs to generate diverse synthesis strategies from a holistic perspective and iteratively refine them based on user feedback. Importantly, LLMs can also be integrated with GNN-based models and search algorithms, offering end-to-end, interactive support for multi-step retrosynthesis. These advances underscore their potential to function as "intelligent synthesis assistants." Thus, the introduction of LLMs not only broadens the methodological scope of retrosynthesis but also lays the foundation for its real-world application.[8]

It is worth emphasizing that retrosynthetic planning involves more than simply identifying potential precursors. A viable and executable synthetic route must also account for reaction conditions and reaction yields (Figure 1). Reaction conditions—including solvent, temperature, pressure, catalysts, and additives—directly influence reaction feasibility and selectivity; in many cases, even minor changes in conditions can result in entirely different products.[13] Therefore, condition prediction is not merely supplementary information but a decisive factor in assessing the practicality of a proposed route. Likewise, reaction yield is arguably the most critical quantitative metric in synthetic chemistry, determining both laboratory feasibility and industrial scalability.[14] A candidate pathway with prohibitively low expected yield, though theoretically valid, is unlikely to be adopted in practice. For this reason, yield prediction is an indispensable component of modern retrosynthetic analysis.

In this review, we define the single-step retrosynthesis module as an integrated system comprising three tightly coupled components—reactant recommendation (the traditional single-step retrosynthesis task), reaction condition prediction, and yield estimation. These modules collectively determine the quality, diversity, and reliability of each synthetic transformation, thereby setting the upper limit for multi-step retrosynthetic planning. A multi-step search algorithm, regardless of its sophistication (e.g., MCTS, A search, or reinforcement learning), can only be as effective as the accuracy and generalization capacity of its underlying single-step predictors.[15] Thus, the interplay between single-step modules and multi-step route exploration is both hierarchical and synergistic: accurate and diverse single-step predictions expand the search space with chemically meaningful options, while multi-step algorithms provide contextual constraints that guide single-step refinements through iterative feedback.[16] Building on this perspective, the present review not only provides a systematic overview of single-step retrosynthesis and multi-step retrosynthesis methods but also highlights recent progress in condition and yield prediction, aiming to offer a more comprehensive and pragmatic evaluation of retrosynthetic feasibility.

In summary, retrosynthetic planning has evolved into both a core challenge in organic synthesis and a prominent application area for AI methodologies. To systematically capture recent progress and emerging trends, this review is organized as follows: Section 2 focuses on single-step retrosynthesis module, covering reactants recommendation (template-based, template-

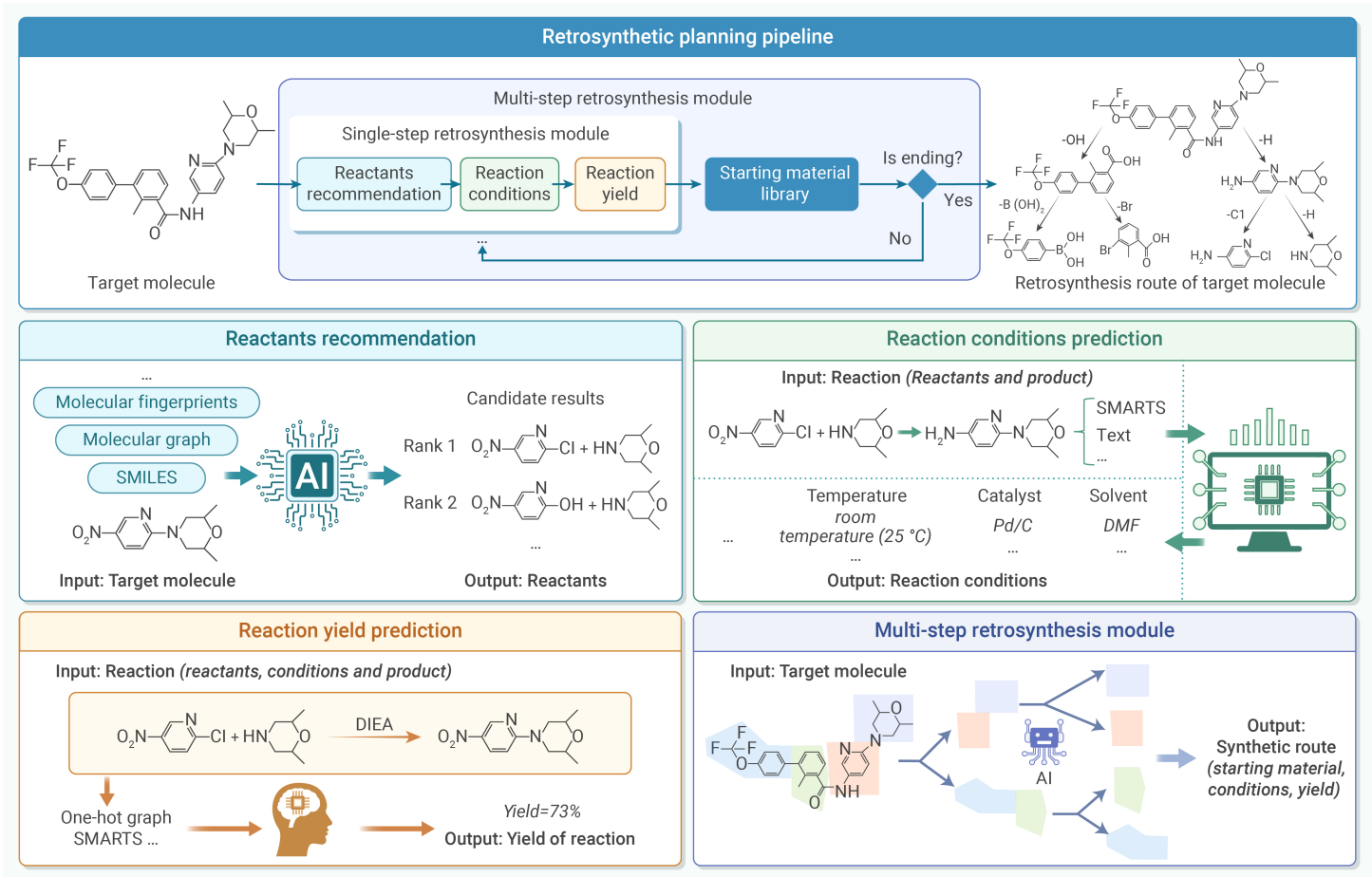


Figure 1. Schematic diagram of the retrosynthesis pipeline The upper panel depicts the overall retrosynthesis workflow, illustrating the interconnections among different tasks in the pipeline. The four modules in the lower panel further detail the key tasks: (A) Reactants Recommendation, which takes a target molecule as input and outputs candidate reactants. (B) Reaction Condition Prediction, which maps a specific reaction to suitable experimental conditions such as reagents, catalysts, solvents, or temperature. (C) Reaction Yield Prediction, which estimates the product yield based on the reactants and predicted conditions; and (D) Multi-step Retrosynthesis Prediction, which recursively decomposes the target molecule into purchasable precursors, thereby generating complete retrosynthetic routes. Together, these tasks provide a systematic framework for efficient and reliable retrosynthetic planning.

free, and semi-template methods), as well as advances in condition and yield prediction, with reference to available datasets and evaluation metrics. Section 3 discusses multi-step retrosynthesis, including commonly used search algorithms, the development of multi-step planning frameworks, associated datasets, and starting material library. Section 4 summarizes and compares existing retrosynthesis platforms and tools, highlighting their value in practice. Section 5 examines the role of large language models (LLMs) in retrosynthetic prediction, which has emerged as a research hotspot in recent years. Section 6 discusses data cleaning and standardization, and Section 7 introduces the evaluation metrics. Finally, Section 8 discusses the current challenges and future prospects of intelligent retrosynthesis, offering valuable insights into potential directions for future research.

SINGLE-STEP RETROSYNTHESIS MODULE (REACTANTS RECOMMENDATION, REACTION CONDITIONS AND YIELDS PREDICTION)

The single-step retrosynthesis module aims to predict one-step reactions for a given target molecule, including reactants recommendation (traditional single-step retrosynthetic prediction), reaction conditions prediction, and expected yields prediction. It forms the foundational layer of multi-step retrosynthetic planning. A critical shortcoming is that most existing single-step retrosynthesis methods primarily focus on the structural mapping between products and reactants, while overlooking the prediction of reaction conditions and yields.[17] This omission directly constrains the feasibility and practical value of proposed synthetic routes. Accordingly, in addition to reviewing recent advances for reactants recommendation, this section places particular emphasis on the current progress in condition prediction and yield prediction. More specifically, this section reviews recent methodological advances, the datasets employed, and the evaluation metrics used for each task (Figure 2).

Prediction method

Reactants recommendation (Traditional single-step retrosynthetic prediction). In recent years, both the accumulation of chemical synthesis data and the growth of deep learning methods have accelerated the rapid development of computer-assisted synthesis processes (CASPs) for reactants recommendation.[17] Recent advances in this area can be broadly categorized into three methodological paradigms: template-based methods, which rely on historical reaction templates for matching and prediction; template-free methods, which use sequence models or graph neural networks to directly generate reactants; and semi-template approaches, which mimic the chemists by decomposing reactants recommendation processes into reaction-center identification and leaving-group recognition. Each of these strategies offers unique advantages in terms of accuracy, diversity, and interpretability.[10] This section provides a summary based on the three scenarios discussed above.

Template-based methods infer the single-step retrosynthesis process of a newly given molecule by finding appropriate reaction templates.[18] These techniques can be categorized into similarity-based, classification-based, and embedding-based approaches. Similarity-based approaches directly leverage similarity metrics to match the target molecule to pre-collected reaction templates.[19] For example, Coley et al.[20] utilized the Tanimoto similarity on Morgan molecular fingerprints to search for suitable retrosynthesis templates. Dai et al.[21] proposed a conditional graphical model constructed upon graph learning networks (GLNs) and performed subgraph pattern matching to acquire candidate templates. Classification-based approaches,

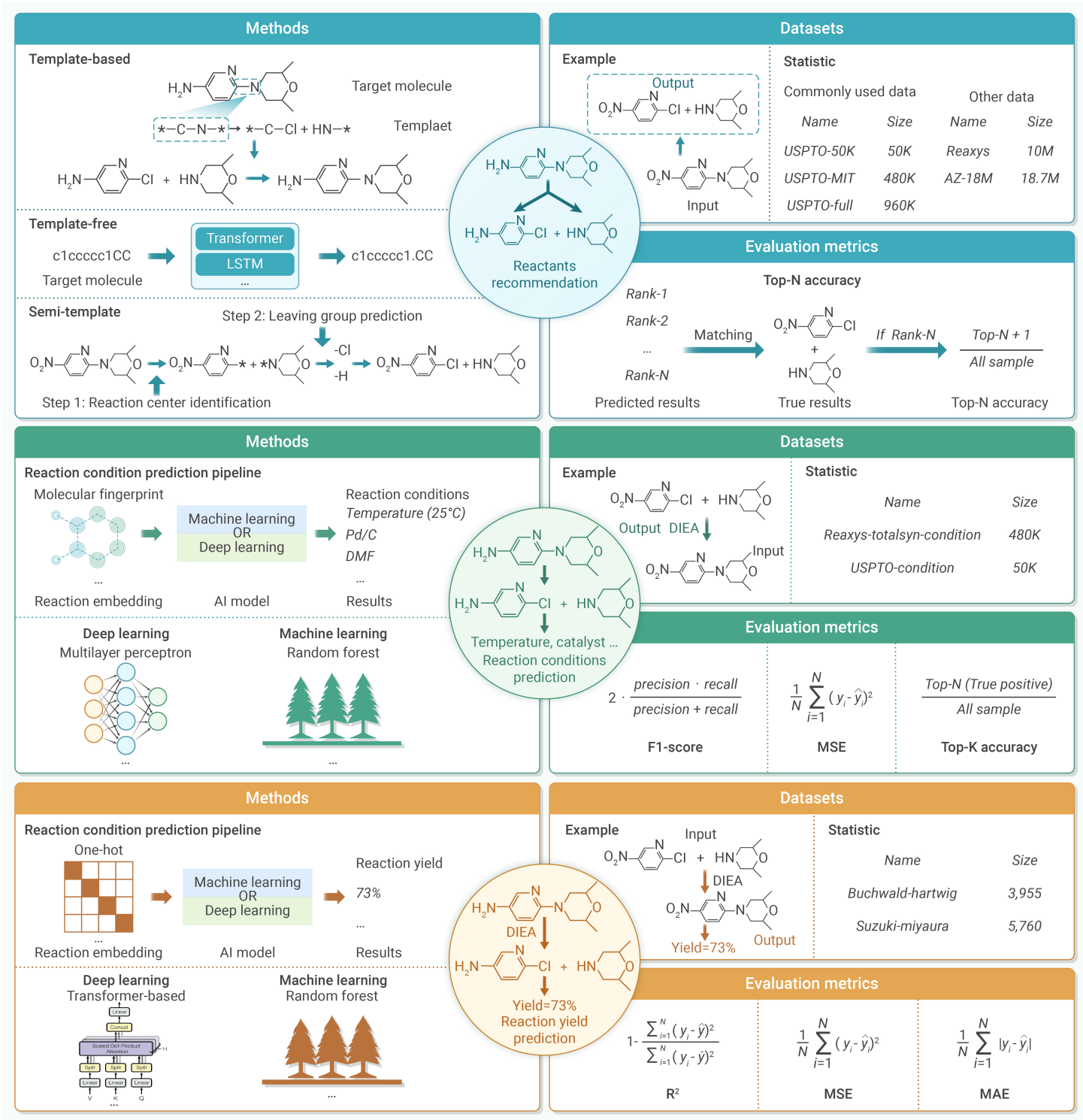


Figure 2. Schematic illustration of the research content in the single-step retrosynthesis module The figure highlights three core tasks: the upper panel shows Reactants Recommendation, where a target molecule is decomposed into potential reactants; the middle panel illustrates Reaction Condition Prediction, mapping reactions to appropriate experimental conditions; and the lower panel presents Reaction Yield Prediction, which estimates the success and efficiency of reactions. Each task is supported by corresponding methods, datasets, and evaluation metrics.

which treat templates as class labels and target molecules as samples, train multiclass classifiers to determine candidate templates for a given target molecule. For example, Segler et al.[22] trained a multiclass deep neural network with extended-connectivity fingerprints (ECFP4) to infer appropriate templates. After organizing templates into reaction type-specific groups, Baylon et al.[23] trained a set of reaction type-specific deep highway networks with Morgan fingerprints to match the correct templates in each template group. Embedding-based approaches map both target molecules and templates into a common embedding space, where a good match is declared if a template is near a target molecule. For instance, Chen et al.[24] designed an MPNN variant (LocalRetro) to encode templates and target molecules. Seidl et al.[25] proposed a modern Hopfield network (MHN) to measure the relevance between templates and the target molecule (initialized by ECFPs). However, template-based methods cannot predict the retrosynthesis results of target molecules with novel synthesis patterns that are outside the synthesis rules contained in the utilized template library. In addition, it is tedious to update template libraries when new synthesis knowledge is discovered.[26] Recent efforts have attempted to alleviate these limitations by generating or

dynamically expanding templates. For instance, Nguyen et al.[27] introduced TempRe, which formulates template generation as a sequence prediction task, enabling the automatic expansion of template libraries and extending applicability to novel reaction patterns. In contrast, both template-free and semi template-based methods can predict single-step retrosynthesis results without a prebuilt template library.

Template-free methods are inspired by machine translation and directly convert the input target molecule into its reactants in the form of strings.[28] The simplified molecular input line entry system (SMILES) is a string notation strategy for representing molecules and reactions. It is a true chemical notation language with a simple vocabulary, including atom symbols, bond symbols, branches, cyclic structures, disconnected structures, and no space.[29] Each unit in the SMILES can be regarded as a word in a machine translation model. This paradigm has boosted the development of template-free methods. In earlier years, Liu et al. proposed an attention-enhanced long short-term memory (LSTM) model[18] to convert the target molecule into its reactants under a sequence-to-sequence encoder-decoder architecture.[30] As new superstars in the natural language processing domain, transformers[31] have been directly applied to retrosynthesis prediction in recent years since a SMILES string is equivalent to a chemical notation sentence.[32,33] However, these methods have led to new issues in which the generated reactants are probably invalid in terms of chemistry. More efforts have been made to address this issue. Zheng et al.[34] designed an extra syntax postchecker based on a transformer to satisfy the chemical validity requirements of generated reactants. Ucak et al.[35] treated molecular substructures (capturing their local atomic environments) as words in a transformer to guarantee the validity of the generated reactants. Similarly, Fang et al.[36] defined a vocabulary of atom-localized substructures to convert a SMILES string into a word sequence under the transformer architecture. However, these methods using SMILES representations cannot effectively capture the rich information hidden in molecular chemical structures (e.g., atomic properties, bond features, and adjacent structure matrices).[37] The subsequently developed methods integrate structural information into transformer-based retrosynthesis prediction frameworks. Mao et al.[38] proposed a graph attention network (GAT) to encode atoms on molecular graphs, enhancing the atom embeddings in transformers. Lin et al.[39] enhanced a self-attention module via node centrality and node positions encodings derived from bipartite molecular graphs. Tu et al.[40] utilized a directed message passing neural network (D-MPNN) on a molecular graph to encode atoms and directly used the decoder of a transformer to generate reactants. To further improve both validity and performance, Deng et al.[41] proposed RSGPT, a large-scale generative pretraining framework that adapts transformer architectures with reinforcement learning to retrosynthesis prediction, achieving substantial gains in Top-1 accuracy on benchmark datasets. In addition, recent works have started to integrate 3D molecular conformer information into template-free models, demonstrating improved handling of stereochemically sensitive molecules and complex ring systems (2025). Furthermore, graph–sequence hybrid architectures have been developed to combine the structure-awareness of GNNs with the sequence-generation capability of transformers, leading to improved validity and diversity of the generated reactants (2025).[42] However, since the process of generating SMILES strings in template-free approaches sequentially outputs individual symbols, the interpretability of the resulting predictions is limited.[43]

Semi template-based methods are inspired by the retrosynthesis inference process employed by chemists, performing two-phase retrosynthesis prediction tasks, including reaction centre prediction and leaving group (LG) prediction. The former subtask finds the reaction centre (i.e., a bond consisting of two reaction sites) where the molecule of interest is split into two synthons (i.e., incomplete reactants).[44] The latter infers appropriate functional groups (i.e., LGs) that attach to two synthons to form reactants. Analogous to template-free methods, some semi template-based methods still adopt translation models to perform two-phase retrosynthesis prediction. For example, Wang et al.[45] formulated the reaction centre prediction and LG prediction tasks as two sequence-to-sequence problems (i.e., target molecules to synthons and synthons to reactants, respectively), which were solved by two independent transformers respectively. To characterize the rich information possessed by molecular structures, some methods treat bonds as samples and utilize graph neural networks to achieve an improved reaction centre identification effect via bond classification. For instance, Shi et al.[46] utilized relational graph convolutional networks (R-GCNs) to recognize reaction centres and converted completed synthons into reactants via a variational graph translation model. Yan et al.[47] applied a GAT variant to predict reaction centres and a sequence-based transformer to convert synthons into reactants. To obtain improved reaction centre predictions with richer chemical meanings, graph editing operations (e.g., node/edge additions and deletions) were leveraged in the follow-up methods. Somnath et al.[48] constructed MPNNs to encode atoms/bonds into embedding spaces, where predefined graph editing operations implemented on atoms/bonds were discriminated to determine the reaction centre, and candidate LGs autoregressively picked from a discrete LG vocabulary were subsequently utilized for determining the corresponding synthons. Similarly, Chen et al.[49] designed a graph message passing network (GMPN) to predict graph editing operations on product molecular graphs to obtain synthons. Furthermore, they found appropriate LGs from a predefined LG list in an autoregressive manner. Moreover, considering the connection between reaction centre prediction and LG prediction, some methods have integrated these tasks into a unified autoregressive reasoning process under graph editing operations. For example, by adopting the encoder-decoder architecture in a transformer,[31] Sacha et al.[50] iteratively started multiple graph editing operations from the target molecule until a STOP operation was encountered (i.e., until its reactants were reached). Zhong et al.[43] adopted a similar architecture but employed more detailed graph editing operations. Building on this paradigm, Ye et al.[51] introduced State2Edits, an end-to-end semi-template framework that directly applies serialized graph-editing operations to transform target molecules into reactants, thus simplifying the two-phase pipeline and improving both accuracy and interpretability. Additionally, Tadanki et al.[52] conducted a systematic error analysis of semi-template methods, highlighting common failure modes and proposing evaluation protocols that inform model refinement in graph-editing based retrosynthesis. To jointly learn the correlation between reaction center recognition and leaving group recognition, Zhao et al.[10] achieved the association through a multi-task learning strategy.

In summary, the study of reactants recommendation (traditional single-step retrosynthesis prediction) has become a highly active research area, bringing new momentum to the advancement of this field. Table 1 presents a summary of the performance of major algorithms on the USPTO-50K dataset.

Reaction condition prediction methods. Reaction condition prediction constitutes a crucial component of the single-step retrosynthesis module, yet it is often overlooked by researchers in the field, despite being an essential bridge between reactants and products.[63] Over the past years, numerous approaches have been developed for predicting reaction conditions, including machine learning-based[13] and deep learning-based models.[64]

Early studies often relied on handcrafted features and shallow models. For example, Gao et al.[63] developed FPRCR, a multi-task classification framework using molecular fingerprints and fully connected neural networks to predict catalysts, solvents, and additives. Afonina et al.[65] employed neural networks to rank candidate condition sets, addressing the challenge posed by the vast reaction condition space.

Subsequently, Maser et al.[66] introduced a multi-label classification framework based on graph convolutional neural networks (GCNNs), modeling condition prediction as a multi-label learning task over a fixed candidate pool. Their results on the USPTO_500MT dataset demonstrated the advantage of GNNs in capturing complex relationships between reaction structures and conditions. Wang et al.[67] proposed the ReaCon model, which combines graph neural networks with reaction templates, achieving a top-3 accuracy of 63.48% on a cleaned USPTO dataset, while significantly improving both the diversity and practicality of candidate conditions. Kwon et al.[68] adopted a graph-enhanced variational autoencoder (VAE)[69] to generate multiple plausible condition combinations, addressing the inherent multiplicity of valid solutions. Schwaller et al.[70] used a Transformer-based generative model to predict both products and conditions simultaneously, and Walker et al.[71] leveraged clustering strategies based on historically similar reactions to recommend solvents, exemplifying the integration of diverse information sources. The latest progress includes large-scale pretraining approaches. For

Table 1. Single-step retrosynthesis methods and performance.

Methods		Top-k Accuracy (%)					
		Reaction Type Unknown			Reaction Type Known		
		1	3	5	1	3	5
Template-Based	MHN[25]	51.8	74.6	81.2	–	–	–
	LocalRetro[24]	53.4	77.6	85.9	63.9	86.8	92.4
	GLN[21]	52.5	69.0	75.6	64.2	79.1	85.2
	NeuralSym[22]	44.4	65.3	72.4	55.3	76.0	81.4
	RetroComposer[19]	54.5	77.2	83.2	65.9	85.8	89.5
	RetroSim[53]	37.3	54.7	63.3	52.9	73.8	81.2
Template-Free	G2GT[39]	54.1	69.9	74.5	–	–	–
	Graph2SMILES[40]	52.9	66.5	70.0	–	–	–
	RetroTRAE[35]	58.3	–	–	–	–	–
	Retroformer[37]	53.2	71.7	76.6	64.0	82.5	86.7
	R-SMILES[54]	56.3	79.2	86.2	–	–	–
	AutoSynRoute[55]	43.1	64.6	71.8	–	–	–
	Tied Transformer[56]	47.1	67.1	73.1	–	–	–
	DualTF[57]	53.6	70.7	74.6	65.7	81.9	84.7
	MEGAN[50]	55.2	74.6	80.5	67.7	84.8	88.9
	SCROP[34]	43.7	60.0	65.2	59.0	74.8	78.1
	NAG2G[58]	55.1	76.9	83.4	67.2	86.4	90.5
	RSGPT[41]	63.4	84.2	89.2	72.8	87.7	91.7
	Seq2Seq[18]	37.4	52.4	57.0	–	–	–
Semi template-Based	EditRetro[59]	60.8	80.6	86.0	–	–	–
	Retro-MTGR[10]	54.3	76.7	90.1	72.2	88.2	92.8
	G2Retro[49]	53.9	74.6	80.7	63.1	84.2	88.5
	Graph2Edits[43]	55.1	77.3	83.4	–	–	–
	RetroPrime[45]	51.4	70.8	74.0	64.8	81.6	85.0
	G2Gs[46]	48.9	67.6	72.5	61.0	81.3	86.0
	GraphRetro Prediction[48]	53.7	68.3	72.2	63.9	81.5	85.2
	RetroExplainer[60]	57.7	79.2	84.8	66.8	88.0	92.5
	RetroCaptioner[61]	54.3	76.3	82.6	67.2	86.0	90.3
	GTA[62]	51.1	67.6	74.8	–	–	–
	GET[38]	44.9	58.5	62.4	57.4	71.3	74.8
	RetroXpert[47]	50.4	61.1	62.3	62.1	75.8	78.5

instance, Wang et al.[72] developed the Parrot model using a Transformer architecture, achieving improvements of 2.64% and 13.44% in top-3 accuracy over Gao et al.'s method on the USPTO-Condition and Reaxys-TotalSyn-Condition datasets, respectively.

Despite these advances, reaction condition prediction remains a challenging task due to the intrinsic heterogeneity and imbalance of available datasets.[63] Public datasets such as USPTO (1976-Sep2016)[73] and Reaxys[74] exhibit substantial bias toward well-optimized or high-yield reactions, leading to severe underrepresentation of low-yield, failed, or exploratory experiments.[75] This data skew limits the model's ability to generalize to real-world laboratory settings, where suboptimal or partial conversions are common. Furthermore, reaction conditions in the literature are often recorded inconsistently, temperature, time, and catalyst loading may be reported in different units or omitted entirely,[17] making it difficult to establish a unified definition or evaluation criterion. Such inconsistencies complicate supervised learning and can introduce systematic noise during training.

In addition, the combinatorial nature of the condition space (e.g., solvent–catalyst–base combinations) poses significant computational and modeling challenges.[63] While multi-label and generative frameworks alleviate this to some extent,[66] they often struggle with capturing conditional dependencies between reagents, or with predicting rare condition sets unseen during training.[75] Moreover, most current benchmarks focus on discrete condition classification rather than continuous parameter optimization (such as temperature or concentration),[76] which further constrains the predictive realism of existing models. Addressing these challenges will require not only more balanced and standardized datasets but also hybrid modeling frame-

Table 2. Summary of reaction condition prediction methods.

Methods	Datasets	Exact Matches (%)		
		Top-1	Top-3	Top-10
Gao et al.[63]	Reaxys[63]	–	50.1	57.3
Wang et al.[67]	USPTO[67]	44.5	63.5	78.6
Wang et al.[72]	USPTO-Condition[72]	26.9	40.4	49.1
	Reaxys-TotalSyn-Condition[72]	37.8	56.9	69.5
Afonina et al.[65]	Reaxys (Prospective test set)[65]	48.0	69.0	82.0
	Reaxys (Retrospective test set)[65]	61.0	77.0	88.0
Maser et al.[66]	Suzuki[66]	77.0	91.3	–
	C–N[66]	63.6	86.5	–
	Negishi[66]	78.0	92.0	–
	PKR[66]	82.0	93.5	–
Walker et al.[71]	Friedel–Crafts [71]	79.0	92.8	–
	Aldol addition[71]	78.0	92.8	–
	Claisen condensation[71]	80.1	94.5	–
	Diels–Alder[71]	79.9	98.0	–
	Wittig[71]	68.8	93.9	–

works capable of incorporating physical chemistry priors and experimental feedback.

Collectively, these studies have expanded the methodological landscape of reaction condition prediction, progressing from classification and ranking to generative and pretraining strategies, providing data-driven solutions and standardized benchmarks for the rational selection of catalysts, solvents, and additives. A summary of these methods is presented in Table 2.

Reaction yield prediction methods. Reaction yield prediction determines whether a given reaction pathway should be adopted, and thus it warrants careful consideration in the retrosynthesis domain.[77] Over the past decade, yield prediction has emerged as a frontier in AI-assisted organic synthesis, with researchers developing a diverse array of approaches, including descriptor-based classical machine learning models (e.g., random forest),[78] as well as deep learning methods such as graph neural networks,[70] sequence-based Transformer architectures,[28] and multi-view pretraining frameworks.[79] For example, in a pioneering 2018 study, Ahneman et al.[78] combined electronic and molecular descriptors with a random forest model to predict the yields of C–N coupling reactions in high-throughput experimentation (HTE), demonstrating the potential of machine learning to navigate complex reaction condition spaces.

Following this, deep learning methods have seen wide adoption. Schwaller et al.[28] introduced the Molecular Transformer, which represents chemical reactions as a "chemical language" and achieves both product prediction and yield uncertainty estimation on large-scale patent datasets, establishing a foundation for applying sequence models to yield prediction. Simultaneously, graph neural networks have shown particular strength in capturing structural features of reactions. For instance, Kwon et al.[68] developed an uncertainty-aware GNN framework, which reliably predicts reaction yields across multiple public datasets and provides chemists with confidence estimates.

With the growing availability of large and heterogeneous reaction datasets, pretraining strategies have also been applied to yield prediction. Shi et al.[79] proposed the ReaMVP framework, which employs multi-view pretraining to substantially improve the model's generalization performance on "out-of-domain" reactions. Additionally, recent graph-Transformer hybrid models, such as those developed by Sato et al.[80] have further advanced performance in low-data regimes while enhancing interpretability.

However, despite these advances, several key challenges persist in reaction yield prediction. A primary limitation arises from the scarcity and imbalance of yield data: public datasets such as USPTO and Reaxys are heavily biased toward successful, high-yield reactions, whereas low-yield or failed reactions are often underreported or missing altogether.[77] This leads to distributional bias during model training, causing overestimation of predicted yields and hindering the model's ability to capture reaction failure modes. Moreover, in many datasets, yield annotations are derived from text-mined or heterogeneous experimental sources, resulting in inconsistent definitions, some reports reflect isolated yield, others conversion yield, and in certain cases, estimated yield ranges rather than precise numerical values.[81] These discrepancies introduce label noise and reduce reproducibility across benchmark studies.

Another emerging issue concerns data domain shift: yield data collected from high-throughput experimentation (HTE) are often quantitatively rich but chemically narrow, whereas patent and literature data are chemically diverse but noisy and incomplete.[82] Integrating these sources for robust yield modeling thus requires advanced domain adaptation or transfer learning strategies. Furthermore, the lack of standardized experimental condition metadata (e.g., concentration, pressure, temperature profiles) complicates the disentanglement of kinetic and thermodynamic influences on yield outcomes, limiting the interpretability of even well-performing models.[83]

Collectively, these studies indicate that reaction yield prediction has evolved from traditional descriptor-based methods to deep learning and pretraining frameworks, forming a multi-tiered landscape. Moving forward, addressing data scarcity, bias correction, and standardized yield representation will be pivotal for improving the robustness and fairness of predictive models. Progress continues to be made in model reliability, cross-domain generalization, and practical applicability. A summary of these approaches is provided in Table 3.

Datasets

Datasets serve as the foundation for model development and operation, as the scale, quality, and diversity of datasets directly determine the upper bound of model performance and generalizability.[86] High-quality datasets not only enable fair benchmarking across different methods but also drive innovation by uncovering new research directions.[17] In this subsection, we provide a systematic summary of the datasets available for reactant recommendation, reaction condition prediction, and reaction yield prediction, with a detailed overview presented in Table 4.

Datasets for reactants recommendation. Public benchmark datasets for reactants recommendation are primarily derived from patent databases. The

Table 3. Summary of reaction yield prediction methods.

Datasets	Methods	R^2	RMSE	MAE
Buchwald-Hartwig[84]	Shi et al.[79]	0.97	4.63	3.11
	Kwon et al.[68]	0.97	4.51	3.01
	Sato et al.[80]	0.97	–	–
	Ahneman et al.[78]	0.92	7.80	–
Suzuki-Miyaura[85]	Shi et al.[79]	0.86	10.37	6.59
	Kwon et al.[68]	0.89	9.52	6.23
	Sato et al.[80]	0.88	–	–

Table 4. Summary of datasets for the single-step retrosynthesis module (reactants recommendation, reaction conditions prediction and reaction yield prediction).

Dataset	Task Type	Source	Records
USPTO-50K	Retrosynthesis prediction	Public patents	~50,000
USPTO-full / PaRoutes-1M	Retrosynthesis prediction	Public patents	~480K–1M
Pistachio	Retrosynthesis prediction	Commercial (NextMove)	~15M
Reaxys	Retrosynthesis prediction	Commercial (Elsevier)	Tens of millions
AZ-18M	Retrosynthesis prediction	Internal (AstraZeneca)	~18.7M
USPTO-Condition	Reaction condition prediction	Public patents	680,741
Reaxys-TotalSyn-Condition	Reaction condition prediction	Literature + patents	180,129
Buchwald–Hartwig HTE	Reaction yield prediction	High-throughput (public)	3,955
Suzuki–Miyaura HTE	Reaction yield prediction	High-throughput (public)	5,760
CJHIF	Reaction yield prediction	Literature (public)	~3.2M
USPTO-Yield	Reaction yield prediction	Public patents	~500,000

most used datasets are USPTO-50K (approximately 50,000 labeled reactions) and large-scale datasets such as USPTO-MIT/USPTO-full (480 K/960 K) and PaRoutes-1M (nearly one million reactions).[28,53,87] These datasets are curated from publicly available patents, covering a wide range of reaction types and providing standardized benchmarks for model evaluation.

In addition, larger commercial reaction databases are available, such as Elsevier's Reaxys, which encompasses hundreds of thousands to millions of chemical reactions extracted from patents and the scientific literature.[74] Based on this, AstraZeneca integrated Reaxys, Pistachio, and internal experimental records to construct the AZ-18M dataset, containing approximately 18.7 million cleaned reactions.[88] Although these commercial datasets are not publicly accessible, they offer broad reaction coverage and are often used to augment training data or serve as internal benchmarks. With the rise of open science, emerging open reaction databases (ORDs) have begun providing limited publicly available reactions, which can supplement training.

Overall, USPTO datasets remain the primary public benchmarks for single-step retrosynthesis due to their accessibility and high standardization, whereas commercial databases provide support for scaling and diversity.

Datasets for reaction condition prediction. In recent years, several public datasets have been developed specifically for reaction condition prediction, covering catalysts, solvents, additives, and other parameters. For example, Wang et al. curated two large condition datasets from USPTO patents and the Reaxys database: USPTO-Condition (approximately 680,741 reactions, without temperature information) and Reaxys-TotalSyn-Condition (approximately 180,129 reactions, including temperature).[72] Each reaction in these datasets is annotated with multiple conditions, including catalyst, solvent, additive, etc., with cases where a condition is not required treated as a "Null" class.[72] Due to its large size and public availability, USPTO-Condition has become a common benchmark for condition prediction. Reaxys-TotalSyn-Condition, although smaller, provides additional details such as temperature. These datasets are characterized by diverse reaction types and comprehensive condition annotations, making them suitable for multi-task learning (e.g., predicting catalysts, solvents, and additives simultaneously). It should be noted that most condition datasets are derived from patents or literature compilations, and their quality depends on text extraction and manual verification. As a result, researchers often perform additional cleaning or balancing before model training.

Datasets for reaction yield prediction. Reaction yield prediction typically relies on high-throughput experimentation (HTE) data and reaction records extracted from patents or literature. Representative HTE datasets include the Buchwald–Hartwig C–N cross-coupling dataset (provided by Ahneman et al.[78]), which employed 1536-well plate experiments combining 15 halogenated arenes, 4 ligands, 3 bases, and 23 isoxazole additives, yielding 3,955 experimental combinations; and the Suzuki–Miyaura C–C cross-coupling dataset (provided by Perera et al.[89]), comprising 15 halogenated arenes, 12 ligands, 8 bases, and 4 solvents, totaling 5,760 experimental combinations. These HTE datasets are high-quality and controllable, making them standard benchmarks for yield prediction modeling.

Large-scale literature-based reaction collections have also been reported, such as Chemical. AI's CJHIF dataset, which contains approximately 3.2 million reactions extracted from publicly available high-impact chemistry journals, annotated with reactants, products, and yields,[79] and millions of reactions automatically extracted from USPTO patents with yield information.[73] Additionally, pharmaceutical companies such as AstraZeneca have released reaction yield data derived from real electronic laboratory notebooks (ELNs), although the scale has not been publicly detailed.[77]

Overall, HTE datasets are commonly used due to their moderate size and rigorous experimental design, while patent- or literature-derived datasets provide much larger volumes (up to tens of millions of reactions) but are noisier and require careful preprocessing.[90]

MULTI-STEP RETROSYNTHESIS

Multi-step retrosynthesis is the process of recursively decomposing a target molecule into purchasable starting materials.[91] Owing to the exponen-

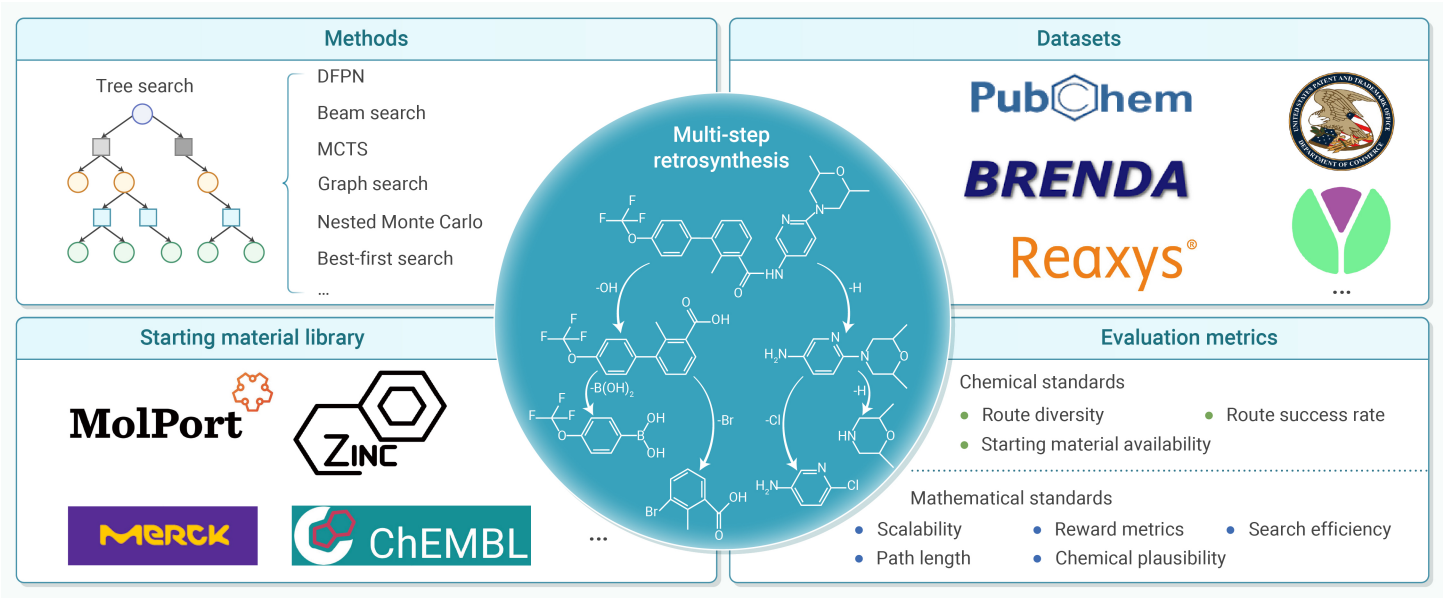


Figure 3. Overview of the content in multi-step retrosynthesis The figure summarizes the major components of multi-step retrosynthesis research. It includes schematic illustrations of representative methods, commonly used datasets, starting material libraries that provide purchasable precursors, and evaluation metrics employed to assess planning performance. Together, these elements form the foundation for developing and benchmarking multi-step retrosynthetic strategies.

tial growth of the combinatorial search space, this task typically requires coupling highly accurate single-step retrosynthesis module with efficient search and planning algorithms to identify chemically feasible and economically reasonable synthetic routes within an acceptable computational budget.[92] In this framework, the performance of multi-step retrosynthesis planning is largely dependent on the reliability of single-step retrosynthesis module, which serve as the fundamental building blocks of route construction.[93] This module generally include reactant recommendation, which proposes candidate precursors for a given product; reaction condition prediction, which ensures the feasibility and selectivity of each proposed transformation; and reaction yield estimation, which evaluates the practical viability and prioritizes higher-probability reaction paths. Together, these modules provide the local chemical reasoning required for each single-step retrosynthetic, while the multi-step planner orchestrates them globally through heuristic, graph-based, or reinforcement learning–based search strategies. However, current research mainly focuses on the coordination between reactant recommendation and search algorithms.[94]

In recent years, leveraging the availability of large-scale data, multi-step retrosynthesis has made significant progress, driven by single-step retrosynthesis modules and AI-assisted search algorithms.[16] In this section, we summarize the current state of multi-step retrosynthesis research from several perspectives, including methodological, datasets, Starting Material Libraries, and evaluation metrics (Figure 3). Methodologically, a diverse array of search strategies has been developed, including Depth-First Proof-Number (DFPN) search,[95] beam search,[16] neural-guided A/Retro*,[96] Monte Carlo Tree Search (MCTS),[22] nested Monte Carlo search,[55] graph-based search,[21] and reinforcement learning approaches.[97] Each method presents trade-offs between exploration and exploitation as well as computational cost.[98] At the data level, the field exhibits a "public small-scale + commercial large-scale" pattern. Public datasets such as USPTO, PaRoutes, and Open Reaction Databases (ORD) coexist with commercial collections like Pistachio and Reaxys. However, these datasets often suffer from noise, atom-mapping errors, missing reaction conditions or yields, and publication bias.[17] In addition, beyond the datasets used to train models, multi-step retrosynthesis requires careful consideration of the availability of Starting Material Libraries, which determine the feasible endpoints of reactions.[99] Specifically, during the recursive process of multi-step retrosynthesis prediction, the program should continuously assess whether to terminate based on the accessibility of starting materials. The choice of Starting Material Library directly affects the practical feasibility of predicted synthetic routes. Yet, these libraries face challenges such as unstable availability, inclusion of virtual or hypothetical compounds, and fragmented access across platforms.[100] Finally, evaluation metrics serve as indicators of the performance of multi-step retrosynthesis algorithms developed by researchers. However, current evaluation standards remain inconsistent, and multidimensional criteria are largely lacking. Existing metrics tend to emphasize success rate or the number of steps, often overlooking factors such as cost, chemical plausibility, and process feasibility.[87]

Available search algorithms and current research status

Search algorithms play a crucial role in multi-step retrosynthesis design because they can assemble single-step retrosynthesis module outputs ("candidate reactions") into complete synthetic routes. By selectively expanding, pruning, and evaluating nodes in a vast solution space, these algorithms can produce executable pathways that lead to purchasable precursors, effectively turning single-step predictions into practical multi-step synthetic plans. Commonly used algorithms include Depth-First Proof-Number (DFPN) search, beam search, neural-guided A/Retro*, Monte Carlo Tree Search (MCTS), nested Monte Carlo, graph-based search, and reinforcement learning approaches. Below, we summarize these strategies in detail:

Depth-first proof-number search (DFPN). DFPN represents the multi-step retrosynthesis problem as an AND–OR tree. Each node maintains a proof number (pn) and disproof number (dn), and expansion follows the "most-proving" path that minimizes the root pn/dn. Nodes are pruned based on thresholds, enabling efficient determination of whether a target molecule is synthesizable with minimal memory usage. By selectively expanding the most "provable" branches, DFPN can quickly identify feasible synthetic routes or efficiently refute infeasible ones, reducing unnecessary single-step module calls. Its depth-first nature allows low memory overhead and easy incorporation of stock availability or chemical constraints at leaf nodes, making it particularly suitable for retrosynthesis scenarios that require provability under limited memory.[95]

Beam search. Beam search starts from an initial sequence or root molecule. At each step, the model generates all possible candidate reactions, retaining the top B candidates as the current beam. Each beam is then expanded to produce B × V new candidates, from which the top B are selected based on scoring. This process repeats until a maximum step limit is reached or all beams terminate. Beam search effectively controls the exponential growth of branches by retaining only the top-k candidates at each step, quickly generating a limited set of high-confidence multi-step routes. It is computationally efficient, easy to implement, and can be combined with neural scoring functions, stock filtering, or cost heuristics. However, its shortsighted nature may overlook optimal routes requiring multi-step foresight, so subsequent refinement (e.g., DFPN, MCTS, or branch-and-bound) is often needed to ensure chemical feasibility and completeness.[28]

Best-first search. This approach treats multi-step retrosynthesis as a heuristic-driven best-first search. Given a target molecule, single-step module, purchasable precursor library, and scoring function, the root node is added to a priority queue. Candidate reactions are generated using the model or policy network, and child nodes are evaluated with cumulative cost and value estimates before being enqueued. The process continues until a solution is found or the computational budget is exhausted.[101] For example, Neural-guided A* prioritizes promising reactions using learned heuristics (policy/value networks), significantly reducing ineffective expansions and single-step calls while increasing the likelihood of finding short-step or low-cost routes. Its performance depends heavily on the quality of heuristics; if inadmissible, global optimality cannot be guaranteed, but it typically produces high-quality, interpretable, and practical candidates within limited budgets.[96]

Monte carlo tree search (MCTS). MCTS operates by iteratively performing selection–expansion–simulation–backpropagation. From the root node, children are selected using UCT or other strategies until an unexpanded node is reached. This node is expanded, and one or more candidate reactions are simulated to a terminal state (e.g., reaching purchasable precursors), generating a reward (e.g., feasibility, cost). Rewards are backpropagated to update node visit counts and average values. MCTS balances exploitation of high-confidence paths and exploration of novel routes, progressively converging toward promising synthetic strategies in large, highly branched search spaces. Single-step model predictions guide expansion and simulation, and backpropagation dynamically adjusts priorities across alternative paths. MCTS is well-suited for tasks with large search spaces and complex path-quality evaluation.[11]

Nested monte carlo. Nested Monte Carlo (NMC) treats multi-step retrosynthesis as a decision sequence. At the lowest level (level 0), fast random or greedy rollouts sample decisions to termination (e.g., reaching purchasable starting materials) to evaluate complete route scores. At higher levels (level ≥1), candidate reactions are evaluated by invoking the next lower level recursively; the best-performing action is fixed, and construction continues iteratively. NMC excels in handling high-branching factors and noisy single-step predictions, effectively capturing long-range dependencies in non-obvious routes. It naturally supports embedding single-step candidate generators, heuristic scoring, or learned policy/value networks to improve efficiency, and enables diverse, highly parallelizable route generation. Its computational cost increases rapidly with nesting depth and rollout number, so practical implementations often use shallow levels and strategy-biased rollouts.[102]

Graph search and reuse. Graph-based approaches represent multi-step retrosynthesis as a directed graph rather than an expanded search tree. Nodes represent intermediates or sub-goals (standardized or hashed to remove duplicates), and edges correspond to possible decomposition reactions or templates. When different paths lead to the same intermediate, they are merged, allowing memoization of previously computed sub-solutions. Graph search facilitates cost propagation, heuristic pruning, Pareto-optimal multi-objective solutions, and parallelization. Advantages include reduced redundant expansions, improved scalability, natural intermediate reuse (enabling shared starting materials or modular synthesis), and support for global optimization and multi-objective evaluation.[103]

Reinforcement learning (RL). In RL-based multi-step retrosynthesis, the decomposition process from the target molecule is formulated as a Markov Decision Process. States represent current unresolved sub-goals (or a single molecule, optionally including step history and accessible stock). Actions correspond to single-step decompositions (reaction templates, candidate reactions, or selecting an intermediate as terminal). Transitions are determined by the single-step model or chemical rules. Rewards are assigned at termination (e.g., positive reward for reaching purchasable starting materials) and may incorporate intermediate shaping for step count, cost, yield, reproducibility, or patent risk. A policy network outputs probabilities for candidate actions, and a value network estimates expected cumulative rewards. Training typically uses Monte Carlo-based rollouts or gradient-based actor-critic methods (PPO, A2C) to reduce variance and improve stability. Trained policies can generate high-quality decompositions directly or serve as heuristics in A*, MCTS, or graph-based searches. Value networks guide pruning and evaluation. RL can learn long-range dependencies and multi-objective "chemical intuition," reducing unproductive branches and improving the probability of identifying low-cost, short-step, and feasible routes. Its main limitations are sensitivity to simulator fidelity and reward design, as well as low sample efficiency.[6]

Research status of multi-step retrosynthesis

Over the past decade, three main lines of development have emerged in multi-step retrosynthesis planning: template-based approaches, template-free generative approaches, and neuro-symbolic or hybrid expert systems. In parallel, the establishment of standardized benchmarks such as PaRoutes[87] and the introduction of quantitative evaluation metrics, including SCScore (synthetic complexity score), coverage, diversity, and route quality scoring, have enabled a more systematic assessment of multi-step retrosynthesis performance and the integration quality between single-step models and global planning algorithms.[104]

Template-based. Template-based methods use manually curated or automatically extracted reaction templates to guide molecular disconnections. These methods initially demonstrated robustness in reactants recommendation: for example, while Segler et al.[16] combined deep policy networks and reaction filter networks within a Monte Carlo Tree Search (MCTS) framework (3N-MCTS), significantly improving both solvable molecule counts and search efficiency. Genheden et al.[105] further open-sourced AiZynthFinder, which integrated MCTS with a template library and introduced the PaRoutes benchmark; systematic comparisons indicated that MCTS generally outperforms several alternative algorithms in terms of route quality and diversity. More recent improvements have focused on automated template extraction and filtering, few-shot template prediction, and reframing template generation as a sequence generation problem to enhance generalizability.[106]

However, template-based strategies inherently depend on the completeness and quality of the reaction template library, which can limit their ability to generalize to novel chemistry or underrepresented reaction types.[107] While they excel in interpretability and can readily integrate human chemical intuition into route planning, their coverage is often constrained by the curated datasets used for template extraction.[107] Moreover, template-matching operations may become computationally expensive when the template pool grows large, leading to efficiency–accuracy trade-offs in large-scale searches.[107] Consequently, template-based systems are particularly effective for well-studied domains such as medicinal or synthetic organic chemistry but may underperform in emerging or unconventional reaction spaces. In contrast, subsequent template-free and generative methods (discussed later) aim to overcome these structural limitations by learning reaction transformations directly from molecular representations, though often at the expense of interpretability.[19] Thus, template-based retrosynthesis can be viewed as a reliable yet domain-constrained foundation upon which more flexible hybrid or neuro-symbolic approaches have been developed.

Template-free. In parallel, template-free (or generative) methods leverage deep generative models to directly predict reactants, thereby avoiding reliance on fixed template libraries. Such methods typically cast retrosynthesis as either a sequence-to-sequence or graph-to-sequence translation problem. Recent research has also split the single-step task into complementary subtasks executed in a loop to improve coverage, or exploited chemical-invariant substructure information via graph neural networks[46] and unsupervised alignment techniques (e.g., UAlign)[108] to further enhance accuracy and chemical plausibility. Moreover, direct route-level generation has been explored. For instance, DirectMultiStep encodes entire synthetic routes as single sequences, thereby reducing search complexity and achieving notable benchmark improvements.[109] Compared to template-based frameworks, template-free generative methods exhibit superior flexibility and scalability, as they can capture a broader reaction space and generalize to unseen transformations without explicit human encoding. Their ability to learn chemical transformations end-to-end from data makes them particularly advantageous in domains with incomplete reaction knowledge or novel compound classes. However, these methods often sacrifice interpretability and controllability—generated reactions may violate chemical constraints or yield unphysical intermediates, necessitating post-hoc validation and filtering. Furthermore, SMILES-based models can be sensitive to syntactic variations, while graph-based approaches, though more chemically grounded, typically

Table 5. Summary of research methods for multi-step retrosynthesis.

Category	Core Idea / Approach	Representative Work / Year
Template-based	Use handcrafted or automatically extracted reaction templates; rank candidates	Segler et al.[16] (3N-MCTS, 2018); Genheden / AiZynthFinder (2020) [105]
Template-free / Generative	Deep generative models (Transformer, GNN→seq) predict reactants directly	Schwaller et al.[28] (Molecular Transformer, 2020); Yu[109] (DirectMultiStep, 2025); Chen[96] (Retro*, 2020); Kishimoto[95] (DFPN-E, 2019); Liu[66](PDVN, 2023); Hong[110](EG-MCTS, 2023); Schreck et al.[111]
Neural-symbolic / Hybrid	Combine symbolic chemical knowledge with data-driven models for interpretability & generalization	Chematica / Synthia[112](classic) ASKCOS (Tu et al.[114]); Former et al.[118]

demand higher computational cost and complex model training pipelines.

Recently, research has focused on embedding single-step models into efficient search strategies to balance computational cost and route quality. Classical and more recent methods include the integration of policy networks with MCTS,[16] the neural-guided A* search algorithm Retro*,[96] improved depth-first proof-number search DFPN-E,[95] and the value-network–driven planner PDVN,[6] which iteratively updates search strategies through online learning of "synthetic feasibility" and "cost" value networks. These integrated frameworks significantly enhance multi-step planning performance by improving route optimality and reducing redundant single-step evaluations. Nevertheless, they remain sensitive to the predictive bias of underlying single-step models, if the base model misjudges reaction feasibility, even sophisticated search algorithms may propagate errors through the route tree. This interdependence highlights the importance of jointly optimizing single-step prediction accuracy and global search efficiency. Experience-guided methods such as EG-MCTS,[10] further enhanced efficiency and route quality by dynamically adjusting scoring functions during search, leveraging historical knowledge to avoid unproductive pathways. Schreck et al.[111] proposed a template-based retrosynthesis framework that formulates synthesis planning as a single-player game, where a reinforcement learning model—guided by a neural network estimating synthesis cost—selects optimal, minimal-step synthetic routes based on a defined cost function, outperforming heuristic methods but requiring high computational resources for exploration. Such adaptive search paradigms represent a promising direction, yet their dependence on extensive prior data and heuristic tuning can pose challenges for generalization across different chemical domains. Overall, while template-free and neural search–integrated methods mark a major leap toward autonomous retrosynthetic reasoning, achieving a stable trade-off between generality, interpretability, and data efficiency remains a central challenge for future research.

Neuro-symbolic and hybrid expert systems. Neuro-symbolic and hybrid expert systems aim to balance the generalization capability of data-driven models with the interpretability and reliability of symbolic chemical knowledge.[22] Early expert systems such as Chematica/Synthia incorporated extensive expert rules and heuristics to propose feasible synthetic pathways.[112] These systems demonstrated unparalleled interpretability and high chemical reliability, as each proposed step could be traced to explicit mechanistic rules or curated reaction templates.[113] However, their dependency on human-coded rules limited scalability and adaptability to new reaction domains, making continuous knowledge expansion costly and time-consuming.[70]

More recently, hybrid systems like ASKCOS (developed by Coley's group) have integrated multiple single-step models, MCTS, feasibility evaluation, and pathway scoring into end-to-end tools now widely used in drug discovery and synthetic chemistry.[114] Such frameworks represent a significant step toward practical deployment, combining the predictive power of neural models with the logical reasoning of symbolic systems.[115] Their major advantage lies in modular flexibility and experimental integration, allowing the entire synthesis planning process—from route generation to lab validation—to operate in a closed loop.[115] Nonetheless, hybrid pipelines remain highly sensitive to the consistency of their internal modules; mismatches between neural predictions and symbolic constraints can lead to route discontinuities or inconsistent feasibility assessments.

Inspired by neuro-symbolic programming, frameworks such as wake–abstraction–dreaming (influenced by DreamCoder[116]) abstract common multi-step patterns from experience, recycle them as templates, and improve planning efficiency and pathway simplicity through retraining and replay, illustrating the potential complementarity of symbolic abstraction and neural networks.[117] Former et al.[118] developed SPARROW, a unified decision-making framework that prioritizes candidate molecules and their synthetic routes by jointly minimizing synthetic cost and maximizing utility through reward–penalty objective functions. This emerging paradigm highlights the potential for continual learning and hierarchical reasoning, neural models extract abstract patterns from data ("wake"), symbolic reasoning generalizes these into reusable knowledge ("abstraction"), and iterative retraining refines them into efficient strategies ("dreaming").[119] However, the main challenge remains how to ensure chemical validity and interpretability during abstraction, since automatically learned rules may not always align with human-understandable chemistry. Moreover, the computational overhead of iterative retraining and replay mechanisms can constrain scalability in large reaction spaces. Overall, neuro-symbolic and hybrid expert systems exhibit the highest potential for real-world application among current multi-step retrosynthesis frameworks, offering a pragmatic balance between accuracy, interpretability, and adaptability. Their success underscores a broader trend toward system-level integration, where learning-based prediction, symbolic reasoning, and experimental feedback are seamlessly coupled in a unified decision framework.[113]

In summary, existing multi-step retrosynthesis methods remain fundamentally limited by the accuracy of single-step modules, making it difficult to fully integrate these independent components. Many published studies have overlooked the intrinsic relationship between single-step and multi-step methods, thereby providing an incomplete picture for building robust retrosynthetic analysis systems.[93] In particular, while neuro-symbolic systems demonstrate promising integrative capacity, their overall performance still depends on the precision and generalization of underlying single-step predictors. Since single-step retrosynthesis module ultimately underpins the multi-step process, future research in multi-step planning should introduce new evaluation metrics that also guide single-step model development.[93] Thus, single-step and multi-step approaches should be regarded not as isolated domains but as tightly interwoven and mutually reinforcing components of retrosynthesis. All multi-step retrosynthesis methods are summarized in Table 5.

Datasets of multi-step retrosynthesis

Currently, datasets for multi-step retrosynthetic planning exhibit a "public small-scale + commercial large-scale" pattern. Academic research typically relies on publicly available USPTO collections (e.g., USPTO-50K, USPTO-MIT/USPTO-full, containing hundreds of thousands of reactions) for reproducible benchmarking.[93] Industrial and larger-scale training often utilizes commercial patent and literature databases such as Pistachio (NextMove) and Reaxys (Elsevier), reaching millions to tens of millions of reaction entries. These datasets are frequently combined with commercial starting material libraries (e.g., eMolecules) to validate purchasability.

While patent and literature sources provide large sample sizes and authentic experimental records, they contain noise (extraction/mapping errors), duplicates, missing or non-standard reagent/condition information, and publication biases (favoring patentable or publishable reactions).[120] These issues directly affect the accuracy of single-step models and are amplified in multi-step planning. Overall, these datasets are sufficient to support training of strong single-step models, but challenges remain in noise (extraction/mapping errors), missing information (conditions, yields, failed reactions), and representativeness (publication/patent bias).[77] Therefore, practical workflows typically require systematic cleaning, atom mapping correction, deduplication,

and integration of purchasable starting material libraries (e.g., eMolecules) for terminal state determination or cost estimation.

USPTO. The USPTO reaction dataset was initially extracted and published by Daniel M. Lowe[33] from full-text US patents (covering 1976–2016), becoming a foundational data source for single-step reaction modeling and retrosynthesis research. Lowe's original extraction has been cleaned and transformed into multiple standardized subsets:

- USPTO-50K (~50,016 reactions) curated by Schneider et al.[121] annotated with 10 reaction classes for classification/benchmark tasks.
- USPTO-480k cleaned variants (~480k high-quality reactions) widely used for multi-step retrosynthesis evaluation.[122]
- The original USPTO extraction contains ~1.7 −1.9M reactions; industry/commercial vendors have even larger extractions (e.g., Pistachio).

USPTO is fundamentally a collection of single-step reaction records (reaction equations or SMILES), not multi-step pathways, multi-step routes are constructed by retrosynthesis algorithms via backtracking and search. Advantages: large scale, coverage of real patented reactions. Limitations: extraction/atom mapping errors, missing/misplaced reagents/solvents/conditions, duplicates and publication/patent bias, sparse yields/failure data. Modern practices use high-quality atom mapping tools (RXNMapper, LocalMapper, GraphormerMapper) combined with rigorous cleaning, deduplication, and temporal splitting to improve usability and reproducibility.

Pistachio. Pistachio, provided commercially by NextMove Software, is a patent reaction database extracted from US, EP, and WIPO patent texts. Reactions can be exported in SMILES/JSON format with a searchable interface. Reported scales vary (~8.3M to ~21M entries), with the number of unique reactions lower than total entries. Advantages: broad patent coverage, inclusion of experimental steps/conditions, valuable for training large-scale models. Limitations: extraction noise, duplicates/redundancy, reagent/solvent/catalyst misclassification, atom mapping errors, and licensing restrictions. Researchers typically apply deduplication, remapping, and filtering before use, or combine it with public datasets for high-coverage augmentation.

Reaxys. Reaxys (Elsevier) is a curated chemistry knowledge base integrating traditional manually indexed resources (Beilstein/Gmelin) with modern literature/patent extraction. It covers tens of millions of reactions and hundreds of millions of substances, with historical records extending to the 18th century. Single-step reactions are the basic unit, but Reaxys provides richer experimental conditions, yields, step descriptions, and literature/patent metadata, enabling reconstruction of multi-step pathways. Advantages: expert curation, comprehensive experimental details, mature search tools. Limitations: commercial access, publication bias, extraction noise, and limited support for large-scale machine learning export.[13]

Open Reaction Database (ORD). ORD is an open chemistry reaction database initiated by MIT, Google Research, and various pharmaceutical/academic partners. It uses Protocol Buffers (protobuf) to store reaction information, including SMILES, temperature, solvent, reagents, time, optimization experiments, and yields. As of 2025, ORD contains ~65,000–70,000 standardized reactions from published literature, patent subsets, and contributor data. Basic units are single-step reactions, though some represent multi-condition experimental clusters. Advantages: fully open, standardized format, richer condition information than USPTO, community-extensible. Limitations: smaller scale than commercial databases and limited pathway coverage, requiring combination with larger datasets for multi-step retrosynthesis.[123]

PaRoutes. Dataset Proposed by Genheden et al.[87] as a unified benchmark for multi-step retrosynthesis. Paths are automatically extracted from USPTO reactions using algorithms like Retro*, yielding ~100,000 feasible synthetic pathways of 3–8 steps. PaRoutes provides full reference routes, including standardized route trees and starting materials, making it suitable for fair algorithm comparison. Limitation: still derived from USPTO, biased toward academic/patent-type synthesis; more suitable as a benchmark than for direct industrial application.[87]

KEGG. Maintained by Kanehisa Laboratory (Kyoto University), KEGG has ~18,000 pathways across 7,000+ species. Path lengths typically 3–20 steps. Strength: comprehensive, standardized, foundational for bioinformatics. Limitation: limited reaction rule formalization; less suitable for ML-driven retrosynthesis modeling.[124]

RetroRules. EPFL team derived this reaction rule set from MetaNetX and BRENDA, containing >100,000 rules for computer-aided retrosynthesis. Supports multi-step pathway generation (1–20+ steps) and is used in MCTS or deep learning models. Strength: structured rules facilitate algorithmic calls. Limitation: coverage limited; less suitable for complex pharmaceutical syntheses compared to commercial databases.[125]

Starting material libraries

In multi-step retrosynthesis, starting material libraries determine whether the generated synthetic routes can be practically executed and provide terminal nodes for the retrosynthesis search. The goal of retrosynthesis is to decompose the target molecule into purchasable starting materials. Without a starting material library, models may produce theoretically valid routes that cannot be realized in practice, reducing the utility of planning. High-quality starting material libraries can significantly improve search accuracy, efficiency, and industrial applicability, ensuring that retrosynthesis results can be closely integrated with experimental procurement and drug discovery.

eMolecules. A commercial database integrating catalogs from over 200 global suppliers, offering more than 50 million purchasable compounds and expandable to 2.8 trillion virtual molecules in chemical space. It includes building blocks, screening compounds, etc., with monthly updates ensuring scale and timeliness. eMolecules provides a free simplified version (basic structure data) and a paid full version (ePLUS), which includes price, supplier, and delivery information.[126]

ZINC. A public database aimed at virtual screening and drug design. Latest versions (ZINC20/ZINC22) include over 700 million commercially purchasable molecules and hundreds of billions of virtual compounds generated via combinatorial expansion. Provides free structure data, supplier info, and download interfaces, with filters for purchasable molecules, drug-likeness, natural product-likeness, etc. Advantages: open access, large scale, closely linked to suppliers; useful both for virtual screening and as a starting material library in multi-step retrosynthesis to determine practical purchasability of intermediates.[127]

ChEMBL. Maintained by EMBL-EBI, ChEMBL is a fully open-access bioactive compound database. The latest version, ChEMBL 33 (2024), contains ~2.4 million compounds with over 2 million activity data points across ~17,000 targets, curated from literature, patents, and clinical trials. Unlike ZINC or eMolecules, ChEMBL emphasizes molecular structure, bioactivity, mechanism, and ADME properties, rather than purchasability. In retrosynthesis research, it serves as a source for target molecules or lead compounds but is not typically used to define terminal purchasable intermediates.[128]

Tokyo Chemical Industry (TCI). A Japanese chemical supplier with hundreds of thousands of commercial chemicals, mainly covering organic synthesis reagents, pharmaceutical intermediates, and specialty chemicals. Database access is limited; full structural and stock information requires official channels or commercial agreements. Used in retrosynthesis for starting material libraries, especially for the Asian market or specific chemicals. (Website: https://www.tcichemicals.com/OP/en/.).

MolPort. A global aggregator of supplier databases, containing ~7 million purchasable compounds from 200+ suppliers. Provides SDF/SMILES downloads; advanced fields (price, stock, procurement info) require registration/payment. Advantages: multi-supplier coverage; limitation: restricted access to full details. (Website: https://pubchem.ncbi.nlm.nih.gov/source/MolPort).

Mcule. A commercial platform with 0.5–0.7 million purchasable molecules, including building blocks, screening compounds, and drug-like molecules. Partial free access with registration; full database, price, and stock info require payment. Strengths: real-time supplier info, reliable molecules; limitations: relatively smaller scale, paid features restrict full access. (Website:https://mcule.com/).

AiZynthFinder Stock. The default stock used by the open-source AiZynthFinder retrosynthesis tool, integrating multiple supplier compounds such as eMolecules and ZINC subsets. Contains millions of molecules, customizable to include new suppliers. Open access, suitable for academic research and algorithm development; used to determine purchasable terminal nodes and can be extended to match specific lab or commercial stocks.[105]

PaRoutes Stock. A starting material set paired with the PaRoutes bench-

mark, used for multi-step algorithm evaluation and training. Based on USPTO reactions and purchasable building blocks, containing tens of thousands of entries. Open-access and suitable for benchmarking, though smaller and less diverse than commercial starting material libraries.[87]

Synthia Stock. Provided by Chematica/Synthia (acquired by Merck & Co.), a commercial stock database for retrosynthesis planning and feasibility assessment. Contains millions of commercially available compounds, covering common building blocks and synthetic intermediates. Closed-source and requires paid access. Strengths: standardized and purchasable compounds; limitation: proprietary access, not fully open (Website: https://www.investing.com/equities/syntheia).

RETROSYNTHESIS PLATFORMS AND TOOLS

Retrosynthesis platforms and tools structure and encode literature, reaction databases, and experimental knowledge to automatically generate, evaluate, and rank executable synthetic routes and operational steps. They enable researchers without deep synthetic chemistry expertise to rapidly design, compare, and implement synthesis plans, reducing the learning curve, improving reproducibility, and accelerating the process from concept to experimental validation. Future directions include stronger multi-step planning based on large language models and graph neural networks, potentially achieving closed-loop integration with automated synthesis platforms. This section highlights the diverse ecosystem of commercial, academic, and open-source retrosynthesis tools, spanning template-based, AI-driven, and hybrid approaches, and indicates trade-offs between openness, coverage, and industrial readiness.

Commercial

Molecule.one. Developed by the Polish startup Molecule.one, this is an end-to-end commercial platform emphasizing route feasibility scoring, large-scale parallel search, and enterprise integration (e.g., with inventory and supplier data).

Pros: High-quality executable routes; scalable enterprise integration.

Cons: Closed-source; limited extrapolation to rare or novel reactions; high cost and integration overhead.

Website: https://molecule.one/

Reaxys Predictive Retrosynthesis. A module within Elsevier's Reaxys, integrated with Iktos AI. Leverages Reaxys literature and reaction databases.

Pros: Combines published routes with AI-predicted ones for traceability and feasibility; API/enterprise integration.

Cons: Subscription required; AI predictions biased by training data; rare or divergent reactions may be missed.

Website: https://www.elsevier.com/products/reaxys/predictive-retrosynthesis

SciFinder CAS SciFinder. SciFinder^{n} includes ChemPlanner for predictive retrosynthesis. Relies on CAS's human-curated literature and patent database.

Pros: High traceability; useful for compliance and patent risk assessment.

Cons: Closed-source; algorithmic details hidden; limited discovery of novel reactions.

Website: https://www.cas.org/solutions/cas-scifinder-discovery-platform

Chematica. Developed by Bartosz A. Grzybowski's team (Chematica) and commercialized by Merck (Synthia). Uses expert rules and heuristic search, suitable for complex natural product and process routes.

Pros: Performs well on complex synthesis; combines rules and heuristics.

Cons: Rule/template-based; coverage depends on manual curation; commercial access limits reproducibility.

Website: https://www.synthiaonline.com/

ICSYNTH. Commercial computer-assisted synthesis planning (CASP) tool. Template/rule-driven interface, focusing on idea generation and alternative routes

Pros: Useful for route brainstorming.

Cons: Template dependence limits generalization; closed-source; academic reproducibility restricted.

Website: https://demo.icsynth.infochem.eu/app/login/auth

Spaya. Developed by Iktos. Commercial AI retrosynthesis tool with real-time route scoring and integration with building block libraries.

Pros: Fast search; easy engineering integration (API).

Cons: Closed-source; performance drops on novel reactions or unlisted starting materials.

Website: https://iktos.ai/solution/spaya

ChemAIRS. Commercial platform by Chemical.AI, emphasizes enterprise synthesis design, impurity prediction, and process optimization.

Pros: Industrial workflow-oriented; integrates route diversity and impurity considerations.

Cons: Closed-source; model/data details unavailable; dependent on vendor support.

Website: https://www.chemical.ai/chemairs

RXN / RoboRXN for Chemistry. Developed by IBM Research. Based on Molecular Transformer and sequence-to-sequence approaches, some versions integrate with automated synthesis (RoboRXN).

Pros: Strong academic/industrial foundation; free educational/demo access; API available.

Cons: Public model performance limited by open data; experimental feasibility requires manual verification.

Website: https://rxn.res.ibm.com/rxn/robo-rxn/welcome

PostEra. Cambridge-based AI pharma company; develops Manifold (chemical search/synthesis design) and Proton (medicinal chemistry platform).

Pros: Focus on drug-design synthesis loop; strong integration with synthesis labs.

Cons: Some tools closed-source or paid; generalization outside pharma chemistry unverified.

Website: https://postera.ai/

Open-source

ASKCOS. Developed at MIT.[17] Open-source/hosted modular synthesis planning suite integrating multiple one-step models, condition prediction, and search strategies.

Pros: Open, modular, reproducible; extensible to custom stock libraries.

Cons: Large-scale workflow optimization may require additional engineering.

Website: https://askcos.mit.edu

AiZynthFinder. Open-source tool from MolecularAI/academic teams. Uses Monte-Carlo Tree Search + reaction templates/neural strategies.

Pros: Open, customizable, fast (seconds to tens of seconds); actively maintained.

Cons: Performance depends on training templates and "buyable" libraries; commercial integration requires engineering.

Website: https://github.com/MolecularAI/aizynthfinder

LillyMol. Internal/open-source chemical informatics library (C++). Used for molecule operations, reaction enumeration, and retrieval.

Pros: Industrial-grade; efficient substructure search and reaction enumeration.

Cons: Low-level library; upper-layer retrosynthesis needs additional strategy/model; installation barrier high.

Website: https://github.com/EliLillyCo/LillyMol

RDChiral. Open-source Python package by Connor W. Coley et al. Focuses on stereochemistry and template application consistency.

Pros: Widely used in CASP/template workflows; ensures stereochemical correctness.

Cons: Only template application layer; highly sensitive to template coverage/quality.

Website: https://github.com/connorcoley/rdchiral

RetroPath2.0. Developed by Jean-Loup Faulon et al. Focused on metabolic engineering and bioretrosynthesis, using rule-driven network search and pathway extraction.

Pros: Strong for chassis → target biodesign; supports pathway network extraction.

Website: https://pubmed.ncbi.nlm.nih.gov/29233745/

Rxnutils/Reaction_utils. Python toolkit by MolecularAI for reaction dataset processing, template extraction, and dataset construction.

Pros: Standardizes large reaction datasets (USPTO, ORD); facilitates reproducible pipelines.

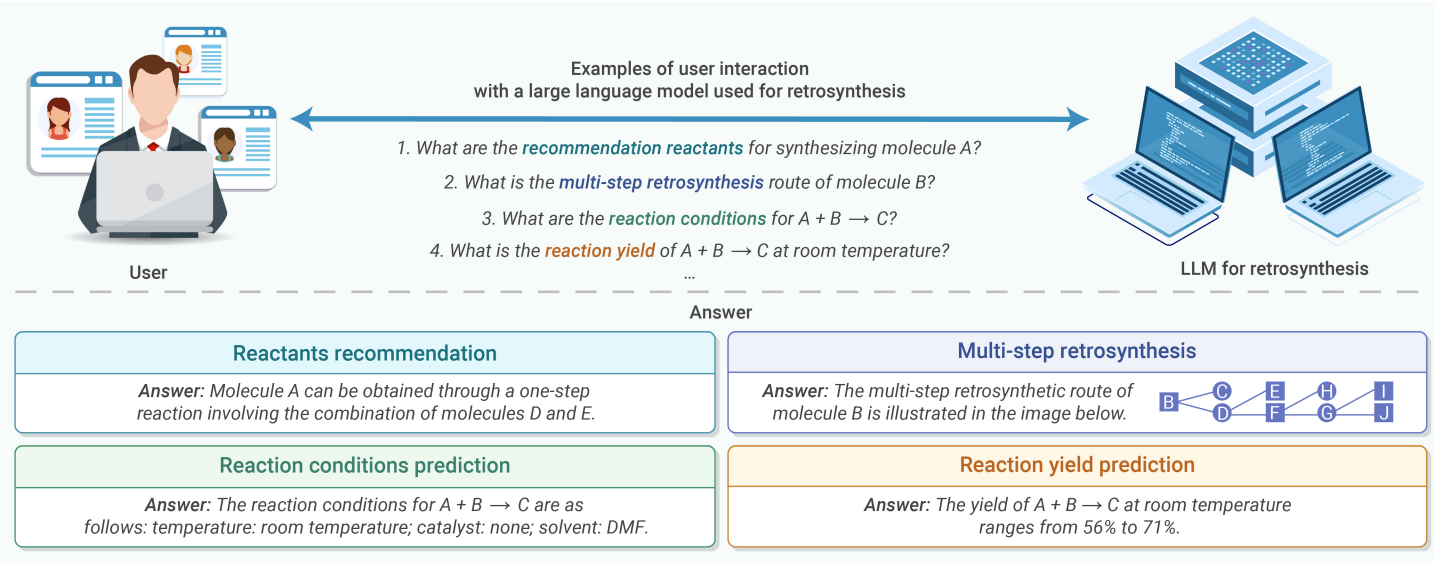


Figure 4. Retrosynthesis assisted by large language models The figure illustrates the interaction between a user and a large language model for retrosynthesis. The upper panel represents examples of user queries, showing how target molecules or retrosynthetic tasks are formulated. The lower panel depicts the model's responses, including recommended reactants, predicted reaction conditions, reaction yield and potential retrosynthetic routes. This interaction highlights the role of large language models in guiding and supporting retrosynthetic planning.

Cons: Primarily for data preprocessing; not a search/planning engine; needs integration with models/search algorithms.

Website: https://github.com/MolecularAI/reaction_utils

LLM-ASSISTED RETROSYNTHESIS PREDICTION

In recent years, Large Language Models (LLMs) have begun to reshape the research landscape for retrosynthesis planning and related predictive tasks. LLMs are generative models based on the Transformer architecture, pretrained on massive corpora and fine-tuned with instruction-based methods. They excel in natural language understanding, cross-task transfer, contextual reasoning, and multi-turn interaction. Unlike conventional deep learning models used in retrosynthesis—which typically rely on reaction templates, graph structures, or molecular fingerprints for single-step prediction—LLMs can flexibly interpret complex constraints in natural language, invoke external tools or knowledge bases, and propose synthetic routes closer to practical laboratory implementation. By understanding chemistry tasks in a natural language interface and integrating tool use and knowledge retrieval, LLMs extend retrosynthesis from a "reaction pathway prediction" problem into an interactive synthesis assistant capable of proposing reaction conditions, estimating yields, and even interfacing with automated experimental platforms to achieve closed-loop optimization (Figure 4).[8,129,130]

The core advantages of LLMs stem from large-scale pretraining and strong language comprehension. Pretraining on extensive patent and literature datasets equips LLMs with rich chemical knowledge, enabling them to generalize to long-tail chemical spaces. Their ability to interpret complex instructions allows them to incorporate experimental constraints and user preferences into synthetic recommendations. Additionally, LLMs can interact with external tools and databases to validate reaction feasibility and filter candidate routes, effectively implementing a "generate-and-verify" loop.[129,131] Nevertheless, a critical limitation remains: most current LLMs still rely on two-dimensional molecular representations such as SMILES or reaction graphs, which fail to capture three-dimensional conformational effects, stereochemistry, and spatial electronic interactions.[132] These 3D structural features play essential roles in determining reaction selectivity, catalyst compatibility, and transition-state energetics—factors directly influencing retrosynthetic route feasibility.[133] Addressing this gap has become a central objective for the next generation of chemistry-oriented LLMs. Based on these capabilities, several LLM-centered retrosynthesis systems and methodologies have emerged.

Agent-based

These systems use a general LLM as the reasoning engine and interface it with chemistry-specific tools (e.g., reaction prediction, name ↔ SMILES conversion, literature search, safety verification, lab interface), enabling chain-of-thought reasoning and action execution. For example, ChemCrow integrates GPT-4 with ~18 specialized tools to iteratively plan reactions, retrieve supporting evidence, and issue experimental instructions, demonstrating a closed-loop workflow from design to experiment.[129] Similar systems, such as SynAsk (based on Qwen-1.5 with LangChain orchestration), convert LLM outputs into tool-executable forms.[134] In a ChEMBL test set of 11,549 small molecules, SynAsk produced retrosynthetic routes for 6,358 compounds (~55%), compared to ~27% for AiZynthFinder,[105] demonstrating the advantage of tool-augmented LLMs in complex molecule planning.[134]

LLMs specialized or fine-tuned for chemical retrosynthesis

These models are pretrained and fine-tuned on large-scale chemical data to enhance chemical semantic understanding and reaction pattern representation. BatGPT-Chem[135] unifies natural language and SMILES representations and is trained on >100 million instruction-style chemical records, achieving superior zero-shot reasoning and multi-task prediction on benchmarks like Suzuki–Miyaura and Buchwald–Hartwig reactions. RSGPT,[41] based on LLaMA2, pretrains on large template-based synthetic reaction datasets and incorporates reinforcement learning from AI feedback (RLAIF), achieving 63.4% Top-1 accuracy on USPTO-50k in zero-template generation tasks. Chemma,[136] fine-tuned on LLaMA-2-7B with >1.28 million reaction Q&A pairs, provides multi-task capabilities including forward prediction, retrosynthesis, condition and yield estimation, and supports active learning with experimental feedback. In open reaction space experiments, Chemma successfully recommended high-efficiency ligands and solvents within 15 experimental cycles, yielding 67%, demonstrating LLM potential for experimental optimization.

Multimodal and structured editing methods

These integrate text, molecular graphs, and 3D information to improve multi-step planning accuracy and controllability. RetroInText[137] incorporates path-level text context into multi-step decisions, encoding ChatGPT-generated textual descriptions alongside product 3D conformations using MolT5 and 3DInfomax, with attention mechanisms fusing text and molecular representations. This improves Top-1 accuracy by 1.8–5% for long-path tasks. Llamole enables switching between text and molecular graph modalities, coupling graph generation (Graph DiT) with GNN-based single-step prediction and LLM-guided A* heuristics, raising retrosynthesis success from ~5.5% to ~35%.[138] ReactSeq explicitly represents atomic-bond-level edits as Molecular Editing Operations (MEO) tokens, allowing a single Transformer to perform end-to-end reaction center identification and synthon-to-reactant

Table 6. Summary of functionalities of large language models for retrosynthesis.

	Reactants Recommendation	Multi-Step Retrosynthesis	Reaction Yield Prediction	Reaction Conditions Prediction
SynAsk[134]	√	√	√	√
BatGPT-Chem[135]	√	√	×	√
RSGPT[41]	√	√	×	×
Chemma[136]	√	√	√	×
RetroInText[137]	√	√	×	×
Llamole[138]	√	√	×	×
ChemCrow[129]	√	√	×	√
LLM-Syn-Planner[141]	×	√	×	×
ReactSeq[139]	√	√	×	×

completion while providing atom mappings and interpretable edit steps.[139] This achieves state-of-the-art top-k accuracy on USPTO-50k with improved interpretability. Recent studies have also started to embed 3D conformational information directly into the LLM architecture. For instance, models like Llamole and Mol3D-GPT adopt multimodal encoders that jointly learn molecular text, 2D graph, and 3D coordinate embeddings, allowing the model to reason about steric effects and geometric constraints in retrosynthetic steps.[140] Similarly, structured editing frameworks such as ReactSeq demonstrate how representing bond-breaking and forming events at the atomic level can provide mechanistic interpretability akin to manual reaction mechanism diagrams.[139] These advances mark a shift from purely symbolic reaction reasoning toward spatially and energetically grounded retrosynthesis modeling, bridging the gap between digital synthesis planning and real chemical reactivity.

Other studies focus on whole-route optimization and synthetically feasible molecule design. Systems like LLM-Syn-Planner/LLM-Syn-Designer[141] serialize synthesis routes and use population/mutation-based evolutionary search, combined with RAG retrieval and SC-score reward, to optimize entire routes. On benchmarks (USPTO, Pistachio), this approach significantly increases solve rates compared with LLMs used solely for single-step prediction plus MCTS/Retro*, achieving ~91% success on USPTO-Easy.

In summary, emerging LLMs are evolving beyond text-based chemistry reasoning toward multimodal, 3D-aware, and mechanistically interpretable paradigms. The integration of spatial molecular representations not only improves the chemical accuracy of predicted transformations but also enhances the interpretability of retrosynthetic reasoning, enabling human chemists to trace the logic behind each proposed transformation. These developments are expected to play a pivotal role in bridging the gap between data-driven retrosynthesis and physical chemistry–grounded synthesis design. Despite notable progress in coverage, top-k accuracy, and experimental validation, challenges remain, including model bias and reproducibility, dataset standardization, output interpretability and safety, and reliable industrial deployment, which are critical directions for future research and engineering. Table 6 provides an overview of the strengths and specialized functions of these models in retrosynthesis tasks.

DATA CLEANING AND STANDARDIZATION

In retrosynthetic route prediction and reaction modeling tasks, data quality is the key factor determining the upper bound of model performance.[142] Unlike computer vision or NLP, where large and relatively clean datasets are available, chemical reaction data are primarily sourced from patent databases (e.g., USPTO), commercial repositories (e.g., Reaxys), and private laboratory collections. Because data extraction involves multiple complex steps, such as text parsing, SMILES conversion, role classification, yield extraction, and atom mapping, errors in any stage can introduce noise and lead to dirty data.[143] In retrosynthesis prediction, such noise not only reduces model accuracy but may also generate chemically implausible synthetic routes, thereby undermining the model's reliability in practical synthesis planning.

Existing studies have revealed three major categories of noise commonly found in reaction databases:

(1) Parsing and structural errors, such as invalid SMILES or missing bond information caused by OCR or regex-based extraction errors. Schwaller et al.[144] reported that 5–10% of the USPTO dataset contains structural inconsistencies or unparseable SMILES.

(2) Reactant/reagent mislabeling, where reagents and solvents are confused with reactants, distorting the identification of the reaction core. The Coley group developed the RDChiral tool, which corrects such errors through template mining and rule-based heuristics.[145]

(3) Success-biased data, where patents predominantly report successful reactions, resulting in the absence of failed or low-yield cases—a phenomenon often referred to as positive publication bias.[146]

Earlier data-cleaning pipelines mainly relied on manual rules and algorithmic standardization, using RDKit for aromaticity and stereochemistry normalization,[144] or removing trivial by-products such as NaCl and H_2O. Coley et al. employed template extraction and heuristic separation of reactants from reagents to minimize noise introduced during template learning.[145] Recently, Chen and Li[147] proposed the AutoTemplate framework, which automates reaction data cleaning and template extraction, substantially reducing atom-mapping errors that often compromise template learning. AutoTemplate standardizes reaction SMILES, filters abnormal reactions, and generates high-quality templates without human intervention, yielding cleaner data representations and improving the reliability of data-driven retrosynthesis models. Schwaller et al.[144] also introduced RXNMapper, which leverages attention mechanisms to automatically infer atom mappings and eliminate traditional graph-matching errors.

More recently, research focus has shifted from merely cleaning data to designing models that can recognize and mitigate noise. One representative approach is multi-view or multimodal learning, which models chemical reactions from multiple complementary perspectives, for example, encoding SMILES strings as text inputs to a Transformer, molecular structures as graphs for a GNN, and reaction conditions or experimental descriptions as language embeddings.[148] Through consistency learning, representations from different modalities are aligned, allowing the model to capture robust and generalizable reaction patterns. Even when one modality contains errors or bias, other modalities can provide corrective signals. This multimodal fusion strategy greatly reduces the impact of noisy single-view data on prediction accuracy.

Another important line of research involves uncertainty-aware modeling, exemplified by the Molecular Transformer.[28] Unlike conventional deterministic models, these approaches output not only a prediction but also its confidence or uncertainty distribution, enabling researchers to identify regions where the model is less certain，often corresponding to noisy or ambiguous reactions. In pharmaceutical applications, this mechanism has been used to filter out low-confidence predictions, thereby reducing experimental trial-and-error and saving both time and resources. Uncertainty-aware models effectively shift retrosynthesis prediction from "providing an answer" to "providing an answer with a reason," significantly improving trustworthiness in real-world settings.

In addition, curriculum learning and self-training strategies have been

introduced to enhance robustness and generalization in retrosynthesis modeling. Following a "from easy to hard" principle, the model is first trained on high-quality or high-confidence data before gradually incorporating more complex or noisy reactions.[149] This progressive approach allows the model to build a stable foundation of reaction knowledge and has been empirically shown to improve Top-1 accuracy. Through curriculum learning, the model evolves from passively consuming noisy data to actively identifying and leveraging it, thereby improving both learning efficiency and predictive reliability.

Beyond noise correction, a systematic success bias remains a critical data quality issue.[150] Public datasets such as USPTO and Reaxys overwhelmingly contain successful reactions, while failed or low-yield experiments are rarely reported. As a result, models lack the ability to learn "what not to do." Recent studies have demonstrated that including failed-reaction data substantially enhances a model's ability to judge reaction feasibility and optimality.[151] With the release of the Open Reaction Database (ORD, 2021), which includes negative and failed experimental outcomes, and initiatives from major pharmaceutical companies such as AstraZeneca to share internal failure data, models can now learn the differences between success and failure. This shift from learning only "optimal routes" to learning "feasible routes" marks a crucial step toward making retrosynthesis systems viable for real-world industrial applications.

EVALUATION METRICS

In this section, we present a comparative overview of the commonly used evaluation metrics in both single-step and multi-step retrosynthesis, followed by a critical analysis of their limitations in guiding practical synthesis decisions. Specifically, in single-step retrosynthesis, reactant prediction methods are primarily evaluated using the top-k accuracy metric, which measures the proportion of cases where the correct reactant set appears within the top-k ranked predictions.[107] For reaction condition prediction, performance is typically assessed through metrics such as precision, recall, and F1-score, reflecting the accuracy and completeness of the predicted reagents, catalysts, or solvents.[152] In reaction yield prediction, the model's performance is commonly evaluated using regression-based indicators such as mean absolute error (MAE), root mean squared error (RMSE), or the coefficient of determination (R^2), which quantify the deviation between predicted and experimental yields.[68]

For multi-step retrosynthesis, the evaluation focuses on route-level performance, encompassing indicators such as route success rate, path length, starting material availability, chemical plausibility, route diversity, computational efficiency, composite scoring metrics, and scalability. These metrics collectively assess not only whether a complete and executable route can be generated, but also its practicality, diversity, and generalizability in realistic synthesis scenarios.[153]

Evaluation metrics for single-step retrosynthesis

Metrics for reactants recommendation. The most commonly used metric for single-step retrosynthesis models is Top-k accuracy, which is widely adopted to evaluate whether a model can correctly predict the true set of reactants given a target molecule.[10,58] Top-k accuracy is defined as follows: if any of the top k predicted reactant sets exactly matches the true reactant set for a given target, the prediction is considered correct. The overall Top-k accuracy is then calculated as the ratio of correctly predicted samples to the total number of target molecules.[58] For instance, in the evaluation of the Retro-MTGR model, the target reactants appeared in the Top-1, Top-3, and Top-5 predictions at rates of 54.3%, 76.7%, and 90.1%, respectively.[10]

Metrics for reaction condition prediction. Reaction condition prediction is typically treated as a multi-label classification problem. Evaluation metrics include Top-k accuracy for each condition (assessing whether catalysts, solvents, additives, etc., are included in the top k predictions) and the F1 score to measure overall performance across all condition labels.[152] When predicting numerical parameters such as temperature, metrics like mean squared error (MSE) or mean absolute error (MAE) are used to quantify prediction accuracy.[152] For example, Chen et al.[152] reported that their model achieved over 50% Top-1 accuracy for catalyst/solvent/additive combinations, with a probability of about 73% that at least one correct condition appears among the top 10 recommendations; temperature predictions achieved an MAE of approximately 8.7°C.

Metrics for reaction yield prediction. Reaction yield prediction is a regression task, commonly evaluated using MSE, root mean square error (RMSE), and MAE to measure the average deviation between predicted and actual yields. The coefficient of determination (R^2) is also used to assess the correlation between predictions and true values.[154] Additionally, scatter plots of predicted versus actual yields are frequently used to visualize model performance. Recent studies have also considered uncertainty estimation, for instance by calculating the Spearman rank correlation coefficient between prediction errors and model uncertainty estimates to quantify the relationship between confidence and accuracy.[68]

Evaluation metrics for multi-step retrosynthesis

Route success / Completion rate. Proposed by Segler et al.[16] in 2018.

Definition: For each target molecule, the algorithm is run, and the proportion of cases generating at least one complete route to purchasable starting materials is calculated.

Significance: Directly measures whether an algorithm can find feasible routes; easy to compare across methods.

Limitation: Does not account for route quality or chemical plausibility, so "successful" routes may still be low-quality.

Path length / Number of steps. Proposed by Coley et al.[17] in 2017.

Definition: Counts the number of reaction steps in the generated route.

Significance: Simple to understand; shorter routes are usually easier to synthesize.

Limitation: Shorter paths are not always optimal; ignores reaction difficulty and chemical feasibility.

Starting material availability / Terminal feasibility. Proposed by Segler et al.[16] in 2018.

Definition: Checks whether the terminal nodes of a route are included in a specified purchasable starting material library.

Significance: Ensures the route is executable in practice, directly connecting to experimental work.

Limitation: Highly dependent on the coverage and quality of the stock database, limiting exploration of new chemical space.

Chemical plausibility / Feasibility. Proposed by Segler et al.[16] in 2018.

Definition: Uses rules or predictive models to evaluate the chemical soundness of each reaction step (e.g., functional group compatibility).

Significance: Ensures routes conform to chemical knowledge, reducing infeasible paths.

Limitation: Depends on the accuracy of rules or models; high computational cost; may not cover unknown reaction types comprehensively.

Route diversity / Coverage. Proposed by Genheden et al.[105] in 2020.

Definition: Measures the number or variability of different routes generated for the same target molecule.

Significance: Evaluates the algorithm's exploration capability and provides multiple alternative routes.

Limitation: High diversity does not guarantee high-quality routes; some may be less feasible.

Search efficiency / Computational cost. Proposed by Segler et al.[16] in 2018.

Definition: Records time, search tree size, or computational resources consumed to generate complete routes.

Significance: Measures practical usability and guides algorithm optimization.

Limitation: Does not assess route quality or chemical plausibility.

Score / Reward metrics. Proposed by Segler et al.[16] in 2018.

Definition: Combines factors such as path length, building block availability, and reaction feasibility into a single numerical score.

Significance: Enables simultaneous optimization of multiple objectives and improves search efficiency.

Limitation: Depends on empirical design; weighting choices influence results and may bias certain metrics.

Scalability / Generalizability. Proposed by Segler et al.[16] in 2018.

Definition: Tests algorithm performance on different target molecules or within larger chemical spaces.

Significance: Measures applicability range and transferability of the method.

Limitation: High experimental cost; difficult to quantify precisely; affected by dataset bias.

Limitations of current evaluation metrics

Although single-step and multi-step metrics provide necessary baselines for method comparison, they are generally biased toward "agreement with literature/annotations" or "existence of a theoretically closed path", and they fail to fully capture the multi-objective decision criteria that practicing chemists care about.[113] Specific issues include:

Top-k/high match rates tend to favor high-frequency, well-reported reactions, potentially suppressing recognition of novel but practically superior routes. This means models optimized for Top-k may reproduce common literature pathways while missing more economical, safer, or scalable alternatives.[113]

Route Success (or Route Solvability) only indicates that a closed sequence of transformations exists, but it does not distinguish economically viable pathways from those that are theoretically possible yet impractical due to reagent or intermediate cost, hazardous/toxic reagents, extreme conditions, or poor scalability. In practice, a route that is synthetically feasible on paper may be unacceptable for scale-up or in a regulated production environment.[153]

Environmental and sustainability aspects—such as atom economy, E-factor, solvent toxicity, and overall process mass intensity—are rarely integrated into standard retrosynthesis benchmarks, despite being central to "green-by-design" decision making in industry. Metrics focusing solely on step count or matching with literature miss these important trade-offs.[155]

Intellectual property (IP) and patent risk are almost never considered in current automated evaluations, yet IP constraints materially affect route selection in industrial contexts. A high-scoring route by Top-k or Route Success may nevertheless be blocked by patent encumbrances, reducing its real-world usefulness. Incorporating simple patent-risk proxies (e.g., overlap with known patent claim coverage) can materially improve practical relevance.[144]

Metric heterogeneity and dataset source inconsistencies (e.g., different purchasable-starting-material libraries, time-dependent price lists, and differing yield definitions) produce poor comparability across studies. Without standardized normalization and clear reporting of dataset versions and supply/pricing assumptions, cross-paper comparisons are of limited value.[87,93]

To bridge this gap, future frameworks should incorporate multi-objective, practice-oriented metrics that complement traditional baselines. These may include:

Cost and scalability proxies estimating reagent prices and process complexity;

Safety and sustainability indices integrating green chemistry principles; and

Novelty/IP awareness scores assessing patent overlap and route uniqueness.

By combining these computable proxies through transparent normalization and reporting, retrosynthesis evaluation can move closer to the real objectives of chemists—efficiency, safety, environmental responsibility, and innovation—rather than mere dataset agreement.

CHALLENGES AND FUTURE DIRECTIONS IN INTELLIGENT RETROSYNTHESIS

In recent years, artificial intelligence–driven retrosynthetic planning has made remarkable progress, yet several fundamental challenges remain unresolved. One major limitation is the insufficient consideration of side reactions, which often occur in parallel with the desired transformation and can significantly reduce yield or even prevent the main reaction from proceeding.[156] Current models generally assume idealized reaction conditions and neglect these competing pathways, thereby overestimating the feasibility of predicted routes. Another critical issue lies in the low quality of reaction datasets. Widely used sources such as USPTO, Reaxys, or Pistachio are known to contain noisy entries, including incorrect atom mappings, missing or inconsistent reaction conditions, and publication bias toward successful reactions. These deficiencies directly hinder the generalization ability of machine learning models and compromise their reliability when applied to real-world synthesis.

Moreover, current retrosynthetic design methods predominantly optimize for synthetic accessibility, i.e., whether a route can be traced back to commercially available starting materials. However, they often fail to integrate reaction yield and experimental conditions into the optimization framework. This omission leads to routes that may be theoretically valid but impractical in the laboratory, where yield, solvent choice, temperature, catalyst, and reagent compatibility critically determine success.[78] On the representational side, most state-of-the-art models rely on two-dimensional molecular graphs or linear string notations such as SMILES.[28] While effective for capturing connectivity, these representations overlook three-dimensional structural features, such as stereochemistry and conformational effects, which play a decisive role in regioselectivity and enantioselectivity. The lack of 3D awareness often causes prediction errors in complex synthetic scenarios.

Looking forward, advancing retrosynthetic prediction requires the integration of side reaction prediction modules, capable of recognizing and quantifying competing pathways to better approximate experimental reality. At the same time, the community must prioritize the development of high-quality, standardized, and richly annotated datasets, ensuring that atom mapping, conditions, and reaction outcomes are consistently recorded. Furthermore, retrosynthesis should evolve from single-objective optimization toward multi-objective frameworks, simultaneously balancing accessibility, yield, cost, and operational feasibility. Finally, progress in molecular representations that incorporate 3D information—potentially through graph neural networks augmented with geometric encoders or multimodal learning strategies—will be essential for capturing stereo electronic effects and improving both accuracy and interpretability. Together, these advances are expected to push retrosynthetic prediction toward greater robustness, practical utility, and seamless integration into automated chemical synthesis platforms.

REFERENCES

1. Zhong Z., Song J., Feng Z., et al. (2024). Recent advances in deep learning for retrosynthesis. *WIREs Comput. Mol. Sci.* **14**:e1694. DOI:10.1002/wcms.1694
2. Armstrong D., Joncev Z., Guo J., et al. (2025). Tango*: Constrained synthesis planning using chemically informed value functions. *Digit. Discov.* **4**:2570–2578. DOI:10.1039/d5dd00130g
3. Corey E.J. and Wipke W.T. (1969). Computer-assisted design of complex organic syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science* **166**:178–192. DOI:10.1126/science.166.3902.178
4. Andronov M., Andronova N., Wand M., et al. (2025). Fast and scalable retrosynthetic planning with a transformer neural network and speculative beam search. *arXiv* preprint. DOI:10.48550/arXiv.2508.01459
5. Jiang Y., Yu Y., Kong M., et al. (2023). Artificial intelligence for retrosynthesis prediction. *Engineering* **25**:32–50. DOI:10.1016/j.eng.2022.04.021
6. Liu G., Xue D., Xie S., et al. (2023). Retrosynthetic planning with dual value networks. *International conference on machine learning. Proc. Mach. Learn. Res.* **202**:22266–22276.
7. Hassen A.K., Torren-Peraire P., Genheden S., et al. (2022). Mind the retrosynthesis gap: Bridging the divide between single-step and multi-step retrosynthesis prediction. *arXiv* preprint. DOI:10.48550/arXiv.2212.11809
8. Zhang S., Li H., Chen L., et al. (2025). Reasoning-Driven retrosynthesis prediction with Large Language Models via reinforcement learning. *arXiv* preprint. DOI:10.48550/arXiv.2507.17448
9. Lee A.A., Yang Q., Sresht V., et al. (2019). Molecular transformerunifies reaction prediction and retrosynthesis across pharma chemical space. *Chem. Commun. (Camb)* **55**:12152–12155. DOI:10.1039/c9cc05122h
10. Zhao P.C., Wei X.X., Wang Q., et al. (2025). Single-step retrosynthesis prediction via multitask graph representation learning. *Nat. Commun.* **16**:814. DOI:10.1038/s41467-025-56062-y
11. Zhao D., Tu S. and Xu L. (2024). Efficient retrosynthetic planning with MCTS exploration enhanced A(*) search. *Commun. Chem.* **7**:52. DOI:10.1038/s42004-024-01133-2
12. Bran A.M., Neukomm T.A., Armstrong D.P., et al. (2025). Chemical reasoning in LLMs unlocks steerable synthesis planning and reaction mechanism elucidation. *arXiv* preprint. DOI:10.48550/arXiv.2503.08537
13. Chen L.Y. and Li Y.P. (2024). Machine learning-guided strategies for reaction conditions design and optimization. *Beilstein J. Org. Chem.* **20**:2476–2492. DOI:10.3762/bjoc.20.212
14. Zhong H., Liu Y., Sun H., et al. (2025). Towards global reaction feasibility and

robustness prediction with high throughput data and bayesian deep learning. *Nat. Commun.* **16**:4522. DOI:10.1038/s41467-025-59812-0
15. Silver D., Schrittwieser J., Simonyan K., et al. (2017). Mastering the game of gowithout human knowledge. *Nature* **550**:354–359. DOI:10.1038/nature24270
16. Segler M.H.S., Preuss M. and Waller M.P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**:604–610. DOI:10.1038/nature25978
17. Coley C.W., Green W.H. and Jensen K.F. (2018). Machine Learning in computer-aidedsynthesis planning. *Acc. Chem. Res.* **51**:1281–1289. DOI:10.1021/acs.accounts.8b00087
18. Liu B., Ramsundar B., Kawthekar P., et al. (2017). Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**:1103–1113. DOI:10.1021/acscentsci.7b00303
19. Yan C., Zhao P., Lu C., et al. (2022). RetroComposer: Composing templates for template-based retrosynthesis prediction. *Biomolecules.* **12**:545. DOI:10.3390/biom12091325
20. Coley C.W., Barzilay R., Jaakkola T.S., et al. (2017). Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **3**:434–443. DOI:10.1021/acscentsci.7b00064
21. Dai H., Li C., Coley C., et al. (2019). Retrosynthesis prediction with conditional graph logic network. *NeurIPS* **32**:354633.
22. Segler M.H.S. and Waller M.P. (2017). Neural-Symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry* **23**:5966–5971. DOI:10.1002/chem.201605499
23. Baylon J.L., Cilfone N.A., Gulcher J.R., et al. (2019). Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. *J. Chem. Inf. Model.* **59**:673–688. DOI:10.1021/acs.jcim.8b00801
24. Chen S. and Jung Y. (2021). Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS. Au.* **1**:1612–1620. DOI:10.1021/jacsau.1c00246
25. Seidl P., Renz P., Dyubankova N., et al. (2022). Improving few- and zero-shot reaction template prediction using modern hopfield networks. *J. Chem. Inf. Model.* **62**:2111–2120. DOI:10.1021/acs.jcim.1c01065
26. Segler M.H.S. and Waller M.P. (2017). Modelling chemical reasoning to predict and invent reactions. *Chemistry* **23**:6118–6128. DOI:10.1002/chem.201604556
27. Xuan-Vu N., Armstrong D.P., Jončev Z., et al. (2025). TempRe: Template generation for single and direct multi-step retrosynthesis. *arXiv* preprint. DOI:10.48550/arXiv.2507.21762
28. Schwaller P., Laino T., Gaudin T., et al. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **5**:1572–1583. DOI:10.1021/acscentsci.9b00576
29. Weininger D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**:31–36. DOI:10.1021/ci00057a005
30. Kalyan K.S., Rajasekharan A. and Sangeetha S. (2022). AMMU: A survey of transformer-based biomedical pretrained language models. *J. Biomed. Inform.* **126**:103982. DOI:10.1016/j.jbi.2021.103982
31. Vaswani A., Shazeer N., Parmar N., et al. (2017). Attention is all you need. *NeurIPS.* **30**:103982. DOI:10.48550/arXiv.1706.03762
32. Karpov P., Godin G. and Tetko I.V. (2019). A transformer model for retrosynthesis. *ICANN. Springer.* **11731**:817–830. DOI:10.1007/978-3-030-30493-5_78
33. Lowe D.M. (2012). Extraction of chemical structures and reactions from the literature. *Apollo–Univ. Camb. Repos.* **1**:103. DOI:10.17863/CAM.16293
34. Zheng S., Rao J., Zhang Z., et al. (2020). Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model* **60**:47–55. DOI:10.1021/acs.jcim.9b00949
35. Ucak U.V., Ashyrmamatov I., Ko J., et al. (2022). Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**:1186. DOI:10.1038/s41467-022-28857-w
36. Fang L., Li J., Zhao M., et al. (2023). Single-step retrosynthesis prediction by leveraging commonly preserved substructures. *Nat. Commun.* **14**:2446. DOI:10.1038/s41467-023-37969-w
37. Wan Y., Hsieh C.-Y., Liao B., et al. (2022). Retroformer: Pushing the limits of end-to-end retrosynthesis transformer. *Int. Conf. Mach. Learn. Proc. Mach. Learn. Res.* **162**:22475–22490.
38. Mao K., Xiao X., Xu T., et al. (2021). Molecular graph enhanced transformer for retrosynthesis prediction. *Neurocomputing.* **457**:193–202. DOI:10.1016/j.neucom.2021.06.037
39. Lin Z., Yin S., Shi L., et al. (2023). G2GT: Retrosynthesis prediction with graph-to-graph attention neural network and self-training. *J. Chem. Inf. Model* **63**:1894–1905. DOI:10.1021/acs.jcim.2c01302
40. Tu Z. and Coley C.W. (2022). Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *J. Chem. Inf. Model* **62**:3503–3513. DOI:10.1021/acs.jcim.2c00321
41. Deng Y., Zhao X., Sun H., et al. (2025). RSGPT: A generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints. *Nat. Commun.* **16**:7012. DOI:10.1038/s41467-025-62308-6
42. Joseph I., Bekeneva Y.A., Adakole O.S., et al. (2025). RetroSynthNet: A sequence-to-sequence modelforretrosynthesis. *Int. Conf. Soft Comput. Meas. (SCM), IEEE.* **2025**:189–193. DOI:10.1109/SCM66446.2025.11060323
43. Zhong W., Yang Z. and Chen C.Y. (2023). Retrosynthesis prediction using an end-to-end graph generative architecture for molecular graph editing. *Nat. Commun.* **14**:3009. DOI:10.1038/s41467-023-38851-5
44. Souza R., Miranda L.S.M. and Bornscheuer U.T. (2017). A retrosynthesisapproachfor biocatalysisin organicsynthesis. *Chemistry.* **23**:12040–12063. DOI:10.1002/chem.201702235
45. Wang X., Li Y., Qiu J., et al. (2021). Retroprime: A diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem. Eng. J.* **420**:129845. DOI:10.1016/j.cej.2021.129845
46. Shi C., Xu M., Guo H., et al. (2020). A graph to graphs framework for retrosynthesis prediction. *Int. Conf. Mach. Learn. Proc. Mach. Learn. Res.* **119**:8818–8827. DOI:10.48550/arXiv.2003.12725
47. Yan C., Ding Q., Zhao P., et al. (2020). Retroxpert: Decomposeretrosynthesis prediction like a chemist. *Adv. Neural Inf. Process. Syst.* **33**:11248–11258. DOI:10.48550/arXiv.2011.02893
48. Somnath V.R., Bunne C., Coley C., et al. (2021). Learning graph models for retrosynthesis prediction. *Adv. Neural Inf. Process. Syst.* **34**:9405–9415. DOI:10.48550/arXiv.2006.07038
49. Chen Z., Ayinde O.R., Fuchs J.R., et al. (2023). G(2)Retro as a two-step graph generative models for retrosynthesis prediction. *Commun. Chem.* **6**:102. DOI:10.1038/s42004-023-00897-3
50. Sacha M., Blaz M., Byrski P., et al. (2021). Molecule editgraphattentionnetwork: Modeling chemical reactions as sequences of graph edits. *J. Chem. Inf. Model.* **61**:3273–3284. DOI:10.1021/acs.jcim.1c00537
51. Ye Z., Yu L. and Ma F. (2025). Semi-template framework for retrosynthesis prediction using graph neural network. *Pattern Recognit.* **3**:111825. DOI:10.1016/j.patcog.2025.111825
52. Tadanki A.S., Rao H.S.P. and Priyakumar U.D. (2025). Dissecting errors in machine learning for retrosynthesis: A granular metric framework and a transformer-based model for more informative predictions. *Digit. Discover.* **4**:831–845. DOI:10.1039/D4DD00263F
53. Coley C.W., Rogers L., Green W.H., et al. (2017). Computer-Assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **3**:1237–1245. DOI:10.1021/acscentsci.7b00355
54. Zhong Z., Song J., Feng Z., et al. (2022). Root-alignedsmiles: A tight representation for chemical reaction prediction. *Chem. Sci.* **13**:9023–9034. DOI:10.1039/D2SC02763A
55. Lin K., Xu Y., Pei J., et al. (2020). Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**:3355–3364. DOI:10.1039/c9sc03666k
56. Kim E., Lee D., Kwon Y., et al. (2021). Valid, plausible, and diverse retrosynthesis using tied two-way transformers with latent variables. *J. Chem. Inf. Model.* **61**:123–133. DOI:10.1021/acs.jcim.0c01074
57. Sun R., Dai H., Li L., et al. (2021). Towards understanding retrosynthesis by energy-based models. *Adv. Neural Inf. Process. Syst.* **34**:10186–10194.
58. Yao L., Guo W., Wang Z., et al. (2024). Node-Aligned graph-to-graph: Elevating template-free deep learning approaches in single-step retrosynthesis. *JACS. Au.* **4**:992–1003. DOI:10.1021/jacsau.3c00737
59. Han Y., Xu X., Hsieh C.Y., et al. (2024). Retrosynthesis prediction with an iterative string editing model. *Nat. Commun.* **15**:6404. DOI:10.1038/s41467-024-50617-1
60. Wang Y., Pang C., Wang Y., et al. (2023). Retrosynthesis prediction with an interpretable deep-learning framework based on molecular assembly tasks. *Nat. Commun.* **14**:6155. DOI:10.1038/s41467-023-41698-5
61. Liu X., Ai C., Yang H., et al. (2024). RetroCapioner: Beyond attention in end-to-end retrosynthesis transformer via contrastively captioned learnable graph representation. *Bioinformatics.* **40**:btae561. DOI:10.1093/bioinformatics/btae561
62. Seo S.W., Song Y.Y., Yang J.Y., et al. (2021). GTA: Graph truncated attention for retrosynthesis. *Proc. AAAI Conf. Artif. Intell.* **35**:531–539. DOI:10.1609/aaai.v35i1.16131
63. Gao H., Struble T.J., Coley C.W., et al. (2018). Using machine learning to predict suitable conditions for organic reactions. *ACS. Cent. Sci.* **4**:1465–1476. DOI:10.1021/acscentsci.8b00357
64. LeCun Y., Bengio Y. and Hinton G. (2015). Deep learning. *Nature* **521**:436–444. DOI:10.1038/nature14539
65. Afonina V.A., Mazitov D.A., Nurmukhametova A., et al. (2021). Prediction of optimal conditions of hydrogenation reaction using the likelihood ranking approach. *Int. J. Mol. Sci.* **23**. DOI:10.3390/ijms23010248
66. Maser M.R., Cui A.Y., Ryou S., et al. (2021). Multilabel classification models for the prediction of cross-coupling reaction conditions. *J. Chem. Inf. Model.* **61**:156–166. DOI:10.1021/acs.jcim.0c01234
67. Wang Z., Lin K., Pei J., et al. (2025). Reacon: A template- and cluster-based framework for reaction condition prediction. *Chem. Sci.* **16**:854–866. DOI:10.1039/d4sc05946h
68. Kwon Y., Lee D., Choi Y.S., et al. (2022). Uncertainty-Aware prediction of chemical reaction yields with graph neural networks. *J. Cheminform.* **14**:2. DOI:10.1186/s13321-021-00579-z
69. Kingma D.P. and Welling M. (2013). Auto-encoding variational bayes. *arXiv* preprint.

DOI:10.48550/arXiv.1312.6114
70. Schwaller P., Petraglia R., Zullo V., et al. (2020). Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**:3316–3325. DOI:10.1039/c9sc05704h
71. Walker E., Kammeraad J., Goetz J., et al. (2019). Learning to predict reaction conditions: Relationships between solvent, molecular structure, and catalyst. *J. Chem. Inf. Model.* **59**:3645–3654. DOI:10.1021/acs.jcim.9b00313
72. Wang X., Hsieh C.Y., Yin X., et al. (2023). Generic Interpretable reaction condition predictions with open reaction condition datasets and unsupervised learning of reaction center. *Research (Wash. D.C.)* **6**:0231. DOI:10.34133/research.0231
73. Lowe D. (2017). Chemical reactions from US patents (1976-Sep2016). DOI:10.6084/m9.figshare.5104873.v1
74. Goodman J. (2009). Computer software review: Reaxys. *J. Chem. Inf. Model.* **49**:2897–2898. DOI:10.1021/ci900437n
75. Strieth-Kalthoff F., Sandfort F., Segler M.H., et al. (2020). Machine learning the ropes: Principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **49**:6154–6168. DOI:10.1039/C9CS00786E
76. Granda J.M., Donina L., Dragone V., et al. (2018). Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**:377–381. DOI:10.1038/s41586-018-0307-8
77. Saebi M., Nan B., Herr J.E., et al. (2023). On the use of real-world datasets for reaction yield prediction. *Chem. Sci.* **14**:4997–5005. DOI:10.1039/d2sc06041h
78. Ahneman D.T., Estrada J.G., Lin S., et al. (2018). Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**:186–190. DOI:10.1126/science.aar5169
79. Shi R., Yu G., Huo X., et al. (2024). Prediction of chemical reaction yields with large-scale multi-view pre-training. *J. Chem.* **16**:22. DOI:10.1186/s13321-024-00815-2
80. Sato A., Asahara R. and Miyao T. (2024). Chemical graph-based transformer models for yield prediction of high-throughput cross-coupling reaction datasets. *ACS Omega* **9**:40907–40919. DOI:10.1021/acsomega.4c06113
81. Voinarovska V., Kabeshov M., Dudenko D., et al. (2024). When yield prediction does not yield prediction: An overview of the current challenges. *J. Chem. Inf. Model.* **64**:42–56. DOI:10.1021/acs.jcim.3c01524
82. Bustillo L. and Rodrigues T. (2023). A focus on the use of real-world datasets for yield prediction. *Chem. Sci.* **14**:4958–4960. DOI:10.1039/d3sc90069j
83. Trinh C., Meimaroglou D. and Hoppe S. (2021). Machine learning in chemical product engineering: The state of the art and a guide for newcomers. *Processes.* **9**:1456. DOI:10.3390/pr9081456
84. Yang B.H. and Buchwald S.L. (1999). Palladium-catalyzed amination of aryl halides and sulfonates. *J. Organomet. Chem.* **576**:125–146. DOI:10.1016/S0022-328X(98)01054-7
85. Miyaura N. and Suzuki A. (1995). Palladium-catalyzed cross-coupling reactions of organoboron compounds. *Chem. Rev.* **95**:2457–2483. DOI:10.1021/cr00039a007
86. Halevy A., Norvig P. and Pereira F. (2009). The unreasonable effectiveness of data. *IEEE Intell. Syst.* **24**:8–12. DOI:10.1109/MIS.2009.36
87. Genheden S. and Bjerrum E. (2022). PaRoutes: Towards a framework for benchmarking retrosynthesis route predictions. *Digit. Discover.* **1**:527–539. DOI:10.1039/D2DD00015F
88. Westerlund A.M., Manohar Koki S., Kancharla S., et al. (2024). Do chemformers dream of organic matter. *Evaluating a transformer model for multistep retrosynthesis. J. Chem. Inf. Model.* **64**:3021–3033. DOI:10.1021/acs.jcim.3c01685
89. Perera D., Tucker J.W., Brahmbhatt S., et al. (2018). A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**:429–434. DOI:10.1126/science.aap9112
90. King-Smith E., Berritt S., Bernier L., et al. (2024). Probing the chemical 'reactome' with high-throughput experimentation data. *Nat. Chem.* **16**:633–643. DOI:10.1038/s41557-023-01393-w
91. Kreutter D. and Reymond J.L. (2023). Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chem. Sci.* **14**:9959–9969. DOI:10.1039/d3sc01604h
92. Mo Y., Guan Y., Verma P., et al. (2020). Evaluating and clustering retrosynthesis pathways with learned strategy. *Chem. Sci.* **12**:1469–1478. DOI:10.1039/d0sc05078d
93. Torren-Peraire P., Hassen A.K., Genheden S., et al. (2024). Models matter: The impact of single-step retrosynthesis on synthesis planning. *Digit. Discover.* **3**:558–572. DOI:10.1039/D3DD00252G
94. Lee H., Ahn S., Seo S.W., et al. (2021). RETCL: A selection-based approach for retrosynthesis via contrastive learning. *arXiv* preprint. DOI:10.48550/arXiv.2105.00795
95. Kishimoto A., Buesser B., Chen B., et al. (2019). Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. *Adv. Neural Inf. Process. Syst.* **32**:357.
96. Chen B., Li C., Dai H., et al. (2020). Retro*: Learning retrosynthetic planning with neural guided A* search. *Proc. Mach. Learn. Res.* **119**:1608–1616. DOI:10.48550/arXiv.2006.15820
97. You J., Liu B., Ying Z., et al. (2018). Graph convolutional policy network for goal-directed molecular graph generation. *Adv. Neural Inf. Process. Syst.* **31**:125. DOI:10.48550/arXiv.1806.02473
98. Roucairol M. and Cazenave T. (2024). Comparing search algorithms on the retrosynthesis problem. *Mol. Inform.* **43**:e202300259. DOI:10.1002/minf.202300259
99. Saigiridharan L., Hassen A.K., Lai H., et al. (2024). AiZynthFinder 4.0: Developments based on learnings from 3 years of industrial application. *J. Cheminform.* **16**:57. DOI:10.1186/s13321-024-00860-x
100. Thakkar A., Kogej T., Reymond J.L., et al. (2020). Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **11**:154–168. DOI:10.1039/C9SC04944D
101. Russell S. and Norvig P. (2021). Artificial intelligence: A modern approach, 4th US ed. **7**:123–133.
102. Roucairol M., Georgiou A., Cazenave T., et al. (2024). DrugSynthMC: An atom-based generation of drug-like molecules with monte carlo search. *J. Cheminform. Inf. Model.* **64**:7097–7107. DOI:10.1021/acs.jcim.4c01451
103. Xie S., Yan R., Han P., et al. (2022). Retrograph: Retrosynthetic planning with graph search. *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Min.* **22**:2120–2129. DOI:10.1145/3534678.3539446
104. Coley C.W., Rogers L., Green W.H., et al. (2018). SCScore: Synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**:252–261. DOI:10.1021/acs.jcim.7b00622
105. Genheden S., Thakkar A., Chadimova V., et al. (2020). AiZynthFinder: A fast, robust and flexible open-source software for retrosynthetic planning. *J. Chem.* **12**:70. DOI:10.1186/s13321-020-00472-1
106. Irwin R., Dimitriadis S., He J., et al. (2022). Chemformer: A pre-trained transformer for computational chemistry. *Mach. Learn. Sci. Technol.* **3**:015022. DOI:10.1088/2632-2153/ac3ffb
107. Dong J., Zhao M., Liu Y., et al. (2022). Deep learning in retrosynthesis planning: Datasets, models and tools. *Brief Bioinform.* **23**:bbab391. DOI:10.1093/bib/bbab391
108. Zeng K., Yang B., Zhao X., et al. (2024). Ualign: Pushing the limit of template-free retrosynthesis prediction with unsupervised SMILES alignment. *J. Cheminform.* **16**:80. DOI:10.1186/s13321-024-00877-2
109. Shee Y., Morgunov A., Li H., et al. (2025). DirectMultiStep: Direct route generation for multistep retrosynthesis. *J. Chem. Inf. Model.* **65**:3903–3914. DOI:10.1021/acs.jcim.4c01982
110. Hong S., Zhuo H.H., Jin K., et al. (2023). Retrosynthetic planning with experience-guided Monte Carlo tree search. *Commun. Chem.* **6**:120. DOI:10.1038/s42004-023-00911-8
111. Schreck J.S., Coley C.W. and Bishop K.J.M. (2019). Learning retrosynthetic planning through simulated experience. *Cent. Sci.* **5**:970–981. DOI:10.1021/acscentsci.9b00055
112. Grzybowski B.A., Szymkuć S., Gajewska E.P., et al. (2018). Chematica: A story of computer code that started to think like a chemist. *Chem.* **4**:390–398. DOI:10.1016/j.chempr.2018.02.024
113. Hastedt F., Bailey R.M., Hellgardt K., et al. (2024). Investigating the reliability and interpretability of machine learning frameworks for chemical retrosynthesis. *Digit. Discover.* **3**:1194–1212. DOI:10.1039/D4DD00007B
114. Tu Z., Choure S.J., Fong M.H., et al. (2025). ASKCOS: An open source software suite for synthesis planning. *arXiv* preprint. DOI:10.48550/arXiv.2501.01835
115. Zeng T., Jin Z., Zheng S., et al. (2024). Developing bioNavi for hybrid retrosynthesis planning. *JACS. Au.* **4**:2492–2502. DOI:10.1021/jacsau.4c00228
116. Ellis K., Wong L., Nye M., et al. (2023). DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep bayesian program learning. *Philos. Trans. A Math. Phys. Eng. Sci.* **381**:20220050. DOI:10.1098/rsta.2022.0050
117. Grzybowski B.A., Badowski T., Molga K., et al. (2023). Network search algorithms and scoring functions for advanced - level computerized synthesis planning. *WIREs Comput. Mol. Sci.* **13**:e1630. DOI:10.1002/wcms.1630
118. Fromer J.C. and Coley C.W. (2024). An algorithmic framework for synthetic cost-aware decision making in molecular design. *Nat. Comput. Sci.* **4**:440–450. DOI:10.1038/s43588-024-00639-y
119. Zhang X., Lin H., Zhang M., et al. (2025). A data-driven group retrosynthesis planning model inspired by neurosymbolic programming. *Nat. Commun.* **16**:192. DOI:10.1038/s41467-024-55374-9
120. Schwaller P., Gaudin T., Lanyi D., et al. (2018). "Found in Translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **9**:6091–6098. DOI:10.1039/c8sc02339e
121. Schneider N., Lowe D.M., Sayle R.A., et al. (2015). Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **55**:39–53. DOI:10.1021/ci5006614
122. Jin W., Coley C., Barzilay R., et al. (2017). Predicting organic reaction outcomes with weisfeiler-lehman network. *Adv. Neural Inf. Process. Syst.* **30**:134.
123. Kearnes S.M., Maser M.R., Wleklinski M., et al. (2021). The open reaction database. *J. Am. Chem. Soc.* **143**:18820–18826. DOI:10.1021/jacs.1c09820
124. Kanehisa M., Sato Y., Furumichi M., et al. (2019). New approach for understanding genome variations in KEGG. *Nucleic. Acids. Res.* **47**:D590–D595. DOI:10.1093/nar/gky962
125. Duigou T., Lac M., Carbonell P., et al. (2019). RetroRules: A database of reaction rules for engineering biology. *Nucleic. Acids. Res.* **47**:D1229–D1235. DOI:10.1093/nar/gky940
126. Neumann A., Marrison L. and Klein R. (2023). Relevance of the trillion-sized

chemical space "explore" as a source for drug discovery. *ACS. Med. Chem. Lett.* **14**:466–472. DOI:10.1021/acsmedchemlett.3c00021
127. Tingle B.I., Tang K.G., Castanon M., et al. (2023). ZINC-22 horizontal line a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.* **63**:1166–1176. DOI:10.1021/acs.jcim.2c01253
128. Zdrazil B., Felix E., Hunter F., et al. (2024). The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic. Acids. Res.* **52**:D1180–D1192. DOI:10.1093/nar/gkad1004
129. Bran M.A., Cox S., Schilter O., et al. (2024). Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**:525–535. DOI:10.1038/s42256-024-00832-8
130. Ruan Y., Lu C., Xu N., et al. (2024). An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.* **15**:10160. DOI:10.1038/s41467-024-54457-x
131. Zhang Y., Yu R., Zeng K., et al. (2024). Text-augmented multimodal llms for chemical reaction condition recommendation. *arXiv* preprint. DOI:10.48550/arXiv.2407.15141
132. Yan B., Chen A. and Cho K. (2025). Inconsistency of llms in molecular representations. *Digit. Discover.* **11**:234. DOI:10.1039/D5DD00176E
133. Li J. and Reid J.P. (2025). Connecting the complexity of stereoselective synthesis to the evolution of predictive tools. *Chem. Sci.* **16**:3832–3851. DOI:10.1039/d4sc07461k
134. Zhang C., Lin Q., Zhu B., et al. (2024). SynAsk: Unleashing the power of large language models in organic synthesis. *Chem. Sci.* **16**:43–56. DOI:10.1039/d4sc04757e
135. Yang Y., Shi R., Li Z., et al. (2025). BatGPT-Chem: A foundation large model for chemical engineering. *Research (Wash D C).* **8**:0827. DOI:10.34133/research.0827
136. Zhang Y., Han Y., Chen S., et al. (2025). Large language models to accelerate organic chemistry synthesis. *Nat. Mach. Intell.,* **7**:1010–1022. DOI:10.1038/s42256-025-01066-y
137. Kang C., Liu X. and Guo F. Retrointext: A multimodal large language model enhanced framework for retrosynthetic planning via in-context representation learning. *Int. Conf. Learn. Represent.*,**13**.
138. Liu G., Sun M., Matusik W., et al. (2024). Multimodal large language models for inverse molecular design with retrosynthetic planning. *arXiv* preprint. DOI:10.48550/arXiv.2410.04223
139. Xiong J., Zhang W., Wang Y., et al. (2025). Bridging chemistry and artificial intelligence by a reaction description language. *Nat. Mach. Intell.* **7**:782–793. DOI:10.1038/s42256-025-01032-8
140. Tang X., Tran A., Tan J., et al. (2024). MolLM: A unified language model for integrating biomedical text with 2D and 3D molecular representations. *Bioinformatics.* **40**:i357–i368. DOI:10.1093/bioinformatics/btae260
141. Wang H., Guo J., Kong L., et al. (2025). LLM-Augmented chemical synthesis and design decision programs. *arXiv* preprint. DOI:10.48550/arXiv.2505.07027
142. Toniato A., Unsleber J.P., Vaucher A.C., et al. (2023). Quantum chemical data generation as fill-in for reliability enhancement of machine-learning reaction and retrosynthesis planning. *Digit. Discov.* **2**:663–673. DOI:10.1039/d3dd00006k
143. Vangala S.R., Krishnan S.R., Bung N., et al. (2024). Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature. *J. Cheminform.* **16**:131. DOI:10.1186/s13321-024-00928-8
144. Schwaller P., Hoover B., Reymond J.L., et al. (2021). Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**:eabe4166. DOI:10.1126/sciadv.abe4166
145. Coley C.W., Green W.H. and Jensen K.F. (2019). RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inf. Model.* **59**:2529–2537. DOI:10.1021/acs.jcim.9b00286
146. Maloney M.P., Coley C.W., Genheden S., et al. (2023). Negative data in data sets for machine learning training. *Org. Lett.* **25**:2945–2947. DOI:10.1021/acs.orglett.3c01282
147. Chen L.Y. and Li Y.P. (2024). AutoTemplate: Enhancing chemical reaction datasets for machine learning applications in organic chemistry. *J. Cheminform.* **16**:74. DOI:10.1186/s13321-024-00869-2
148. Gao W. and Coley C.W. (2020). The synthesizability of molecules proposed by generative models. *J. Chem. Inf. Model.* **60**:5714–5723. DOI:10.1021/acs.jcim.0c00174
149. Sheshanarayana R. and You F. (2025). Rethinking retrosynthesis: Curriculum learning reshapes transformer-based small-molecule reaction prediction. *J. Chem. Inf. Model.* **65**:11047–11063. DOI:10.1021/acs.jcim.5c01508
150. Strieth-Kalthoff F., Sandfort F., Kuhnemund M., et al. (2022). Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem. Int. Ed. Engl.* **61**:e202204647. DOI:10.1002/anie.202204647
151. Toniato A., Vaucher A.C., Laino T., et al. (2025). Negative chemical data boosts language models in reaction outcome prediction. *Sci. Adv.* **11**:eadt5578. DOI:10.1126/sciadv.adt5578
152. Chen L.Y. and Li Y.P. (2024). Enhancing chemical synthesis: A two-stage deep neural network for predicting feasible reaction conditions. *J. Cheminform.* **16**:11. DOI:10.1186/s13321-024-00805-4
153. Choe J., Kim H., Chok Y.T., et al. (2025). Retrosynthetic crosstalk between single-step reaction and multi-step planning. *J. Cheminform.* **17**:130. DOI:10.1186/s13321-025-01088-z
154. Schwaller P., Vaucher A.C., Laino T., et al. (2021). Prediction of chemical reaction yields using deep learning. *Mach. Learn. Sci. Technol.* **2**:015016. DOI:10.1088/2632-2153/abc81d
155. Mikolajczyk A., Zhdan U., Antoniotti S., et al. (2023). Retrosynthesis from transforms to predictive sustainable chemistry and nanotechnology: A brief tutorial review. *Green Chem.* **25**:2971–2991. DOI:10.1039/D2GC04750K
156. Kisla M.M., Hassan M.A.K., Osman H.M., et al. (2023). Incorporation of protecting groups in organic chemistry: A mini-review. *Curr. Org. Synth.* **20**:491–503. DOI:10.2174/1570179419666220820152723

FUNDING AND ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (62372375) and the Beijing Natural Science Foundation (L259021). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank previous team members for their assistance during the early stages of the project.

AUTHOR CONTRIBUTIONS

P.C.Z. and H.Z.Z. contributed equally to this work. P.C.Z. and J.Y.S. conceived and designed the research. P.C.Z., H.Z.Z., Z.Y.W., and Q.W. collected and analyzed the literature. P.C.Z. and H.Z.Z. drafted the manuscript. H.Y. and J.Y.S. revised the manuscript. All authors contributed to the manuscript and approved the final version.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DATA AND CODE AVAILABILITY

No new data were created or analyzed in this study.