26 March 2026

The Syntax of Matter: Synthesis Planning as the Foundation of Generative Chemistry

Anton Morgunov, Yu Shee, Alexander V Soudackov, Victor S Batista

Abstract

Recent advances in deep learning have improved benchmark performance for chemical property prediction, yet reliable transfer to new chemical domains remains limited. A contributing factor is that many models treat molecules primarily as static graphs, ignoring the causal logic of how they are constructed. This review surveys multistep synthesis planning (2020-2026) and argues that the field is undergoing a fundamental transition: from an Era of Navigability (2018-2023), focused on the computational feasibility of finding any route through combinatorial search space, to an Era of Validity (2024-Present), focused on the chemical correctness of those routes. We organize the literature around two dominant paradigms, search-based planning and direct sequence generation, and analyze how their design choices relate to different notions of validity. To resolve the ambiguity of current "solvability" metric, which frequently exceeds 99% by measuring only topological connectivity, we introduce a formalized Hierarchy of Chemical Validity (Solv-N). This framework distinguishes between syntactic (Solv-0) and topological (Solv-1) success, which are largely solved, and the higher-order constraints of selectivity (Solv-2) and executability (Solv-3), which remain open challenges. We critically examine how legacy benchmarks and inflated virtual inventories obscure this distinction, and we conclude with a roadmap for synthesis-aware foundation models evaluated under explicit Tier 2-3 constraints.

Keywords

retrosynthesis, synthesis planning, chemical AI, artificial chemical intelligence, foundation models, benchmarking, synthetic accessibility, graph search, sequence generation, reaction templates, validity hierarchy

The Syntax of Matter: Synthesis Planning as the Foundation of Generative Chemistry

Anton Morgunov,[*,†] Yu Shee,[†] Alexander V. Soudackov,[†] and Victor S. Batista[*,†,‡]

[†]*Department of Chemistry, Yale University, New Haven, CT, 06511*
[‡]*Yale Quantum Institute, Yale University, New Haven, CT 06511, USA*

E-mail: anton@ischemist.com; victor.batista@yale.edu

Abstract

Recent advances in deep learning have improved benchmark performance for chemical property prediction, yet reliable transfer to new chemical domains remains limited. A contributing factor is that many models treat molecules primarily as static graphs, ignoring the causal logic of how they are constructed. This review surveys multistep synthesis planning (2020–2026) and argues that the field is undergoing a fundamental transition: from an *Era of Navigability* (2018–2023), focused on the computational feasibility of finding *any* route through combinatorial search space, to an *Era of Validity* (2024–Present), focused on the chemical correctness of those routes. We organize the literature around two dominant paradigms, search-based planning and direct sequence generation, and analyze how their design choices relate to different notions of validity. To resolve the ambiguity of current "solvability" metric, which frequently exceeds 99% by measuring only topological connectivity, we introduce a formalized *Hierarchy of Chemical Validity (Solv-N)*. This framework distinguishes between syntactic (Solv-0) and topological (Solv-1) success, which are largely solved, and the higher-order constraints of selectivity (Solv-2) and executability (Solv-3), which remain open challenges. We critically examine how legacy benchmarks and inflated virtual inventories obscure this distinction, and we conclude with a roadmap for synthesis-aware foundation models evaluated under explicit Tier 2—3 constraints.

Contents

Acronyms

ACANet	activity cliff-awareness network.
ACES-GNN	activity cliff explanation supervised GNN.
ACI	artificial chemical intelligence.
CGR	chemistry-guided reasoning.
DFPN-E	depth-first proof-number search (enhanced).
DL	deep learning.
GCN	graph convolutional network.
GLN	graph logic network.
GNN	graph neural network.
KA-GNN	Kolmogorov-Arnold graph neural network.
KMN	kernel metric network.
LLM	large language model.
MCTS	Monte Carlo tree search.
ML	machine learning.
MLF-GNN	multi-level fusion graph neural network.
MLM	masked language modeling.
MOSAIC	multiple optimized specialists for AI-assisted chemical prediction.
MPNN	message passing neural network.
NLP	natural language processing.
QSAR	quantitative structure-activity relationship.
RAscore	retrosynthetic accessibility score.
RMSE	root mean square error.
RSFP	reaction specific fingerprints.
SAscore	synthetic accessibility score.
SCScore	synthetic complexity score.
SMARTS	SMILES arbitrary target specification.
SMILES	simplified molecular input line entry system.
Solv-N	solvability hierarchy.
SSP	successful synthesis probability.
STR	stock-termination rate.
SVM	support vector machines.
SYBA	synthetic Bayesian accessibility.

USPTO	United States Patent and Trademark Office.

Glossary

A* Search A graph traversal and path-finding algorithm that is widely used in computer science. It finds the least-cost path from a given initial node to any goal node by maintaining a priority queue of paths to visit, ordered by a heuristic cost function, $f(n) = g(n) + h(n)$.

activity cliff In medicinal chemistry, pairs of structurally similar molecules that exhibit a large and unexpected difference in biological activity. These instances represent sharp discontinuities in the structure-activity landscape and are challenging for predictive models to capture.

AND/OR Graph A directed bipartite graph used to represent the search space in retrosynthesis. Molecule nodes connect to alternative reaction nodes (OR-choice), while reaction nodes connect to the complete set of precursors required for that reaction (AND-constraint).

Artificial Chemical Intelligence (ACI) A proposed paradigm for chemical AI characterized by generalizable reasoning capabilities derived from foundational pre-training on the causal logic of molecular transformation, in contrast to specialized models trained on static structure-property correlations.

Beam Search A heuristic search algorithm that explores a graph by expanding the most promising nodes in a limited set. It is an optimization of best-first search that reduces memory requirements by keeping only a predetermined number (the "beam width") of best candidates at each step.

Direct Sequence Generation An architectural paradigm for synthesis planning where a full multistep synthetic route is generated directly as a single, serialized sequence of tokens, typically using a transformer-based model. This approach contrasts with search-based methods that build routes step-by-step.

Era of Navigability A term proposed in this review to characterize the period of synthesis planning research (c. 2018–2023) primarily focused on the computational challenge of finding any topologically valid path from a target molecule to a set of starting materials through a vast combinatorial search space.

Era of Validity A term proposed in this review to characterize the current and future phase of synthesis planning research (c. 2024–Present), where the focus shifts from finding any path (navigability) to ensuring the chemical and experimental correctness of proposed routes.

Foundation Model In machine learning, a large-scale model pre-trained on a broad, self-supervised objective that develops versatile and robust internal representations. These representations enable strong performance on a wide range of downstream tasks, often with zero-shot or few-shot learning.

Graph Neural Network (GNN) A class of neural networks designed to operate directly on graph-structured data. GNNs learn node representations by iteratively aggregating information from their neighbors, making them well-suited for molecular property prediction and reaction modeling.

Hierarchy of Chemical Validity (Solv-N) A formal framework introduced in this review to disambiguate the term "solvability" by classifying proposed synthetic routes into four tiers of increasing chemical rigor: Syntactic (Solv-0), Topological (Solv-1), Selectivity (Solv-2), and Executability (Solv-3).

In-context Learning A capability of large language models where the model learns to perform a new task by conditioning on a few examples provided in the prompt, without any updates to its weights. This allows for rapid adaptation to new problems.

Large Language Model (LLM) A deep learning model with a very large number of parameters (typically billions), pre-trained on vast amounts of text data. LLMs, such as those based on the Transformer architecture, exhibit remarkable capabilities in understanding and generating natural language.

Masked Language Modeling (MLM) A self-supervised pre-training objective for language models where some tokens in the input are randomly masked, and the model is trained to predict the original identity of the masked tokens based on their context. A common variant in chemistry involves masking atoms or substructures.

Mean Reciprocal Rank of Validity (MRR-V) An evaluation metric proposed in this review that measures the mean reciprocal rank of the first valid route proposed by a planner. Validity is defined with respect to a specific tier in the Solv-N hierarchy (e.g., MRR-V_2 for selectivity-valid routes), rewarding planners that rank valid solutions highly.

Monte Carlo Tree Search (MCTS) A probabilistic and heuristic search algorithm for finding optimal decisions in a given domain by taking random samples in the decision space and building a search tree according to the results. It is widely used in search-based planners to navigate the retrosynthetic AND/OR graph.

Neurosymbolic AI An approach to artificial intelligence that combines neural networks with symbolic reasoning. In synthesis planning, this often involves using a neural model to guide or propose steps within a formal, rule-based search algorithm, blending data-driven intuition with logical rigor.

Reaction Template A generalized subgraph transformation rule, often encoded in a format like SMARTS, that defines a chemical reaction by specifying the change in bonding patterns at the reaction center. Template-based planners use these rules to propose retrosynthetic disconnections.

Search-Based Planning An architectural paradigm for synthesis planning that frames the problem as a search for a valid path on an implicit AND/OR graph. It decouples the proposal of single-step transformations (via a single-step model) from the algorithmic traversal of the search tree.

Single-Step Model A model that predicts a set of candidate precursors for a given product molecule in a single retrosynthetic step. Also known as an expansion model, it serves as the core chemical knowledge engine within a search-based planner.

Solv-0 (Syntactic Solvability) The lowest tier of the validity hierarchy, requiring only that all molecules in a proposed route are syntactically valid graphs (i.e., obey rules of valency and aromaticity).

Solv-1 (Topological Solvability) The second tier of the validity hierarchy, requiring that a route successfully connects a target molecule to a defined inventory of starting materials via a series of topologically valid reaction steps. This is equivalent to the traditional stock-termination rate (STR) metric.

Solv-2 (Selectivity Solvability) The third tier of the validity hierarchy, requiring that each transformation in a route is chemically plausible by satisfying constraints of chemoselectivity, regioselectivity, diastereoselectivity, enantioselectivity, and stoichiometry.

Solv-3 (Executability Solvability) The highest tier of the validity hierarchy, requiring that a route is experimentally viable under realistic laboratory conditions, accounting for factors such as yield, purification, safety, cost, and reagent availability.

Stock-Termination Rate (STR) An evaluation metric that measures the fraction of target molecules for which a planner successfully finds at least one complete synthetic route terminating exclusively in commercially available starting materials from a defined stock. This metric quantifies Solv-1 validity.

Synthetic Accessibility A heuristic-based measure designed to provide a rapid, quantitative estimate of the synthetic feasibility of a molecule. Such scores typically rely on statistical analyses of molecular substructures and complexity features; their limitations are discussed in Section 4.1.

Topological Planning The task of identifying a valid sequence of chemical transformations that connects a target molecule to starting materials, where molecules and reactions are represented as graphs and graph edits, respectively. This level of planning typically abstracts away quantitative details like reaction conditions and yields.

Transformer A neural network architecture based on self-attention mechanisms, which allows it to weigh the importance of different parts of the input sequence. It has become the dominant architecture for natural language processing and is widely used for sequence-to-sequence tasks in chemistry.

Value Function In reinforcement learning and search algorithms, a function that estimates the expected utility or cost of being in a particular state. In retrosynthesis, a value function may predict the difficulty or cost of synthesizing a given intermediate, guiding the search toward more promising routes.

1. Introduction

Artificial intelligence in chemistry has historically focused on predicting scalar molecular properties from molecular structure, a field established as quantitative structure-activity relationship (QSAR) by the pioneering work of Corwin Hansch.[1] While foundational, early QSAR models were often limited by assumptions of linearity, a constraint that prompted the shift toward classical machine learning (ML) techniques in the 1990s and 2000s, such as support vector machiness (SVMs)[2] and random forests,[3] which applied to expert-designed features like molecular fingerprints.[4,5] The deep learning era ushered in end-to-end learning with graph neural networks (GNNs)[6–8] and transformers,[9] which automatically extract hierarchical features from raw molecular data. While architectures such as graph convolutional networks (GCNs),[10–12] message passing neural networks (MPNNs),[13,14] and Kolmogorov-Arnold graph neural networks (KA-GNNs)[15,16] have progressively improved predictive accuracy on benchmarks,[17–19] the efficiency of small-molecule drug discovery has not seen a commensurate improvement. Comprehensive surveys indicate that despite the proliferation of complex multimodal architectures, generalization to unseen chemical domains remains limited.[20–23]

We propose that this fragility (manifesting in both the proposal of inaccessible candidates and the misprediction of activity cliffs) may partly reflect a mismatch of priorities: many models optimize for the semantics of function (what a molecule does) before learning the syntax of construction (how a molecule is made). This review argues that synthesis planning is a strong candidate for the foundational pre-training objective for generative chemistry. By analogy with large language models, which acquired generalizable reasoning capabilities in part by mastering the structural grammar of text, we hypothesize that artificial chemical intelligence (ACI) may acquire more robust physical reasoning by training on the causal logic of molecular transformation.

This review analyzes the literature on multistep synthesis planning from January 2020 to February 2026, a period of rapid progress built on decades of foundational work. While early expert systems proved that retrosynthesis could be computationally formalized,[24–29] their reliance on hand-curated rules limited their scalability. The transition to learnable models began with data-driven approaches to single-step prediction,[30–33] setting the stage for the first generation of modern, multistep planners. Foundational methods from 2018–2019, including Monte Carlo tree search (MCTS),[34] DFPN-E,[35] and the Molecular Transformer,,[36] receive detailed treatment as the architectural precursors against which subsequent progress is measured. We characterize the period from 2018 to roughly 2023 as the *Era of Navigability*, where the primary scientific objective was to demonstrate that algorithms could effectively navigate the combinatorial explosion of the retrosynthetic tree. By the primary metric of this era, stock-termination rate (STR), modern planners now routinely achieve success rates exceeding 99% against large inventories. We argue that this saturation signals the end of the navigability phase and necessitates a transition to the *Era of Validity*, where evaluation must pivot from finding *a* path to verifying the chemical correctness of the proposed route.

In this review, we center on multistep *topological planning*: identification of structures of starting materials and intermediates involved in the synthetic procedure. We distinguish this from *quantitative planning*, which involves prediction of specific reaction conditions (catalysts, solvents, temperature) and outcomes (yield). While quantitative variables are critical for experimental success, this review focuses on the topological problem, which most current generative frameworks prioritize. The frameworks discussed may require domain-specific adaptation for other areas such as catalysis, materials chemistry, or biocatalytic synthesis.[37–39]

The review is organized as follows. Section 3 establishes the conceptual foundation by examining failures in static structure-property mapping and motivating synthesis planning as a pre-training objective. Section 4 highlights the limitations of scalar accessibility proxies, and Section 5 formalizes the planning problem. Section 6 analyzes the architectural distinction between explicit graph search and direct sequence generation. Section 7 critically evaluates navigability-era benchmarks, exposing how stock set inflation and conditioned target selection obscure planning failures. Section 8.1 introduces the *solvability hierarchy* (Solv-N), a new framework that separates topological connectivity (Solv-1) from selectivity (Solv-2) and executability (Solv-3). Section 9 briefly reviews advances in quantitative planning. Finally, Section 10 outlines the path toward a chemical foundation model that integrates physical constraints into the generative process, a necessary transition visualized in Table 2.

Table 2: Paradigm comparison across domains illustrating the proposed blueprint for artificial chemical intelligence (ACI). Multistep synthesis planning is proposed as the foundational pre-training objective for chemistry, evaluated via the Solv-N hierarchy.

Domain	**Pre-Training Objective**		**Ground Truth Data**		**Gold Standard Benchmark**		**Result**
Natural Language	Next-token prediction	⟶	Internet-scale text	⟶	GLUE/MMLU	⟶	Emergent reasoning
Structural Biology	3D structure prediction	⟶	PDB & evolutionary constraints	⟶	CASP	⟶	Zero-shot folding & design
Chemistry (status quo)	SMILES masking; static graph reconstruction	⟶	Static molecular graphs (ZINC, GDB)	⟶	MoleculeNet, TDC	⟶	Struggles to generalize (activity cliffs)
Chemistry (blueprint)	**Multistep synthesis planning**	⟶	**Causal reaction trajectories (experimental, QM-filtered)**	⟶	**Solv-N hierarchy**	⟶	**Artificial Chemical Intelligence**

2. The Review at a Glance

This review examines multistep data-driven retrosynthetic planning as a central component of artificial chemical intelligence, with a primary focus on small-molecule organic synthesis as represented in contemporary reaction corpora and benchmark settings.[40] Rather than revisiting static structure-property modeling in detail, we focus on how multistep planning systems construct routes from commercially available starting materials to target molecules and how this process can inform chemistry-aware representation learning.

A recurring theme is *synthetic accessibility.* Across a range of benchmarks,[15,20–23] modern GNN- and transformer-based models have improved scalar property prediction while still struggling to generalize reliably to new chemical domains, often proposing candidates whose experimental realization is unclear. Therefore, we treat multistep synthesis planning not only as a downstream application but as a candidate organizing objective for aligning generative models with the causal logic of molecular transformations (Table 2).

Within this scope, the main objectives of this review are:

- formalization of the multistep retrosynthetic planning problem and clarification of the distinction between *topological planning* (identifying reactants, intermediates, and overall route structure) and *quantitative planning* (conditions, yields, and related variables);
- overview of the major planning architectures, including search-based systems such as Monte Carlo tree search (MCTS)[34] and neural value-guided alternatives like Retro*,[41] and the emerging direct sequence generation approaches exemplified by the Molecular Transformer and DirectMultiStep;[36,42]
- analysis of the evaluation practices with an emphasis on how stock-set design, target selection, and benchmark construction can inflate apparent performance in the *Era of Navigability*, where stock-termination rate is near saturation;

To organize existing and emerging methods, we introduce the *solvability hierarchy* (Solv-N). Solv-1 focuses on topological reachability (existence of a complete route from stock to target); Solv-2 incorporates basic chemical plausibility and selectivity; and Solv-3 addresses executability, including conditions and outcomes where data permit. We use this hierarchy to structure the transition from the navigability-centric phase of method development toward an *Era of Validity*, in which evaluation is increasingly tied to chemical realism, and to articulate how multistep synthesis planning may function as a foundational pre-training task for artificial chemical intelligence.

3. The Limitations of Static Structure-Activity Mapping

3.1. Out-of-Distribution Failure in Molecular Property Prediction

To motivate synthesis planning as a candidate pre-training objective, we first examine a recurring limitation in models that learn exclusively from static structure-property correlations: a sharp failure to generalize beyond the training distribution. This issue is particularly well-documented in bioactivity prediction, where the objective is to map molecular topology (structure) to functional outcomes such as binding affinity, inhibitory potency, or toxicity. Despite the proliferation of increasingly sophisticated deep learning architectures, rigorous benchmarking studies[22] indicate that improvements in performance on familiar chemical space do not carry over when models are tested on structurally novel molecules outside of the distribution of the training set.

For instance, several benchmarking studies evaluating graph neural networks and transformers find that, despite reporting state-of-the-art performance on global metrics like root mean square error (RMSE), these models frequently fail to outperform classical machine learning baselines under rigorous scaffold splitting.[43–45] Gaussian processes and support vector machines utilizing fixed fingerprints often match or exceed the predictive accuracy of more complex neural architectures, suggesting that increases in model complexity do not reliably translate into improved generalization to novel chemical space. The disparity is most acute at *activity cliff*: when minor structural modifications lead to disproportionate shifts in biological potency.

van Tilborg et al. demonstrated that on these critical edge cases, deep learning models often exhibit comparable or worse predictive accuracy than descriptor-based methods.[46] This failure likely has a mechanistic origin: standard optimization objectives (e.g., RMSE) encourage models to smooth the structure-activity landscape, effectively treating the sharp discontinuities characteristic of specific molecular recognition events as noise rather than signal.[47] Consequently, the subtle electronic or steric syntax that distinguishes a therapeutic from a toxic analogue is often lost in the learned representation.[48,49]

The inadequacy of purely topological learning is further evidenced by the architectural adaptations developed to mitigate these failures. Recent state-of-the-art frameworks frequently augment end-to-end learning with explicit handcrafted features or auxiliary supervision. For example, multi-level fusion graph neural network (MLF-GNN)[50] fuses learned graph representations with traditional Morgan fingerprints; activity cliff explanation supervised GNN (ACES-GNN)[51] incorporates auxiliary attention constraints derived from manually curated activity cliff pairs; and GraphCliff[49] employs gating mechanisms to preserve local features. Similarly, Shi et al. apply Group Lasso regularization to enforce explicit separation of scaffold from decoration substructures,[52] while activity cliff-awareness network (ACANet)[53] applies a special training objective (contrastive triplet loss) that simultaneously pulls together metric representations of molecules with similar activity and pushes apart those with different activity, enforcing the metric structure and shaping the learned latent space to separate active and inactive compounds. Collectively, these design choices imply that standard end-to-end training on static topology is insufficient for robust navigation of the activity landscape.

This fragility extends to structure-based drug design, where models explicitly incorporate the three-dimensional geometry of the target protein. Multimodal architectures demonstrate measurable improvements over ligand-only baselines,[54,55] but remain vulnerable to the same generalization failures. Zhang et al. documented a narrow evaluation trap defined by benchmark memorization rather than physical generalization,[56] and Wang et al. showed empirically that complex 3D convolutional networks frequently fail to enrich active binders in realistic decoy scenarios.[57] Even alternative pre-training strategies have not resolved this deficit. Standard self-supervised objectives such as motif prediction or masked atom reconstruction can induce detrimental bias on activity cliffs by encouraging models to memorize global scaffold patterns at the expense of local functional group sensitivity.[58,59] When data leakage is removed via strict structural splitting, the performance of many deep learning models drops substantially, in some cases to levels approaching simple nearest-neighbor heuristics.[60] Taken together, these persistent limitations across diverse architectures suggest that the bottleneck is not a lack of model complexity, but the fundamental insufficiency of learning chemical reasoning solely through static structure-property correlations.

3.2. Parallels in Natural Language and Computer Vision

The recent history of natural language processing (NLP) and computer vision offers a compelling parallel. Early NLP systems mirrored the current state of QSAR, relying on supervised learning for specific scalar tasks, such as sentiment classification or entailment. These models functioned as narrow experts, achieving high performance within their training distribution but failing when faced with novel phrasing or context.[61] A fundamental shift occurred when the field adopted a generative objective: predicting the next token in a sequence. By optimizing for the grammar of text rather than a specific label, models acquired internal representations capable of generalized reasoning, eventually outperforming supervised baselines on tasks they were never explicitly trained to solve.[61–63]

A similar progression occurred in computer vision, which transitioned from classifiers trained on discrete categories (e.g., ImageNet classes) to foundation models trained on the correspondence between images and text.[64,65] Cherti et al. found that the robustness of visual models scales with this form of pre-training, decoupling the learned representation from any single classification task.[66] In both domains, the move from specialized label prediction to broad structural learning resulted in representations that were more durable under distribution shift.

By analogy, current chemical AI appears to occupy a developmental stage similar to pre-generative NLP. While the term *foundation model* is frequently applied in chemistry, many architectures lack the defining characteristic of their linguistic counterparts: an *emergence* of capacity to transfer knowledge to entirely new tasks without any task-specific training, a property known as zero-shot generalization.[63,67] The implication is that the field must identify its chemical equivalent of next-token prediction: an objective that forces models to internalize the rules of molecular transformation rather than merely correlating static graphs with properties.

3.3. Synthesis Planning as a Pre-training Objective

Identifying the chemical equivalent of next-token prediction requires distinguishing between the syntax of *notation* and the syntax of *matter*. Early chemical language models treated the string representation of molecules (SMILES[68]) as the grammar to be learned. However, models trained on static masked language modeling (MLM) frequently fail to outperform simple regression baselines on downstream property tasks.[69,70] While larger models show improved prediction of simple physicochemical properties such as lipophilicity, their performance on complex bioactivity tasks tends to plateau, suggesting that learning the grammar of molecular notation is not sufficient for genuine chemical understanding.[71] Furthermore, it has been demonstrated[72] that scaling SMILES-based pre-training even to 1.1 billion molecules yields diminishing returns, consistent with models learning statistical regularities of the notation rather than internalized chemical rules.

We propose that the true syntax of chemistry is the transformation of matter through reactivity. Consequently, the chemical analogue of next-token prediction is synthesis planning: the step-by-step prediction of how a molecule is constructed. This objective bifurcates into forward synthesis (reactants to products) and retrosynthesis (products to reactants). We propose that multistep retrosynthesis is the candidate for foundational pre-training most likely to yield robust physical reasoning.

This preference is grounded in both data quality and computational tractability. Retrosynthesis can be supervised directly by the vast accumulated data of published multistep routes, providing training trajectories grounded in successful experimental execution. In contrast, multistep forward training typically relies on algorithmically generated routes, which must assume perfect separability of byproducts and lack experimental validation. Furthermore, the search space for retrosynthesis is bounded by the structural complexity of the target, whereas forward search branches with the size of the starting material inventory, rendering unconstrained exhaustive exploration computationally prohibitive.

Recent literature provides preliminary support for this reactivity-centric hypothesis. Chen et al. found that framing activity prediction as conditional structure generation improved performance on activity cliffs relative to scalar regression.[73] Similarly, architectures like REMO[74] and HiCLR[75] demonstrate that pre-training objectives requiring the reconstruction of reaction centers or reactants yield representations that capture functional group nuances better than static baselines. By requiring the model to predict missing atoms from their chemical environment, these training objectives encourage the model to internalize local reactivity rules rather than simply memorize structural patterns.

While encouraging, these results represent only the initial validation of the hypothesis. Current reactivity-aware models demonstrate improved transfer to related tasks, but they have not yet exhibited the *emergence*, zero-shot reasoning on unrelated problems, characteristic of foundation models in NLP. We postulate that scaling multistep retrosynthesis forces models to internalize electronic constraints, functional group compatibility, and selectivity boundaries, providing the necessary inductive bias to bridge this gap. Whether this approach is capable of yielding a true chemical reasoner remains a critical empirical question for the field.

3.4. Alternative foundation objectives and competing interpretations

The argument of this review is not that synthesis planning should replace all other learning objectives in chemistry, nor that it has already been established as the uniquely correct foundation-model target. A narrower and, in our view, better-supported claim is that synthesis planning may be a particularly valuable organizing objective for tasks that depend on reactivity, synthetic accessibility, and multistep planning, with multistep retrosynthesis serving as a pragmatically well-posed starting point in implementation due to data avilability and computational tractability. This objective likely coexists with other objectives in a broader chemical foundation-model stack. Different pretraining targets encode different aspects of chemical intelligence, and there is little reason to expect that a single objective will be uniformly optimal across discovery, synthesis, and deployment settings.[37,76,77]

The duality of forward and retrosynthetic planning. Within the broader objective of synthesis planning, the forward and retrosynthetic directions offer complementary perspectives. Forward reaction prediction learns the mapping from reactants and, when available, reagents or conditions, to products. This objective is naturally aligned with the causal direction of laboratory execution and is often better suited than retrosynthesis for modeling local reaction realism, including chemoselectivity, regioselectivity, stereochemical outcome and uncertainty in product formation.[78] In contrast, retrosynthesis is more naturally aligned with inverse design and route construction: it emphasizes strategic disconnection, intermediate accessibility, and long-horizon decomposition of a target into purchasable or preparable precursors. The two objectives therefore probe different abstractions of chemical knowledge. Forward models may be better at validating whether a proposed step is likely to succeed, whereas retrosynthesis may be better at organizing the search over multistep route topologies. In this sense, they are better viewed as complementary than as mutually exclusive pretraining targets.

From topological plans to executable workflows. Condition prediction shifts the learning problem from *which transformation* to *under what circumstances* that transformation is likely to succeed. This objective moves beyond the purely topological scope of many route planning benchmarks to capture procedural variables that are often decisive experimentally: catalysts, solvents, bases, stoichiometric additives, temperature, and, in broader workflow settings, purification or order-of-addition decisions.[79–81] These details are often central to practical feasibility, selectivity and reproducibility, especially when multiple topologically plausible routes exist but only a narrow subset is experimentally robust. While essential for executability, condition prediction is typically a local, single-step problem. It complements the global, multistep logic of synthesis planning, which captures broader route-level regularities that condition models do not by themselves provide, such as convergency, protecting-group strategy, and whether a target can be decomposed into sensible intermediates at all. A realistic synthesis system will likely require both: a synthesis planner to define the topological route structure, and condition or workflow models to annotate and refine those routes toward actual executable chemical process.

Multimodal structure–property pretraining. Multimodal molecular pretraining integrates chemical structure with additional modalities such as text, assay measurements, spectra, images, or omics-derived phenotypes. These objectives may capture biological, semantic, and task-level information that is only weakly coupled to synthetic route structure, and they may therefore transfer more directly than retrosynthesis to downstream tasks in drug discovery, molecular retrieval, or property prediction.[82] In particular, multimodal models can exploit textual and experimental context that is invisible to graph-only or reaction-only objectives. Retrosynthesis, by comparison, may better encode transformation logic and route feasibility priors, but it does not automatically provide the richer cross-modal grounding needed for many endpoint-focused applications. The most plausible interpretation is therefore not that retrosynthesis subsumes structure–property learning, but that the two objectives provide distinct and potentially synergistic inductive biases: retrosynthesis contributes a reactivity and planning prior, while multimodal pretraining contributes a broader property and semantics prior.

3D and physics-informed objectives. Objectives grounded in 3D geometry, conformational ensembles, quantum chemistry, docking, or other physics-informed signals address limitations of purely 2D symbolic learning. Relative to retrosynthesis, they are often better positioned to encode steric effects, noncovalent interactions, conformational preferences, and energetic constraints that strongly influence molecular behavior but may be difficult to infer from graph transformations alone.[83] This is particularly relevant in domains where geometry and energetics are first-order determinants of outcome, such as structure-based drug design, catalysis, and many materials problems. Retrosynthesis, however, captures aspects of synthetic decision-making that physics-based objectives do not naturally supply, including precedent-conditioned disconnection logic, route convergence, and the combinatorics of planning over discrete intermediates. Here again, the strongest view is hybrid rather than exclusive: 3D or physics-informed objectives may improve mechanistic fidelity, while retrosynthesis may supply the strategic scaffold on which synthetic reasoning operates.

Lab-in-the-loop and active-learning approaches. Closed-loop experimental systems optimize not only prediction accuracy on historical data, but also decision-making under real resource constraints. Relative to retrosynthesis, lab-in-the-loop and active-learning approaches better capture adaptation, uncertainty, and empirical correction under distribution shift, because the model is evaluated by the quality of the experiments it chooses and the information it gains from execution.[84] This is a different notion of competence from route planning alone. At the same time, retrosynthesis can remain valuable inside such systems as a planning prior: it defines candidate routes, intermediates, or intervention points that a closed-loop platform can then test, reject, or refine. Rather than displacing retrosynthesis, active-learning frameworks may situate it within a larger decision-theoretic architecture in which planning, validation, and experimentation continuously inform one another.

Taken together, these alternatives support a pluralistic rather than adversarial interpretation. Synthesis planning may be a high-value organizing objective for some classes of chemical representation learning, especially those involving reactivity, synthetic accessibility, and multistep compositional reasoning. But that does not imply that it replaces forward objectives, condition prediction, multimodal structure–property learning, physics-informed supervision, or experimental feedback. A more realistic outlook is that chemical foundation models will require several interacting objectives: retrosynthesis for route-level structure, forward and condition models for executable transformation detail, multimodal objectives for broad transfer across endpoints and modalities, physics-informed objectives for geometric and mechanistic fidelity, and closed-loop experimentation for empirical grounding and adaptation.

4. Evaluating Synthetic Accessibility

Generative models that lack explicit synthesis constraints frequently propose structures that are topologically valid but chemically unsynthesizable.[85] Assessing whether a candidate molecule can be made is therefore a prerequisite for practical molecular design. For the past decade, the field has relied on heuristic scores to estimate this feasibility without the computational cost of full synthesis planning. This section reviews the growing evidence that these heuristic approximations diverge from experimental reality, motivating the transition toward explicit route generation.

4.1. Intrinsic Limitations of Heuristic Accessibility Scores

The prevailing evaluation paradigm treats synthetic accessibility as an intrinsic molecular property. Real-world feasibility, by contrast, is highly context-dependent: it relies on the specific inventory of starting materials, the operational scope of available reactions, and purification constraints. Compressing this context-dependent feasibility into a single numerical score creates a metric that often correlates poorly with experimental success.

This reductionist approach originated with the synthetic accessibility score (SAscore),[86] which estimates difficulty by quantifying the statistical prevalence of substructures in public databases and applying penalties for complexity features such as non-standard ring fusions or stereocenters. Although this simple approach was a reasonable practical compromise when full retrosynthetic planning was computationally intractable in 2009, it rests on the flawed assumption that visual structural complexity is a reliable proxy for synthetic effort. A complex scaffold may be accessible via a single transformation like a Diels-Alder cycloaddition, whereas a simple structure may be elusive due to subtle stereochemical constraints.[86] Recent evaluations confirm that SAscore frequently penalizes valid complex structures, such as PROTACs, while failing to flag difficult chiral centers.[87] On realistic datasets, Li and Chen demonstrated that SAscore performance collapses to near-random guessing,[88] and Liu et al. found the mean scores for feasible and infeasible candidates to be statistically indistinguishable.[89] Topological complexity metrics, however rigorously formalized, often fail to capture synthetic difficulty when they ignore reaction-specific constraints; for instance, Flamm et al. showed that assembly theory metrics assign optimal complexity scores to pathways that delay ring closure until the final step—a strategy that is topologically efficient but synthetically implausible.[90]

Learned replacements for these heuristics have demonstrated similar limitations. Coley et al. introduced synthetic complexity score (SCScore)[91] to derive complexity directly from reaction data, aiming to capture the "synthetic gradient" from simple reactants to complex products. However, the model functions primarily as a measure of reactant popularity rather than mechanistic difficulty; Parrot et al. quantified this by showing that SCScore exhibits effectively zero correlation with the solvability determinations of explicit retrosynthetic search.[92] Addressing the out-of-distribution fragility of neural models, Voršilák et al. proposed SYBA,[93] a Bayesian classifier based on substructure frequency differences between synthesized and unsynthesized molecules. Despite this statistical grounding, blind evaluations indicate that SYBA similarly fails to reliably distinguish solvable targets from impossible ones.[94] Even retrosynthetic accessibility score (RAscore), which is explicitly supervised by the outcomes of retrosynthetic planning software,[95] retains the artifacts of structural pattern matching. Chen and Jung observe that RAscore frequently assigns high accessibility probabilities to unsynthesizable analogs solely due to their topological similarity to training examples.[96]

Beyond structural insensitivity, heuristic scores cannot account for shifting inventory constraints. Calvi et al. demonstrate that static scorers assign identical values regardless of whether key intermediates are commercially available.[97] This invariance to supply chain realities means that optimizing for general synthesizability yields a fundamentally different chemical space than optimizing for in-house inventories.[98] While heuristics correlate with solvability in drug-like space, this relationship collapses for functional materials such as organic semiconductors.[99]

The consequences of this divergence are most acute in generative optimization. When accessibility is defined by a proxy, reinforcement learning agents optimize the metric rather than physical feasibility.[100] Gao and Coley identified that unconstrained generators routinely assign high feasibility scores to impossible structures.[85] Gao et al. subsequently showed that agents optimizing SCScore generate simple long-chain structures to minimize complexity penalties, while those optimizing SAscore produce repetitive fused scaffolds.[101] Seo et al. quantified this failure: fragment-based models trained to maximize SAscore achieved a 0.00% success rate when validated by a rigorous retrosynthesis oracle.[102] Similarly, Koziarski et al. found that maximizing SAScore improved the proxy metric without improving actual synthesizability,[103] and Gao et al. observed genetic algorithms drifting almost exclusively into unsynthesizable chemical space.[104]

These limitations reflect the historical necessity of estimating feasibility when explicit planning was computationally intractable. While heuristic scores retain utility as coarse pre-filters for high-throughput screening, they are unreliable as primary evaluation metrics for generative chemistry.

4.2. The Necessity of Explicit Route Generation

To address the disconnect between heuristic estimation and experimental reality, Parrot et al. argue that the most operationally reliable metric of synthesizability is the explicit construction of a route terminating in available starting materials.[92] Empirical support for this definition is found in the performance of reaction-based generative models. By constructing molecules via explicit reaction templates rather than atom-by-atom assembly, these architectures inherently constrain the output to the logic of available chemistry. In direct comparisons, reaction-based models achieve synthesis validation rates between 56% and 100%, whereas shape-first models relying on post-hoc heuristic filtering achieve rates as low as 23%.[102,103,105]

We adopt this perspective as the foundation for the subsequent analysis, with the additional requirement that every transformation should satisfy selectivity constraints (Tier 2, Section 5.2.1). By enforcing explicit route generation, this framework also resolves the ambiguity of inventory constraints, as a molecule is deemed accessible only if the planner can connect it to the defined stock set. This results in a shift in objective: from training classifiers to *estimate* synthesizability toward developing planners that *demonstrate* it.

5. Problem Formulation and Definitions

5.1. Formalizing Retrosynthetic Logic

Retrosynthetic analysis, introduced in 1963 by Vleduts[106] and formalized by E. J. Corey in 1969,[107] is the logical deconstruction of a target molecule into progressively simpler precursors. The fundamental operation is the *disconnection*: a conceptual cleavage of a strategic bond that implies a forward chemical reaction capable of forming it. This operation transforms the target structure into a set of immediate precursors or *synthons*. The analysis is recursive; each precursor becomes a subsequent target for disconnection, generating a branching tree of potential pathways. This process terminates only when a branch reaches a *starting material*: a compound present in the chemist's available inventory. Early computational implementations, such as LHASA[24] and SECS,[25] established this logic but relied on hand-coded heuristics that could not scale to the full diversity of organic chemistry.[27]

Mathematically, retrosynthesis can be formulated as a search over a directed bipartite AND/OR graph $G = (V_M \cup V_R, E)$, in which two types of nodes alternate. Molecule nodes V_M represent OR choices: the planner selects one disconnection from several reactions that could produce the molecule in question. Reaction nodes V_R represent AND constraints: once reaction is selected, *all* of its required precursors must be obtained, with each becoming a new molecule node to be solved. A solved synthetic route is a subgraph of this graph with every terminal node belonging to the available starting material stock set $\mathcal{S}_{\text{stock}}$.

Within this framework, it is helpful to distinguish two computational problems. The *search feasibility* problem asks whether *any* valid route exists that connects the target to $\mathcal{S}_{\text{stock}}$. The *optimality* problem asks which of these feasible routes is best under a chosen cost function (e.g., step count, price, or safety). In the current literature, most benchmarks primarily measure search feasibility, often termed "solvability", because defining a universally valid cost function for chemical optimality remains an open challenge.

The central difficulty in planning is that the AND/OR graph is implicit. The graph is too large to precompute; it must be built incrementally during search. Since the number of plausible disconnections grows exponentially with depth (see Table 3), exhaustive enumeration is intractable. The planner's task, therefore, is to allocate a limited computational budget to expand only the most promising branches. This is complicated by the sparse reward signal: a precursor set may appear chemically sound at step 1 but fail to connect to stock at a later step.

Table 3: Tree Search Size of Exhaustive Retrosynthetic Search as a function of route length (depth, d) and templates per step (branching factor, b). Exhaustive enumeration scales as $O(b^d)$, rendering unguided search computationally intractable even at moderate depths .

		Templates Per Step		
		5	**50**	**100**
Route Length	**1**	5	50	100
	2	25	2500	10^4
	3	125	10^5	10^8
	5	3125	10^{8*}	10^{10}
	10	10^7	10^{16**}	10^{20}
	15	10^{10}	10^{25}	10^{30}

[*] 5 days CPU time; 2 hours on 64 cores; 10 GB memory.

[**] 1 M years CPU time; 147 years on 10,752 cores; 10^6 TB memory.

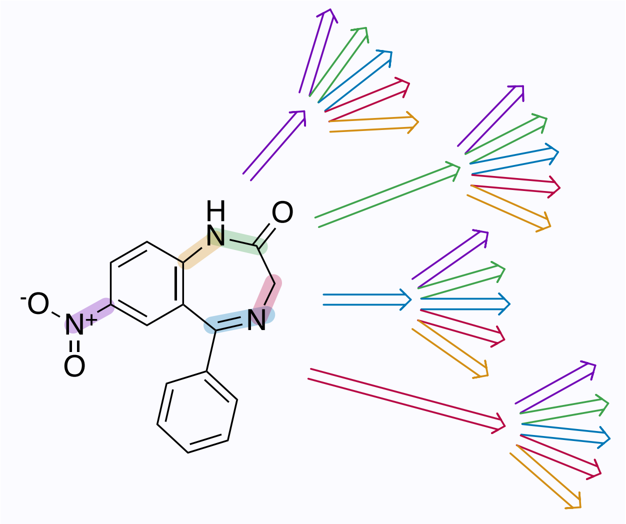


Figure 1: Schematic diagram of the retrosynthetic tree search. A target molecule (e.g., nitrazepam) presents multiple viable strategic bond disconnections. Each disconnection (marked with a different color) yields a new set of required precursors, initiating a recursive branching process.

5.2. Reaction Templates and Chemical Rules

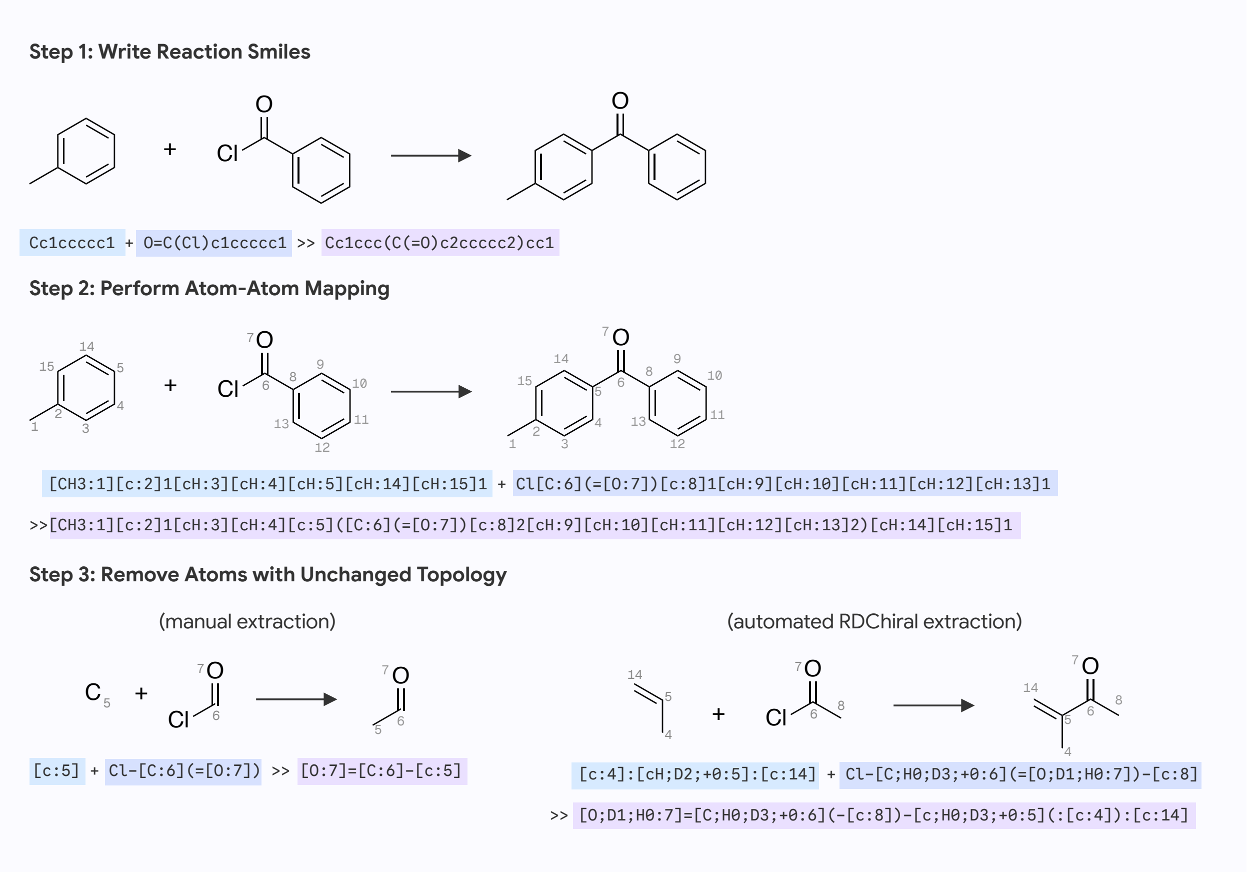
In template-based planning, the legal moves are defined by *reaction templates*. A template is a subgraph transformation rule: it specifies the atoms whose bonds change (the reaction center) and a minimal neighborhood of context, typically encoded as reaction SMARTS. For example, an amide hydrolysis template describes the transformation of `[C:1](=[O:2])[N:3]` into `[C:1](=[O:2])[OH] + [N:3]`.

Crucially, templates are local: they define the change at the reaction site but contain no information about the rest of the molecule. For instance, a template for Grignard addition may match a ketone substructure perfectly, even if an unprotected carboxylic acid elsewhere in the molecule would quench the reaction immediately. Consequently, applying a template guarantees only *syntactic* and *topological* validity (meaning the graph edit is structurally legal) but does not guarantee *selectivity*, *i.e.*, that the reaction will actually proceed as intended in the presence of competing functional groups.

Templates originate from two sources. Expert systems like Chematica[28,29] rely on hundreds of hand-coded rules[109–112] augmented with steric and electronic guards. Conversely, modern deep learning approaches automatically extract templates from reaction databases by atom-mapping reactants to products and identifying the changed core[108,113] (Fig. 2). While automated extraction scales to tens to hundreds of thousands of reactions, it often produces noisy rule sets that lack the rigorous context guards of expert systems.

5.2.1. Hierarchy of Chemical Validity

The term *validity* in retrosynthetic planning frequently obscures the distinction between graph-theoretic connectivity and experimental feasibility. To address this ambiguity, we define four levels of constraints (Table 4) that a proposed transformation must satisfy.



Step 1: Write Reaction Smiles

```
Cc1ccccc1 + O=C(Cl)c1ccccc1 >> Cc1ccc(C(=O)c2ccccc2)cc1
```

Step 2: Perform Atom-Atom Mapping

```
[CH3:1][c:2]1[cH:3][cH:4][cH:5][cH:14][cH:15]1 + Cl[C:6](=[O:7])[c:8]1[cH:9][cH:10][cH:11][cH:12][cH:13]1
>>[CH3:1][c:2]1[cH:3][cH:4][c:5]([C:6](=[O:7])[c:8]2[cH:9][cH:10][cH:11][cH:12][cH:13]2)[cH:14][cH:15]1
```

Step 3: Remove Atoms with Unchanged Topology

(manual extraction)

```
[c:5] + Cl-[C:6](=[O:7]) >> [O:7]=[C:6]-[c:5]
```

(automated RDChiral extraction)

```
[c:4]:[cH;D2;+0:5]:[c:14] + Cl-[C;H0;D3;+0:6](=[O;D1;H0:7])-[c:8]
>> [O;D1;H0:7]=[C;H0;D3;+0:6](-[c:8])-[c;H0;D3;+0:5](:[c:4]):[c:14]
```

Figure 2: **Extraction and Representation of Reaction Templates.** The sequential process of deriving local graph-transformation rules from reaction data (Section 5.2). **(Step 1)** A chemical transformation is represented as a Reaction SMILES string. **(Step 2)** Atom-to-atom mapping establishes a rigorous correspondence between reactant and product atoms to identify the reaction center (bonds formed and cleaved). **(Step 3)** Unchanged molecular topology is discarded to isolate the generalized rule. Manual extraction typically yields broad, minimal templates, whereas automated algorithms (e.g., RDChiral[108]) retain explicit local environment guards (e.g., atom degree, hydrogen counts) to constrain applicability. By definition, these templates strictly enforce Syntactic and Topological validity (Tiers 0–1, Section 5.2.1) but cannot guarantee molecule-wide Selectivity (Tier 2).

Table 4: **Hierarchy of Chemical Validity in Retrosynthetic Planning.** A proposed transformation must satisfy constraints at multiple levels to be experimentally realizable. Specific failure modes are visualized in Figures 3–5. Template-based methods guarantee syntactic and topological validity (Tiers 0–1) by construction, but provide no formal control over selectivity (Tier 2). Sequence-based methods lack formal guarantees at any tier, requiring explicit post-hoc validation.

Tier	**Definition**	**Treatment in Current Planners**
0. Syntactic	Obeys graph-theoretic rules (valency, aromaticity, charge balance). See Fig. 3 (Rxn 3).	Templates: enforced Sequence models: no guarantee
1. Topological	Correct reaction center modification per mechanistic template. See Fig. 3 (Rxn 2).	Templates with applicability check: enforced
2. Selectivity	Correct outcome among multiple chemically plausible pathways. See Figs. 4 and 5.	
– *Chemoselectivity*	Correct functional group reacts; no incompatible FG conflicts.	Human-curated rules: partial Learned policies: statistical
– *Regioselectivity*	Correct site among non-equivalent positions.	Learned policies: statistical QM: rarely integrated
– *Diastereoselectivity*	Correct relative stereochemistry.	Learned policies: statistical QM: rarely integrated
– *Enantioselectivity*	Correct absolute stereochemistry.	Largely ignored
– *Stoichiometry*	Control of single vs. multiple equivalent transformations.	Largely ignored; single-equivalent assumed
3. Executability	Lab-realistic conditions (yield, purification, safety, scale).	Condition predictors: single-step Route-level: rarely integrated

Syntactic (Tier 0) and *Topological* (Tier 1) validity refer to the construction of a well-formed molecular graph and a legal reaction center modification. Template-based methods enforce these constraints by definition, whereas template-free sequence models must learn them from the training data. As a result, unconstrained generation can yield proposals that violate valence rules or posit chemically impossible bond migrations, as illustrated in Fig. 3.

Selectivity (Tier 2) requires that the transformation be chemically plausible in the presence of competing functional groups and stereochemical requirements. While learned policies implicitly capture some of these constraints from training data, the standard template formalism does not guarantee them. A template defined by a local graph edit may be topologically applicable but fail to encode the global molecular context required to prevent unintended reactions at competing functional groups, *i.e.* ensure that the intended transformation occurs selectively at the target site rather than across chemically similar sites. (Fig. 4). Similarly, while reaction SMARTS can encode stereochemistry, automated extraction frequently yields non-specific rules. Consequently, a planner may satisfy Tier 1 validity by applying a generic template that discards the necessary stereochemical information (Fig. 5). Without explicit verification, these selectivity constraints are never directly enforced—whether they are satisfied depends entirely on the statistical quality of the policy rather than on any structural guarantee.

Validating Tier 2 constraints requires distinguishing reactive environments that standard fingerprints often fail to differentiate. Kogej et al. address this with SMARTS-RX, a curated vocabulary of functional group patterns (e.g., distinguishing heteroaryl from phenyl halides) that captures the electronic context necessary to predict reaction failure.[114] Integrating such granular definitions into the planning loop is likely a prerequisite for automated Tier 2 verification.

Executability (Tier 3) demands that the step be viable under specific laboratory conditions, including yield, purification, and safety.

5.3. Inventory Definitions and Search Boundaries

Retrosynthetic search terminates only when all required precursors lie in the chosen inventory of starting materials. The definition of that stock therefore acts as a difficulty dial: expanding the inventory reduces the depth the planner must reach and increases success rates. In practice, evaluations often treat two distinct inventory types as interchangeable. The *physical tier* comprises genuinely in-stock, rapidly deliverable compounds (typically $\sim 10^5$–10^6 entries). The *virtual tier* consists of make-on-demand listings that are purchasable in name but typically require vendor synthesis (often $\sim 10^7$–10^9 entries).

Allowing termination in the virtual tier relaxes the planning task by permitting routes to stop at complex intermediates whose remaining synthesis is simply outsourced. This shifts the operational burden from the algorithm to the vendor, often incurring lead times that are incompatible with iterative screening cycles. Consequently, high success rates against make-on-demand inventories reflect a different, less constrained objective than delivering actionable routes from physical stock.

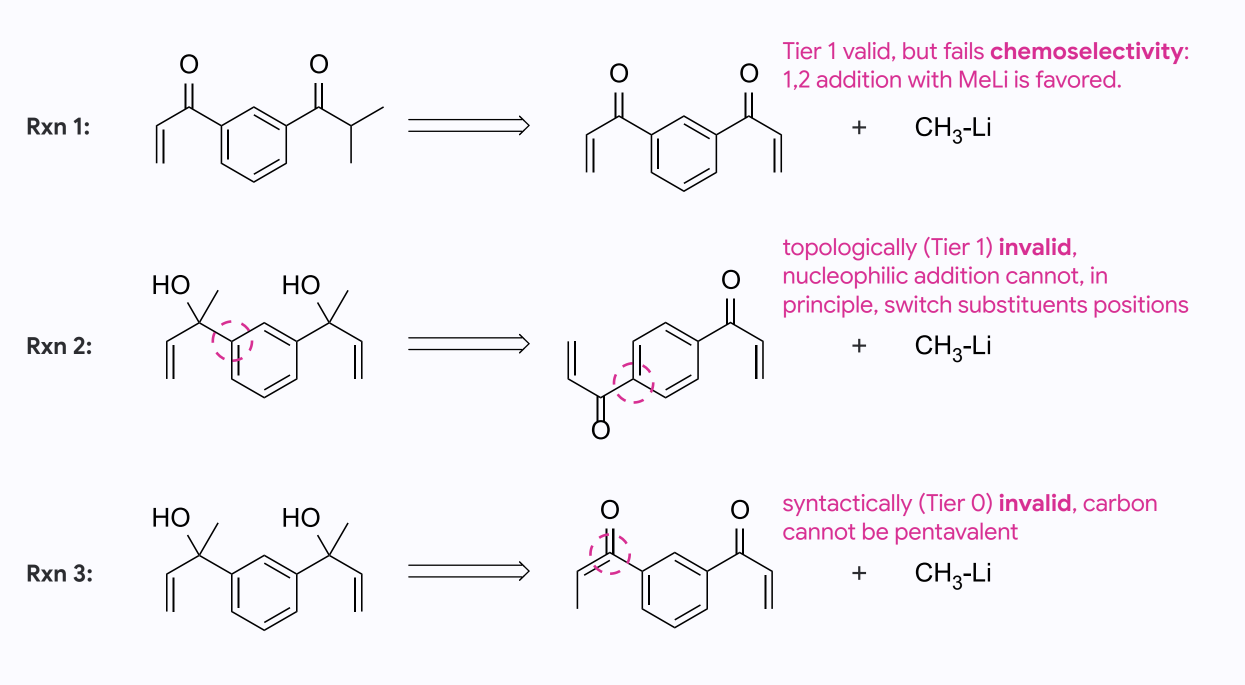

Rxn 1:
Tier 1 valid, but fails **chemoselectivity**: 1,2 addition with MeLi is favored.
+ CH_3-Li
Rxn 2:
topologically (Tier 1) **invalid**, nucleophilic addition cannot, in principle, switch substituents positions
+ CH_3-Li
Rxn 3:
syntactically (Tier 0) **invalid**, carbon cannot be pentavalent
+ CH_3-Li
HO
O

Figure 3: **Illustrative Failure Modes of Template-Free Sequence Policies Across the Solv-N Hierarchy.** Unconstrained autoregressive models (Section 6.2.2) may propose transformations that violate fundamental chemical and topological constraints (Section 5.2.1). **(Rxn 1)** A Tier 2 (Selectivity) violation: the proposed disconnection is topologically valid (Solv-1) but chemically implausible. Methyllithium strongly favors 1,2-addition over the implied conjugate addition (Solv-2C). Furthermore, even if substituted with a soft nucleophile (e.g., Me_2CuLi) to force 1,4-addition, the reaction violates regioselectivity constraints (Solv-2R), as conjugate addition occurs at the β-carbon, not the α-carbon depicted. **(Rxn 2)** A Tier 1 (Topological) violation: the model hallucinates a non-physical migration of the aryl substituents from a *para*- to a *meta*-relationship during the addition step. **(Rxn 3)** A Tier 0 (Syntactic) violation: the generation of a pentavalent carbon atom, violating basic valency rules. While template-based policies prevent Tier 0 and 1 errors by construction, sequence models require rigorous post-hoc sanitization to identify such structural anomalies.

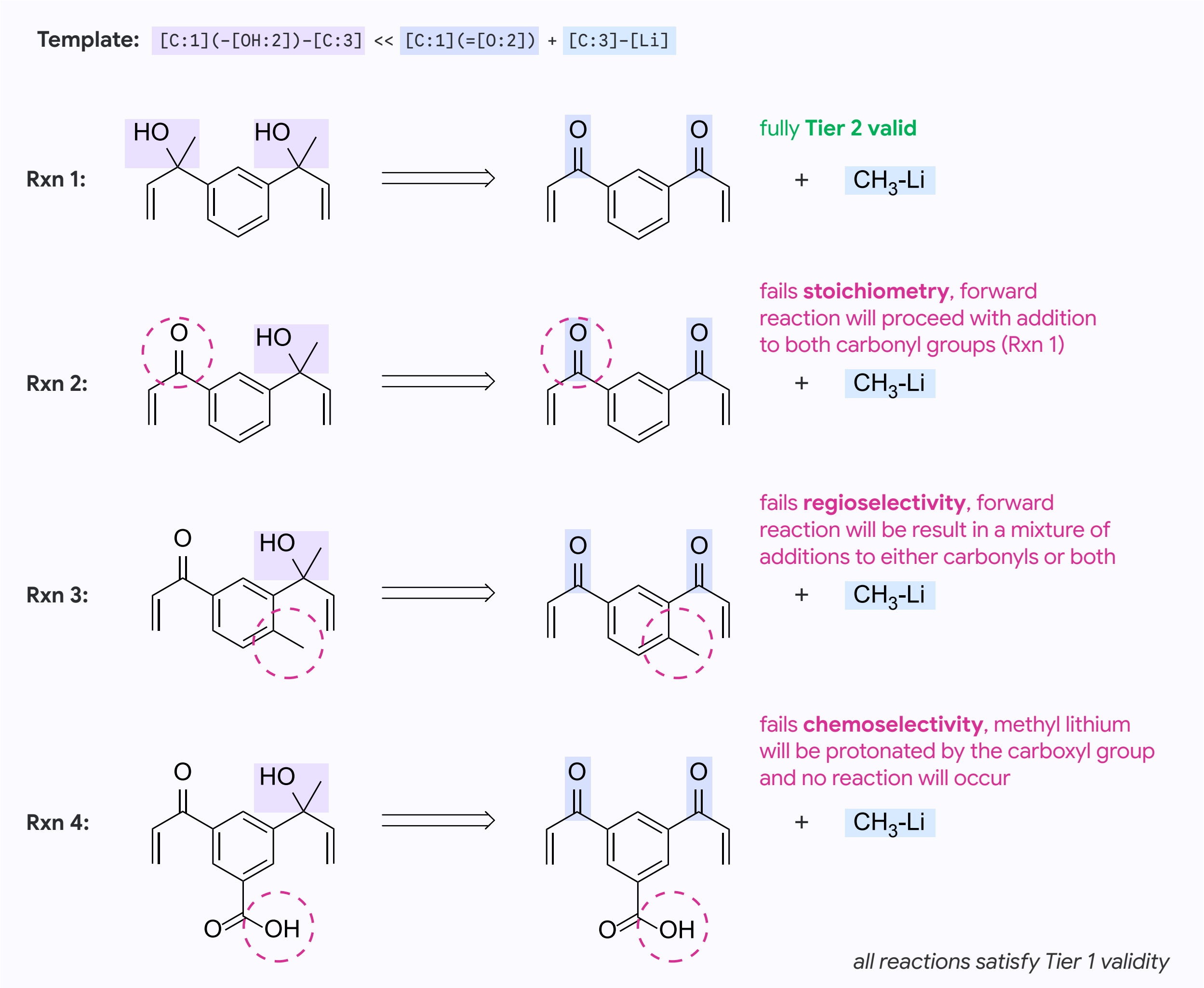
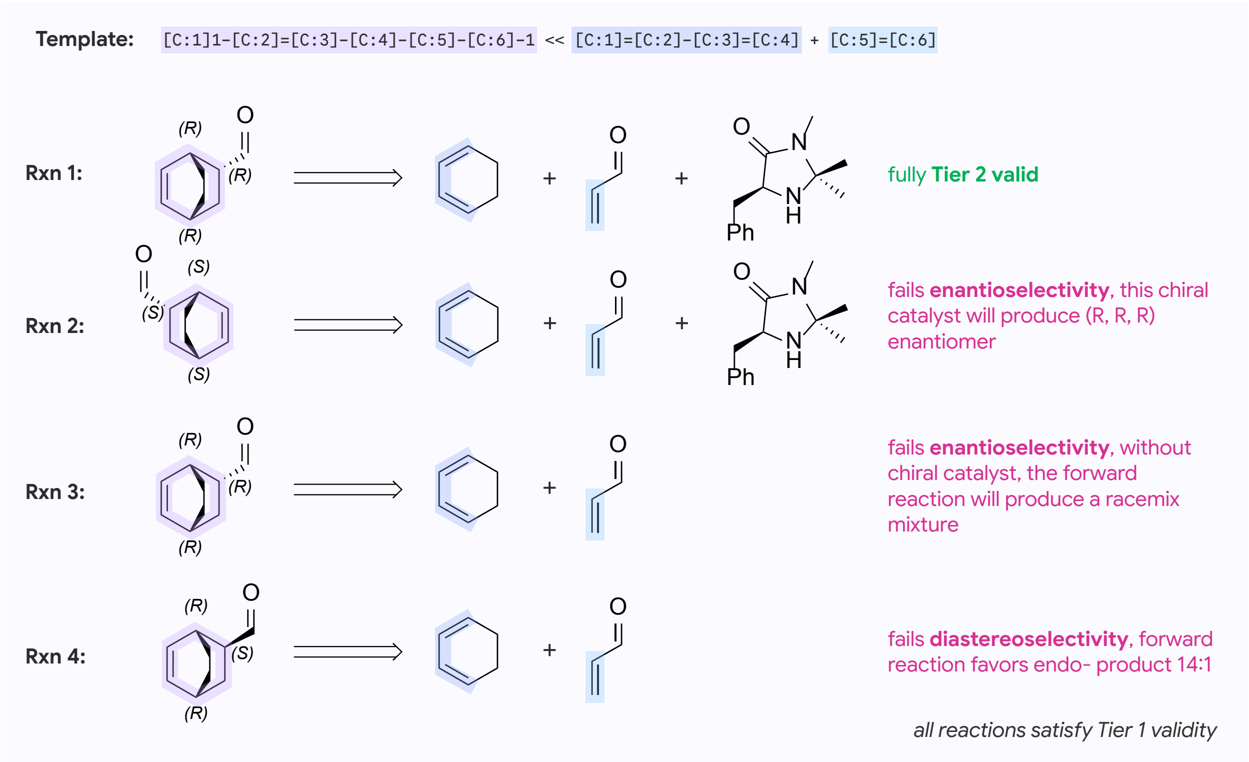

Template: [C:1](-[OH:2])-[C:3] << [C:1](=[O:2]) + [C:3]-[Li]
Rxn 1:
HO
HO
O
O
fully Tier 2 valid
+ CH_3-Li
Rxn 2:
O
HO
O
O
fails stoichiometry, forward reaction will proceed with addition to both carbonyl groups (Rxn 1)
+ CH_3-Li
Rxn 3:
O
HO
O
O
fails regioselectivity, forward reaction will be result in a mixture of additions to either carbonyls or both
+ CH_3-Li
Rxn 4:
O
HO
O
OH
O
O
O
OH
fails chemoselectivity, methyl lithium will be protonated by the carboxyl group and no reaction will occur
+ CH_3-Li
all reactions satisfy Tier 1 validity

Figure 4: **The Insufficiency of Topological Validity: Tier 2 Failures in Template Application.** While reaction templates guarantee Syntactic (Tier 0) and Topological (Tier 1) validity (Section 5.2.1) by enforcing valid local graph edits, they are inherently blind to the global molecular context governing Selectivity (Tier 2). All four proposed disconnections perfectly match the methyllithium addition template, but only one is experimentally viable. **(Rxn 1)** A fully valid (Solv-2) transformation, correctly depicting the exhaustive alkylation of two equivalent carbonyls. **(Rxn 2)** A Solv-2S (Stoichiometric) violation: proposing mono-addition to a symmetric dicarbonyl lacking a control mechanism, which would inevitably over-react to form the Rxn 1 product. **(Rxn 3)** A Solv-2R (Regioselective) violation: attempting selective addition to one of two competing electrophilic sites. The intrinsic reactivity difference is insufficient, yielding a complex mixture. **(Rxn 4)** A Solv-2C (Chemoselective) violation: the strongly basic organolithium reagent will be immediately quenched by the unprotected carboxylic acid via proton transfer, precluding the intended nucleophilic addition. These failure modes demonstrate why planners must integrate explicit selectivity verification (Section 8.1) rather than relying on local template applicability.


Template: `[C:1]1-[C:2]=[C:3]-[C:4]-[C:5]-[C:6]-1 << [C:1]=[C:2]-[C:3]=[C:4] + [C:5]=[C:6]`

fully **Tier 2 valid**

fails **enantioselectivity**, this chiral catalyst will produce (R, R, R) enantiomer

fails **enantioselectivity**, without chiral catalyst, the forward reaction will produce a racemix mixture

fails **diastereoselectivity**, forward reaction favors endo- product 14:1

all reactions satisfy Tier 1 validity

Figure 5: **Stereochemical Blind Spots in Topological Planning: Tier 2 Failures.** Diels-Alder cycloadditions highlight the inability of standard 2D reaction templates (Section 5.2) to enforce 3D spatial constraints. All four proposed disconnections perfectly match the [4+2] cycloaddition template (Solv-1), but three fail critical experimental constraints (Section 5.2.1). **(Rxn 1)** A fully valid (Solv-2) transformation: the inclusion of a specific chiral organocatalyst (e.g., MacMillan imidazolidinone[115]) correctly maps to the enantiopure *endo* product. **(Rxn 2)** A Solv-2E (Enantioselective) violation: the planner proposes the (S,S,S) enantiomer, but the specified catalyst strictly induces the (R,R,R) geometry. **(Rxn 3)** A Solv-2E violation: attempting to synthesize an enantiopure target without a source of chiral induction. The forward reaction will yield a racemic mixture. **(Rxn 4)** A Solv-2D (Diastereoselective) violation: the proposed disconnection targets the *exo* isomer, but the unconstrained forward reaction intrinsically favors the *endo* transition state via secondary orbital interactions. These examples emphasize that physical executability requires models to internalize geometric and kinetic control, not merely graph connectivity.

5.4. Evaluation Metrics

The literature relies on three primary metrics, each probing a different aspect of validity. *Solvability* measures the fraction of targets for which a planner finds *any* route terminating in the stock set. Because this metric strictly evaluates topological connectivity (Tier 1), we adopt the term stock-termination rate (STR).[116] This redefinition clarifies that the metric assesses the capacity to navigate the search graph rather than the chemical correctness of the result.

To approximate higher-tier validity, studies typically employ two proxies. *Route reconstruction* (Top-K accuracy) assesses whether the planner recovers known experimental routes. While reconstructed steps inherit the validity of the historical data (Tier 2–3), this metric is conservative; it penalizes valid, novel routes that differ from the reference. *Round-trip accuracy* evaluates self-consistency by applying a forward reaction predictor to the proposed precursors. While useful for filtering syntactic errors, this check is model-dependent. If the forward predictor shares the training distribution or architectural biases of the planner, a successful round-trip confirms consistency rather than independent chemical correctness.

5.5. Data Sources and Reaction Databases

Data-driven planners are fundamentally bounded by the quality of their training datasets. Most modern systems rely on reactions extracted from patent literature (USPTO),[40,117] which introduces distinct biases. First, because the patents overwhelmingly report successful transformations, the model is trained exclusively on reactions that worked, with no exposure to failed attempts or undesired outcomes. Models learn feasible disconnections but receive no direct supervision regarding failure modes. This absence of negative data weakens the model's ability to identify infeasibility and selectivity boundaries. Second, automated extraction frequently obscures reaction roles. Datasets often represent reactions as unordered mixtures, treating structural reactants, auxiliary reagents, catalysts, and solvents as interchangeable participants rather than distinguishing their roles in the transformation. This forces models to infer chemical roles from co-occurrence statistics rather than explicit labels, occasionally leading to incoherent proposals where solvents or bases are treated as stoichiometric building blocks. While proprietary databases (e.g., Reaxys, Pistachio) offer cleaner curation, their licensing restrictions limit their utility for reproducible benchmarking. Consequently, open-source development is still limited by the noise and ambiguity of raw patent text.

6. Algorithmic Architectures for Planning

Computational approaches to retrosynthesis have converged on two distinct paradigms: *search-based planning* (verify-then-search) and *direct route generation* (generate-then-verify). In the former, retrosynthesis is cast as a discrete optimization problem over an AND/OR graph. A single-step model proposes local disconnections, which a search algorithm then assembles into a complete route under explicit constraints (e.g., inventory availability and depth limits). In the latter, the route is serialized as a token sequence, and transformer architectures model the conditional probability of the entire pathway. This section traces the evolution of both approaches and makes explicit their central trade-off: the formal validity guarantees of explicit search (Section 5.2.1, Tiers 0–1) versus the global conditioning learned implicitly by sequence models.

6.1. Graph-Based Search Strategies

Search-based planners decouple chemical logic from algorithmic traversal. The operational pipeline typically consists of three distinct modules: (1) an *expansion model* that maps a product to candidate precursors; (2) a *feasibility filter* that prunes invalid or out-of-scope transformations; and (3) a *scoring function* that estimates the cost or probability of completing the route from a given state. By separating these components, search-based architectures allow for the modular improvement of chemical reasoning without altering the underlying search logic (Fig. 6).

6.1.1. Probabilistic Exploration (MCTS)

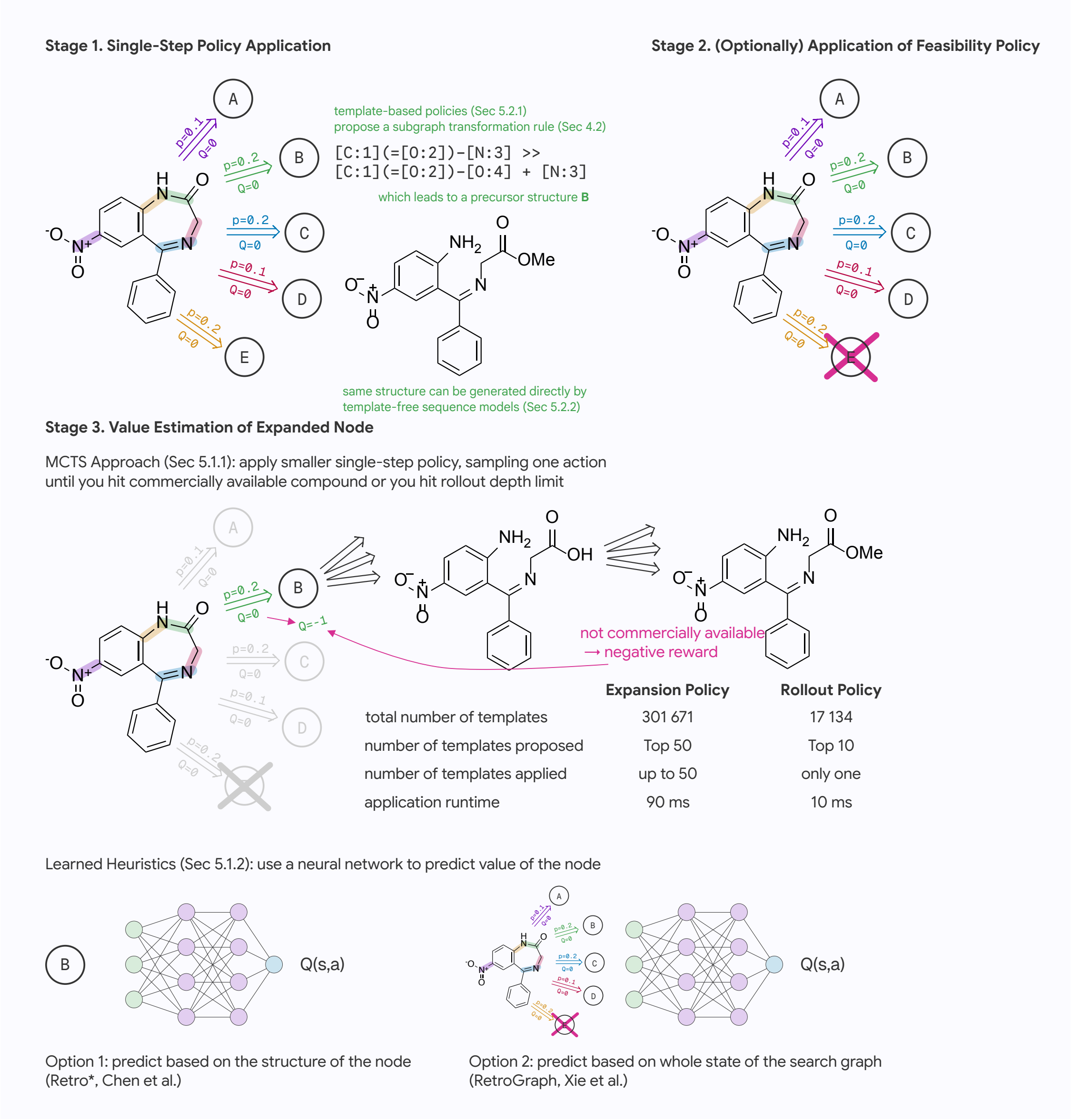
The historical analogy between retrosynthesis and combinatorial games[134] was computationally realized when Segler et al. adapted Monte Carlo tree search (MCTS) to chemical planning.[34] This adaptation, 3N-MCTS, established the modern paradigm by replacing hand-engineered heuristics with learned components within the search loop. In this framework, the expansion model ranks transformation rules, a feasibility filter screens the proposed reactions, and stochastic simulations (rollouts) estimate the value of the node by probing how likely a given molecule node (intermediate) is to lead to a completed route terminated with a purchasable starting material within a fixed number of steps.

To train the feasibility filter without experimental failure data, the authors[34] employed *algorithmic negatives*: applying templates to known reactants and labeling any unreported products as invalid (Fig. 7). While effective on dense proprietary databases like Reaxys, this closed-world assumption degrades on sparser open databases such as USPTO. Conflating undocumented products with chemically impossible ones creates significant false negatives, systematically penalizing valid but novel routes. Consequently, most (but not all[135]) subsequent open-source planners[118] have abandoned explicit feasibility filters embedding feasibility assessments implicitly into the expansion model.

A persistent limitation of MCTS is the high variance induced by the branching factor of chemical synthesis. At any state, numerous plausible disconnections exist (OR nodes), but each chosen disconnection may yield multiple precursors that must all be solved (AND constraints). Kishimoto et al. formalized this asymmetry, demonstrating that stochastic simulations can be dominated by shallow branches favored by the model's initial score, thereby failing to explore longer-horizon routes that terminate only after many steps.[35] To address this exploitation bias, Wang et al. introduced dynamic exploration schedules (mUCT) to broaden coverage of low-scoring but chemically plausible branches, demonstrating that such mechanisms can also incorporate auxiliary objectives like green solvent selection.[136] More recently, Tripp et al. reframed the planning objective: rather than identifying a single optimal route, the goal becomes selecting a portfolio of routes that together maximize the successful synthesis probability (SSP), accounting for the possibility that individual steps may fail in practice.

Table 5: **Taxonomy of Retrosynthetic Planning Architectures (2018–2026).** Models are classified by their planning paradigm: Explicit Graph Search (constructing a proof tree via discrete steps), Direct Sequence Generation (generating full routes as conditional distributions), and Hybrid/Neurosymbolic (integrating semantic reasoning into search loops). The *Tier Guarantees* column states which levels of the validity hierarchy (Section 5.2.1) are enforced by construction; all remaining tiers are delegated to learned policies or post-hoc validation. None formal indicates that even Tier 0–1 validity must be checked post-hoc via SMILES parsing and sanitization.

Model	Reference	Policy Architecture	Search Strategy	Tier Guarantees	Key Contribution
I. Explicit Graph Search (Verify-then-Search)					
3N-MCTS	Segler et al.[34]	MLP (Templates)	MCTS (Rollout)	0–1 by construction	First data-driven MCTS; the AlphaGo moment.
DFPN-E	Kishimoto et al.[35]	MLP (Templates)	Proof-Number Search	0–1 by construction	Addresses lopsided chemical trees via heuristic edge initialization.
Retro*	Chen et al.[41]	MLP (Templates)	Neural A*	0–1 by construction	Replaces MCTS rollouts with a learned value function $V(s)$.
AiZynthFinder	Genheden et al.[118]	MLP (Templates)	MCTS (UCB)	0–1 by construction	Open-source industrial standard; emphasizes software engineering.
GRASP	Yu et al.[119]	Actor-Critic (TD3)	MCTS (Goal-driven)	0–1 by construction	Pioneered starting-material-constrained planning via RL.
RetroGraph	Xie et al.[120]	GNN (Graph-Aware)	AND-OR Graph Search	0–1 by construction	Merges redundant intermediates to reduce search space.
MEEA*	Zhao et al.[121]	MLP (Templates)	Hybrid MCTS + A*	0–1 by construction	Balances exploration (MCTS) and exploitation (A*) via path consistency.
Higher-Level	Roh et al.[122]	MLP (Abstract Templates)	MCTS	0; abstract Tier 1	Decouples strategy (synthon) from tactics (FGI) to bypass horizon effects.
SynPlanner	Akhmetshin et al.[123]	GCN (Templates)	MCTS / A*	0–1 by construction	Standardized open-source implementation for benchmarking.
InterRetro	Wang and Montana[124]	Graph2Edits	Greedy (Compiled)	None formal	Search-free inference via self-imitation learning (policy compilation).
II. Direct Sequence Generation (Generate-then-Verify)					
AutoSynRoute	Lin et al.[125]	Transformer	MCTS (Heuristic)	None formal	Early integration of Transformers as policy priors.
DirectMultiStep	Shee et al.[42]	Transformer (Seq2Seq)	Beam Decoding	None formal	First to generate full retrosynthetic trees as a single sequence; explored constrained generation (target+SM).
SynLlama	Sun et al.[126]	Llama-3 (LLM)	Greedy Decoding	None formal	Fine-tuned LLM on synthetic data
RetroSynFormer	Granqvist et al.[127]	Decision Transformer	Beam Search	None formal	Models synthesis as a sequence of (State, Action, Reward) tuples.
III. Hybrid & Neurosymbolic Architectures					
Llamole	Liu et al.[128]	LLM (Heuristic)	A* Search	0–1 (template steps)	Uses LLM reasoning to estimate synthetic cost $h(n)$ for search.
DESP	Yu et al.[129]	MLP + Distance Net	Bidirectional Search	0–1 by construction	Interleaves top-down retro and bottom-up forward search.
RetroChimera	Maziarz et al.[130]	GNN + Transformer	Retro*	0–1 (template branch)	Ensembles template-based validity with transformer generalization.
LARC	Baker et al.[131]	MEEA* + LLM Critic	MCTS (Agentic)	0–1; partial Tier 2 via LLM	Uses LLM agents to prune hazardous intermediates (semantic constraints).
AOT*	Song et al.[132]	LLM (Generative)	AND-OR Tree	None formal (LLM macro-steps)	LLM proposes full macro-actions (sequences) verified by tree search.
TempRe	Xuan-Vu et al.[133]	Transformer (Template)	MCTS / Direct	0–1 (constrained mode)	Generates reaction templates autoregressively rather than SMILES.


Stage 1. Single-Step Policy Application
Stage 2. (Optionally) Application of Feasibility Policy
A
B
C
D
E
p=0.1
Q=0
p=0.2
Q=0
p=0.2
Q=0
p=0.1
Q=0
p=0.2
Q=0
template-based policies (Sec 5.2.1)
propose a subgraph transformation rule (Sec 4.2)
which leads to a precursor structure B
same structure can be generated directly by
template-free sequence models (Sec 5.2.2)
Stage 3. Value Estimation of Expanded Node
Q=-1
not commercially available
→ negative reward
Q(s,a)
Q(s,a)

```
[C:1](=[O:2])-[N:3] >>
[C:1](=[O:2])-[O:4] + [N:3]
```

MCTS Approach (Sec 5.1.1): apply smaller single-step policy, sampling one action until you hit commercially available compound or you hit rollout depth limit

	Expansion Policy	Rollout Policy
total number of templates	301 671	17 134
number of templates proposed	Top 50	Top 10
number of templates applied	up to 50	only one
application runtime	90 ms	10 ms

Learned Heuristics (Sec 5.1.2): use a neural network to predict value of the node

Option 1: predict based on the structure of the node (Retro*, Chen et al.)

Option 2: predict based on whole state of the search graph (RetroGraph, Xie et al.)

Figure 6: **Mechanics of Explicit Graph Search in Retrosynthetic Planning.** The operational pipeline of verify-then-search architectures (Section 6.1). **(Stage 1)** The expansion phase utilizes a single-step policy (Section 6.2)—either a template-based classifier (Section 6.2.1) or a template-free sequence generator (Section 6.2.2)—to propose candidate precursor sets (Nodes A–E) and assign prior expansion probabilities (p). **(Stage 2)** An optional feasibility policy explicitly prunes invalid or out-of-scope moves prior to evaluation. **(Stage 3)** Value estimation updates the expected utility (Q) of the expanded node. Probabilistic Exploration approaches (e.g., MCTS, Section 6.1.1) rely on stochastic rollouts using a lightweight, latency-optimized policy to estimate termination probability, assigning negative rewards for unpurchasable dead ends. Value-guided optimization frameworks (e.g., Retro*, Section 6.1.2) replace rollouts with learned heuristics, directly predicting the cost-to-go either from the isolated target node or by aggregating context across the entire AND/OR search graph (e.g., RetroGraph).

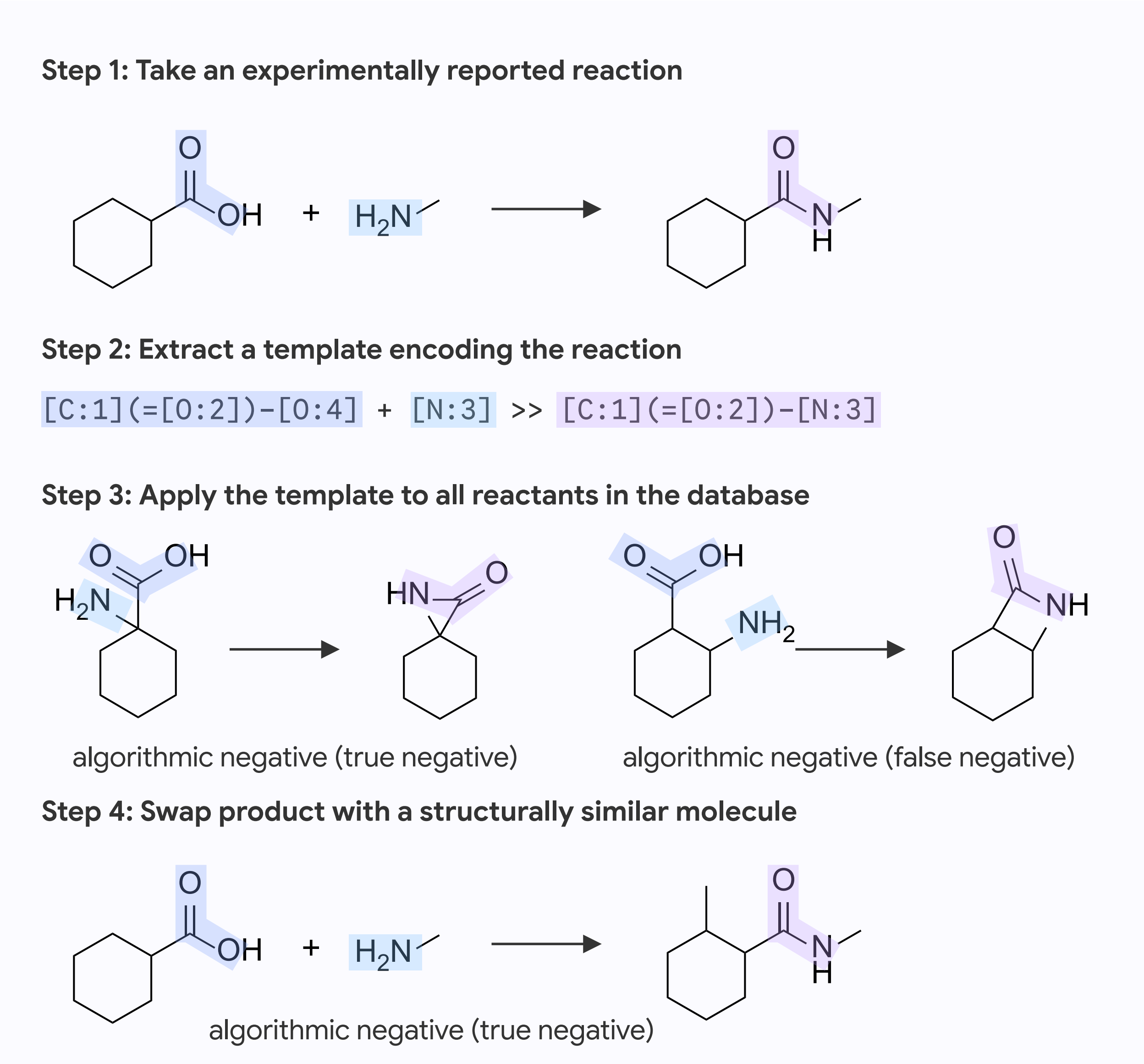

Step 1: Take an experimentally reported reaction
O
OH
+
H_2N
O
N
H
Step 2: Extract a template encoding the reaction
[C:1](=[O:2])-[O:4] + [N:3] >> [C:1](=[O:2])-[N:3]
Step 3: Apply the template to all reactants in the database
O
OH
H_2N
HN
O
algorithmic negative (true negative)
O
OH
NH_2
O
NH
algorithmic negative (false negative)
Step 4: Swap product with a structurally similar molecule
O
OH
+
H_2N
O
N
H
algorithmic negative (true negative)

Figure 7: **Generation of Algorithmic Negatives for Feasibility Policies.** The construction of synthetic negative data, a standard technique for training in-scope filters and feasibility classifiers (e.g., 3N-MCTS[34] and RetroGFN[135]). **(Step 1–2)** A literature reaction is abstracted into a reaction template (Section 5.2). **(Step 3)** Template Misapplication: the rule is applied exhaustively to other reactants in the database. Any generated product not explicitly recorded in the positive corpus is labeled an algorithmic negative. While this captures genuine chemical impossibilities (True Negatives), the closed-world assumption—that unreported equals impossible—systematically generates False Negatives. For example, the intramolecular lactamization shown is chemically viable but penalized simply for lacking precedent. **(Step 4)** Product Swapping: true reactants are paired with a structurally similar but incorrect product to train discriminators. The reliance on algorithmic negatives highlights the epistemic limits of positive-only patent databases (Section 5.5) and motivates the integration of explicit physics-based supervision.

6.1.2. Best-First Search with Learned Heuristics

To reduce the variance and computational cost of stochastic simulations, a complementary line of work replaces rollout-based evaluation with learned estimates of synthetic cost. Schreck et al. pioneered this by framing planning as a single-player game, training a value function to predict the expected remaining synthesis cost from any given intermediate.[138] At inference time, this yields a deterministic planner that selects disconnections by minimizing the immediate reaction cost plus the predicted cost of the resulting precursors. Retro* subsequently formalized this principle within a neural-guided A* framework, decomposing node evaluation into accumulated cost and a learned heuristic for future difficulty on an AND/OR tree.[41] Xie et al. extended this logic to graph-based structures (RetroGraph), merging identical intermediates to share value estimates across redundant branches.[120] Finally, recent efforts such as InterRetro[124] demonstrate that the search process itself can be distilled into a single-step policy via self-imitation learning, allowing for greedy inference that approximates the results of MCTS without the runtime cost of tree expansion.

However, controlled comparisons suggest that the choice of traversal algorithm yields only incremental gains relative to the quality of the chemical model. Within the AiZynthFinder framework,[118] Roucairol and Cazenave found that nested Monte Carlo search and greedy best-first search improved solvability only modestly over standard MCTS, concluding that performance is primarily bounded by the expansion model's ability to propose valid steps.[139]

6.1.3. Horizon Effects and Strategic Control

A central limitation of explicit search is the *horizon effect*: when node scoring is optimized for short-term objectives (e.g., maximizing single-step likelihood), the planner systematically penalizes steps whose utility is only realized later in the route. This includes strategic disconnections[140] that enable convergent assembly, as well as auxiliary operations like protection/deprotection that temporarily increase molecular complexity. MEEA* addresses these limitations by combining best-first expansion with short exploratory simulations a few steps ahead before committing to a node, while training the value function with a path-consistency regularizer to stabilize cost estimates across the search trajectory.[121]

Complementary approaches enforce long-range intent explicitly. ReTReK[141] injects curated chemical knowledge, such as preferences for ring disconnections or convergent steps, directly into the selection rule, biasing exploration toward strategies that pure likelihood models often neglect. Similarly, Westerlund et al. demonstrate that in medicinal chemistry, satisfying user intent (e.g., preserving a specific scaffold) often requires forcing the planner to break or freeze specific bonds, a constraint effectively implemented via multi-objective MCTS.[142]

Structural modifications to the search process have also been proposed to capture dependencies that span many steps ahead. DESP[129] handles the constraint of reaching specific starting materials by combining top-down retrosynthesis with bottom-up forward expansion – growing the route from both ends simultaneously and using a learned metric of synthetic proximity to guide the convergence of the two frontiers.. Beyond the single target, Picazo et al. introduced MultiAiZ[143] to exploit shared intermediates across batches of molecules, dynamically updating the available inventory to encourage convergent plans – routes that reduce overall synthetic effort by routing multiple targets through common key intermediates. A complementary approach elevates the planning problem to a higher level of abstraction entirely. Roh et al. replace atom-level reaction templates with rules that operate on generalized synthons, effectively decoupling strategic bond disconnections from the tactical implementation of specific functional group interconversions. This abstraction allows the planner to bypass the tactical complexity of protecting group sequences, which often creates local minima that trap myopic search algorithms. While this approach yields dramatic improvements in solvability on complex targets, it does so by changing the objective: instead of delivering a fully specified and executable route, it produces a high-level plan whose individual steps must still be instantiated before the synthesis can be carried out.

6.2. Single-Step Reaction Prediction

For search-based planners, the multistep algorithm serves only to assemble routes from the local disconnections proposed by its *expansion model*: a predictor that maps a target product P to a set of candidate precursors $\{R\}$, typically generating dozens or hundreds of suggestions per step and ranking them according to criteria such as predicted reaction feasibility, chemical similarity to known precedents, estimated yield, or learned heuristics from training data. The expansion models are very diverse and often based on templates (RetroSym,[32] NeuralSym,[144] GLN,[145] RetroPath2.0,[146] LocalRetro[147]), neural networks (Seq2Seq,[31] MEGAN,[148] Chemformer[70]), or hybrid approaches (RetroXpert,[149] GraphRetro,[150] BioNavi[151]). The ranking is crucial, as it determines the order in which disconnections are evaluated during tree expansion; for instance, neural network-based expansion models often output a sorted list or probabilistic scores (e.g., via softmax distributions over templates or reactants), prioritizing those deemed most synthetically viable based on patterns extracted from reaction databases. While differences in traversal algorithms (e.g., MCTS vs. A*) alter how computational resources are allocated, such as through biased sampling in MCTS or heuristic-guided queuing in A*, they cannot compensate for a deficit in chemical knowledge within the expansion model itself. With limited computational resources, a route is effectively unreachable if its critical disconnection is never ranked highly enough by the expansion model, as search algorithms typically limit branching to the top-k proposals (where k is a hyperparameter like 10 or 50), potentially overlooking rare but optimal disconnections buried in lower ranks due to model biases, incomplete training, or overemphasis on common reaction motifs.

6.2.1. Template-Based Prediction

The canonical expansion model ranks a finite library of reaction templates (Section 5.2) and applies the top-scoring rules to generate precursors. This classification-based approach underlies 3N-MCTS[34] and open-source standards like AiZynthFinder,[118] typically employing lightweight neural networks to score on the order of 10^4–10^5 possible transformations based on molecular fingerprints.

The primary advantage of template-based models is their structural discipline: they enforce local syntactic and topological constraints by construction (Section 5.2.1, Tiers 0–1). Because every proposed step results from applying a pre-validated graph edit, the output is guaranteed to be a valid molecular graph. Architectural refinements have largely focused on the ranking problem itself: GNNs (e.g., LocalRetro[147] and GLNs[145]) improve accuracy by identifying reaction centers directly on the molecular graph, rather than relying solely on global fingerprint vectors.

However, these advantages are accompanied by an intrinsic limitation in coverage: the model's chemical vocabulary is restricted to transformations represented in the predefined template library. As a result, the model's performance degrades dramatically with the distribution shift when the target molecules or required reactions in a new scenario deviate significantly from those in the training data. In such cases, the model simply lacks the templates needed to propose valid disconnections, leading to incomplete or failed route predictions, as it cannot generalize beyond its hardcoded rules. For instance, AiZynthFinder achieves a solvability rate of 70.9% on ChEMBL but drops to 10.1% on the enumerated GDB MedChem set.[152] This reduction suggests that fixed template libraries struggle to generalize to chemical space outside the historical reaction space. To address this, approaches like RetroGFN[135] and generative template models[153] replace discrete classification over fixed templates with sequential template construction, which builds reaction templates step-by-step (through autoregressive generation[135] or sampling from the latent space of an autoencoder[153]). This enhances the diversity of proposed disconnections by enabling the creation of novel transformations beyond predefined libraries but at the cost of higher inference latency since each proposed disconnection requires multiple iterative model evaluations rather than a single, parallel classification pass.

6.2.2. Template-Free Sequence Generation

Template-free models bypass the fixed library by directly generating precursor SMILES strings, typically formulating prediction as a sequence-to-sequence translation task. The Molecular Transformer established this paradigm for forward reaction prediction,[154] and subsequent work extended it to retrosynthesis by coupling transformer-based generation with explicit search procedures.[36] AutoSynRoute integrated these sequence models into MCTS, using likelihood scores to prioritize node expansion,[125] while Chemformer utilized BART-style pre-training to improve robustness on smaller datasets.[70]

The central benefit of template-free generation is coverage: the model is not bounded by the discretization artifacts of template extraction and can, in principle, propose any chemically valid string. The trade-off is the loss of formal guarantees (Section 5.2.1). Because the model emits tokens rather than applying a graph edit, it may generate invalid SMILES or chemically impossible bond changes (Tier 0 failures). These errors necessitate rigorous post-hoc validation and correction. Furthermore, RetroRanker identifies a "frequency bias" in these models, where high-confidence proposals often reflect the statistical prevalence of common reactants rather than the specific structural logic of the target.[155] This has motivated the development of ensemble architectures such as RetroChimera,[130] which combine the coverage of sequence models with the structural constraints of graph-based editors.

6.2.3. Throughput as a Planning Constraint

When an expansion model is embedded within a search loop, its inference speed becomes a structural constraint on planning depth. MCTS and A* algorithms require hundreds to thousands of expansions to explore a tree effectively. Template-based classifiers (MLPs/GNNs) can be queried in milliseconds, whereas autoregressive sequence models require computationally expensive token-by-token decoding, often taking seconds per expansion.

This latency differential creates a *speed-accuracy frontier*. Maziarz et al. report that under a fixed time budget (e.g., 10 minutes per target), transformer-based planners execute so few expansions that they fail to solve complex targets, despite having higher single-step accuracy.[156] Similarly, Hassen et al. observed that Chemformer, despite superior top-1 predictive performance, underperformed the faster LocalRetro model in multistep solvability (53.4% vs. 80.6%) simply because the search could not explore deep enough within the practical time limit.[157]

Consequently, accelerating inference is not merely an engineering detail but a prerequisite for the utility of sequence models in planning. Recent work has applied speculative decoding to increase transformer throughput, recovering some of the performance gap under strict time limits.[158] The broader implication is that for retrosynthesis, model throughput is a component of chemical capability: a marginally less accurate but orders-of-magnitude faster model may be the superior planner.

6.3. Hybrid and Neurosymbolic Approaches

The coverage and latency constraints identified in Section 6.2 motivated the development of hybrid architectures. These systems retain explicit search as the scaffold for compositional planning but inject learned modules (rankers, scoring functions, and heuristics) to steer which steps and partial routes are expanded. By integrating sequence models and large language models (LLMs) as guidance mechanisms rather than primary planners, these approaches aim to combine the rigor of tree search with the semantic reasoning of generative models.

6.3.1. Ensemble and Re-Ranking Architectures

A direct form of integration combines models with complementary inductive biases to improve candidate quality. Maziarz et al. exemplify this with RetroChimera,[130] which combines a graph-based editing component (NeuralLoc) with a sequence-based generator (R-SMILES 2) and trains a learning-to-rank model to prioritize proposals from both sources. By fusing graph-based logic, which preserves rule-constrained edits (Tier 1 validity, Sec 5.2.1), with sequence-based generation that expands coverage into rare transformations, the architecture mitigates the specific failure modes of each paradigm. RetroChimera further validates this approach through expert preference evaluations, where chemists frequently rated model proposals as superior to historical ground truth, suggesting that the ensemble improves chemical plausibility even when it diverges from specific reference routes.[130]

A modular alternative keeps the expansion model fixed but adds a learned verifier to score the plausibility of proposed bond changes. RetroRanker[155] implements this strategy by encoding the reaction center via a graph neural network, outputting a re-ranking score designed to suppress high-confidence but chemically implausible proposals attributed to frequency bias. Although the reported multistep gains are modest, likely because the re-ranking is practically constrained to the first expansion step, the work demonstrates that verification models conditioning on reaction-center structure provide a control mechanism that token-likelihood models lack. Effectively, these methods use structural constraints to filter the syntactic errors common to sequence models, thereby reinforcing Tier 0 validity.

Other hybrid approaches impose preferences at the route level rather than the reaction step. Zipoli et al. steer exploration by comparing the evolving sequence of reactions to embeddings of successful patent routes.[159] This strategy retains explicit search for compositional correctness while supplying an external retrieval signal that encourages the model to mimic the structural patterns of known chemistry, effectively mitigating the tendency of local scoring functions to miss strategic long-term dependencies.

Hybrid architectures can also address specific chemical deficits in pure search. Westerlund et al. developed a post-hoc graph modification framework that wraps the AiZynthFinder planner with a physics-guided selectivity module.[160] The system first diagnoses potential chemoselectivity conflicts, such as competing nucleophiles, using graph neural networks to predict condensed Fukui coefficients: quantum chemical reactivity descriptors that quantify how susceptible each atom in a molecule is to nucleophilic or electrophilic attack. Upon detecting a valid conflict, it automatically "repairs" the route by inserting protection and deprotection steps generated by a specialized transformer, overriding the planner's tendency to favor shorter, unprotected pathways. This mechanism explicitly bridges the gap between topological connectivity (Solv-1, Section 5.2.1) and experimental selectivity (Solv-2), proving that chemical validity often requires structural complexity that purely data-driven, length-minimizing policies will systematically avoid.

6.3.2. Large Language Models as Heuristic Guides

General-purpose LLMs provide a semantic layer often absent in purely structural planners. Rather than serving as primary expansion models, they function effectively as heuristic guides that inject non-topological constraints, such as safety, material availability, or procedural complexity, into the search loop. Liu et al. (Llamole) implement this by embedding an LLM within A* search to compute the heuristic cost-to-go, $h(n)$, directly from textual descriptions of synthetic feasibility.[128] Baker et al. (LARC) similarly deploy the LLM as a "critic" agent within the MEEA* framework,[121,131] pruning hazardous or impractical intermediates that topologically valid policies might otherwise pursue.

Beyond scalar scoring, Song et al. utilize LLMs for macro-expansion via AOT* (And-Or Tree search).[132] Here, the model proposes coherent multi-step route fragments which are subsequently verified and grafted into the search tree by template-based logic. This neurosymbolic design effectively treats the LLM as a source of strategic intuition while retaining the symbolic engine for rigorous Tier 1 graph validation. However, the stochastic nature of text generation introduces significant reproducibility challenges. Unlike standard discriminative models that provide deterministic outputs, LLM-based guidance depends heavily on decoding strategies and prompt formulation, requiring rigorous standardization of the inference context to guarantee consistent planning behavior.

6.4. Direct Sequence Generation

While hybrid systems retain an explicit search procedure assembling a route step by step through tree search, the direct sequence paradigm shifts the entire burden of multistep planning to a single generative model. In this framework, retrosynthetic routes are not assembled by the recursive expansion of a reaction tree, but are instead generated directly as a sequence of tokens produced via autoregressive decoding (the same way a language model generates a sentence). This architectural shift replaces the modular complexity of heuristic search algorithms with scalable representation learning, effectively relocating the computational effort from inference-time tree exploration to the training stage, distilling the strategic knowledge into model parameters once and then appling it during the inference.

6.4.1. Autoregressive Template and Trajectory Modeling

Intermediate approaches in this domain focus on predicting the sequence of reaction templates rather than the molecular graph itself. This strategy aims to preserve Tier 1 structural constraints while enabling the model to condition on the entire planning history.

Xuan-Vu et al. introduced TempRe to explore this middle ground, utilizing a transformer to generate reaction templates (SMARTS) autoregressively from the product rather than emitting reactant SMILES directly.[133] This representation expands the effective chemical vocabulary while ensuring that every proposed transformation remains chemically valid by construction. Since each generated token corresponds to a predefined graph edit, the model cannot produce reactions that violate basic chemical rules. Crucially, empirical evaluations demonstrate that constrained generation (filtering outputs against a template library) yields significantly higher route reconstruction fidelity than unconstrained decoding.

A complementary strategy models the planning trajectory itself. Granqvist et al.[127] train a decision transformer using the records of complete planning episodes of multistep retrosynthesis (the current intermediate molecule, the disconnection applied, and the quality of the resulting route) drawn from the PaRoutes dataset.[40] However, this approach reveals a computational bottleneck: to achieve solvability rates competitive with standard search algorithms, the model requires a large beam width (e.g., 50), resulting in inference throughput significantly lower than efficient MCTS implementations. This suggests that without the pruning logic of a search tree, sequence models must implicitly recreate the search process during decoding to ensure valid termination.

6.4.2. Full-Route Sequence Prediction

Full-route generators push this paradigm to its logical conclusion: the model emits the entire retrosynthetic tree (including intermediates, branching points, and termination leaves) as a single structured sequence. This formulation addresses a key limitation of the step-by-step greedy search: the model plans the entire route at once and thus it can recognize when two separate branches of a synthesis should converge on a shared intermediate, and account for the fact that an early disconnection may only make chemical sense in light of what comes several steps later. These strategic considerations are often overlooked by a planner working one step at a time.

Recent implementations frame retrosynthetic planning as a translation task, where the target molecule is mapped directly to a complete synthetic route represented as a structured sequence. Shee et al. developed DirectMultiStep,[42] a family of transformer-based models trained from scratch to generate multistep retrosynthetic routes as a single string, leveraging a mixture-of-experts approach to improve efficiency and accuracy. The architecture employs a classical encoder-decoder transformer setup[9] with additional gated mixture-of-expert blocks,[161] with the encoder processing the input sequence (the target SMILES string and, optionally, the desired route length or starting material) and the decoder predicting the output route. The mixture-of-experts elements significantly improve coverage across the diverse reaction types encountered in multistep planning, routing different chemical subproblems to specialized sub-networks within the model. This design allows DirectMultiStep to predict full routes, including branches and termination leaves, in a single pass, outperforming iterative search methods by capturing long-range dependencies and convergent strategies. The flagship variant, DMS Explorer XL, which requires only the target structure as input, outperformed contemporary search-based methods on the PaRoutes benchmark, achieving 1.9-fold and 3.1-fold improvements in Top-1 route reconstruction accuracy on the n_1 and n_5 evaluation sets, respectively, and demonstrated generalization to FDA-approved drugs absent from the training data.

More recently, Sun et al. introduced SynLlama,[126] a distinct specialized model derived from Meta's Llama-3 large language models Llama-3.1-8B (8 billion parameters) and Llama-3.2-1B (1 billion parameters) through supervised fine-tuning on retrosynthetic data constructed by applying pre-defined templates to the Enamine library.[162] This approach enables the direct generation of linear route descriptions from target molecular structures represented as SMILES strings. By leveraging the pre-trained linguistic capabilities of the base models, SynLlama bypasses the recursive tree expansion, instead focusing on deconstructing targets into commercially available precursors. The model achieves high reconstruction rates (up to 74% on unseen Enamine sets) with reduced training data compared to prior generative baselines.

A similar approach was proposed by Wang et al. Their model, LLM-Syn-Planner,[163] adapts general-purpose LLMs such as GPT-4o or DeepSeek-V3 without additional fine-tuning to produce linear retrosynthetic routes also directly from target SMILES strings as inputs. The model outputs sequential decision lists, encompassing rationales, products, reactions, and reactants for each step, and terminating when all precursors are purchasable from databases like eMolecules.[164] Their design employs an evolutionary optimization algorithm that initializes, evaluates, and mutates full routes, drawing on examples from similar historical syntheses retrieved via molecular fingerprints. The result is enhanced performance on benchmarks like USPTO and Pistachio, with solve rates exceeding 90% on simpler sets.

This architectural shift towards the full-route sequence prediction results in a characteristic performance profile: while explicit search algorithms (e.g., MCTS) excel at finding *any* topological path to a starting material (high navigability), sequence-based generators typically demonstrate superior fidelity in recovering the specific convergent logic of experimental reference routes (high validity). By optimizing for the joint probability of the entire sequence, these models avoid the locally valid but strategically incoherent decisions often made by step-by-step planners. However, end-to-end generation introduces a structural rigidity. In explicit planners, the chemical logic (expansion model) and search constraints (inventory, forbidden reactions) are modular; one can swap stock lists without retraining the policy. In direct sequence generators, these boundary conditions are implicitly baked into the model weights during training, requiring fine-tuning or complex constrained decoding to adapt to new inventories.

6.5. Comparative Overview of Multistep Planning Method Families

To summarize the qualitative comparisons developed in this section, Table 6 organizes the major families of methods side by side. It highlights representative systems, their dominant training signals, the extent to which explicit search is required, and their characteristic strengths and limitations. The final columns relate each family to the Solv-N hierarchy introduced in Section 5.2.1, indicating the level at which current practice most naturally supports robust evaluation, and list common failure modes that motivate the transition from navigability-focused benchmarks to validity-oriented assessment.

7. From Navigability to Validity: A Critical Analysis of Benchmarking

From 2018 to 2023, the primary challenge in computational retrosynthesis was demonstrating that algorithms could navigate the combinatorial explosion of the search tree (Fig. 1). This period, which we characterize as the *Era of Navigability*, focused on finding *any* topological path connecting a target to a starting material. By the primary metric of this era, stock-termination rate (STR), modern planners have largely solved the navigation problem, routinely achieving success rates exceeding 99% on standard benchmarks. Further improvements (e.g. from 99.5% to 99.8%) are statistically marginal and often reflect hyperparameter tuning rather than algorithmic progress. This necessitates a transition into the *Era of Validity*, where the objective is no longer to find a path, but to verify its chemical correctness.

Fig. 8 illustrates the conceptual distinction between topological connectivity and experimental feasibility. While a planner optimizing for Solv-1 identifies a dense graph of potential connections, experimental constraints such as selectivity and purification requirements impose a filter that likely prunes this graph. Currently, standard evaluation metrics do not quantify this reduction; they treat all topologically valid routes as equal. Consequently, the field currently lacks the instrumentation to distinguish chemically sound proposals from those that are merely graph-theoretically connected.

7.1. Inventory Size as a Difficulty Dial

Because the STR measures only whether a route ends, it is strictly dependent on the definition of the available inventory. The size of the stock set functions as a difficulty dial: expanding the inventory increases the density of termination points, statistically shortening the required search depth and increasing the probability of success. Consequently, STR values are not portable across studies unless the inventory is fixed.

In practice, evaluations often treat two distinct inventory types as interchangeable. As defined in Section 5.3, the **physical tier** ($\sim 10^5$–10^6 compounds) forces the planner to deconstruct targets into simple commodity chemicals. The **virtual tier** ($\sim 10^7$–10^9 compounds), essentially a list of make-on-demand targets, relaxes the problem by allowing termination at complex intermediates. This effectively transforms the computational task from deep route planning to intermediate retrieval.

Table 6: Comparison of multistep retrosynthetic planning method families. Columns summarize representative systems, dominant training signals, the role of explicit search, and characteristic strengths and limitations. The final columns relate each family to the Solv-N hierarchy (Section 5.2.1) and list common failure modes that motivate a shift from navigability-focused to validity-oriented assessment.

Method family	**Representative systems**	**Dominant training signal**	**Explicit search**	**Characteristic strength**	**Characteristic limitation**	**Solv-N level**	**Failure mode**
Search-based planners	MCTS planners,[34] DFPN-E,[35] related tree-search systems	Supervised single-step policies / values; rollout-based returns	Yes (MCTS or heuristic tree search)	Systematic exploration of large route spaces; explicit route trees; flexible scoring and constraints	High computational cost; strong dependence on heuristics, stock set, and search budget; sensitivity to reaction-space shift	Solv-1 to lower Solv-2: topological reachability with partial plausibility via heuristics / filters	Routes that terminate but are circuitous or fragile; overuse of rare precedents; high STR masking low route quality
Template-based planners	Knowledge- and rule-based systems (LHASA-like), curated template libraries, template-augmented policies	Supervised learning over expert- or corpus-derived templates; applicability / frequency modeling	Typically yes (tree or graph search over templates)	High interpretability; direct linkage to precedent; easy integration of domain knowledge and hard constraints at rule level	Coverage bounded by template library; costly to extend; difficulty representing genuinely novel disconnections or mechanisms	Solv-1: connectivity within encoded chemistry; limited reach into Solv-2 unless selectivity is encoded in rules	No route when target requires chemistry outside rule set; misapplication of overgeneral templates; brittleness on underrepresented scaffolds
Template-free (single-step) models	Seq2seq models (e.g., Molecular Transformer),[36] graph-to-graph / edit-based models	Supervised reaction prediction or single-step retrosynthesis; sequence or graph-edit likelihoods	Yes in practice (global route built via search over single-step model)	Flexible reaction-space representation; can capture patterns not easily encoded as templates; scales with large corpora	Reduced interpretability; hard to enforce strict constraints; vulnerable to dataset biases and spurious correlations	Solv-1 to mid Solv-2: connectivity plus some implicit selectivity / plausibility	Locally plausible yet globally inconsistent routes; implausible disconnections with high confidence; cumulative per-step errors over long routes
Hybrid / neurosymbolic planners	Neural policies over symbolic rule sets; constraint- or theorem-proving backends guided by learned scores	Mixed: supervised template / policy learning plus symbolic reasoning; sometimes RL over symbolic states	Yes (symbolic search guided by neural components)	Combines interpretability and hard constraints with data-driven priors; can enforce feasibility filters while exploiting learned ranking	Architectural complexity; dependence on both rule quality and score calibration; nontrivial engineering of neural-symbolic interface	Upper Solv-2: topology plus more explicit selectivity / plausibility; extensible toward Solv-3 with condition / outcome models	Over- or under-constrained search; viable routes pruned by miscalibrated priors or strict filters; dead-ends from inconsistent neural-symbolic interaction
Direct full-route generators	End-to-end models generating full routes as sequences; route-graph generators outputting multistep pathways in one shot	Supervised learning on complete trajectories (full routes); sequence or graph likelihoods over plans	Not required (search only for sampling / reranking)	Fast inference; global modeling of route structure; can learn trajectory-level regularities (e.g., step count patterns)	Challenging credit assignment across steps; difficult to guarantee stepwise executability; limited control over individual disconnections without auxiliary checks	Solv-1 with emerging Solv-2: global connectivity; route-level plausibility typically needs external filters or validators	Syntactically valid but chemically inconsistent routes; missing or incompatible intermediates; gradual drift into unrealistic chemistry not captured by navigability metrics

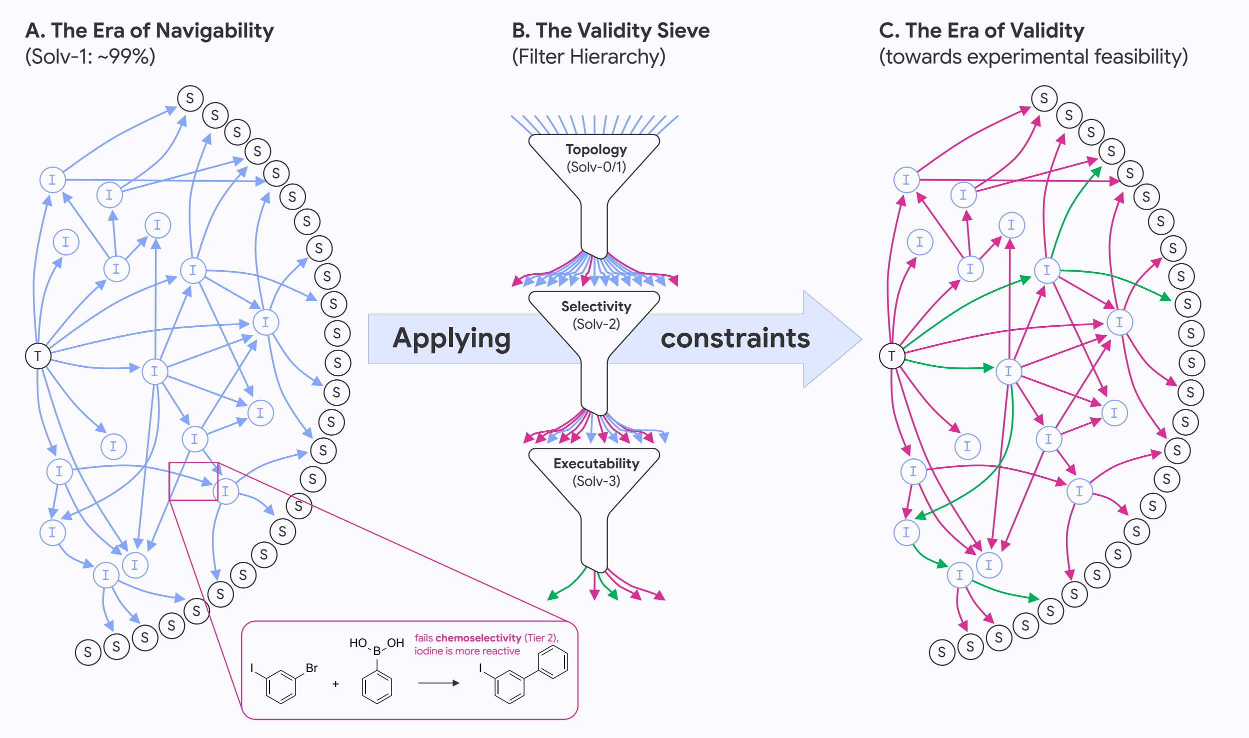


Figure 8: **The Phase Transition from Navigability to Validity. (A) The Era of Navigability:** When evaluated strictly on topological stock-termination (Solv-1, Section 8.1), planners routinely achieve ∼99% success, perceiving a dense, highly connected graph of plausible routes (blue). **(B) The Validity Sieve:** Transitioning to experimental reality requires filtering proposed routes through higher-order chemical constraints. The inset illustrates a characteristic Tier 2 (Chemoselectivity) failure hidden within Solv-1 graphs: a planner proposes a Bromo-Suzuki coupling, ignoring that the more reactive iodide will undergo preferential oxidative addition. **(C) The Era of Validity:** Upon applying Selectivity (Solv-2) and Executability (Solv-3) constraints, the vast majority of topologically valid routes are rendered chemically non-viable (red). The true experimentally actionable search space (green) is drastically sparser than legacy metrics suggest, underscoring the necessity of the rigorous benchmarking framework proposed in Section 5.2.1.

Table 7: **Impact of Stock Set Scope on Reported Solvability.** A comparison of reported success rates across recent literature reveals a strong dependence of algorithmic performance on the chosen inventory boundary conditions. Models evaluated against extended virtual libraries ($> 10^8$ million compounds) frequently report near-perfect stock-termination (Solv-1), as the expanded termination criteria shorten the requisite search depth. Conversely, evaluations constrained to physically in-stock inventories ($\sim 10^5$ compounds) yield substantially lower termination rates, particularly for complex (PaRoutes $n = 5$) or out-of-distribution (ChEMBL) targets. This disparity underscores the necessity of standardizing boundary conditions to isolate algorithmic capability from inventory scale.

Reference	**Model**	**Benchmark**	**Stock Source**	**Inventory**	**Solv-1**
I. Virtual Tier (make-on-demand / extended)					
Chen et al.[41]	Retro*	USPTO-190	eMolecules	~231 M[a]	86.84%
Zhao et al.[121]	MEEA*	USPTO-190	eMolecules	~231 M	100.0%
Wang and Montana[124]	InterRetro	Retro*-190[b]	eMolecules	$> 10^8$ M	100.0%
Xie et al.[120]	RetroGraph	USPTO-190	eMolecules	~231 M	99.47%
Wang et al.[163]	LLM-Syn-Planner	USPTO-190	eMolecules	~23 M	92.6%
Liu et al.[165]	PDVN	USPTO-190	eMolecules	~23.1 M	99.47%
Blackshaw et al.[166]	Enh-MCTS	ChEMBL (Rand)	ZINC +eMolecules	~35 M	99.2%
Shee et al.[42]	DirectMultiStep	ChEMBL-5000	eMolecules	~23.1 M	75.58%
Guo et al.[167]	ReSynZ	Retro*-190[b]	Sigma + eMol	~18 M	73.54%
Torren-Peraire et al.[168]	Chemformer (Search)	Caspyrus10k	ZINC Full	~17.4 M	94.1%
II. Physical Tier (off-the-shelf / in-stock)					
Shee et al.[42]	DirectMultiStep	ChEMBL-5000	ASKCOS Buyables	~330 k	68.66%
Guo et al.[167]	ReSynZ	Retro*-190[b]	Sigma-Aldrich	~85 k	50.26%
Wang et al.[136]	μMCT-dc	Reaxys (Rand)	Sigma + eMol	~107 k	76.2%
Akhmetshin et al.[123]	SynPlanner (MCTS)	PaRoutes ($n = 5$)	ASKCOS Buyables	~186 k	56.24%
Sun et al.[126]	SynLlama	ChEMBL	Enamine BB	~230 k	19.7%
Akhmetshin et al.[123]	SynPlanner (MCTS)	SAScore > 5	ASKCOS Buyables	~186 k	18.71%

[a]Reported inventory size. The data file in the original repository contains ~23.1M compounds, a number consistent with more recent literature. [b]The USPTO-190 and Retro*-190 benchmarks refer to the same set of 190 molecules, pre-filtered for high single-step model performance, introduced in Chen et al.[41]

Empirical comparisons confirm that varying the inventory alters apparent performance more than the choice of search algorithm (Table 7). For example, Guo et al. showed that for a fixed planner, expanding the inventory from a physical subset to a large virtual set increased STR from 73.5% to 87.3%. Similarly, recent state-of-the-art results, such as the 100.0% STR for MEEA*[121] and 99.5% for PDVN,[165] rely on virtual catalogs exceeding 230 million entries. When the inventory is restricted to physically deliverable buyables, success rates drop sharply; standard MCTS planning achieves only 18.7% on high-difficulty targets under strict stock constraints.[123] High success rates against virtual libraries therefore reflect the breadth of the inventory rather than the depth of the planning logic.

Furthermore, even nominally identical stock sources do not guarantee stable experimental conditions. Studies citing eMolecules or ZINC inventories often differ in whether they employ full catalogs, screening subsets, or merged composites.[120,163,166,168] For instance, Retro* evaluations utilized an eMolecules inventory of ∼231 million entries,[41] whereas subsequent work reported substantially smaller sets (∼23 million) or hybrid lists (∼35 million for ZINC+eMolecules).[163,166] This variance confirms that STR values are not portable across the literature unless the stock definition is rigorously standardized.

Table 8: **The Validity Gap and the Complexity Cliff.** The disconnect between Tier 1 success (topological stock-termination) and higher-tier criteria related to chemical plausibility (Tiers 2–3). **(A)** Stock-termination can trade off against ground-truth route recovery and computational efficiency. In the *Models Matter* benchmark,[168] Chemformer achieves near-perfect termination but low route recovery and high inference latency (∼8 hours/target), whereas AiZynthFinder exhibits the opposite trade-off under the same evaluation. **(B)** Stratified analysis on *RetroCast* (`mkt-lin-500`)[116] reveals a complexity-dependent drop in reconstruction: explicit search dominates on shallow targets but degrades rapidly by Length 4, while the direct sequence model (DMS) maintains higher performance on deeper routes.

Panel A: The Inverse Correlation of Solvability, Accuracy, and Speed
Benchmark: Caspyrus / PaRoutes (Models Matter[168])

Model	**Policy Type**	**Solvability** *(Stock-Termination)*	**Route Accuracy** *(Ground Truth Top-50)*	**Search Time** *(Seconds/Target)*
Chemformer	Direct Sequence	**99.7%**	11.9%	∼28,000
LocalRetro	Explicit Graph	86.0%	36.1%	∼160
AiZynthFinder	Explicit Graph	66.3%	**61.8%**	**∼160**

Panel B: The Complexity Cliff (Horizon Effect)
Benchmark: RetroCast mkt-lin-500[116]

		Top-10 Reconstruction Accuracy		
Model	**Search Strategy**	**Short (Len 2)**	**Medium (Len 4)**	**Deep (Len 6)**
AiZynthFinder	MCTS (Search)	**81.0%**	28.0%	9.0%
Retro*	A* (Search)	40.0%	17.0%	17.0%
DMS Explorer XL	Beam (Sequence)	67.0%	**54.0%**	**50.0%**

7.2. The Saturation and Fragmentation of Test Sets

The second major confounder in current benchmarking is target selection. The field's reliance on a small number of historical test sets, particularly USPTO-190 (Retro*-190), has led to a saturated evaluation environment. This benchmark was constructed by explicitly filtering for targets whose synthetic steps were ranked highly by a baseline model, a process designed to isolate and test graph search algorithms in 2020.[41] While instrumental in the navigability era, this pre-conditioning means the set is not representative of novel or challenging chemical space. As a result, modern planners now routinely achieve STRs between 93% and 100%,[120–122,124,129,132,137,163,165,167,169–173] rendering the benchmark non-discriminative for state-of-the-art systems.

Outside this saturated standard, the evaluation landscape is highly fragmented. Studies frequently employ bespoke test sets, including custom subsets from ChEMBL or Reaxys, hazard-filtered targets, or case studies selected for specific chemical features.[125,131,136,139,141,174–177] While useful for specific investigations, this practice prevents the accumulation of shared knowledge about model strengths and weaknesses, and makes it difficult to perform meaningful cross-paper comparisons.

To resolve these issues, evaluation must move from reporting aggregate success to performing category-specific analysis on representative benchmarks. The *PaRoutes* dataset established a standard for this by binning targets by difficulty (n_1 vs. n_5).[40] An even more granular analysis using reference route length (as a proxy for planning difficulty) has revealed a *complexity cliff*: STR remains high for short routes with 2-4 steps but collapses on the longer syntheses with more than 6 steps.[116] Ultimately, verifying true generalization requires that this category-specific analysis be paired with rigorous hold-out sets defined by scaffolds, reaction classes, or temporal splits, thereby disentangling chemical reasoning from the memorization of training patterns.[133,156,178]

7.3. The Divergence Between Topological Success and Chemical Validity

Beyond issues of target selection, the stock-termination rate metric is insensitive to the distinction between topological connectivity and chemical plausibility. A high STR certifies that a planner can find *a* path, but provides no information about whether that path corresponds to a viable experimental procedure.

This divergence is apparent in benchmarks that report both STR and Top-K route reconstruction. For example, on the PaRoutes n_5 set, planners with near-identical STR show measurable differences in route reconstruction accuracy.[40] In the Torren-Peraire et al. audit, one planner achieved 99.7% STR yet recovered only 11.9% of ground-truth routes (Table 8, Panel A).[168] Performance also degrades with route complexity; as shown in Table 8 (Panel B), the reconstruction accuracy of explicit search planners often collapses as route length increase. The issue is particularly acute on the USPTO-190 benchmark, where despite near-universal STR, audits find that Top-10 route reconstruction is in the low single digits.[116]

A common explanation for this disparity is that planners may find valid routes that differ from the historical reference, which strict Top-K matching penalizes. While this is possible, the validity of such algorithmically generated alternatives is difficult to verify without experimental follow-up, and recent analyses of the underlying single-step models provide reason for skepticism. An audit by Tran et al., for instance, reveals a systemic bias in these models toward proposing simpler transformations than those recorded experimentally.[179] This bias manifests as frequent errors in stereochemistry, leaving group assignment, and an underestimation of reaction complexity, suggesting that "novel" routes may often be chemically naïve artifacts rather than viable alternatives. Audits of "solved" multistep routes confirm this ambiguity, revealing proposals such as single steps with seven distinct reactants.[116] Consequently, until the field develops automated Tier 2/3 metrics, route reconstruction, however conservative, remains a primary proxy for grounding evaluation in experimental reality.[179]

Addressing the limitations of relying on a single reference route, Guo et al. proposed a learned scoring function that predicts a route's similarity to a hypothetical expert-designed path, even for novel targets.[180] By fine-tuning this predictor on human ratings, they developed a metric that correlates with chemical intuition better than binary solvability. However, the model architecture treats the reactions in a route as an unordered collection, a design choice that limits its ability to evaluate syntheses where the precise sequence of transformations is critical, such as those involving protecting groups.

7.4. Evaluator Dependence and Metric Fragmentation

Finally, evaluation is compromised when the metric is dependent on the model being evaluated. A prominent example is round-trip accuracy, in which a forward reaction predictor is used to validate the retrosynthetic disconnections proposed by the planner. While useful for filtering syntactic errors, these checks do not provide an independent measure of validity. When the forward model shares the same training distribution and architectural biases as the planner, a successful round-trip primarily confirms internal consistency rather than objective chemical correctness.

This challenge is further compounded by metric fragmentation, in which new methods are routinely introduced together with custom evaluation criteria. Metrics such as novel diversity scores or composite "feasibility scores"[173] can effectively highlight the strengths of a particular architecture, yet they hinder direct comparisons across the literature. Composite metrics, in particular, obscure the underlying causes of performance gains by combining multiple distinct factors, such as termination rate and the confidence of a learned classifier, into a single scalar value.

Progress in the Era of Validity requires that method development be decoupled from the definition of evaluation metrics. Rigorous evaluation is most informative when conducted within independent community-standardized frameworks that fix the evaluator, the stock set, and the target distribution (e.g., Syntheseus,[156] RetroCast[116]). Only by standardizing the measurement protocols can the field reliably attribute performance gains to algorithmic advances rather than to choices in metric design.

8. A Framework for Validity-Centric Evaluation

As retrosynthesis planners have matured, the limitations of evaluating them with a single aggregate success rate have become apparent. The saturation of stock-termination metrics on standard benchmarks (Section 7) necessitates a move toward more granular and chemically meaningful evaluation protocols. To facilitate this transition and enable more rigorous comparison of model capabilities, we propose a framework for validity-centric evaluation. This framework is designed to differentiate between topological connectivity and experimental plausibility, assess the quality of ranked route suggestions, and encourage standardized and reproducible benchmarking practices.

8.1. The Solvability Hierarchy (Solv-N)

The cornerstone of this framework is a tiered classification of solvability designed to resolve the ambiguity of current metrics. We propose expanding the term "solvability" (currently used to refer to the stock-termination rate) into a formal series, Solv-N (Table 9), where each level corresponds to a progressively stricter set of chemical validity constraints, as previously outlined in Section 5.2.1 and Table 4.

- **Solv-0 (Syntactic Solvability):** The route consists of syntactically valid molecular graphs.
- **Solv-1 (Topological Solvability):** The route connects the target to the stock set via topologically valid transformations. This is equivalent to the current stock-termination rate (STR) metric.
- **Solv-2 (Selectivity Solvability):** The route's transformations are chemically plausible, satisfying selectivity constraints.
- **Solv-3 (Executability Solvability):** The route is experimentally viable under realistic laboratory conditions.

This tiered system allows for more precise reporting. For example, a planner achieving near-perfect stock termination without verified chemical correctness would be accurately described as having a high Solv-1 rate. This notation clarifies that higher-order chemical constraints remain unverified, preventing the conflation of graph connectivity with experimental feasibility.

Achieving Solv-2 is particularly challenging, as it requires satisfying all sub-criteria—chemoselectivity (C), regioselectivity (R), diastereoselectivity (D), enantioselectivity (E), and stoichiometry (S)—simultaneously for every step. A proposed route is only fully validated at this level if it meets the Solv-2C, -2R, -2D, -2E, and -2S constraints concurrently. Given the difficulty of building automated verifiers for all these aspects, we suggest that as an interim measure of progress, individual sub-tier success rates (e.g., a Solv-2C rate) can serve as valuable diagnostics for specific model capabilities. The development of open-source community-standardized Solv-2 verifiers, such as the recently proposed ChemCensor,[181] would therefore be a critical step toward enabling large-scale validity-centric benchmarking.

8.2. Toward operational Solv-2/3 benchmarks

The higher tiers of the Solv-N hierarchy extend evaluation beyond graph connectivity and therefore require judgments about chemical plausibility and experimental feasibility. At

Table 9: **Operational Benchmark Scaffold for the Hierarchy of Chemical Validity (Solv-N).** Each tier corresponds to a stronger notion of success, moving from syntactic well-formedness (Solv-0) and stock-terminated topological planning (Solv-1) to chemically plausible, selective routes (Solv-2) and experimentally executable routes under realistic constraints (Solv-3). The table is intended as a minimal operational scaffold rather than a finalized community standard. As the hierarchy ascends, evaluation requires richer metadata, more heterogeneous validators, and increasing expert involvement.

Tier	Core Question	Minimum Required Inputs	Validator Type	Benchmark Output	Typical Failure Modes	Label-Noise Sources	Human Adjudic.
I. Representation and Topology							
S0	Syntactic validity	Canonical molecular representation (e.g., SMILES, SELFIES, graph); reaction format; atom mapping if relevant	Deterministic parser, sanitizer, valence checker, grammar validator	Validity rate; parsable fraction; invalid-string frequency	Invalid strings or graphs, valence errors, malformed reaction records, atom-mapping corruption	Toolkit disagreement, canonicalization differences, parser behavior, representation-conversion artifacts	No
S1	Stock-terminated topological route	Target; one-step model or reaction network; search policy; stock definition; stopping rule; evaluator protocol	Route-connectivity checker, stock-membership checker, search/evaluator audit	Stock-termination rate; topological solvability; route depth/length; success under fixed budget	Circular routes, evaluator-dependent success, termination in inflated inventories, shortcut artifacts	Inventory leakage, virtual-stock inflation, stock normalization differences, target overlap with stock or training data	Usually no
II. Chemical and Practical Validity							
S2	Chemically plausible / selective route	Ordered route; stereochemistry; reaction context; optional reagents/roles; precedent or forward-model scores	Rule-based filter, forward model, stereo/selectivity checker, precedent matching, expert panel	Step plausibility; route pass/fail; all-steps-plausible fraction; selectivity-aware success	Chemoselectivity or regioselectivity errors, stereo loss/inversion, incompatible functional groups, missing protecting-group logic	Missing stereochemistry, incomplete metadata, noisy atom mapping, uneven precedent coverage, model miscalibration	Often yes
S3	Executable route under realistic constraints	Route order; reagents/conditions; stoichiometry if available; protecting-group strategy; purification assumptions; scale/cost/time constraints	Condition model, workflow/rule engine, availability/cost checker, process filter, lab record or expert review	Executable-route rate; constrained success; expected cost/time/yield; fraction passing all constraints	Condition incompatibility, unavailable reagents, cumulative yield collapse, purification bottlenecks, unsafe or unscalable steps	Underspecified conditions, missing yield/scale data, supplier drift, undocumented purification burden, lab-to-lab variability	Frequently yes

present, these tiers lack universally accepted benchmark protocols. We do not claim that Solv-2 and Solv-3 are presently available as universally standardized labels; rather, we propose a minimal operational scaffold that could make validity-centric benchmarking progressively more reproducible across datasets and model classes.[38,40,182]

Minimal metadata requirements. Evaluating Solv-2 plausibility requires more information than a bare reaction graph. At minimum, benchmarks should provide an ordered sequence of steps, explicit stereochemical annotations, atom mappings or equivalent atom correspondence information, and, where available, reagent or reaction-context fields. Additional metadata such as reaction-class labels, precedent links, or forward-model confidence scores can support automated plausibility checks, but should not be treated as mandatory if absent from the underlying corpus .[78,182] Solv-3 evaluation requires richer metadata still, including reaction conditions (e.g., solvent, catalyst, temperature when known), starting-material availability, and coarse workflow constraints such as protecting-group strategy, purification assumptions, and route-level resource constraints.[38,79]

Stepwise versus routewise evaluation. Solv-2 is most naturally evaluated at the *step level*, since selectivity, stereochemical fidelity, and functional-group compatibility are local properties of individual transformations. In practice, however, the benchmarked quantity should still be reported at the *route level*: a route passes Solv-2 only if all constituent steps satisfy a defined plausibility criterion. This mirrors the logic of route-benchmark frameworks such as PaRoutes, where route quality is ultimately assessed over complete multistep plans rather than isolated disconnections .[40] Stepwise validation is therefore the mechanism; routewise success is the reporting unit.

Treatment of stereochemistry. Stereochemistry should be treated as an explicit constraint rather than an optional annotation. A step should count as stereochemically valid only if the proposed transformation preserves, removes, or induces stereochemical information in a manner consistent with known precedent, mechanistic expectation, or a calibrated forward predictor.[78,183] If stereochemistry is missing or underspecified in the benchmark record, this should be recorded as ambiguity rather than silently defaulted to a pass, since incomplete stereochemical metadata is a known source of reaction-data noise.[182]

Stoichiometry and reagent roles. Where available, stoichiometric roles (reactant, reagent, catalyst, solvent) should be retained and checked for consistency with the proposed transformation. In many public reaction corpora these fields are incomplete or noisy, so stoichiometry may need to be treated as a soft constraint at Solv-2 rather than a hard requirement.[182] For Solv-3, however, missing or implausible reagent-role assignments should count against executability, because condition selection and workflow feasibility depend directly on them.[38,79]

Protecting-group logic. Protecting-group strategy is a common source of hidden complexity in multistep routes. At Solv-2, missing or incompatible protecting-group logic can be counted as a plausibility failure when functional-group conflicts are evident locally. At

Solv-3, protecting-group handling must be assessed in the context of the entire route, including the feasibility and cost of installation and removal steps, compatibility with neighboring transformations, and the downstream purification burden they impose.[38,184]

Stepwise versus route-level executability. Solv-3 is inherently a *route-level* property. Individual steps can be screened for condition compatibility, reagent availability, or likely failure modes, but true executability depends on cumulative effects such as yield attrition, purification bottlenecks, scheduling constraints, and cross-step incompatibilities.[40,184] Stepwise filters are therefore useful as preliminary screens, but final Solv-3 assessment should be defined over the full route.

Reconciling validator disagreement. In practice, rule-based filters, predictive models, and expert judgment will sometimes disagree. Rather than forcing a single authority, benchmarks should record validator outputs separately and treat disagreement as structured uncertainty. One pragmatic protocol is to use a consensus or weighted-consensus rule for binary pass/fail reporting while preserving the underlying validator scores for downstream analysis. Expert adjudication can then be reserved for disputed or high-impact cases. This is especially important because reaction-data preparation, atom mapping, and condition annotation are themselves nontrivial sources of label noise .[182,185] A validity-centric benchmark should therefore report not only aggregate success but also the provenance and uncertainty of the labels used to define that success.

8.3. Ranking Beyond Termination: MRR-V

Binary solvability metrics are insufficient for planners that generate multiple candidate routes. A system that outputs 99 chemically invalid routes and one valid route achieves 100% solvability, but is practically inferior to a system that consistently ranks the valid route first. To capture this, we propose the *Mean Reciprocal Rank of Validity (MRR-V_i)*, defined as the mean reciprocal rank of the first route that satisfies Tier-i validity.

$$\text{MRR-V}_i = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{\text{rank}_i(q)}$$

Here, Q is the set of test targets, and $\text{rank}_i(q)$ is the rank of the first route proposed for target q that passes the Tier-i validity check. Adopting MRR-V_2, for example, would incentivize the development of models that not only find a selectivity-valid route but also rank it highly, directly aligning algorithmic optimization with practical laboratory utility.

8.4. A Call for Independent and Standardized Benchmarking

The history of machine learning in other domains suggests that rigorous progress requires separating method development from benchmarking. In the navigability era, it was common for papers to introduce a new search algorithm and a new custom benchmark simultaneously. This practice of coupled method-and-metric development introduces structural confounding, where apparent algorithmic gains cannot be isolated from relaxed boundary conditions or favorable test set selection. To ensure that progress is transparent and reproducible, the field must move toward evaluation by independent, standardized protocols.

A significant practical barrier to this goal is the difficulty of conducting fair, head-to-head comparisons. A truly rigorous comparison of two algorithms requires retraining both on identical datasets, using the same reaction templates and stock definitions. However, the lack of publicly available training scripts and the high computational cost of retraining often make such comparisons infeasible. As a result, the literature contains few direct, controlled studies of planner performance, impeding the community's ability to discern which algorithmic innovations are most impactful.

To address this challenge, we advocate for the dual-track evaluation model formalized by Morgunov and Batista,[116] which distinguishes between two complementary evaluation goals:

1. **The Developer Track:** This protocol is designed for the rigorous assessment of algorithmic novelty. It requires that method creators demonstrate the advantages of a new approach through fair, retrained comparisons against established baselines under fixed boundary conditions.

2. **The Chemist Track:** This protocol addresses the needs of practical application by facilitating the evaluation of pre-trained, off-the-shelf models as-is, without the requirement of retraining. A practicing chemist is often less concerned with theoretical algorithmic superiority and more with which available tool provides the most reliable routes for a given target.

By distinguishing between the assessment of algorithmic novelty and practical utility, this dual-track framework allows for both rigorous validation and pragmatic, application-focused assessment. Adopting such a standard would create a clearer path for both foundational research and the development of tools that serve the daily needs of the chemistry community.

8.5. Enabling Progress Through Shared Data and Outputs

A primary obstacle to automating higher-tier validity checking (Solv-2 and Solv-3) is the scarcity of large-scale, annotated datasets of chemically plausible but invalid routes. The patent literature provides an abundance of positive examples but offers no explicit supervision on why alternative synthetic paths fail.

To bootstrap the development of the next generation of automated validity models, we propose that the community adopt a standard of open route reporting. If future retrosynthesis studies were to publish their full, raw generated route trees in a standardized, machine-readable format (e.g., JSON), it would create an invaluable community resource. This shared data would enable researchers to crowdsource the auditing process, progressively building the ground-truth datasets of both successful and failed proposals required to train robust automated Solv-2 verifiers. The existence of a public infrastructure for sharing such outputs, like SynthArena,[116] demonstrates that this practice is technically feasible and would significantly accelerate the field's transition into the Era of Validity.

9. Beyond Topology: Toward Generation of Reliable Synthetic Procedures

This review has focused on topological planning: the construction of a stock-terminated sequence of graph edits (Tier 1, Solv-1). However, laboratory success is governed by *executability* (Solv-3), which requires that each step $A \to B$ admits a workable experimental procedure: choice of reagents and catalysts, solvents, temperature and time profile, and a purification strategy that yields material suitable for subsequent steps. This layer is difficult to learn from patent-derived datasets, where conditions are often unreported, inconsistent, or implicit.

Recent efforts to bridge this gap have gone beyond simple condition regression toward LLMs and agentic frameworks, a shift documented in recent comprehensive surveys[23,96,186] and general capability evaluations.[187–189] These frameworks range from retrieval-augmented generators of standard operation procedures grounded in equipment manuals and safety documents for compliant laboratory workflows,,[190] to robotic agents that translate natural language instructions into executable hardware controls for autonomous experimentation.[191–193] Specialized LLM models such as ChemCrow,[194] ChemActor,[195] ReactXT,[196] and ReactGPT[197] that translate between reaction SMILES representations and natural language descriptions of experimental protocols, use pretraining and in-context fine-tuning of the base LLM models to improve yield prediction and condition optimization. The experimental success rates include 67% yields in novel Suzuki-Miyaura couplings via human-AI collaboration (Chemma[198]), 94.5% yields in optimizations and scale-ups (LLM-RDF[199]), product confirmation in autonomous runs,[193] and 97% execution success in robotic tasks (CLAIRify[191]). Other training strategies that use reinforcement learning from verifiable rewards, such as the scientific reasoning model QFANG,[80] further push the boundary by producing chemically consistent step-by-step synthetic workflows that sometimes even improve verified literature protocols.[80] Collectively, these efforts show a shift from static text prediction toward tool-augmented and agentic LLM architectures that bridge high-level reaction plans and executable synthetic procedures.

Overall, demonstrated advantages include substantial acceleration of synthesis planning and documentation and clear evidence that fine-tuned and tool-augmented LLMs can outperform prior methods on tasks like procedure prediction and condition recommendation. However, despite these capabilities, generalist models frequently suffer from two distinct failure modes: regression to the global mode (predicting average conditions for specific chemistry) and semantic hallucination (generating fluent but chemically incoherent procedures).

9.1. Overcoming Regression to the Mode: The Specialist Approach

To address the limitations of global averaging, Li et al. introduced multiple optimized specialists for AI-assisted chemical prediction (MOSAIC):[81] a framework that abandons the single-model paradigm in favor of an ensemble of local experts (see Fig. 1 in Ref. 81).

The central object of MOSAIC is the *reaction universe*: a learned metric space of reaction specific fingerprints (RSFP). It is generated by a kernel metric network (KMN) from concatenated Morgan fingerprints of the reactants and products of chemical reactions in the training set containing natural language descriptions of the reaction procedures along with the reaction conditions and reaction yields. Training ensures that similar reactions are represented by neighboring points in the metric space of the reaction universe. The reaction classes are then defined by partitioning the entire reaction space into ∼2,500 distinct Voronoi cells using the FAISS clustering algorithm.[200] Clustering is purely metric-driven and does not enforce any traditional reaction type labels, allowing the system to learn and exploit similarities among chemical transformations directly from RSFP space and not being biased by the reaction classification adopted in the literature. Therefore, each Voronoi cell is a cluster of transformations with empirically similar fingerprints, often spanning multiple closely related named reactions, reflecting similarity in conditions, reagents and synthetic procedures. These cells are interpreted as *domains of chemical knowledge* and serve as training sets for fine-tuning ∼2,500 low-rank adaptation (LoRA[201]) models (dubbed "chemical experts"[81]) that capture region-specific statistics and precise reagent and condition profiles associated with the chemical subclass rather than regressing to a global average.

MOSAIC operates by routing queries to the most relevant expert model based on chemical similarity identified by the nearest Voronoi cell. The expert prediction yields a reproducible and human-readable experimental protocol that includes full details such as reagents, order of addition, reaction conditions, purification steps, and yield of the procedure. Furthermore, the distances from a query to the centroid of the nearest Voronoi cell or nearest training databease example within the cell provide explicit confidence metrics, allowing the system to distinguish confident interpolations from out-of-distribution predictions. These metrics can also serve as scoring functions for assessing novelty and feasibility of queried reactions: small distances indicate high confidence in protocol executability, as the query lies near the heart of a well characterized domain or close to a well characterized and validated reaction, while larger distances signal potential extrapolation into underrepresented or unexplored chemical territories. Potential applications of these metrics include prioritizing synthesis candidates in drug discovery or integrating with tree search algorithms in retrosynthetic computational pipelines.

In wet-lab validation on 37 de novo compounds spanning pharmaceuticals and agrochemicals, MOSAIC achieved a 71% success rate across 52 attempted transformations. Notably, the system successfully proposed executable protocols for challenging Buchwald-Hartwig aminations and olefin metathesis reactions that were structurally distinct from the training examples, suggesting that the learned metric space effectively clusters chemically related transformations. However, this performance relies on training and maintaining thousands of independent disjoint models, trading simplicity for the precision of extreme specialization.

9.2. Grounding Procedural Generation

Procedure generation demands more than predicting catalysts and solvents; it requires orchestrating a chemically coherent sequence of operations (addition, quench, workup, isolation) consistent with the intended transformation. Unconstrained language models frequently produce procedures that are stylistically plausible yet chemically invalid because the generated text is decoupled from the underlying structural change. Liu et al.[80] demonstrate that even advanced models like GPT-5 can misinterpret reaction intent—for example, misidentifying a benzylic oxidation as a benzoylation—resulting in fluent protocols that synthesize the wrong molecule. Similarly, models often suggest catalytic hydrogenation for intermediates containing reducible motifs that must be preserved (e.g., hydrogenolysis of C-N bonds).

QFANG[80] addresses these failure modes by explicitly grounding procedural generation in the atom-mapped graph edit ensuring that generated experimental protocols remain chemically consistent with the underlying molecular transformations. The approach is based on the chemistry-guided reasoning (CGR) framework, a two-stage process designed to produce high-quality reasoning datasets at scale. In the first stage, a "factual scaffold" is extracted from the atom-mapped reaction SMILES, capturing the essential functional group changes, bond formations, and disconnections that define the core topological logic of the reaction. This scaffold acts as a set of graph transformation constraints, enforcing rules such as atom conservation, valence adherence, and stereochemical fidelity to prevent hallucinatory deviations. In the second stage, an LLM expands this set of constraints into a detailed procedural narrative, incorporating contextual elements such as reagent quantities and reaction conditions while remaining tethered to the ground-truth graph edits. By conditioning the text generation on these explicit graph-based constraints, the model aligns the procedural steps with the verifiable topology, demonstrating superior generalization to out-of-domain reactions and adaptability to user-specified parameters such as scale and temperature. This architecture confirms that for generative chemistry, textual fluency is a liability unless strictly constrained by the syntax of the graph transformation while unconstrained models often produce plausible but chemically invalid outputs.

9.3. The Non-Markovian Nature of Purity

Finally, a persistent blind spot in both synthesis planning and procedure generation is the assumption of modularity. Current systems typically treat purification as a solved abstraction, assuming that each individual reaction step yields sufficiently pure material to serve as the input for the subsequent step without interference. In practice, however, organic synthesis is inherently non-Markovian: the outcome of a given step depends not only on its immediate precursors but also on the accumulated history of the route. Impurities, residual catalysts, and inseparable byproducts can propagate through multiple steps, leading to cascading failures. For instance, a copper-catalyzed cross-coupling reaction achieving a 90% isolated yield may appear successful in isolation, but it becomes functionally untenable if trace copper carryover poisons a downstream palladium-catalyzed cycle or interferes with a biological assay of the final product. Effects of such propagation underscore the need for models that explicitly account for intermolecular interactions and contaminant persistence across the entire synthetic sequence.

Leading platforms like ASKCOS[202] have begun to address this challenge by integrating explicit impurity prediction modules, which filter topologically valid but chemically undesirable steps during retrosynthetic analysis. These modules leverage data-driven approaches, such as template-free forward prediction models trained on large reaction databases, to anticipate impurities arising from five primary modes: minor products, side reactions, dimerizations, solvent adducts, and reactions involving subsets of reactants. By simulating these side processes, ASKCOS can identify and discard proposals that would introduce problematic contaminants, thereby enhancing the practical feasibility of suggested routes. However, these impurity checks predominantly operate as local constraints, applied step-wise without fully considering downstream consequences. True Solv-3 planning, which emphasizes executability under realistic laboratory conditions, including yield optimization, purification requirements, safety considerations, and scalability, requires a more holistic approach. This could involve an introduction of the route-level objective function of the synthetic sequence, that penalizes not only individual step inefficiencies but also cumulative factors such as separation complexity and impurity propagation. Formally, this function might incorporate terms for predicted impurity profiles, chromatographic separability scores, and compatibility assessments across steps, shifting the optimization paradigm from mere step-wise yield maximization to route-wise material quality and overall process robustness.

10. Outlook: Toward a Chemical Foundation Model

This review has charted the evolution of synthesis planning through a phase transition. The challenges of the *Era of Navigability*, which focused on finding any valid path through a combinatorial search space, have largely been solved by modern algorithms. The field now enters the *Era of Validity*, a period defined by the pursuit of chemical correctness. The central task of the coming decade will be to build systems that not only connect molecular graphs but do so with the causal logic and experimental reliability of a trained chemist.

Achieving robust performance on Solv-2 (Selectivity) and Solv-3 (Executability) metrics will require more than improved evaluation alone; it necessitates concerted efforts in infrastructure design, data generation, and the fundamental architecture of the models themselves. We conclude by outlining these key areas, which together form a path toward bridging the gap between topological search and physical reality.

10.1. Infrastructure as a Scientific Instrument

The advancement of synthesis planning is an inherently interdisciplinary endeavor, relying on distinct expertise from both organic chemistry and computer science. In the field's formative years, progress was driven by framing retrosynthesis in terms that were most amenable to established computational techniques: specifically, as a search problem on an exceptionally large graph. This formulation was a necessary and productive abstraction, as it allowed for the direct application of powerful search algorithms and spurred rapid innovation in solving the topological connectivity challenge.

This early focus, however, also illustrates a crucial principle for the future of computational chemistry. The software we build is not a neutral tool; it is the scientific instrument through which we probe a problem. The design of that instrument profoundly influences the questions we ask and the answers we obtain. An instrument architected to optimize graph traversal will naturally orient the field toward measuring success with graph-based metrics. To ask the deeper chemical questions of selectivity and experimental feasibility, the instrument must be designed with those principles as its foundational logic. This requires a paradigm of co-design that moves beyond consultation to active architectural contribution. The most significant advances in this new era will likely be driven by a new generation of researchers fluent in both reaction mechanisms and performant software design. The success of integrated platforms like ASKCOS[202] provides a compelling demonstration that when this deep collaboration is achieved, the research focus naturally shifts from computational benchmarks to practical, experimental utility.

10.2. Physics-Based Supervision and Automated Experimentation

A prevailing narrative in chemical AI suggests that progress is fundamentally constrained by the scale of available experimental data, with automated laboratories often presented as the primary solution. While automated experimentation is an invaluable tool for generating ground-truth data, this perspective may underutilize the vast predictive power of established theoretical and computational chemistry. Decades of research have yielded robust physical models that can, in principle, predict the very outcomes – selectivity, stability, and reactivity – that are most critical for building valid planners.

Historically, however, these powerful theoretical tools have been deployed as artisanal, single-molecule calculations rather than as at-scale data generation engines. The critical bottleneck, therefore, may not be a deficit in scientific theory, but rather a deficit in the engineering required to transform these physical models into high-throughput supervision pipelines. Instead of relying solely on the sparse and biased data from patent literature, the field can generate its own high-fidelity labels.

10.3. Search-Augmented Generation

The tension between explicit graph search and direct sequence generation is likely a transient phase in the field's development. A recurring observation in computationally intensive sciences is that general-purpose architectures capable of scaling with computation eventually outperform systems that rely on complex, hand-engineered heuristics.[203] This suggests a powerful, symbiotic path forward for synthesis planning.

In this paradigm, explicit search, guided by rigorous physical constraints (e.g., automated Solv-2 filters), acts as the "teacher." It can explore the vast combinatorial space of synthesis to generate large, high-fidelity datasets of valid routes. High-capacity sequence models can then act as the "student," distilling this complex physical and strategic logic into a fast, generalizable policy. This approach amortizes the immense computational expense of search into the inference step, combining the rigor of symbolic methods with the speed and pattern-recognition capabilities of deep learning.

10.4. Synthesis Planning as a Pre-training Objective

We return to the central thesis of this review: that synthesis planning is the chemical analogue of next-token prediction. Current chemical foundation models, largely trained on static graph masking or SMILES reconstruction, often fail to generalize to activity cliffs (Section 2). We hypothesize that this fragility arises because they learn the syntax of representation rather than the syntax of transformation.

Retrosynthesis is a uniquely demanding generative objective because it forces the model to internalize the structural grammar of transformation under explicit validity constraints, rather than to correlate static motifs with labels. This is also why synthesis planning is a plausible route to emergence. The same electronic-structure determinants that govern reactivity—functional-group electronics, polarization, steric accessibility, and conformational preferences—also shape many downstream properties by controlling how a molecule interacts with its environment.

The implication is not that property prediction disappears, but that it becomes a readout of a representation shaped by planning. If the community adopts the validity-centric framework proposed here—pairing this objective with auditable routes and rigorous Solv-2 metrics—synthesis planning offers a path to grounded chemical representation learning. Under this view, a chemical foundation model—or, more cautiously, a program toward artificial chemical intelligence—becomes a concrete research agenda rather than a branding term: pre-train on planning as the chemical analogue of next-token prediction, and evaluate progress by whether planning competence transfers to new chemistry and new functional questions.

10.5. What evidence would challenge this thesis?

The hypothesis of this review is that pre-training on the causal logic of synthesis planning will yield more robust and generalizable chemical representations than objectives based on static molecular structures. This thesis would be substantially weakened if empirical results provided any of the following outcomes:

- **A failure of learned planned skills to generalize to broader chemical reasoning.** The thesis would be falsified if a model becomes an expert at route-finding, yet its learned representations provide no significant advantage for unrelated chemical tasks, such as reasoning about structure-property relationships and discerning activity cliffs. This would demonstrate that learning the "syntax of matter" yields no more general chemical intelligence than the "syntax of notation", ultimately failing to surpass the known performance plateaus of static pre-training. Synthesis planning would thus be revealed as a narrow, specialized skill, not a foundational one.

- **The data requirements for effective pre-training prove prohibitive.** The thesis is predicated on the existence of a sufficiently large and diverse corpus of valid synthetic routes to learn from. It would be practically falsified if the performance of planning-based models saturates at a low level of competence due to the inherent limitations of available experimental data, and if the physics-based data generation (e.g., high-throughput QM) proves unable to bridge this gap at a reasonable computational cost.

- **The analogy to natural language processing proves to be flawed.** The success of large language models may depend critically on post-training alignment techniques like reinforcement learning from human feedback,[23] which are used to refine raw predictive models into useful assistants. This thesis would be significantly weakened if a similar massive-scale feedback loop from expert chemists is the true bottleneck for achieving artificial chemical intelligence, and that such a data-generating process is not scalable. In this scenario, the pre-training objective alone would be insufficient.

In summary, reproducible evidence that alternative pretraining objectives or validators consistently outperform synthesis planning centered approaches on synthesis and property-transfer tasks would motivate rethinking the centrality of synthesis planning in chemical foundation modeling.

11. Conclusions

Multistep synthesis planning has advanced rapidly in recent years, but the meaning of reported progress depends strongly on what is being measured. Across much of the recent literature, benchmark success has often reflected improvements in *navigability*: the ability to find a stock-terminated route through a large combinatorial search space. That progress is real and important. At the same time, this review has argued that navigability alone is an incomplete proxy for practical synthetic competence, because topologically valid routes may still fail at the level of selectivity, stereochemical fidelity, protecting-group logic, condition compatibility, or overall executability.

To clarify this distinction, we introduced the *Hierarchy of Chemical Validity (Solv-N)*, which separates syntactic validity (Solv-0), topological solvability (Solv-1), chemically plausible and selective routing (Solv-2), and executable route construction under realistic constraints (Solv-3). We view this hierarchy not as a finalized standard, but as a minimal scaffold for organizing evaluation and for making explicit which levels of validity a given model, benchmark, or claim actually addresses. In particular, the review has highlighted that many widely reported near-saturation results pertain primarily to Solv-1, and that these results can depend strongly on stock definition, evaluator design, and inventory scope.

Within the domain considered here, recent work suggests a shift in emphasis from route finding alone toward stronger notions of route validity. Search-based systems, direct sequence generators, and hybrid or neurosymbolic architectures all contribute differently to this transition, but none yet resolves the full Solv-2/3 problem in a standardized and experimentally grounded way. For this reason, we argue that future benchmarking should place greater weight on chemical plausibility, selectivity, and execution constraints, and should report these dimensions separately rather than collapsing them into a single notion of solvability.

A broader interpretive claim of this review is that synthesis planning may be a valuable organizing objective for learning chemistry-aware representations. More specifically, multistep retrosynthesis appears to be a strong candidate objective for tasks that depend on reactivity, synthetic accessibility, and route-level compositional reasoning. We have deliberately framed this as a hypothesis supported by converging evidence, rather than as a settled conclusion. Forward reaction modeling, condition and workflow prediction, multimodal structure–property learning, 3D or physics-informed objectives, and lab-in-the-loop systems each capture aspects of chemical intelligence that retrosynthesis alone does not. The most plausible path forward is therefore not a single universal objective, but a broader foundation-model stack in which retrosynthesis provides one important organizing prior among several complementary learning signals.

We emphasize that the conclusions of this review are grounded primarily in multistep small-molecule organic retrosynthesis, especially database-driven planning systems built on patent and reaction-corpus precedent. Whether the same framework transfers unchanged to catalysis, inorganic and organometallic synthesis, polymer and materials chemistry, electrochemistry, or peptide and biocatalytic synthesis remains unresolved. These domains differ substantially in representation, data quality, mechanistic structure, and criteria for experimental success, and will likely require domain-specific extensions of both the Solv-N hierarchy and the modeling conclusions developed here.

Taken together, the literature supports a more validity-centric view of progress in contemporary retrosynthesis. If the field can move from topological route finding toward reproducible evaluation of chemical plausibility and executability, then synthesis planning may become not only a benchmark task, but also a useful organizing framework for more general chemical machine learning. The central challenge ahead is therefore not simply to find more routes, but to measure—and ultimately learn—which routes are chemically credible, experimentally actionable, and robust under realistic constraints.

Author Information

Corresponding Authors

Author Contributions

A.M. conceived the central thesis of the review, conducted the primary literature survey, and wrote the original manuscript. Y.S. contributed to the initial conceptualization and framing of the core arguments. A.V.S. provided critical review and editing. V.S.B. supervised the project, provided extensive revisions, and secured the invitation for the review. All authors read and approved the final manuscript.

Conflict of Interest

A.M. is the creator of RetroCast and SynthArena and lead developer of DirectMultiStep, which are discussed in this review. Y.S. conceived the idea and is a co-developer of DirectMultiStep. V.S.B. is a co-author on studies cited in this review (Refs 55, 81, 116, 42, 153). The authors declare no other competing financial interests.

Biographies

Dr. Anton Morgunov received his S.B. in Chemistry and Biology from the Massachusetts Institute of Technology (MIT) and earned his Ph.D. in Chemistry from Yale University in 2026 under the supervision of Victor S. Batista. His doctoral research established structural foundations for artificial chemical intelligence, focusing on the intersection of generative models and rigorous software infrastructure. He is the lead architect of DirectMultiStep and the creator of RetroCast and SynthArena. His contributions to generative planning were recognized with Second Place in the Standard Industries Chemical Innovation Challenge.

Dr. Yu Shee received his B.S. in Chemical Biology and Computational Chemistry from the University of California, Berkeley. He earned his Ph.D. in Chemistry from Yale University in 2025 under the supervision of Victor S. Batista. His doctoral research focused on the application of machine learning methods to organic reactions and retrosynthetic planning. His work on generative models was recognized with Second Place in the Standard Industries Chemical Innovation Challenge.

Dr. Alexander V. Soudackov received his B.Sc. in chemistry and M.Sc. in quantum chemistry from Lomonosov Moscow State University and his Ph.D in physics and mathematics from the Karpov Institute of Physical Chemistry in Moscow, Russia. He was a recipient of an Alexander von Humboldt Research Fellowship and conducted his postdoctoral research in quantum chemistry in the University of Hanover in Germany. His other postdoctoral appointments include University of Notre Dame and University of Utah. He has contributed to the studies of electronic structure of transition metal complexes, dynamics of charge transfer reactions in polar media, and theory of proton-coupled electron transfer reactions in complex environments. Currently, he is a research scientist in the Chemistry Department at Yale University and his research is focused on quantum computing and machine learning methods applied to various problems in physics and chemistry.

Prof. Victor S. Batista is the John Gamble Kirkwood Professor of Chemistry at Yale University, where he has been a faculty member since 2001. He earned his B.Sc. in Chemistry from the University of Buenos Aires in 1989 and his Ph.D. in Theoretical Chemistry from Boston University in 1996. After completing postdoctoral research at the University of California, Berkeley, and the University of Toronto, he joined Yale's Department of Chemistry. Professor Batista's research focuses on theoretical and computational chemistry, particularly on developing semiclassical and quantum dynamics methods to study photoinduced reactions and catalytic processes. He has published over 425 articles in peer-reviewed journals, contributing significantly to the understanding of photosynthetic systems and quantum control of chemical dynamics. As the Director of the NSF Center for Quantum Dynamics on Modular Quantum Devices, Professor Batista leads efforts to develop new paradigms for quantum simulations of complex chemical systems using programmable Kerr-cat platforms. The Center aims to demonstrate the unique capabilities of bosonic modular devices in simulating chemical dynamics and correlated many-body systems. Throughout his career, Professor Batista has received numerous accolades. He is a Fellow of the Royal Society of Chemistry and an elected member of the Connecticut Academy of Science and Engineering. His recent publications include work on quantum machine learning applications in drug discovery as well as the development of quantum algorithms for simulating non-Markovian dynamics in chemical systems.

Acknowledgements

The authors acknowledge support from the National Science Foundation under Grant No. CHE-2124511.

References

(1) Hansch, C.; Fujita, T. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *Journal of the American Chemical Society* **1964**, *86*, 1616–1626.

(2) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/Nondrug Classification. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1882–1889.

(3) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 1947–1958.

(4) Zupan, J.; Gasteiger, J. Neural Networks: A New Method for Solving Chemical Problems or Just a Passing Phase? *Analytica Chimica Acta* **1991**, *248*, 1–30.

(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 983–996.

(6) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems* **2015**, *28*.

(7) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

(8) Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Chemical Society Reviews* **2019**, *48*, 2665–2679.

(9) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. 2017; https://arxiv.org/abs/1706.03762.

(10) Kipf, T. N.; Welling, M. Semi-supervised classification with graph convolutional networks. 2016; https://arxiv.org/abs/1609.02907.

(11) Sun, M.; Li, P. Graph convolutional networks for computational drug development and discovery. *Briefings in Bioinformatics* **2020**, *21*, 919–935.

(12) Kensert, A.; Alvarsson, J.; Norinder, U.; Spjuth, O. Graph Convolutional Networks for Improved Prediction and Interpretability of Chromatographic Retention Data. *Analytical Chemistry* **2021**, *93*, 12678–12688.

(13) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. 2017; https://arxiv.org/abs/1704.01212.

(14) Jo, J.; Lee, S.; Lee, S.; Hwang, S. J.; Hwang, H. The message passing neural networks for chemical property prediction on SMILES. *Methods* **2020**, *179*, 65–72.

(15) Li, L.; Zhang, Y.; Wang, G.; Xia, K. Kolmogorov–Arnold graph neural networks for molecular property prediction. *Nature Machine Intelligence* **2025**, *7*, 1346–1354.

(16) Hasebe, T. Knowledge-Embedded Message-Passing Neural Networks: Improving Molecular Property Prediction with Human Knowledge. *ACS Omega* **2021**, *6*, 25953–25964.

(17) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; others Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling* **2019**, *59*, 3370–3388.

(18) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, B.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph neural networks for materials science and chemistry. *Communications Materials* **2022**, *3*, 93.

(19) Rittig, J. G.; Gao, K.; Ihme, M.; Coley, C. W.; Mitsos, A. Graph neural networks for the prediction of molecular structure-property relationships. 2022; https://arxiv.org/abs/2208.04852.

(20) Berry, K.; Cheng, L. A Survey of Graph Neural Networks for Drug Discovery: Recent Developments and Challenges. 2025; https://arxiv.org/abs/2509.07887.

(21) Wang, R.; Zhuang, C. Graph neural networks driven acceleration in drug discovery. *Acta Pharmaceutica Sinica B* **2025**, *15*, 6163–6177.

(22) Xia, J.; Zhang, L.; Zhu, X.; Liu, Y.; Gao, Z.; Hu, B.; Tan, C.; Zheng, J.; Li, S.; Li, S. Z. Understanding the Limitations of Deep Models for Molecular Property Prediction: Insights and Solutions. Advances in Neural Information Processing Systems. 2023.

(23) Alampara, N.; Aneesh, A.; Ríos-García, M.; Mirza, A.; Schilling-Wilhelmi, M.; Aghajani, A. A.; Sun, M.; Prastalo, G.; Jablonka, K. M. General-Purpose Models for the Chemical Sciences: LLMs and Beyond. *Chemical Reviews* **2026**, *126*, 2484–2549.

(24) Corey, E. J.; Wipke, W. T.; Cramer, R. D. I.; Howe, W. J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *Journal of the American Chemical Society* **1972**, *94*, 421–430.

(25) Wipke, W. T.; Dyott, T. M. Simulation and evaluation of chemical synthesis. Computer representation and manipulation of stereochemistry. *Journal of the American Chemical Society* **1974**, *96*, 4825–4834.

(26) Gelernter, H.; Rose, J. R.; Chen, C. Building and refining a knowledge base for synthetic organic chemistry via the methodology of inductive and deductive machine learning. *Journal of Chemical Information and Computer Sciences* **1990**, *30*, 492–504.

(27) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Organic Process Research & Development* **2015**, *19*, 357–368.

(28) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* **2016**, *55*, 5904–5937.

(29) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L.; Grzybowski, B. A. Efficient Syntheses of

Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem* **2018**, *4*, 522–532.

(30) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Central Science* **2016**, *2*, 725–732.

(31) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Nguyen, Q. L.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Central Science* **2017**, *3*, 1103–1113.

(32) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Central Science* **2017**, *3*, 1237–1245.

(33) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chemistry – A European Journal* **2017**, *23*, 6118–6128.

(34) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(35) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning. Advances in Neural Information Processing Systems. 2019.

(36) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Häuselmann, R.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11*, 3316 – 3325.

(37) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1603.

(38) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive chemistry: machine learning for reaction deployment, reaction development, and reaction discovery. *Chemical science* **2023**, *14*, 226–244.

(39) Pyser, J. B.; Chakrabarty, S.; Romero, E. O.; Narayan, A. R. State-of-the-art biocatalysis. *ACS central science* **2021**, *7*, 1105–1116.

(40) Genheden, S.; Bjerrum, E. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery* **2022**, *1*, 527–539.

(41) Chen, B.; Li, C.; Dai, H.; Song, L. Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search. 2020; https://arxiv.org/abs/2006.15820.

(42) Shee, Y.; Morgunov, A.; Li, H.; Batista, V. S. DirectMultiStep: Direct Route Generation for Multistep Retrosynthesis. *Journal of Chemical Information and Modeling* **2025**, *65*, 3903–3914.

(43) Green, J.; Diaz, C. C.; Jakobs, M. A. H.; Dimitracopoulos, A.; van der Wilk, M.; Greenhalgh, R. D. Current Methods for Drug Property Prediction in the Real World. 2023; https://arxiv.org/abs/2309.17161.

(44) Deng, J.; Yang, Z.; Wang, H.; Ojima, I.; Samaras, D.; Wang, F. Unraveling Key Elements Underlying Molecular Property Prediction: A Systematic Study. 2023; https://arxiv.org/abs/2209.13492.

(45) Liyaqat, T.; Ahmad, T.; Saxena, C. Advancements in Molecular Property Prediction: A Survey of Single and Multimodal Approaches. *Archives of Computational Methods in Engineering* **2024**, *33*, 613 – 643.

(46) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *Journal of Chemical Information and Modeling* **2022**, *62*, 5938–5951, PMID: 36456532.

(47) Szostek, T.; Szulczyk, D. From obstacle to design advantage: activity cliff aware modeling for small-molecule drug discovery. *Drug Discovery Today* **2026**, *31*, 104589.

(48) Cheng, Z.; Xiang, H.; Ma, P.; Zeng, L.; Jin, X.; Yang, X.; Lin, J.; Deng, Y.; Song, B.; Feng, X.; Deng, C.; Zeng, X. MaskMol: knowledge-guided molecular image pre-training framework for activity cliffs with pixel masking. *BMC Biology* **2024**, *23*.

(49) Kim, H.; Park, J.; Choe, J.; Baek, S.; Hwang, H.; Kang, J. GraphCliff: Short-Long Range Gating for Subtle Differences but Critical Changes. 2025; https://arxiv.org/abs/2511.03170.

(50) Liu, X.; biao Li, H.; Liu, Y.; Fan, C. Multi-Level Fusion Graph Neural Network for Molecule Property Prediction. *Journal of chemical information and modeling* **2025**,

(51) Chen, X.; Yu, D.; Zhao, L.; Liu, F. ACES-GNN: can graph neural network learn to explain activity cliffs? *Digital Discovery* **2025**, *4*, 2062–2074.

(52) Shi, Z.; Wang, Y.; Weerawarna, P.; Zhang, J.; Richardson, T.; Wang, Y.; Huang, K. Structure-Aware Compound-Protein Affinity Prediction via Graph Neural Network with Group Lasso Regularization. 2025; https://arxiv.org/abs/2507.03318.

(53) Shen, W.; Cui, C.; Su, X.; Zhang, Z.; Arce, A. V.; Wang, J.; Shi, X.; Zhang, Y.; Wu, J.; Chen, Y. Z.; Zitnik, M. Activity Cliff-Informed Contrastive Learning for Molecular Property Prediction. *Research Square* **2024**, rs.3.rs–2988283.

(54) Shu, Z.; Deng, Y.; Zhang, H.; Nie, Z.; Chen, J. MTPNet: Multi-Grained Target Perception for Unified Activity Cliff Prediction. International Joint Conference on Artificial Intelligence. 2025.

(55) Kyro, G. W.; Smaldone, A. M.; Shee, Y.; Xu, C.; Batista, V. S. T-ALPHA: A Hierarchical Transformer-Based Deep Neural Network for Protein–Ligand Binding Affinity Prediction with Uncertainty-Aware Self-Learning for Protein-Specific Alignment. *Journal of Chemical Information and Modeling* **2025**, *65*, 2395–2415, PMID: 39965912.

(56) Zhang, Y.; Li, S.; Meng, K.; Sun, S. Machine Learning for Sequence and Structure-Based Protein-Ligand Interaction Prediction. *Journal of chemical information and modeling* **2024**,

(57) Wang, D. D.; Wu, W.; Wang, R. Structure-based, deep-learning models for protein-ligand binding affinity prediction. *Journal of Cheminformatics* **2024**, *16*.

(58) Wu, F. A Semi-supervised Molecular Learning Framework for Activity Cliff Estimation. International Joint Conference on Artificial Intelligence. 2024.

(59) Wan, Y.; Wu, J.; Hou, T.; Hsieh, C.-Y.; Jia, X. Multi-channel learning for integrating structural hierarchies into context-dependent molecular representation. *Nature Communications* **2023**, *16*.

(60) Tamura, S.; Miyao, T.; Bajorath, J. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *Journal of Cheminformatics* **2023**, *15*.

(61) Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. 2019.

(62) Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018.

(63) Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; Amodei, D. Language Models are Few-Shot Learners. 2020; https://arxiv.org/abs/2005.14165.

(64) Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. 2021; https://arxiv.org/abs/2103.00020.

(65) Wang, J.; Liu, Z.; Zhao, L.; Wu, Z.; Ma, C.; Yu, S.; Dai, H.; Yang, Q.; Liu, Y.; Zhang, S.; Shi, E.; Pan, Y.; Zhang, T.; Zhu, D.; Li, X.; Jiang, X.; Ge, B.; Yuan, Y.; Shen, D.; Liu, T.; Zhang, S. Review of Large Vision Models and Visual Prompt Engineering. 2023; https://arxiv.org/abs/2307.00855.

(66) Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible Scaling Laws for Contrastive Language-Image Learning. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2023; pp 2818–2829.

(67) OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.;

Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Kaiser, L.; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Kondraciuk, L.; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; Zoph, B. GPT-4 Technical Report. 2024; https://arxiv.org/abs/2303.08774.

(68) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* **1988**, *28*, 31–36.

(69) Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. 2020; https://arxiv.org/abs/2010.09885.

(70) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology* **2022**, *3*, 015022.

(71) Ahmad, W.; Simon, E.; Chithrananda, S.; Grand, G.; Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. 2022; https://arxiv.org/abs/2209.01712.

(72) Singh, R.; Barsainyan, A. A.; Irfan, R.; Amorin, C. J.; He, S.; Davis, T.; Thiagarajan, A.; Sankaran, S.; Chithrananda, S.; Ahmad, W.; Jones, D.; McLoughlin, K.; Kim, H.; Bhutani, A.; Sathyanarayana, S. V.; Viswanathan, V.; Allen, J. E.; Ramsundar, B. ChemBERTa-3: an open source training framework for chemical foundation models. *Digital Discovery* **2026**, *5*, 662–685.

(73) Chen, H.-G.; Vogt, M.; Bajorath, J. DeepAC - Conditional transformer-based chemical language model for the prediction of activity cliffs formed by bioactive compounds. *Digital Discovery* **2022**, *1*, 898–909.

(74) Tang, H.; Feng, S.; Lin, B.; Ni, Y.; Liu, J.; Ma, W.-Y.; Lan, Y. Contextual Molecule Representation Learning from Chemical Reaction Knowledge. 2024; https://arxiv.org/abs/2402.13779.

(75) Wu, J.; Zhu, Y.; Wang, X.; Li, Y.; Yin, M.; Wang, T.; Han, Y.; Kang, Y.; Deng, Y.; Wu, J.; Hsieh, C.-Y.; Hou, T. HiCLR: Knowledge-Induced Hierarchical Contrastive Learning with Retrosynthesis Prediction Yields a Reaction Foundation Model. *JACS Au* **2025**, *5*, 3140 – 3155.

(76) Choi, J.; Nam, G.; Choi, J.; Jung, Y. A perspective on foundation models in chemistry. *JACS Au* **2025**, *5*, 1499–1518.

(77) Zhang, S.-Q.; Xu, L.-C.; Li, S.-W.; Oliveira, J. C.; Li, X.; Ackermann, L.; Hong, X. Bridging chemical knowledge and machine learning for performance prediction of organic synthesis. *Chemistry–A European Journal* **2023**, *29*, e202202834.

(78) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science* **2019**, *5*, 1572–1583.

(79) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS central science* **2018**, *4*, 1465–1476.

(80) Liu, G.; Li, J.; Zhao, Z.; Inanc, E.; Maziarz, K.; Torres, J. G.; Satorras, V. G.; Ueda, S.; Bishop, C. M.; Segler, M. A Scientific Reasoning Model for Organic Synthesis Procedure Generation. 2025; https://arxiv.org/abs/2512.13668.

(81) Li, H.; Sarkar, S.; Lu, W.; Loftus, P. O.; Qiu, T.; Shee, Y.; Cuomo, A. E.; Webster, J.-P.; Kelly, H. R.; Manee, V.; Sreekumar, S.; Buono, F. G.; Crabtree, R. H.; Newhouse, T. R.; Batista, V. S. Collective intelligence for AI-assisted chemical synthesis. *Nature* **2026**,

(82) Liu, S.; Nie, W.; Wang, C.; Lu, J.; Qiao, Z.; Liu, L.; Tang, J.; Xiao, C.; Anandkumar, A. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence* **2023**, *5*, 1447–1457.

(83) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-mol: A universal 3d molecular representation learning framework. The eleventh international conference on learning representations. 2023.

(84) Szymanski, N. J.; Rendy, B.; Fei, Y.; others An Autonomous Laboratory for the Accelerated Synthesis of Inorganic Materials. *Nature* **2023**, *624*, 86–91.

(85) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *Journal of chemical information and modeling* **2020**,

(86) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009**, *1*, 8.

(87) Neeser, R. M.; Correia, B.; Schwaller, P. FSscore: A Machine Learning-based Synthetic Feasibility Score Leveraging Human Expertise. 2024; https://arxiv.org/abs/2312.12737.

(88) Li, B.; Chen, H. Prediction of Compound Synthesis Accessibility Based on Reaction Knowledge Graph. *Molecules* **2021**, *27*.

(89) Liu, S.; Zhang, D.; Tu, Z.; Dai, H.; Liu, P. Evaluating Molecule Synthesizability via Retrosynthetic Planning and Reaction Prediction. 2025; https://arxiv.org/abs/2411.08306.

(90) Flamm, C.; Merkle, D.; Stadler, P. F. Assembly in Directed Hypergraphs. 2025; https://arxiv.org/abs/2505.22826.

(91) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling* **2018**, *58*, 252–261, PMID: 29309147.

(92) Parrot, M.; Tajmouati, H.; da Silva, V. B. R.; Atwood, B. R.; Fourcade, R.; Gaston-Mathé, Y.; Huu, N. D.; Perron, Q. Integrating synthetic accessibility with AI-based generative drug design. *Journal of Cheminformatics* **2021**, *15*.

(93) Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics* **2020**, *12*, 35.

(94) Skoraczyński, G.; Kitlas, M.; Miasojedow, B.; Gambin, A. Critical assessment of synthetic accessibility scores in computer-assisted synthesis planning. *Journal of Cheminformatics* **2023**, *15*.

(95) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science* **2020**, *12*, 3339 – 3349.

(96) Chen, S.; Jung, Y. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *Journal of Cheminformatics* **2024**, *16*.

(97) Calvi, A.; Gaudin, T.; Miketa, D.; Sydow, D.; Wilbraham, L. Leap: molecular synthesisability scoring with intermediates. 2024; https://arxiv.org/abs/2403.13005.

(98) Hassen, A. K.; Šícho, M.; van Aalst, Y. J.; Huizenga, M. C. W.; Reynolds, D. N. R.; Luukkonen, S.; Bernatavicius, A.; Clevert, D.-A.; Janssen, A. P. A.; van Westen, G. J. P.; Preuss, M. Generate what you can make: achieving in-house synthesizability with readily available resources in de novo drug design. *Journal of Cheminformatics* **2025**, *17*.

(99) Guo, J.; Schwaller, P. Directly optimizing for synthesizability in generative molecular design using retrosynthesis models. *Chemical Science* **2024**, *16*, 6943 – 6956.

(100) Karwowski, J.; Hayman, O.; Bai, X.; Kiendlhofer, K.; Griffin, C.; Skalse, J. Goodhart's Law in Reinforcement Learning. 2023; https://arxiv.org/abs/2310.09144.

(101) Gao, S.; Zhou, X.; Liang, L.; Lin, J. RetroScore: graph edit distance-guided retrosynthesis for accessibility scoring with route metrics. *Journal of Cheminformatics* **2025**, *18*.

(102) Seo, S.; Kim, M.; Shen, T.; Ester, M.; Park, J.; Ahn, S.; Kim, W. Y. Generative Flows on Synthetic Pathway for Drug Design. 2025; https://arxiv.org/abs/2410.04542.

(103) Koziarski, M.; Rekesh, A.; Shevchuk, D.; van der Sloot, A.; Gaiński, P.; Bengio, Y.; Liu, C.-H.; Tyers, M.; Batey, R. A. RGFN: Synthesizable Molecular Generation Using GFlowNets. 2024; https://arxiv.org/abs/2406.08506.

(104) Gao, W.; Luo, S.; Coley, C. W. Generative AI for navigating synthesizable chemical space. *Proceedings of the National Academy of Sciences of the United States of America* **2025**, *122*.

(105) Jocys, Z.; Zhu, Z.; Willems, H. M. G.; Farrahi, K. SynthFormer: Equivariant Pharmacophore-based Generation of Synthesizable Molecules for Ligand-Based Drug Design. 2025; https://arxiv.org/abs/2410.02718.

(106) Vleduts, G. Concerning one system of classification and codification of organic reactions. *Information Storage and Retrieval* **1963**, *1*, 117–146.

(107) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science* **1969**, *166*, 178–192.

(108) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *Journal of Chemical Information and Modeling* **2019**, *59*, 2529–2537.

(109) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de novo Design of Bioactive Compounds. *PLOS Computational Biology* **2012**, *8*, 1–12.

(110) Button, A.; Merk, D.; Hiss, J. A.; Schneider, G. Automated de novo molecular design by hybrid machine intelligence and rule-driven chemical synthesis. *Nature Machine Intelligence* **2019**, *1*, 307–315.

(111) Gao, W.; Mercado, R.; Coley, C. W. Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. 2022; https://arxiv.org/abs/2110.06389.

(112) Gao, W.; Luo, S.; Coley, C. W. Generative Artificial Intelligence for Navigating Synthesizable Chemical Space. 2024; https://arxiv.org/abs/2410.03494.

(113) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, eabe4166.

(114) Kogej, T.; Kannas, C.; Genheden, S.; Caldeweyher, E.; Kabeshov, M. SMARTS-RX: A SMARTS-Based Representation of Chemical Functions for reactivity analysis. *ChemRxiv* **2025**, *2025*.

(115) Ahrendt, K. A.; Borths, C. J.; MacMillan, D. W. C. New Strategies for Organic Catalysis: The First Highly Enantioselective Organocatalytic Diels-Alder Reaction. *Journal of the American Chemical Society* **2000**, *122*, 4243–4244.

(116) Morgunov, A.; Batista, V. S. Procrustean Bed for AI-Driven Retrosynthesis: A Unified Framework for Reproducible Evaluation. 2025; https://arxiv.org/abs/2512.07079.

(117) Lowe, D. Chemical reactions from US patents (1976-Sep2016). 2017; https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.

(118) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics* **2020**, *12*, 70.

(119) Yu, Y.; Wei, Y.; Kuang, K.; Huang, Z.; Yao, H.; Wu, F. GRASP: Navigating Retrosynthetic Planning with Goal-driven Policy. Advances in Neural Information Processing Systems. 2022; pp 10257–10268.

(120) Xie, S.; Yan, R.; Han, P.; Xia, Y.; Wu, L.; Guo, C.; Yang, B.; Qin, T. RetroGraph: Retrosynthetic Planning with Graph Search. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2022; pp 2120–2129.

(121) Zhao, D.; Tu, S.; Xu, L. Efficient retrosynthetic planning with MCTS exploration enhanced A* search. *Communications Chemistry* **2024**, *7*.

(122) Roh, J.; Joung, J. F.; Yu, K.; Tu, Z.; Bartholomew, G. L.; Santiago-Reyes, O. A.; Fong, M. H.; Sarpong, R.; Reisman, S. E.; Coley, C. W. Higher-level Strategies for Computer-Aided Retrosynthesis. *ChemRxiv preprint* **2025**,

(123) Akhmetshin, T.; Zankov, D.; Gantzer, P.; Babadeev, D.; Pinigina, A.; Madzhidov, T.; Varnek, A. SynPlanner: An End-to-End Tool for Synthesis Planning. *Journal of Chemical Information and Modeling* **2025**, *65*, 15–21.

(124) Wang, M.; Montana, G. Retrosynthesis Planning via Worst-path Policy Optimisation in Tree-structured MDPs. 2025; https://arxiv.org/abs/2509.10504.

(125) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic retrosynthetic route planning using template-free models. *Chemical Science* **2020**, *11*, 3355 – 3364.

(126) Sun, K.; Bagni, D.; Cavanagh, J. M.; Wang, Y.; Sawyer, J. M.; Zhou, B.; Gritsevskiy, A.; Zhang, O.; Head-Gordon, T. SynLlama: Generating Synthesizable Molecules and Their Analogs with Large Language Models. *ACS Central Science* **2025**, *11*, 2108–2120.

(127) Granqvist, E.; Mercado, R.; Genheden, S. Retrosynformer: planning multi-step chemical synthesis routes via a decision transformer. *Digital Discovery* **2026**, *5*, 348–362.

(128) Liu, G.; Sun, M.; Matusik, W.; Jiang, M.; Chen, J. Multimodal Large Language Models for Inverse Molecular Design with Retrosynthetic Planning. 2024; https://arxiv.org/abs/2410.04223.

(129) Yu, K.; Roh, J.; Li, Z.; Gao, W.; Wang, R.; Coley, C. W. Double-Ended Synthesis Planning with Goal-Constrained Bidirectional Search. 2024; https://arxiv.org/abs/2407.06334.

(130) Maziarz, K.; Liu, G.; Misztela, H.; Tripp, A.; Li, J.; Kornev, A.; Gaiński, P.; Hoefling, H.; Fortunato, M.; Gupta, R.; Segler, M. Chemist-aligned retrosynthesis by ensembling diverse inductive bias models. 2025; https://arxiv.org/abs/2412.05269.

(131) Baker, F. N.; Adu-Ampratwum, D.; Averly, R.; Yu, B.; Sun, H.; Ning, X. LARC: Towards Human-level Constrained Retrosynthesis Planning through an Agentic Framework. 2025; https://arxiv.org/abs/2508.11860.

(132) Song, X.; Pan, X.; Zhao, X.; Ye, H.; Zhang, S.; Tang, J.; Yu, T. AOT*: Efficient Synthesis Planning via LLM-Empowered AND-OR Tree Search. 2025; https://arxiv.org/abs/2509.20988.

(133) Xuan-Vu, N.; Armstrong, D. P.; Jončev, Z.; Schwaller, P. TempRe: Template generation for single and direct multi-step retrosynthesis. 2025; https://arxiv.org/abs/2507.21762.

(134) Todd, M. Computer-Aided Organic Synthesis. *Chemical Society reviews* **2005**, *34*, 247–66.

(135) Gaiński, P.; Koziarski, M.; Maziarz, K.; Segler, M.; Tabor, J.; Śmieja, M. RetroGFN: Diverse and Feasible Retrosynthesis using GFlowNets. 2025; https://arxiv.org/abs/2406.18739.

(136) Wang, X.; Qian, Y.; Gao, H.; Coley, C. W.; Mo, Y.; Barzilay, R.; Jensen, K. F. Towards efficient discovery of green synthetic pathways with Monte Carlo tree search and reinforcement learning. *Chem. Sci.* **2020**, *11*, 10959–10972.

(137) Tripp, A.; Maziarz, K.; Lewis, S.; Segler, M.; Hernández-Lobato, J. M. Retro-fallback: retrosynthetic planning in an uncertain world. 2024; https://arxiv.org/abs/2310.09270.

(138) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Central Science* **2019**, *5*, 970–981.

(139) Roucairol, M.; Cazenave, T. Comparing search algorithms on the retrosynthesis problem. *Molecular Informatics* **2024**, *43*.

(140) Gajewska, E. P.; Szymkuć, S.; Dittwald, P.; Startek, M.; Popik, O.; Mlynarski, J.; Grzybowski, B. A. Algorithmic Discovery of Tactical Combinations for Advanced Organic Syntheses. *Chem* **2020**, *6*, 280–293.

(141) Ishida, S.; Terayama, K.; Kojima, R.; Takasu, K.; Okuno, Y. AI-Driven Synthetic Route Design Incorporated with Retrosynthesis Knowledge. *Journal of Chemical Information and Modeling* **2022**, *62*, 1357–1367.

(142) Westerlund, A. M.; Saigiridharan, L.; Genheden, S. Human-guided synthesis planning via prompting. *Chemical Science* **2025**, *16*, 14655 – 14667.

(143) Picazo, P. I.; Voronov, A.; Genheden, S.; Westerlund, A. M. Joint synthesis planning by leveraging common intermediates. *ChemRxiv* **2026**, *2026*.

(144) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry – A European Journal* **2017**, *23*, 5966–5971.

(145) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. Advances in Neural Information Processing Systems. 2019; pp 8870–8880.

(146) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metabolic Engineering* **2018**, *45*, 158–170.

(147) Chen, S.; Jung, Y. Deep Retrosynthetic Reaction Prediction using Local Reactivity and Global Attention. *JACS Au* **2021**, *1*, 1612–1620.

(148) Sacha, M.; Błaż, M.; Byrski, P.; Dąbrowski-Tumański, P.; Chromiński, M.; Loska, R.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. *Journal of Chemical Information and Modeling* **2021**, *61*, 3273–3284.

(149) Yan, C.; Ding, Q.; Zhao, P.; Zheng, S.; Yang, J.; Yu, Y.; Huang, J. RetroXpert: Decompose Retrosynthesis Prediction like a Chemist. 2020.

(150) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learning Graph Models for Retrosynthesis Prediction. 2021; https://arxiv.org/abs/2006.07038.

(151) Zeng, T.; Jin, Z.; Zheng, S.; Yu, T.; Wu, R. Developing BioNavi for Hybrid Retrosynthesis Planning. *JACS Au* **2024**, *4*, 2492–2502.

(152) Saigiridharan, L.; Hassen, A. K.; Lai, H.; Torren-Peraire, P.; Engkvist, O.; Genheden, S. AiZynthFinder 4.0: developments based on learnings from 3 years of industrial application. *Journal of Cheminformatics* **2024**, *16*.

(153) Shee, Y.; Li, H.; Zhang, P.; Nikolic, A. M.; Lu, W.; Kelly, H. R.; Manee, V.; Sreekumar, S.; Buono, F. G.; Song, J. J.; Newhouse, T. R.; Batista, V. S. Site-specific template generative approach for retrosynthetic planning. *Nature Communications* **2024**, *15*, 7818.

(154) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5*, 1572–1583.

(155) Li, J.; Fang, L.; Lou, J.-G. RetroRanker: leveraging reaction changes to improve retrosynthesis prediction through re-ranking. *Journal of Cheminformatics* **2023**, *15*.

(156) Maziarz, K.; Tripp, A.; Liu, G.; Stanley, M.; Xie, S.; Gaiński, P.; Seidl, P.; Segler, M. H. S. Re-evaluating retrosynthesis algorithms with Syntheseus. *Faraday Discuss.* **2025**, *256*, 568–586.

(157) Hassen, A. K.; Torren-Peraire, P.; Genheden, S.; Verhoeven, J.; Preuss, M.; Tetko, I. Mind the Retrosynthesis Gap: Bridging the divide between Single-step and Multi-step Retrosynthesis Prediction. 2022; https://arxiv.org/abs/2212.11809.

(158) Andronov, M.; Andronova, N.; Wand, M.; Schmidhuber, J.; Clevert, D.-A. Fast and scalable retrosynthetic planning with a transformer neural network and speculative beam search. 2025; https://arxiv.org/abs/2508.01459.

(159) Zipoli, F.; Baldassari, C.; Manica, M.; Born, J.; Laino, T. Growing strings in a chemical reaction space for searching retrosynthesis pathways. *npj Computational Materials* **2024**, *10*, 1–14.

(160) Westerlund, A. M.; Sigmund, L. M.; Kannas, C.; Genheden, S.; Kabeshov, M. Toward lab-ready AI synthesis plans with protection strategies and route scoring. *ChemRxiv* **2025**, *2025*.

(161) Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. 2017; https://arxiv.org/abs/1701.06538.

(162) Enamine US Inc. Enamine. Building block catalogs. https://enamine.net/.

(163) Wang, H.; Guo, J.; Kong, L.; Ramprasad, R.; Schwaller, P.; Du, Y.; Zhang, C. LLM-Augmented Chemical Synthesis and Design Decision Programs. 2025; https://arxiv.org/abs/2505.07027.

(164) eMolecules Inc. eMolecules Catalog. https://www.emolecules.com/.

(165) Liu, G.; Xue, D.; Xie, S.; Xia, Y.; Tripp, A.; Maziarz, K.; Segler, M. H. S.; Qin, T.; Zhang, Z.; Liu, T.-Y. Retrosynthetic Planning with Dual Value Networks. International Conference on Machine Learning. 2023.

(166) Blackshaw, T. M.; Davies, J. C.; Spoerer, K. T.; Hirst, J. D. Enhancing Monte Carlo Tree Search for Retrosynthesis. *Journal of Chemical Information and Modeling* **2025**, *65*, 6537 – 6546.

(167) Guo, J.; Yu, C.; Li, K.; Zhang, Y.; Wang, G.; Li, S.; Dong, H. Retrosynthesis Zero: Self-Improving Global Synthesis Planning Using Reinforcement Learning. *Journal of chemical theory and computation* **2024**,

(168) Torren-Peraire, P.; Hassen, A. K.; Genheden, S.; Verhoeven, J.; Clevert, D.-A.; Preuss, M.; Tetko, I. V. Models Matter: the impact of single-step retrosynthesis on synthesis planning. *Digital Discovery* **2024**, *3*, 558–572.

(169) Hong, S.; Zhuo, H. H.; Jin, K.; Shao, G.; Zhou, Z. Retrosynthetic planning with experience-guided Monte Carlo tree search. *Communications Chemistry* **2023**, *6*, 120.

(170) Kim, J.; Ahn, S.; Lee, H.; Shin, J. Self-Improved Retrosynthetic Planning. 2021; https://arxiv.org/abs/2106.04880.

(171) Han, P.; Zhao, P.; Lu, C.; Huang, J.; Wu, J.; Shang, S.; Yao, B.; Zhang, X. GNN-Retro: Retrosynthetic Planning with Graph Neural Networks. AAAI Conference on Artificial Intelligence. 2022.

(172) Zhang, X.; Lin, H.; Zhang, M.; Zhou, Y.; Ma, J. A data-driven group retrosynthesis planning model inspired by neurosymbolic programming. *Nature Communications* **2025**, *16*, 192.

(173) Choe, J.; Kim, H.; Chok, Y. T.; Gim, M.; Kang, J. Retrosynthetic crosstalk between single-step reaction and multi-step planning. *Journal of Cheminformatics* **2025**, *17*.

(174) Sun, X.; Liu, K.; Lin, Y.; Wu, L.; Xing, H.; Gao, M.; Liu, J.; Tan, S.; Ni, Z.; Han, Q.; Wu, J.; Fan, J. ChemiRise: a data-driven retrosynthesis engine. 2021; https://arxiv.org/abs/2108.04682.

(175) Zhang, Y.; He, X.; Gao, S.; Zhou, A.; Hao, H. Evolutionary Retrosynthetic Route Planning [Research Frontier]. *IEEE Computational Intelligence Magazine* **2023**, *19*, 58–72.

(176) Mrugalla, F.; Franz, C.; Alber, Y.; Mogk, G.; Villalba, M.; Mrziglod, T.; Schewior, K. Generating diversity and securing completeness in algorithmic retrosynthesis. *Journal of Cheminformatics* **2025**, *17*.

(177) Kreutter, D.; Reymond, J.-L. Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chemical Science* **2023**, *14*, 9959 – 9969.

(178) Bradshaw, J.; Zhang, A.; Mahjour, B.; Graff, D. E.; Segler, M. H. S.; Coley, C. W. Challenging Reaction Prediction Models to Generalize to Novel Chemistry. *ACS Central Science* **2025**, *11*, 539–549.

(179) Tran, S. B. A.; Roh, J.; Coley, C. W. Quantifying the Failure Modes of Current One-step Retrosynthesis Models. *ChemRxiv* **2026**, *2026*.

(180) Guo, Y.; Le, T. H. D.; Genheden, S.; Mijangos, M. V.; Voinarvoska, V.; Bergonzini, G.; Engkvist, O.; Kabeshov, M.; Kaski, S. An Expert-Augmented Deep Learning Approach for Interpretable Synthesis Route Evaluation. *ChemRxiv* **2026**, *2026*.

(181) Zagribelnyy, B.; Ilin, I.; Kuznetsov, M.; Bondarev, N.; Schutski, R.; MacDougall, T.; Shayakhmetov, R.; Miftakhutdinov, Z.; Mizera, M.; Aladinskiy, V.; Aliper, A.; Zhavoronkov, A. When Single Answer Is Not Enough: Rethinking Single-Step Retrosynthesis Benchmarks for LLMs. 2026; https://arxiv.org/abs/2602.03554.

(182) Wigh, D. S.; Arrowsmith, J.; Pomberger, A.; Felton, K. C.; Lapkin, A. A. Orderly: data sets and benchmarks for chemical reaction data. *Journal of Chemical Information and Modeling* **2024**, *64*, 3790–3798.

(183) Chen, S.; Babazade, R.; Kim, T.; Han, S.; Jung, Y. A large-scale reaction dataset of mechanistic pathways of organic reactions. *Scientific Data* **2024**, *11*, 863.

(184) Coley, C. W.; Thomas III, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; others A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, eaax1566.

(185) Chen, L.-Y.; Li, Y.-P. AutoTemplate: enhancing chemical reaction datasets for machine learning applications in organic chemistry. *Journal of Cheminformatics* **2024**, *16*, 74.

(186) Ramos, M. C.; Collison, C. J.; White, A. D. A Review of Large Language Models and Autonomous Agents in Chemistry. 2024; https://arxiv.org/abs/2407.01603.

(187) Nascimento, C. M. C.; Pimentel, A. S. Do Large Language Models Understand Chemistry? A Conversation with ChatGPT. *Journal of Chemical Information and Modeling* **2023**, *63*, 1649–1655.

(188) Zhang, D.; Liu, W.; Tan, Q.; Chen, J.; Yan, H.; Yan, Y.; Li, J.; Huang, W.; Yue, X.; Ouyang, W.; Zhou, D.; Zhang, S.; Su, M.; Zhong, H.-S.; Li, Y. ChemLLM: A Chemical Large Language Model. 2024; https://arxiv.org/abs/2402.06852.

(189) Liu, H.; Yin, H.; Luo, Z.; Wang, X. Integrating Chemistry Knowledge in Large Language Models via Prompt Engineering. 2024; https://arxiv.org/abs/2404.14467.

(190) Mallouhy, R. E.; Alqahtani, H.; Lardhi, S. AI-Powered Standard Operating Procedure Generation and Optimization Using Large Language Models for Chemical Laboratory Applications. *IEEE Access* **2026**, *14*, 19383–19395.

(191) Yoshikawa, N.; Skreta, M.; Darvish, K.; Arellano-Rubach, S.; Ji, Z.; Kristensen, L. B.; Li, A. Z.; Zhao, Y.; Xu, H.; Kuramshin, A.; Aspuru-Guzik, A.; Shkurti, F.; Garg, A. Large language models for chemistry robotics. *Autonomous Robots* **2023**, *47*, 1057–1086.

(192) He, J.; Lai, H.; Saigiridharan, L.; Ghiandoni, G. M.; Jenei, K.; Gokalp, U.; Nuković, A.; Engkvist, O.; Janet, J. P.; Genheden, S. Democratising real-world drug discovery through agentic AI. *Drug Discovery Today* **2026**, *31*, 104605.

(193) Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578.

(194) Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, *6*, 525–535.

(195) Zhang, Y.; Yu, R.; Tian, J.; Zhu, F.; Liu, J.; Yang, X.; Jin, Y.; Xu, Y. ChemActor: Enhancing Automated Extraction of Chemical Synthesis Actions with LLM-Generated Data. 2025; https://arxiv.org/abs/2506.23520.

(196) Liu, Z.; Shi, Y.; Zhang, A.; Li, S.; Zhang, E.; Wang, X.; Kawaguchi, K.; Chua, T.-S. ReactXT: Understanding Molecular "Reaction-ship" via Reaction-Contextualized Molecule-Text Pretraining. 2024; https://arxiv.org/abs/2405.14225.

(197) Chen, Z.; Fang, Z.; Tian, W.; Long, Z.; Sun, C.; Chen, Y.; Yuan, H.; Li, H.; Lan, M. ReactGPT: Understanding of Chemical Reactions via In-Context Tuning. *Proceedings of the AAAI Conference on Artificial Intelligence* **2025**, *39*, 84–92.

(198) Zhang, Y.; Han, Y.; Chen, S.; Yu, R.; Zhao, X.; Liu, X.; Zeng, K.; Yu, M.; Tian, J.; Zhu, F.; Yang, X.; Jin, Y.; Xu, Y. Large language models to accelerate organic chemistry synthesis. *Nature Machine Intelligence* **2025**, *7*, 1010–1022.

(199) Ruan, Y.; Lu, C.; Xu, N.; He, Y.; Chen, Y.; Zhang, J.; Xuan, J.; Pan, J.; Fang, Q.; Gao, H.; Shen, X.; Ye, N.; Zhang, Q.; Mo, Y. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications* **2024**, *15*, 10160.

(200) Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; Jégou, H. The Faiss library. 2025; https://arxiv.org/abs/2401.08281.

(201) Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. 2021; https://arxiv.org/abs/2106.09685.

(202) Tu, Z.; Choure, S. J.; Fong, M. H.; Roh, J.; Levin, I.; Yu, K.; Joung, J. F.; Morgan, N.; Li, S.-C.; Sun, X.; Lin, H.; Murnin, M.; Liles, J. P.; Struble, T. J.; Fortunato, M. E.; Liu, M.; Green, W. H.; Jensen, K. F.; Coley, C. W. ASKCOS: Open-Source, Data-Driven Synthesis Planning. *Accounts of Chemical Research* **2025**, *58*, 1764–1775.

(203) Sutton, R. S. 2019; http://www.incompleteideas.net/IncIdeas/BitterLesson.html.