
Procrustean Bed for AI-Driven Retrosynthesis: A Unified Framework for Reproducible Evaluation

Anton Morgunov*
Yale University
anton@ischemist.com

Victor S. Batista
Yale University
victor.batista@yale.edu

Abstract

Progress in computer-aided synthesis planning (CASP) is obscured by the lack of standardized evaluation infrastructure and the reliance on metrics that prioritize topological completion over chemical validity. We introduce RetroCast, a unified evaluation suite that standardizes heterogeneous model outputs into a common schema to enable statistically rigorous, apples-to-apples comparison. The framework includes a reproducible benchmarking pipeline with stratified sampling and bootstrapped confidence intervals, accompanied by SynthArena (syntharena.ischemist.com), an interactive platform for qualitative route inspection. We utilize this infrastructure to evaluate leading search-based and sequence-based algorithms on a new suite of standardized benchmarks. Our analysis reveals a divergence between "solvability" (stock-termination rate) and route quality; high solvability scores often mask chemical invalidity or fail to correlate with the reproduction of experimental ground truths. Furthermore, we identify a "complexity cliff" in which search-based methods, despite high solvability rates, exhibit a sharp performance decay in reconstructing long-range synthetic plans compared to sequence-based approaches. We release the full framework, benchmark definitions, and a standardized database of model predictions to support transparent and reproducible development in the field

1 Introduction

We distinguish between two fundamental classes of scientific problems to which machine learning is applied: quantitative and structural. Quantitative problems, such as predicting drug toxicity, are defined by scalar targets and often constrained by data scarcity, analogous to early NLP challenges like sentiment analysis. In contrast, structural problems, like language modeling or protein folding, require generating complex objects governed by an underlying grammar. The most transformative successes of AI, from large language models [1–4] to AlphaFold [5–8], have occurred in these structural domains; foundation models trained on the structure of language, for example, now excel at sentiment analysis with little to no task-specific fine-tuning. We contend that mastery of structure is a prerequisite for solving downstream quantitative tasks. In organic chemistry, the paramount structural challenge is designing a valid synthetic pathway to a molecule of interest. This capability, retrosynthesis, is the key to unlocking critical quantitative problems like predicting a molecule's synthetic feasibility, a significant bottleneck in synthesis-aware virtual screening. Current accessibility heuristics, however, bypass the core structural challenge, relying on learned patterns that correlate with accessibility without ever generating the synthetic pathway itself. This, we argue, is a fundamental limitation: a model cannot judge the difficulty of a journey it cannot first articulate.

The dominant paradigm for computational retrosynthesis follows a two-part framework: a single-step model proposes disconnections, and a search algorithm explores the resulting pathway space [9].

*Work performed in adherence with isChemist Protocol

While both components have seen rapid progress [10–20], the field’s primary measure of success, traditionally called *solubility*, creates a disconnect between reported performance and practical utility. A route is deemed "solved" if all its terminal nodes exist in a predefined commercial stock, but this is a purely topological check that provides no guarantee of chemical validity for the intermediate steps. This represents a methodological departure from the field’s early best practices, which incorporated dedicated networks to filter infeasible reactions [15]. The field now operates on an implicit and unevaluated assumption that single-step predictors have learned all complex rules of chemical feasibility. Consequently, high scores can be achieved for routes containing chemically nonsensical steps, rewarding any topological path regardless of its plausibility. To avoid the misleading implication of "solving" a chemical problem, we will henceforth refer to this metric by a more precise term: the *Stock-Termination Rate (STR)*.

Attempts to address this validity gap with proxy metrics, such as forward-prediction confidence or round-trip accuracy [21–23], substitute direct validation with a reliance on auxiliary models, inheriting their biases without establishing a standardized measure of plausibility. Beyond these issues, the utility of STR is further undermined by a profound lack of standardization in its most critical component: the starting material stock. Our survey of prominent CASP tools reveals that the stock sets used for evaluation vary by over three orders of magnitude, from physically in-stock compounds to massive "make-on-demand" virtual libraries (Table S1). This thousand-fold disparity makes direct comparison of reported STR scores between models unreliable and can obscure the true signal of a model’s chemical intelligence.

The PaRoutes benchmark [24], the first large-scale dataset of experimental routes extracted from patent literature, sought to bridge this validity gap by measuring a model’s ability to reproduce known syntheses. It exposed a stark disparity: models reporting over 97% STR could find ground-truth routes with only 35–50% top-10 accuracy. This motivated new architectures, such as the sequence-to-sequence DirectMultiStep model, which demonstrated an inverse performance profile: substantially improved route reproduction at the cost of a slightly lower STR [25]. However, interpreting PaRoutes as a universal gold standard is challenging because its reference routes are constructed from reactions reported within single patents. The endpoints of these patent-derived syntheses are not necessarily commercial precursors; to quantify this, we compared the n5 evaluation stock against the ASKCOS Buyables set, a collection of compounds available for purchase from eMolecules, Sigma-Aldrich, Mcule, LabNetwork, and ChemBridge [26–28]. Of the 13,306 unique leaf molecules, only 7,513 (56%) are present in the Buyables set; as a result only 4,279 of the 10,000 reference routes in n5 have all their leaves within this realistic commercial stock. Optimizing for exact reproduction may therefore incentivize learning patent-specific patterns rather than the general principles of synthesis from commercial precursors.

More fundamentally, a focus on exact reproduction is inherently conservative, penalizing the discovery of novel yet plausible pathways. The field is therefore caught in a methodological bind: an STR metric that rewards novelty at the expense of plausibility, and a reproduction metric that ensures plausibility at the expense of novelty. This leaves no reliable way to distinguish a plausible new route from a chemically unfeasible artifact, creating a critical measurement gap where the most desirable outcome—a novel, valid synthesis—cannot be properly evaluated.

This methodological challenge is compounded by a practical infrastructure gap that blocks rigorous, large-scale comparison. Any attempt at a meta-analysis requires developing bespoke parsers for the fundamentally incompatible output formats of different CASP tools (Fig. 1). This heterogeneity is exacerbated by the high computational cost of generating predictions. This constraint is not hypothetical: a recent transformer-based model, for instance, was evaluated on only a 240-molecule subset of PaRoutes due to "high computing time"[21]. Faced with both unreliable metrics and these significant practical hurdles, the field’s leading groups often bypass large-scale automated evaluation entirely, resorting instead to small-scale, costly human expert validation for their final assessments[29]. Collectively, these issues make direct model comparison impractical, if not impossible, leaving the field without a clear way to measure progress.

To resolve these issues, we introduce a unified framework that transforms ad-hoc evaluation into a systematic, scalable, and community-driven effort. First, we present *RetroCast*, an open-source software package that provides both a universal translation layer for heterogeneous model outputs and an automated pipeline for performing statistically robust comparative analysis, all while ensuring auditable data provenance with cryptographic manifests. Second, using this framework, we perform

the first rigorous, apples-to-apples comparison of the field’s leading models on a suite of new, curated benchmarks designed to provide high diagnostic signal at a fraction of the computational cost. Our analysis exposes how high STR scores can mask chemically implausible predictions and reveals the divergent architectural signatures that emerge when models are evaluated with a more chemically meaningful, multi-ground-truth protocol. Finally, we release not only our tools and benchmarks but also the complete, standardized prediction database as a reusable community asset, accompanied by *SynthArena*, an interactive web platform for qualitative inspection of route predictions. This provides the shared infrastructure needed to catalyze a shift in the field’s evaluation criteria: from merely checking for stock termination to the far more meaningful challenge of generating chemically plausible synthetic plans.

2 Results

2.1 A Unified Framework for Reproducible Evaluation

To resolve the field’s infrastructure gap and enable rigorous comparison, we developed RetroCast, an open-source evaluation suite. Its foundation is a universal translation layer that addresses the heterogeneity of model outputs (Fig. 1). Using an adapter-based architecture, RetroCast parses native formats into a single, standardized interchange schema, providing the necessary lingua franca for any cross-model analysis. We provide ready-to-use adapters for a comprehensive suite of tools, including search-based planners (AiZynthFinder [30], Retro* [16], ASKCOS [26], Syntheseus [31], RetroChimera [29], DreamRetroEr [32], MultiStepTTL [21], SynPlanner [33]), sequence-based models (DirectMultiStep [25], SynLLaMa [23]), and standard dataset formats like PaRoutes [24]. Beyond simple translation, the software provides a complete, automated pipeline for executing robust evaluations. Given a set of standardized predictions and a benchmark definition, RetroCast calculates key metrics, including Stock-Termination Rate and Top-K accuracy, with bootstrapped 95% confidence intervals, and provides results stratified by route properties such as length and topology. To ensure all comparisons are auditable and reproducible, the entire workflow is accompanied by a system of cryptographic manifests; each stage of processing generates a manifest recording the SHA256 hashes of all inputs and outputs, creating a computationally verifiable chain of data provenance.

Building on this standardized data layer, we designed a principled evaluation protocol to address two distinct but often conflated goals: assessing a model’s immediate practical utility versus enabling fair algorithmic comparison. To this end, we introduce two evaluation tracks, each with a corresponding suite of curated benchmarks derived from the PaRoutes n5 dataset (Table S2). The **Chemist-Aligned (mkt-)** series is designed to answer the practical question, "Which model is most useful today?" Models in this track have no training data restrictions and are evaluated against the ASKCOS Buyables stock, a realistic set of 300k commercially available compounds. In contrast, the **Developer-Aligned (ref-)** series facilitates controlled comparison by asking, "Which algorithm is superior?" This track requires models to be trained on a common open-source dataset. To prevent data leakage, we propose that the entire PaRoutes n1 and n5 evaluation sets be excluded from any training corpus. Evaluation for this series uses the original patent-derived stocks from PaRoutes to isolate algorithmic performance from stock availability. All benchmarks were constructed using a stratified design based on route length and topology, providing high diagnostic signal at a fraction of the original’s computational cost (details in Sec. 4.1).

A critical ambiguity arises in the calculation of rank-dependent metrics like Top-K accuracy: which population of predictions should be ranked? The set of all raw model outputs, or only the subset that satisfies task-specific constraints? The choice can dramatically alter results and conclusions about model performance. Our framework resolves this by adopting a user-centric perspective: for a practicing chemist, a model’s output is only useful if it represents a valid, coherent answer to the query. A route that fails to meet basic structural or user-specified constraints is not, by definition, a solution. This philosophy leads to a clear, sequential filtering protocol that must be applied before any ranking is performed: 1) all raw outputs are first filtered for structural integrity (e.g., valid SMILES, forming a directed acyclic graph); 2) the structurally sound pool is then filtered against all explicit task constraints, such as stock-termination; 3) finally, Top-K accuracy is calculated on this fully validated set, strictly preserving the model’s original output order. This protocol ensures a model is judged on its ability to both produce and correctly rank valid solutions: the only relevant measure of performance from a user’s perspective. This provides a stable and extensible foundation. As the field

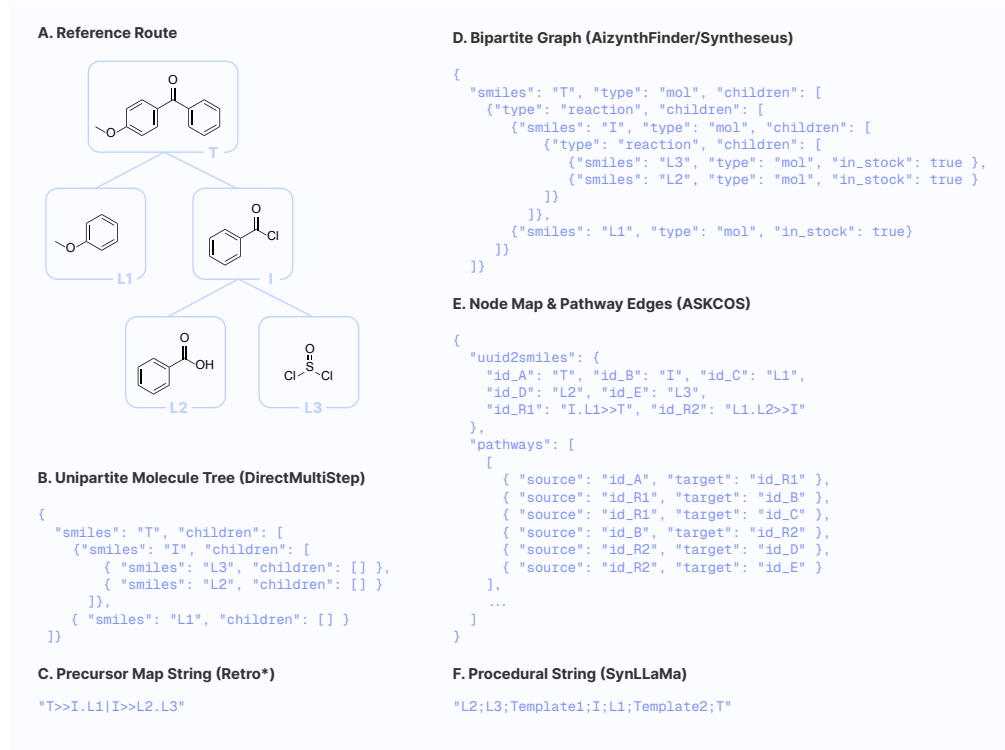


Figure 1: The babel of retrosynthesis formats. Illustration of five fundamentally incompatible output formats for a single reference synthetic route (A). Placeholders correspond to the target (T), an intermediate (I), and purchasable leaves (L1, L2, L3). Formats range from verbose, explicit graph structures (B, D, E) to concise, implicit string-based representations (C, F). (B) a simpler nested json of only molecule nodes, where reactions are implicit. (C) a declarative string mapping products to reactants. (D) a nested json where molecule nodes alternate with explicit reaction nodes. (E) a schema where a route is defined as a list of edges that reference a separate map of nodes. (F) a linear "recipe" string where the product of one step becomes an implicit reactant in the next. This heterogeneity necessitates a translation layer like RetroCast for any comparative analysis.

advances toward more complex queries, such as specifying starting materials [25, 34] or forbidden reactions, our user-centric maxim offers a consistent path forward: first, filter for all constraints; then, and only then, evaluate the quality and ranking of the valid solutions.

To complement aggregate statistics with qualitative insight, our framework includes SynthArena, an open-source platform for interactive route inspection. SynthArena ingests standardized outputs from RetroCast and provides an interface for side-by-side route comparison, difference overlays, and annotation of commercial availability for all leaf nodes. This capacity for direct inspection allowed us to discover that some routes, marked incorrect by Top-K accuracy, are in fact shorter sub-routes of the reference pathway that terminated at commercially available intermediates. This valid outcome that was incorrectly penalized by a strict single ground-truth metric directly motivated the development of our more chemically meaningful multi-ground-truth evaluation protocol. To serve the community, we host a public instance at syntharena.ischemist.com as a central "arena" and living leaderboard. As the platform is fully open-source, developers can also deploy it locally as a powerful tool for day-to-day model development, such as comparing checkpoints or diagnosing the effects of architectural changes. Our goal is to transform evaluation from a static, periodic exercise into a dynamic, ongoing process of collective error analysis and model improvement.

2.2 Stock-Termination Rate is a Misleading Signal of Chemical Validity

Our unified analysis of prominent open-source models on the public USPTO-190 benchmark (evaluated with ASKCOS Buyables as a stock set) reveals that a high Stock-Termination Rate (STR) can

be a misleading signal of chemical validity. While a conventional analysis would suggest a clear performance hierarchy, with the original Retro* model achieving a dominant 73.2% STR (Table 1), systematic qualitative inspection of the underlying routes reveals a critical flaw in the metric. Because STR validates only the commercial availability of a route’s terminal nodes, it provides no guarantee of chemical plausibility for the intermediate steps.

Model	STR (%)	Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
Retro* (High)	73.2 [66.8, 79.5]	10.0 [5.8, 14.2]	10.0 [5.8, 14.2]	35.3
Retro*	44.7 [37.9, 51.6]	7.9 [4.2, 12.1]	7.9 [4.2, 12.1]	11.0
AiZynF Retro* (High)	36.8 [30.0, 43.7]	0.5 [0.0, 1.6]	2.1 [0.5, 4.2]	133.0
AiZynF MCTS (High)	33.7 [26.8, 40.5]	1.6 [0.0, 3.7]	2.1 [0.5, 4.2]	41.8
AiZynF Retro*	32.1 [25.3, 38.4]	1.1 [0.0, 2.6]	2.1 [0.5, 4.2]	35.0
Syntheseus LocalRetro	33.7 [26.8, 40.5]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	18.4
DMS Explorer XL	29.5 [23.2, 35.8]	0.5 [0.0, 1.6]	1.1 [0.0, 2.6]	20.5
AiZynF MCTS	24.7 [18.4, 31.1]	1.1 [0.0, 2.6]	2.1 [0.5, 4.2]	11.04
ASKCOS	17.4 [12.1, 23.2]	0.5 [0.0, 1.6]	1.6 [0.0, 3.7]	30.2

Table 1: **High stock-termination scores on the USPTO-190 benchmark often mask underlying challenges in chemical validity.** Performance of major retrosynthesis models on the 190-target USPTO test set, evaluated using the ASKCOS Buyables stock. Metrics include stock termination rate (STR, the fraction of targets for which a route to purchasable starting materials was found) and Top-K accuracy (the fraction of targets for which a reference route was found). (High) denotes a 500-iteration search vs. the 100-iteration default. The evaluated Retro* implementation returns a single route, making its Top-1 and Top-10 accuracy values identical. Values in brackets indicate bootstrapped 95% confidence intervals. The full, interactive leaderboard for this benchmark is available on SynthArena: [link](#). See Sec. 4.3 for hardware specifications.

This measurement gap is exemplified by the analysis of target USPTO-082 (Fig. 2A-C). The top-performing model’s "solved" route is predicated on a chemically implausible seven-reactant combination. This is not a model hallucination but an exact reproduction of a flawed transformation present in the benchmark’s own reference route. The model is thus rewarded for accurately pattern-matching corrupted data, while newer models that avoid this nonsensical step are penalized, failing to achieve a "solved" status.

The failure to penalize invalid chemistry is a systemic issue, not an isolated artifact. Fig. 2D-H presents a catalog of similarly unsound transformations extracted from five other "solved" routes generated by the same model. These examples document apparent violations of fundamental chemical principles, including mass balance errors and chemically nonsensical conversions. By incentivizing the discovery of any topological path to a commercial stock, STR systematically fails to capture a model’s understanding of chemistry. Consequently, high STR scores alone—whether on this benchmark or any other—cannot be reliably interpreted as a signal of a model’s ability to generate chemically sound synthetic plans.

2.3 Multi-Ground-Truth Evaluation Reveals Divergent Architectural Signatures

The machine learning concept of a single "ground truth" is an oversimplification in organic synthesis, where multiple valid pathways to a target often exist. Current benchmarks, however, rigidly evaluate against a single patent-derived reference, incorrectly penalizing models for identifying shorter, more efficient syntheses that terminate at commercially available intermediates—a frequent occurrence that became apparent during qualitative inspection with SynthArena.

To create a more chemically meaningful benchmark, we expand the set of acceptable solutions. Our approach includes not only the full experimental sequence but also any of its constituent sub-routes that terminate exclusively in commercially available precursors (details Sec. 4.2.2). This represents a pragmatic first step toward a more comprehensive evaluation. While it cannot yet reward entirely novel valid pathways, it provides a principled way to expand the reference set without sacrificing the chemical plausibility guaranteed by the original experiment. We refer to this as a Multi-Ground-Truth (MGT) evaluation, reflecting this expansion of the target set.

We applied this MGT protocol to our Chemist-Aligned benchmarks, which are evaluated against the ASKCOS Buyables stock. On the mkt-cnv-160 benchmark of convergent routes, this reveals

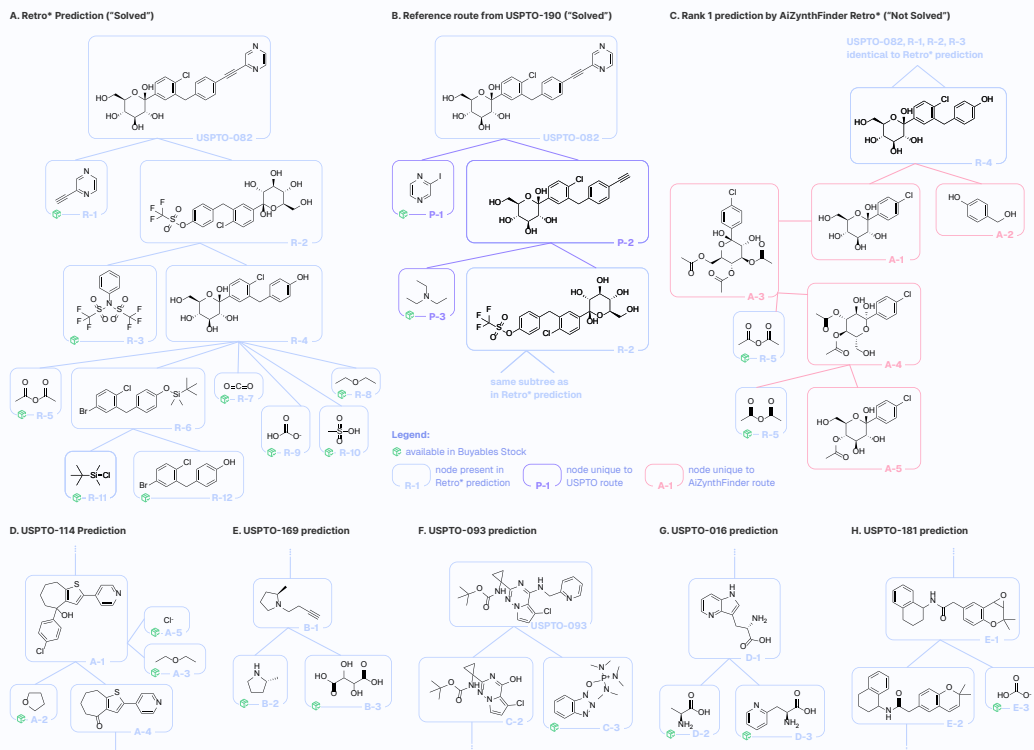


Figure 2: High stock-termination rate rewards chemically invalid routes. Analysis reveals how metrics blind to chemical validity can mislead. (A-C) The case of target USPTO-082. **A**, A "solved" route from the top-performing model hinges on a chemically implausible seven-reactant step. **B**, The official reference route contains the identical flawed transformation, showing the model's success is an artifact of pattern-matching corrupted data. **C**, A route from a newer model that avoids the nonsensical step is penalized for failing to find a "solved" path. (D-H) A catalog of chemical hallucinations from other "solved" routes, demonstrating the systemic nature of the issue. Violations include: **D**, mass balance error (un-sourced chloro-phenyl group); **E**, implausible transformation (tartaric acid as a propargyl source); **F**, mass balance error (un-sourced pyridylmethylamine); **G**, implausible reaction (amino acids to tryptophan core); **H**, unspecified reagent (epoxidation with carbonic acid). These cases show that naive stock termination rate fails to capture fundamental chemical principles. Interactive versions are on SynthArena: USPTO-082, USPTO-114, USPTO-169, USPTO-93, USPTO-16, USPTO-181.

two distinct and often opposing performance profiles (Table 2). Search-based models consistently achieve near-perfect stock-termination rates. This, however, is an intrinsic property of their design; the algorithm is engineered to explore until a stock set is reached, making high STR a satisfaction of a stopping condition rather than an independent measure of the route's quality. In contrast, the sequence-based DirectMultiStep model, for which stock enforcement is a post-processing step, exhibits the inverse profile: slightly lower STR but substantially higher accuracy in matching the structural patterns of known, plausible syntheses. This divergence underscores that stock termination and route reproduction are measuring fundamentally different capabilities.

This performance dichotomy is not an artifact of convergent topologies; an identical pattern is observed on the linear routes of the mkt-1in-500 benchmark (Table 3). The results consistently show that models optimized for topological search and models that learn holistic route structure are being driven toward different solutions, a critical distinction obscured by previous evaluation paradigms.

Model	STR (%)	Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
DMS Explorer XL	96.3 [93.1, 98.8]	33.8 [26.9, 41.3]	57.5 [50.0, 65.0]	17.9
AiZynF MCTS	98.1 [95.6, 100.0]	21.3 [15.0, 27.5]	41.3 [33.8, 48.8]	6.7
AiZynF MCTS (High)	99.4 [98.1, 100.0]	20.6 [14.4, 26.9]	38.8 [31.3, 46.3]	25.8
Retro* (High)	100.0 [100.0, 100.0]	33.8 [26.3, 41.3]	33.8 [26.3, 41.3]	1.2
Retro*	98.8 [96.9, 100.0]	33.1 [25.6, 40.6]	33.1 [25.6, 40.6]	0.8
AiZynF Retro*	98.8 [96.9, 100.0]	8.1 [4.4, 12.5]	27.5 [20.6, 34.4]	32.7
AiZynF Retro* (High)	99.4 [98.1, 100.0]	6.9 [3.1, 11.3]	23.8 [17.5, 30.6]	122.8
Syntheseus LocalRetro	95.6 [92.5, 98.8]	8.8 [4.4, 13.1]	20.6 [14.4, 26.9]	15.8
ASKCOS	93.1 [88.8, 96.9]	6.3 [3.1, 10.0]	19.4 [13.1, 25.6]	30.0

Table 2: **MGT evaluation reveals divergent architectural signatures on convergent routes.** Performance on the `mkt-cnv-160` benchmark, where the reference set is expanded to include all valid, commercially-terminated sub-routes. The results expose two distinct profiles: search-based models achieve near-perfect stock-termination rates (STR), while the sequence-based model shows substantially higher route-matching accuracy. The evaluated Retro* implementation returns a single route, making its Top-1 and Top-10 accuracy values identical. Brackets indicate 95% confidence intervals. The full, interactive leaderboard is available on SynthArena: [link](#)

Model	STR (%)	Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
DMS Explorer XL	97.2 [95.6, 98.6]	31.4 [27.4, 35.6]	55.4 [51.0, 59.8]	14.5
AiZynF MCTS	97.6 [96.2, 98.8]	17.8 [14.4, 21.2]	35.8 [31.6, 40.0]	6.1
AiZynF MCTS (High)	98.4 [97.2, 99.4]	17.4 [14.2, 20.8]	33.4 [29.2, 37.6]	20.8
AiZynF Retro*	98.0 [96.6, 99.2]	10.4 [7.8, 13.2]	28.6 [24.6, 32.6]	26.3
AiZynF Retro* (High)	99.0 [98.0, 99.8]	9.2 [6.8, 11.8]	25.4 [21.6, 29.2]	103.1
Retro* (High)	99.8 [99.4, 100.0]	22.2 [18.6, 25.8]	22.2 [18.6, 25.8]	0.6
Retro*	99.8 [99.4, 100.0]	22.2 [18.6, 25.8]	22.2 [18.6, 25.8]	0.5
Syntheseus LocalRetro	94.0 [91.8, 96.0]	8.6 [6.2, 11.2]	18.8 [15.4, 22.4]	11.9
ASKCOS	94.4 [92.4, 96.4]	6.2 [4.2, 8.4]	16.2 [13.2, 19.6]	29.5

Table 3: **The performance dichotomy persists on linear routes.** Performance on the `mkt-lin-500` benchmark using MGT evaluation. The results confirm the findings from the convergent set, with models optimized for topological search (high STR) and models that learn holistic route structure (high accuracy) exhibiting inverse performance profiles. This demonstrates the robustness of the observation. The evaluated Retro* implementation returns a single route, making its Top-1 and Top-10 accuracy values identical. Brackets indicate 95% confidence intervals. The full, interactive leaderboard is available on SynthArena: [link](#)

Model	Length 2	Length 3	Length 4	Length 5
AiZynF MCTS	75.0 [62.5, 87.5]	50.0 [35.0, 65.0]	25.0 [12.5, 40.0]	15.0 [5.0, 27.5]
AiZynF MCTS (High)	80.0 [67.5, 92.5]	37.5 [22.5, 52.5]	25.0 [12.5, 40.0]	12.5 [2.5, 22.5]
AiZynF Retro*	65.0 [50.0, 80.0]	17.5 [7.5, 30.0]	12.5 [2.5, 22.5]	15.0 [5.0, 27.5]
AiZynF Retro* (High)	50.0 [35.0, 65.0]	20.0 [7.5, 32.5]	12.5 [2.5, 25.0]	12.5 [2.5, 22.5]
ASKCOS	37.5 [22.5, 52.5]	20.0 [7.5, 32.5]	5.0 [0.0, 12.5]	15.0 [5.0, 27.5]
DMS Explorer XL	75.0 [60.0, 87.5]	57.5 [42.5, 72.5]	45.0 [30.0, 60.0]	52.5 [37.5, 67.5]
Retro*	50.0 [35.0, 65.0]	37.5 [22.5, 52.5]	25.0 [12.5, 40.0]	20.0 [7.5, 32.5]
Retro* (High)	50.0 [35.0, 65.0]	37.5 [22.5, 52.5]	27.5 [15.0, 42.5]	20.0 [7.5, 32.5]
Syntheseus LocalRetro	40.0 [25.0, 55.0]	22.5 [10.0, 35.0]	10.0 [2.5, 20.0]	10.0 [2.5, 20.0]

Table 4: **Accuracy decays with complexity, exposing a measurement crisis.** Top-10 route-matching accuracy on the `mkt-cnv-160` benchmark, stratified by reference route length. Search-based models excel on short routes but their performance collapses as complexity increases, while the sequence-based model (DMS) remains more robust. This sharp decay reveals the fundamental limit of reference-based evaluation: it penalizes the discovery of novel routes as failures, making it impossible to distinguish a failed search from a creative success. The evaluated Retro* implementation returns a single route, making its Top-1 and Top-10 accuracy values identical. Brackets indicate 95% confidence intervals. Detailed statistics are available for interactive exploration on SynthArena: [link](#)

2.4 Stratified Analysis Uncovers a "Complexity Cliff"

Stratifying the analysis by route length exposes the architectural signatures that aggregate metrics obscure. On the mkt-cnv-160 benchmark, search-based models excel at matching short reference routes, but their accuracy decays sharply as synthetic complexity increases (Table 4). In contrast, the sequence-based DirectMultiStep model maintains more consistent performance on longer routes. This trend culminates in a "complexity cliff" on the ref-1ng-84 benchmark, a stress test of exclusively long routes (lengths 8-10), where the route-matching accuracy of all evaluated search-based models collapses to near-zero (Table S7).

This result presents two non-exclusive interpretations. The first is algorithmic: iterative tree search may be inherently ill-suited for the combinatorial complexity of long-range planning. The second, however, is a measurement limitation: a search algorithm succeeding in finding a novel, plausible route is penalized as a failure by any reference-based metric. This ambiguity exposes the ultimate limitation of the current evaluation paradigm. It is fundamentally incapable of distinguishing a failed search from a creative success, creating a critical measurement gap where the field’s most important goal—the discovery of novel, valid syntheses—cannot be rewarded.

2.5 The Cost-Performance Frontier in Synthesis Planning

To provide a practical dimension to our analysis, we augment accuracy with computational cost, measured in USD based on cloud compute pricing (details in Sec. 4.3). The resulting Pareto plot for the mkt-cnv-160 benchmark establishes a cost-performance frontier, quantifying the trade-off between predictive accuracy and economic cost (Fig. 3).

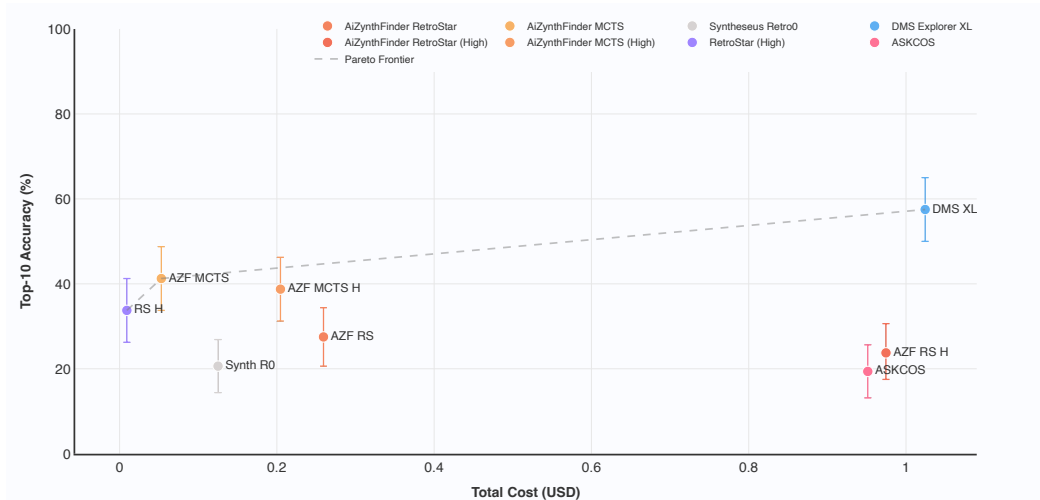


Figure 3: The Economic Trade-offs of Synthesis Planning. Pareto plot of Top-10 route-matching accuracy versus computational cost (USD) on the mkt-cnv-160 benchmark. The efficient frontier (dashed line) illustrates the apparent optimal trade-off, but this landscape is defined by the field’s measurement limitations: models in the low-cost region can be predicated on chemically implausible steps, while the accuracy metric itself penalizes the discovery of valid novel routes. Error bars represent 95% bootstrapped CIs.

This frontier, however, must be interpreted through the lens of our prior findings. First, the low-cost, high-STR region is occupied by models whose performance can be predicated on chemically implausible steps. Second, the accuracy metric itself penalizes novel discovery, potentially misclassifying exploratory models as inefficient when they are merely operating outside the benchmark’s conservative definition of success.

These qualifications notwithstanding, this analysis establishes computational cost as a critical and quantifiable axis for evaluation. Existing models already show order-of-magnitude differences in cost, which at the scale of a synthesis-aware virtual screening campaign is the difference between a

feasible project and an impossible one. As this cost-performance landscape is robust across route topologies (Fig. S1), it provides a necessary, if incomplete, framework for assessing the practical utility of synthesis planning tools.

3 Discussion

The evidence presented in this work demonstrates that the stock-termination rate is an incomplete and potentially misleading proxy for progress in retrosynthesis. Our analysis revealed that because the metric ignores the chemical validity of intermediate steps, it can reward the generation of chemically implausible routes. For the dominant class of search-based models, high STR scores are not an independent measure of a route’s quality but rather a reflection of the algorithm’s success in satisfying its core search objective. An over-reliance on such a metric may inadvertently steer the field’s focus toward topological pathfinding at the expense of chemical plausibility. We therefore suggest the community reconsider its role, shifting its use from a primary performance objective to that of a necessary but insufficient post-hoc filter.

While route reproduction accuracy is a clear improvement over naive stock termination, we must acknowledge its primary limitation: it is inherently conservative. By defining correctness as adherence to a known experimental path, Top-K accuracy provides a crucial proxy for chemical plausibility but cannot, by definition, reward the discovery of a novel and potentially more efficient synthetic route. This presents a critical measurement challenge. As our stratified analysis showed, the sharp decline in Top-K accuracy for search-based models on complex targets could represent either a failure to find the known path or a success in discovering a valid alternative that our benchmark cannot recognize.

Disambiguating these outcomes—distinguishing a failed prediction from a creative one—is perhaps the central challenge for the next generation of retrosynthesis evaluation. Our public release of the complete, standardized prediction database is intended to catalyze this effort. By creating a reusable community asset, we decouple the development of novel plausibility metrics from the significant computational cost of running planning algorithms, allowing researchers to rapidly prototype and validate new scoring functions on a comprehensive set of state-of-the-art predictions. Furthermore, our SynthArena platform can be extended beyond a simple viewer into a system for distributed, expert annotation. By enabling chemists to flag and categorize invalid reaction steps, we can transform our static data release into a living, community-curated dataset of "chemical bugs," moving beyond passive metrics to an active, adversarial process of chemical bug hunting.

This focus on developing better plausibility metrics points toward a more rigorous standard for the field. Our work also demonstrates that large-scale benchmarks are not always necessary; smaller, carefully stratified benchmarks can provide greater diagnostic power by revealing performance boundaries invisible in a single, top-line number. To ensure these improved practices lead to sustainable progress, we advocate for the formal recognition of two distinct research tracks: one for methods and another for evaluation. This separation of concerns would discourage the practice of proposing new metrics alongside the methods they are designed to evaluate and ensure that progress is always measured against a stable, community-vetted yardstick.

Establishing a reliable measure of chemical plausibility is the key that unlocks a more sophisticated, multi-faceted evaluation of synthesis planning. The utility of secondary metrics, such as route diversity or average length, is currently limited; these measures are meaningless when calculated over a pool of predictions that includes chemically invalid pathways. Establishing a baseline of chemical validity, however, would transform these currently noisy measures into high-signal diagnostics. Only then can the field begin to ask more nuanced questions about the creativity, efficiency, and elegance of the chemically sound plans that different models produce.

4 Methods

4.1 Benchmark and Stock Set Curation

Our evaluation framework is built upon a combination of curated benchmark subsets and precisely defined starting material stocks. For the Chemist-Aligned (mkt-) series, designed to assess practical utility, we standardized on the ASKCOS "Buyables" set [27, 28]. This stock contains 313 458 compounds available from eMolecules, Sigma-Aldrich, LabNetwork, Mcule, and ChemBridge,

representing a realistic set of commercially acquirable starting materials. For the Developer-Aligned (ref-) series, designed for fair algorithmic comparison, we used the original stock sets from the PaRoutes publication, defined as the set of all unique leaf nodes from the corresponding evaluation set (n1 or n5). This follows the original protocol to isolate algorithmic performance from variations in stock availability.

All curated benchmarks were derived from the 10 000 experimental routes in the PaRoutes n5 evaluation set, except for ref-lng-84 which used routes from n1 set as well. To construct the mkt-benchmarks, the full n5 set was first filtered to retain only routes where all leaf nodes are present in the ASKCOS Buyables stock, yielding a commercially relevant subset of 4 279 routes. The ref-benchmarks were derived from the full, unfiltered n5 set.

From these source sets, benchmarks were created using stratified random sampling based on route length and topology (linear vs. convergent). For instance, the mkt-lin-500 benchmark was constructed by randomly sampling 100 routes from each length category (2, 3, 4, 5, and 6) from the commercially relevant linear route subset. This stratified design ensures that performance evaluation is not skewed by the natural prevalence of shorter routes in the source data and enables a direct, unweighted analysis of how model performance varies with synthetic complexity. To incorporate our multi-ground-truth evaluation paradigm while ensuring absolute reproducibility, the route pruning and expansion algorithm (described in Section 4.2.2) was executed as a pre-computation step. The complete, expanded set of all valid ground-truth routes for each target is stored directly within the final benchmark definition file. This transforms the benchmark into a static, verifiable artifact and ensures that all subsequent evaluations are performed against an identical set of solutions

Recognizing that a single random sample could yield unrepresentative results, we implemented a rigorous seed selection protocol to ensure the statistical stability of our benchmarks. For each proposed benchmark, we generated 15 candidate subsets using 15 distinct random seeds. We then evaluated a reference model (DMS Explorer XL) on each of the 15 subsets to obtain performance metrics (Solvability, Top-1, and Top-10 accuracy). A deviation score was calculated for each seed, defined as the sum of the squared Z-scores for the three metrics relative to the mean performance across all 15 seeds. This score quantifies how much a given seed’s resulting benchmark deviates from the average behavior. The seed with the lowest deviation score was selected to generate the final, canonical benchmark used in this study, providing strong evidence that our results are robust to sampling variance (See Fig. S2-S4). Finally, the USPTO-190 benchmark was used as published, with evaluation performed against the ASKCOS Buyables stock to maintain consistency with our mkt-series analysis.

4.2 Statistical Analysis and Metrics

4.2.1 Confidence Intervals and Significance Testing

All reported metrics are accompanied by 95% confidence intervals (CIs) calculated using a non-parametric bootstrap procedure with 10,000 resamples. For a given metric and a set of N targets, we generated 10,000 bootstrap samples by resampling the N target outcomes with replacement. The reported CI represents the 2.5th and 97.5th percentiles of the distribution of the means of these bootstrap samples. To guard against misleading CIs from small or skewed samples, we implemented a reliability check flagging estimates derived from sample sizes $N < 30$ or where the number of positive or negative outcomes was less than 5, in line with standard statistical practice for proportions.

To determine if the performance difference between two models is statistically significant, we employed a paired bootstrap difference test. For a set of targets evaluated by both models, we first created a vector of the paired differences in their outcomes (e.g., 1 if model B succeeded and A failed, -1 if A succeeded and B failed, 0 otherwise). This vector of differences was then bootstrapped 10,000 times to construct a 95% CI for the mean difference. A difference was considered statistically significant if the resulting 95% CI did not contain zero.

4.2.2 Refined Top-K Accuracy (Multi-Ground-Truth Evaluation)

The standard Top-K accuracy metric is overly rigid, incorrectly penalizing models for identifying valid, economically superior routes that terminate early at commercially available intermediates. To address this limitation, we developed a more chemically meaningful multi-ground-truth (MGT) evaluation protocol. For a given reference route from a benchmark and a specified commercial stock

set, our algorithm generates an expanded set of valid ground-truth pathways through the following procedure:

1. **Identify Pruning Points:** All intermediate molecules within the reference route that are also present in the commercial stock are identified as potential pruning points.
2. **Generate Valid Sub-routes:** An expanded set of acceptable ground-truth routes is generated by treating these intermediates as alternative starting materials. To avoid generating redundant or invalid sub-routes (e.g., pruning at both an intermediate and its own precursor), the generation is constrained to combinations of intermediates that form an *antichain* within the route’s directed acyclic graph. An antichain is a set of nodes where no node is an ancestor of another, ensuring that each pruning point is independent.
3. **Validate Stock Termination:** Each newly generated "pruned" route is validated to ensure all of its terminal leaf nodes are present in the commercial stock.
4. **Evaluate Match:** A model’s prediction is scored as a Top-K success if it achieves an exact topological match with *any* route in this expanded set of valid ground truths (i.e., the original reference route plus all valid, solvable pruned variants).

4.3 Model Selection and Execution

The models evaluated in this study comprise all major open-source, self-deployable retrosynthesis planners available at the time of writing: AiZynthFinder, ASKCOS, Retro*, Syntheseus, SynPlanner, and DirectMultiStep. All model execution was automated via scripts provided in the RetroCast repository. To ensure full computational reproducibility, the environment was managed by the uv package manager, with a committed uv.lock file guaranteeing byte-for-byte identical versions of all dependencies.

To establish a baseline reflecting a standard installation, model configurations were set to their published defaults; all configuration files are committed to the repository for complete auditability. The sole deviation from default settings was an increase in the number of search iterations to 500 for all applicable models, aligning our protocol with the high-effort configuration used in the original PaRoutes study [24]. Execution was performed on cloud compute resources. Search-based planners were run on AWS EC2 c7i.xlarge instances (4 vCPUs, 8 GB RAM, \$0.1785/hr), with the more demanding ASKCOS planner run on a c7i.4xlarge instance (16 vCPU, 32 GB RAM, \$0.714/hr). The sequence-based DirectMultiStep model was run on NVIDIA A100 GPUs (40 GB SXM4, \$1.29/hr) provided by Lambda, Inc.

Code and Data Availability

All code, configuration files, and analysis scripts are organized in the ‘RetroCast’ Python package, which is open-source and publicly available on GitHub at github.com/ischemist/project-procrustes under an MIT license.

The complete dataset generated and analyzed in this study—including all raw model outputs, standardized route data, and aggregated statistical results—is permanently archived and publicly accessible at files.ischemist.com/retrocast/publication-data/. The integrity of the entire data archive can be computationally verified by executing the `retrocast verify --all` command.

The interactive web platform, SynthArena, is also open-source, with its code available at github.com/ischemist/syntharena. A public instance of the platform, hosting the results presented in this paper, is accessible at syntharena.ischemist.com.

Author Contributions

Anton Morgunov: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing – Original Draft, Visualization.

Victor S. Batista: Supervision, Funding Acquisition, Writing – Review & Editing, Project Administration.

Conflict of Interest

A.M. is the primary author of the DirectMultiStep (DMS) model, one of the methods evaluated in this study. To ensure objectivity, the evaluation framework and all associated data are fully open-source, and all results are computationally reproducible via the provided scripts, allowing for independent verification of the findings. V.S.B. declares no competing interests.

Acknowledgement

A.M. thanks Dr. Bogdan Zagribelnyy for insightful discussions. This research was supported by the National Science Foundation under Grant No. CHE-2124511. This research was also supported in part by Lambda, Inc.

References

- [1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- [2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, Sep 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL <https://doi.org/10.1038/s41586-025-09422-z>.
- [3] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen

He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kopic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL <https://arxiv.org/abs/2303.08774>.

- [4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrlke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023. URL <https://arxiv.org/abs/2303.12712>.
- [5] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL <https://doi.org/10.1038/s41586-021-03819-2>.
- [6] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O’Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Židek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL <https://doi.org/10.1038/s41586-024-07487-w>.
- [7] Oleg Kovalevskiy, Juan Mateos-Garcia, and Kathryn Tunyasuvunakool. Alphafold two years on: Validation and impact. *Proceedings of the National Academy of Sciences*, 121(34):e2315002121, 2024. doi: 10.1073/pnas.2315002121. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2315002121>.
- [8] Letícia M. F. Bertoline, Angélica N. Lima, Jose E. Krieger, and Samantha K. Teixeira. Before and after alphafold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, Volume 3 - 2023, 2023. ISSN 2673-7647. doi: 10.3389/fbinf.2023.1120370. URL <https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2023.1120370>.

- [9] E. J. Corey and W. Todd Wipke. Computer-Assisted Design of Complex Organic Syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. *Science*, 166(3902):178–192, 1969. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.166.3902.178. URL <https://www.science.org/doi/10.1126/science.166.3902.178>.
- [10] Connor W. Coley, Regina Barzilay, Tommi S. Jaakkola, William H. Green, and Klavs F. Jensen. Prediction of organic reaction outcomes using machine learning. *ACS Central Science*, 3(5):434–443, 2017. doi: 10.1021/acscentsci.7b00064. URL <https://doi.org/10.1021/acscentsci.7b00064>. PMID: 28573205.
- [11] Wengong Jin, Connor W. Coley, Regina Barzilay, and Tommi S. Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2607–2616, 2017. URL <https://proceedings.neurips.cc/paper/2017/hash/ced556cd9f9c0c8315cfbe0744a3baf0-Abstract.html>.
- [12] Songtao Liu, Zhengkai Tu, Minkai Xu, Zuobai Zhang, Lu Lin, Rex Ying, Jian Tang, Peilin Zhao, and Dinghao Wu. Fusionretro: molecule representation fusion via in-context learning for retrosynthetic planning. In *International Conference on Machine Learning*, pages 22028–22041. PMLR, 2023.
- [13] Hanjun Dai, Chengtao Li, Connor Coley, Bo Dai, and Le Song. Retrosynthesis prediction with conditional graph logic network. In *Advances in Neural Information Processing Systems*, pages 8870–8880, 2019.
- [14] Yu Shee, Haote Li, Pengpeng Zhang, Andrea M. Nikolic, Wenxin Lu, H. Ray Kelly, Vidhyadhar Manee, Sanil Sreekumar, Frederic G. Buono, Jinhua J. Song, and et al. Site-specific template generative approach for retrosynthetic planning. *Nature Communications*, 15(1), Sep 2024. doi: 10.1038/s41467-024-52048-4.
- [15] Marwin H. Segler, Mike Preuss, and Mark P. Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, Mar 2018. doi: 10.1038/nature25978.
- [16] Binghong Chen, Chengtao Li, Hanjun Dai, and Le Song. Retro*: Learning retrosynthetic planning with neural guided a* search. In *The 37th International Conference on Machine Learning (ICML 2020)*, 2020.
- [17] John S. Schreck, Connor W. Coley, and Kyle J. M. Bishop. Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970–981, 2019. doi: 10.1021/acscentsci.9b00055. URL <https://doi.org/10.1021/acscentsci.9b00055>. PMID: 31263756.
- [18] Yemin Yu, Ying Wei, Kun Kuang, Zhengxing Huang, Huaxiu Yao, and Fei Wu. Grasp: Navigating retrosynthetic planning with goal-driven policy. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 10257–10268. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/42beaab8aa8da1c77581609a61eced93-Paper-Conference.pdf.
- [19] Siqi Hong, Hankz Hankui Zhuo, Kebing Jin, Guang Shao, and Zhanwen Zhou. Retrosynthetic planning with experience-guided monte carlo tree search. *Communications Chemistry*, 6(1), Jun 2023. doi: 10.1038/s42004-023-00911-8.
- [20] Shufang Xie, Rui Yan, Peng Han, Yingce Xia, Lijun Wu, Chenjuan Guo, Bin Yang, and Tao Qin. Retrograph: Retrosynthetic planning with graph search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD ’22, page 2120–2129, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539446. URL <https://doi.org/10.1145/3534678.3539446>.
- [21] David Kreutter and Jean-Louis Reymond. Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search. *Chem. Sci.*, 14:9959–9969, 2023. doi: 10.1039/D3SC01604H. URL <http://dx.doi.org/10.1039/D3SC01604H>.
- [22] Piotr Gaiński, Michał Koziarski, Krzysztof Maziarz, Marwin Segler, Jacek Tabor, and Marek Śmieja. Retrogn: Diverse and feasible retrosynthesis using gflownets, 2025. URL <https://arxiv.org/abs/2406.18739>.
- [23] Kunyang Sun, Dorian Bagni, Joseph M. Cavanagh, Yingze Wang, Jacob M. Sawyer, Bo Zhou, Andrew Gritsevskiy, Oufan Zhang, and Teresa Head-Gordon. Synllama: Generating synthesizable molecules and their analogs with large language models. *ACS Central Science*, 0(0):null, 0. doi: 10.1021/acscentsci.5c01285. URL <https://doi.org/10.1021/acscentsci.5c01285>.

- [24] Samuel Genheden and Esben Bjerrum. PaRoutes: towards a framework for benchmarking retrosynthesis route predictions. *Digital Discovery*, 1(4):527–539, 2022. ISSN 2635-098X. doi: 10.1039/D2DD00015F. URL <http://xlink.rsc.org/?DOI=D2DD00015F>.
- [25] Yu Shee, Anton Morgunov, Haote Li, and Victor S. Batista. Directmultistep: Direct route generation for multistep retrosynthesis. *Journal of Chemical Information and Modeling*, 65(8):3903–3914, 2025. doi: 10.1021/acs.jcim.4c01982. URL <https://doi.org/10.1021/acs.jcim.4c01982>. PMID: 40197023.
- [26] Zhengkai Tu, Sourabh J. Choure, Mun Hong Fong, Jihye Roh, Itai Levin, Kevin Yu, Joonyoung F. Joung, Nathan Morgan, Shih-Cheng Li, Xiaoqi Sun, Huiqian Lin, Mark Murnin, Jordan P. Liles, Thomas J. Struble, Michael E. Fortunato, Mengjie Liu, William H. Green, Klavs F. Jensen, and Connor W. Coley. Askcos: Open-source, data-driven synthesis planning. *Accounts of Chemical Research*, 58(11):1764–1775, 2025. doi: 10.1021/acs.accounts.5c00155. URL <https://doi.org/10.1021/acs.accounts.5c00155>. PMID: 40397546.
- [27] Jihye Roh, Joonyoung F. Joung, Kevin Yu, Zhengkai Tu, G. Logan Bartholomew, Omar A. Santiago-Reyes, Mun Hong Fong, Richmond Sarpong, Sarah E. Reisman, and Connor W. Coley. Higher-level strategies for computer-aided retrosynthesis. *ChemRxiv*, 2025. doi: 10.26434/chemrxiv-2025-21zvt-v2.
- [28] Yu Shee and Anton Morgunov. Data for “directmultistep: Direct route generation for multistep retrosynthesis”. https://figshare.com/articles/dataset/Data_for_DirectMultiStep_Direct_Route_Generation_for_Multistep_Retrosynthesis_/28629470, 3 2025. Accessed: 20xx-xx-xx.
- [29] Krzysztof Maziarz, Guoqing Liu, Hubert Misztela, Austin Tripp, Junren Li, Aleksei Kornev, Piotr Gaiński, Holger Hoefling, Mike Fortunato, Rishi Gupta, and Marwin Segler. Chemist-aligned retrosynthesis by ensembling diverse inductive bias models, 2025. URL <https://arxiv.org/abs/2412.05269>.
- [30] Samuel Genheden, Amol Thakkar, Veronika Chadimová, Jean-Louis Reymond, Ola Engkvist, and Esben Bjerrum. Aizynthfinder: a fast, robust and flexible open-source software for retrosynthetic planning. *Journal of Cheminformatics*, 12(1):70, Nov 2020. ISSN 1758-2946. doi: 10.1186/s13321-020-00472-1. URL <https://doi.org/10.1186/s13321-020-00472-1>.
- [31] Krzysztof Maziarz, Austin Tripp, Guoqing Liu, Megan Stanley, Shufang Xie, Piotr Gaiński, Philipp Seidl, and Marwin H. S. Segler. Re-evaluating retrosynthesis algorithms with syntheseus. *Faraday Discuss.*, 256: 568–586, 2025. doi: 10.1039/D4FD00093E. URL <http://dx.doi.org/10.1039/D4FD00093E>.
- [32] Xuefeng Zhang, Haowei Lin, Muhan Zhang, Yuan Zhou, and Jianzhu Ma. A data-driven group retrosynthesis planning model inspired by neurosymbolic programming. *Nature Communications*, 16(1):192, Jan 2025. ISSN 2041-1723. doi: 10.1038/s41467-024-55374-9. URL <https://doi.org/10.1038/s41467-024-55374-9>.
- [33] Tagir Akhmetshin, Dmitry Zankov, Philippe Gantzer, Dmitry Babadeev, Anna Pinigina, Timur Madzhidov, and Alexandre Varnek. Synplanner: An end-to-end tool for synthesis planning. *Journal of Chemical Information and Modeling*, 65(1):15–21, 2025. doi: 10.1021/acs.jcim.4c02004. URL <https://doi.org/10.1021/acs.jcim.4c02004>. PMID: 39739735.
- [34] Kevin Yu, Jihye Roh, Ziang Li, Wenhao Gao, Runzhong Wang, and Connor W. Coley. Double-ended synthesis planning with goal-constrained bidirectional search, 2024. URL <https://arxiv.org/abs/2407.06334>.
- [35] Akihiro Kishimoto, Beat Buesser, Bei Chen, and Adi Botea. Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/4fc28b7093b135c21c7183ac07e928a6-Paper.pdf.
- [36] Junsu Kim, Sungsoo Ahn, Hankook Lee, and Jinwoo Shin. Self-improved retrosynthetic planning, 2021. URL <https://arxiv.org/abs/2106.04880>.

Supplementary Notes

This document provides supplementary information to the main text. It includes detailed results tables, figures, and the full methods section.

Supplementary Tables S1 and S2 Table S1, referenced in the main text Introduction, details the wide variance in starting material stock sets used across prominent retrosynthesis models. Table S2 provides a summary of the curated benchmarks introduced in this work, including their size, description, and the stock set used for evaluation.

Supplementary Tables S3 and S4 Tables S3 and S4 provide the baseline evaluation results on the `mkt-cnv-160` and `mkt-lin-500` benchmarks, respectively, using a rigid single-ground-truth (SGT) paradigm. As discussed in the main text, these results systematically underestimate model performance by penalizing the discovery of valid, shorter sub-routes. They serve as the control against which the improved multi-ground-truth (MGT) evaluation (main text Tables 2 and 3) is compared.

Supplementary Tables S5, S6, and S7 These tables present the full results for the Developer-Aligned (`ref-`) series of benchmarks. These benchmarks use the original patent-derived stocks from the PaRoutes dataset to facilitate fair algorithmic comparison, isolating model performance from the choice of commercial stock. Table S6 (`ref-lin-600`) and Table S5 (`ref-cnv-400`) corroborate the findings from the Chemist-Aligned series. Table S7 (`ref-lng-84`) presents the results of the "complexity cliff" stress test on exclusively long routes, as discussed in the main text.

Caveat on Training Data Standardization: The primary objective of the Developer-Aligned (`ref-`) series is to isolate algorithmic performance by controlling for the training corpus. However, to provide an immediate operational baseline, the results in Tables S5–S7 utilize the official pre-trained weights for each model. Since these models were originally trained on varying datasets, performance differences may currently reflect data discrepancies as well as architectural ones. We present these results as a provisional snapshot; we expect that as the community adopts this framework and populates the leaderboard with models retrained on the standardized split, these benchmarks will evolve into a pure measure of algorithmic superiority.

Supplementary Figure S1 Figure S1 shows the Pareto plot of accuracy versus cost for the `mkt-lin-500` benchmark, demonstrating that the cost-performance trade-offs identified in the main text are robust across different route topologies.

Supplementary Figures S2-S5 These figures (S2, S3, S4, S5) detail the statistical stability analysis performed to select the final seed for each curated benchmark. By evaluating a reference model on 15 candidate subsets for each benchmark, we selected the seed that produced a benchmark with performance metrics closest to the multi-seed average, ensuring our results are robust against sampling artifacts.

S1 Supplementary Tables

Model	Stock Source	Approx. Size	Notes
MCTS[15]	Curated	423 000	Compounds from ZINC15 and Reaxys
DFPN[35]	USPTO	977 000	All molecules from the patent training data
Retro*[16]	eMols	231 000 000	size of eMolecules screening in 2020
Self-Improved Retro[36]	eMols	231 000 000	
EG-MCTS[19]	eMols	231 000 000	
GRASP[18]	eMols	231 000 000	
RetroGraph[20]	eMols	231 000 000	
DreamRetroEr[32]	eMols	231 000 000	
SynLLaMa[23]	Enamine	230 000	
MultiStepTTL[21]	Curated	534 000	Enamine + MolPort
RetroChimera[29]	eMols	23 100 000	size of eMolecules screening in 2025

Table S1: **Inconsistency in Starting Material Stocks Across Major Retrosynthesis Models.** The size and composition of stock sets used to define a "solved" route vary by over 1000x, ranging from curated commercial catalogs to massive, speculative screening libraries. This disparity makes direct comparison of reported solvability scores between models unreliable.

Benchmark	Description	N Targets	Stock Set
<i>Chemist-Aligned Series (mkt-)</i>			
mkt-lin-500	Linear routes (len 2-6)	500	ASKCOS Buyables
mkt-cnv-160	Convergent routes (len 2-5)	160	ASKCOS Buyables
<i>Developer-Aligned Series (ref-)</i>			
ref-lin-600	Linear routes (len 2-7)	600	PaRoutes n5
ref-cnv-400	Convergent routes (len 2-5)	400	PaRoutes n5
ref-lng-84	Long routes (len 8-10)	84	PaRoutes n1+n5

Table S2: **Curated Benchmarks for Targeted Retrosynthesis Evaluation.** The mkt- series assesses practical utility using a commercial stock, while the ref- series enables fair algorithmic comparison using patent-derived stocks. The lin- and cnv- benchmarks contain an equal number of routes sampled from each length category, enabling a direct, unweighted analysis of performance that is not skewed by the natural prevalence of shorter routes.

Model	Stock Termination (%)	Top-1 Acc. (%)	Top-10 Acc. (%)
DMS Explorer XL	96.3 [93.1, 98.8]	21.9 [15.6, 28.7]	44.4 [36.9, 51.9]
AiZynF MCTS	98.1 [95.6, 100.0]	3.1 [0.6, 6.3]	8.8 [4.4, 13.1]
AiZynF MCTS (High)	99.4 [98.1, 100.0]	3.1 [0.6, 6.3]	7.5 [3.8, 11.9]
AiZynF Retro* (High)	99.4 [98.1, 100.0]	0.0 [0.0, 0.0]	4.4 [1.3, 7.5]
AiZynF Retro*	98.8 [96.9, 100.0]	1.3 [0.0, 3.1]	4.4 [1.3, 8.1]
Retro*	98.8 [96.9, 100.0]	4.4 [1.3, 8.1]	4.4 [1.3, 8.1]
Retro* (High)	100.0 [100.0, 100.0]	4.4 [1.3, 8.1]	4.4 [1.3, 8.1]
Syntheseus LocalRetro	95.6 [92.5, 98.8]	0.0 [0.0, 0.0]	1.3 [0.0, 3.1]
ASKCOS	93.1 [88.8, 96.9]	0.6 [0.0, 1.9]	1.3 [0.0, 3.1]

Table S3: **Baseline evaluation on mkt-cnv-160 using a rigid single-ground-truth paradigm.** These results correspond to an evaluation where only the full, original patent route is considered correct. As discussed in the main text, this rigid definition unfairly penalizes models for finding valid shorter routes, leading to a significant underestimation of accuracy. This table serves as the baseline for the corrected results presented in main text Table 2. Brackets indicate 95% confidence intervals. The interactive leaderboard is available on SynthArena: [link](#)

Model	Stock Termination (%)	Top-1 Acc. (%)	Top-10 Acc. (%)
DMS Explorer XL	97.2 [95.6, 98.6]	27.6 [23.6, 31.6]	50.2 [45.8, 54.6]
AiZynF MCTS	97.6 [96.2, 98.8]	4.2 [2.4, 6.0]	14.0 [11.0, 17.2]
AiZynF MCTS (High)	98.4 [97.2, 99.4]	4.4 [2.6, 6.2]	11.6 [8.8, 14.6]
AiZynF Retro*	98.0 [96.6, 99.2]	1.6 [0.6, 2.8]	9.0 [6.6, 11.6]
AiZynF Retro* (High)	99.0 [98.0, 99.8]	1.2 [0.4, 2.2]	7.2 [5.0, 9.6]
Retro*	99.8 [99.4, 100.0]	9.0 [6.6, 11.6]	9.0 [6.6, 11.6]
Retro* (High)	99.8 [99.4, 100.0]	9.0 [6.6, 11.6]	9.0 [6.6, 11.6]
Syntheseus LocalRetro	94.0 [91.8, 96.0]	1.0 [0.2, 2.0]	5.0 [3.2, 7.0]
ASKCOS	94.4 [92.4, 96.4]	0.8 [0.2, 1.6]	3.8 [2.2, 5.6]

Table S4: **Baseline evaluation on mkt-lin-500 using a rigid single-ground-truth paradigm.** Performance on the linear route benchmark using the original patent route as the sole definition of correctness. Consistent with the convergent set, these scores systematically underestimate the performance of models that identify shorter, valid pathways. This table serves as the baseline for the corrected results presented in main text Table 3. Brackets indicate 95% confidence intervals. The interactive leaderboard is available on SynthArena: [link](#)

Model	Stock Termination (%)	Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
DMS Explorer XL	77.5 [73.3, 81.5]	35.8 [31.0, 40.5]	46.8 [41.8, 51.5]	19.8
AiZynF MCTS (High)	87.3 [84.0, 90.5]	19.0 [15.3, 23.0]	32.5 [28.0, 37.3]	29.6
Retro*	96.5 [94.5, 98.3]	31.5 [27.0, 36.0]	31.5 [27.0, 36.0]	1.8
Retro* (High)	99.5 [98.8, 100.0]	31.5 [27.0, 36.0]	31.5 [27.0, 36.0]	2.5
AiZynF MCTS	81.3 [77.3, 85.0]	17.3 [13.5, 21.0]	27.8 [23.5, 32.3]	9.2
AiZynF Retro*	87.5 [84.3, 90.8]	8.0 [5.5, 10.8]	21.0 [17.0, 25.0]	47.9
AiZynF Retro* (High)	94.0 [91.5, 96.3]	5.8 [3.5, 8.0]	15.3 [11.8, 18.8]	153.7
Syntheseus LocalRetro	74.8 [70.3, 79.0]	6.0 [3.8, 8.5]	13.3 [10.0, 16.8]	15.6

Table S5: **Controlled comparison on convergent routes confirms architectural performance signatures.** Performance on the ref-cnv-400 benchmark. As a developer-aligned evaluation, this set uses the original PaRoutes n5 stock. The results are highly consistent with those on the chemist-aligned mkt-cnv-160 set (Table 2), reinforcing the central observation of a performance divergence. Search-based methods excel at satisfying the stock termination condition, while the end-to-end model is more proficient at reproducing the structure of known synthetic plans. Brackets indicate 95% confidence intervals. The full, interactive leaderboard is available on SynthArena: [link](#)

Model	Stock Termination (%)			Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
DMS Explorer XL	76.8	[73.3,	80.2]	30.3 [26.7, 34.2]	44.8 [40.8, 48.8]	18.5
Retro* (High)	97.8	[96.7,	99.0]	25.8 [22.3, 29.3]	25.8 [22.3, 29.3]	3.5
Retro*	95.8	[94.2,	97.3]	25.5 [22.0, 29.0]	25.5 [22.0, 29.0]	1.7
AiZynF MCTS	82.8	[79.8,	85.8]	10.2 [7.8, 12.7]	21.5 [18.2, 24.8]	8.2
AiZynF MCTS (High)	88.8	[86.3,	91.3]	9.7 [7.3, 12.2]	21.5 [18.3, 24.8]	25.1
AiZynF Retro*	88.2	[85.5,	90.7]	7.3 [5.3, 9.5]	21.7 [18.3, 25.0]	29.4
AiZynF Retro* (High)	95.0	[93.2,	96.7]	6.2 [4.3, 8.2]	16.5 [13.7, 19.5]	109.9
Syntheseus LocalRetro	67.2	[63.3,	71.0]	5.0 [3.3, 6.8]	13.0 [10.3, 15.7]	15.4

Table S6: **Controlled comparison on linear routes confirms the divergence between stock termination and accuracy.** Performance on the `ref-lin-600` benchmark. This developer-aligned set uses the original PaRoutes `n5` stock, isolating algorithmic performance from the choice of commercial stock. The results corroborate the findings from the main text (Table 3): search-based models achieve high stock termination rate, while the sequence-based DirectMultiStep model exhibits substantially higher Top-K accuracy. This demonstrates that the observed divergence is a robust architectural signature, not an artifact of a specific stock set. Brackets indicate 95% confidence intervals. The full, interactive leaderboard is available on SynthArena: [link](#)

Model	Stock Termination (%)			Top-1 Acc. (%)	Top-10 Acc. (%)	Time/Target (s.)
DMS Explorer XL	72.6	[63.1,	82.1]	45.2 [34.5, 56.0]	50.0 [39.3, 60.7]	31.3
Retro* (High)	94.0	[89.3,	98.8]	8.3 [2.4, 14.3]	8.3 [2.4, 14.3]	8.6
Retro*	86.9	[79.8,	94.0]	8.3 [2.4, 14.3]	8.3 [2.4, 14.3]	3.9
AiZynF Retro* (High)	76.2	[66.7,	84.5]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	134.2
AiZynF Retro*	61.9	[51.2,	72.6]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	31.7
AiZynF MCTS (High)	59.5	[48.8,	70.2]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	27.2
AiZynF MCTS	46.4	[35.7,	57.1]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	8.9
Syntheseus LocalRetro	36.9	[27.4,	47.6]	0.0 [0.0, 0.0]	0.0 [0.0, 0.0]	16.6

Table S7: **Stress test on long routes provides stark evidence of the "complexity cliff."** Performance on the `ref-lng-84` benchmark, a challenging set composed of all routes of length 8-10 from the PaRoutes evaluation sets. This benchmark is designed to probe model performance at the limits of planning complexity. The results provide the clearest evidence of the architectural trade-offs discussed in the main text: the Top-K accuracy of all search-based models collapses to near-zero. In contrast, the sequence-based DirectMultiStep model retains substantial accuracy, demonstrating its robustness on long-range planning tasks. Brackets indicate 95% confidence intervals. The full, interactive leaderboard is available on SynthArena: [link](#)

S2 Supplementary Figures

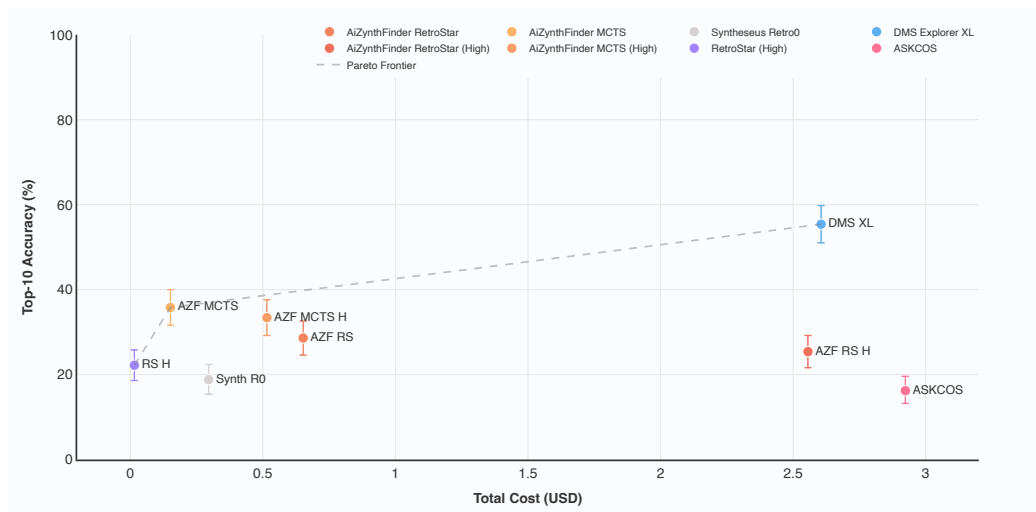


Figure S1: The Cost of Accuracy on Linear Routes. A Pareto plot of Top-10 Accuracy versus Total Cost (USD) on the `mkt-1in-500` benchmark. The analysis confirms the trade-off structure seen in Figure 3 is generalizable to linear routes. While absolute costs differ due to the larger benchmark size, the relative cost-performance profiles and the shape of the efficient frontier are consistent, demonstrating a robust relationship between accuracy and computational cost. Error bars are 95% bootstrapped CIs.

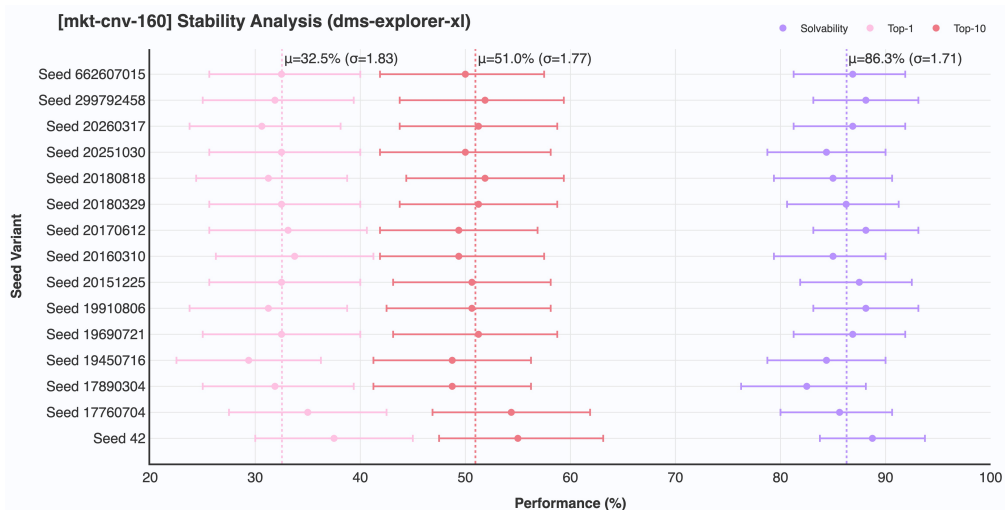


Figure S2: **Selection of a statistically representative benchmark seed for mkt-cnv-160.** The plot shows the performance variation of a reference model (DMS Explorer XL) across 15 candidate benchmarks, each generated with a different random seed. Points indicate the mean accuracy (Solvability, Top-1, and Top-10), with horizontal bars representing the bootstrapped 95% confidence intervals. Dashed vertical lines mark the grand mean performance across all seeds. This stability analysis allows us to quantify the variance introduced by subset sampling and select a seed (in this case, 20180329) that yields a benchmark whose metrics are demonstrably close to the central tendency, ensuring our evaluations are robust against sampling artifacts.

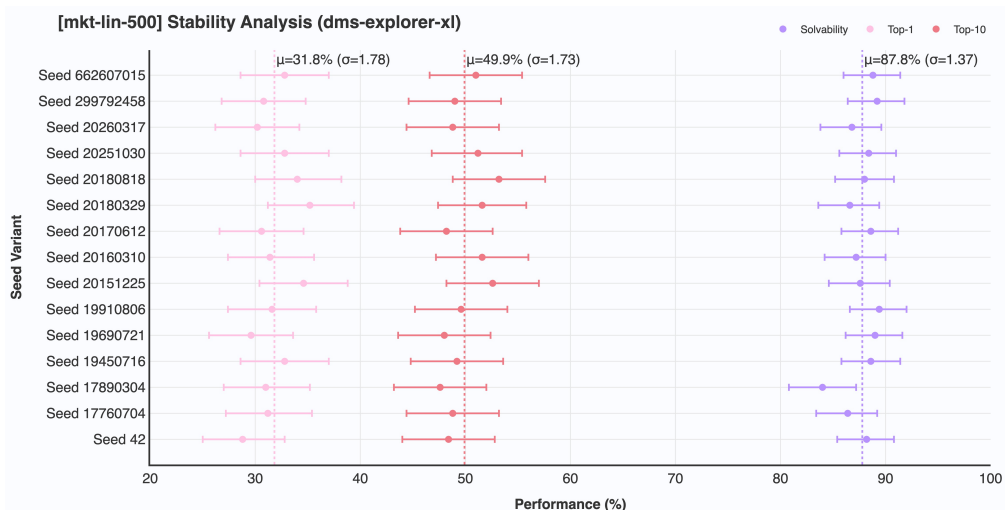


Figure S3: **Benchmark stability analysis for mkt-lin-500.** Performance of a reference model is shown for 15 candidate benchmarks generated from different random seeds. The analysis confirms that while sampling introduces variance, a representative seed (19450716) can be selected by minimizing the deviation from the grand mean (dashed lines), thereby ensuring the benchmark is not an outlier.

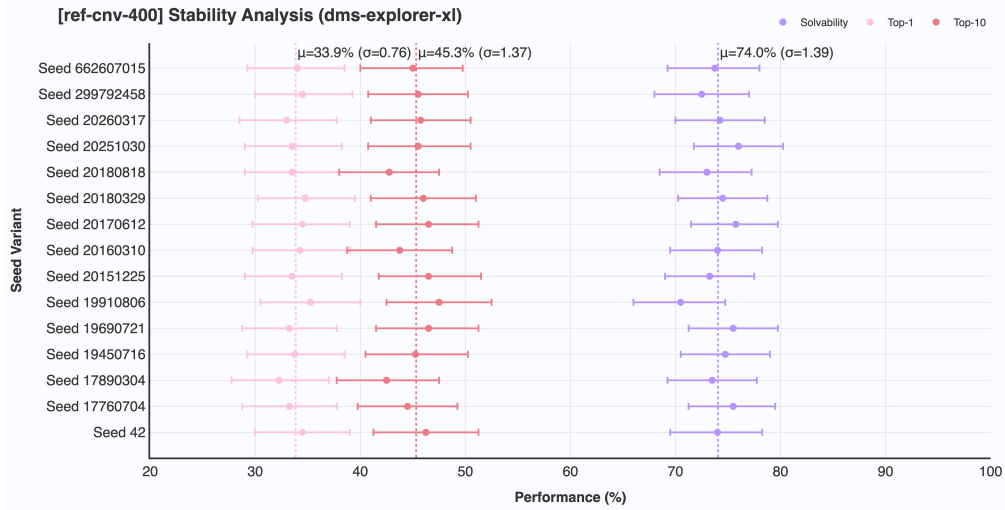


Figure S4: **Benchmark stability analysis for ref-cnv-400.** As with other benchmarks, we evaluated 15 candidate subsets to characterize the variance from random sampling. The selected seed (662607015) yields a benchmark with performance metrics near the central tendency of all possible samples, providing a stable and fair basis for model comparison.

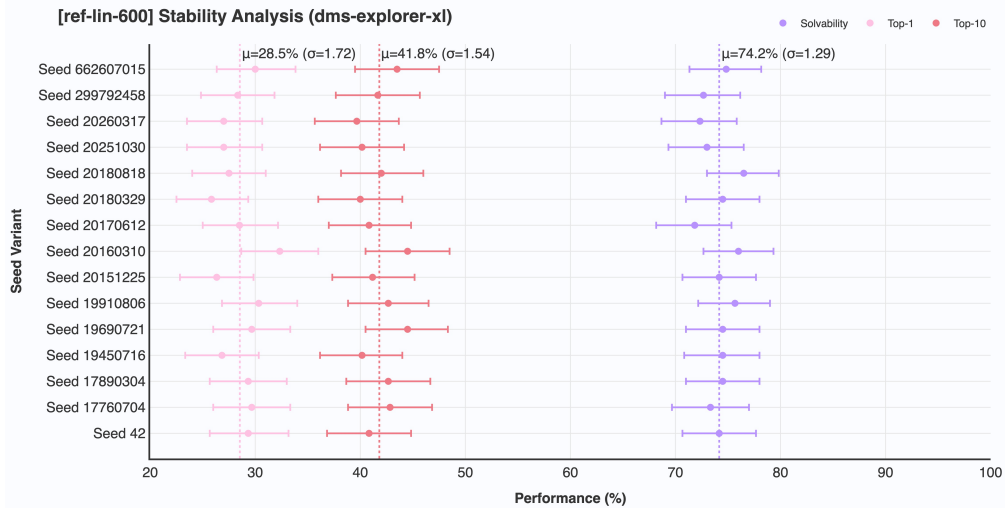


Figure S5: **Benchmark stability analysis for ref-lin-600.** The plot demonstrates the range of performance outcomes resulting purely from the choice of random seed in benchmark sampling. Our selection of a seed (17890304) with a low Z-score deviation from the multi-seed average ensures that our reported results are generalizable and not an artifact of a fortuitous or pessimistic random sample.