

Guiding large language models to predict edit sequences for molecular synthesizability optimization

In the format provided by the
authors and unedited

Contents

1	Methodological Details	2
1.1	Source-Molecule Generation Procedure	2
1.2	Chemistry-Quality Profile of the Source Pools	2
1.3	Definition of the Editing Action Space	3
1.4	Graph-Matching Edit-Path Extraction (Iterative MCS)	4
1.5	System Prompt	5
1.6	Few-Shot Exemplar Construction	5
1.7	Multi-LLM Endpoint Configuration	6
1.8	SimpRetro Validation Protocol	6
1.9	Interaction-Aware Constraints Pipeline	7
2	Ablation Studies	8
2.1	Experimental Setup	8
2.2	Results	8
2.3	Discussion	8
2.3.1	Superiority of Edit Prediction over Direct Generation	8
2.3.2	Importance of Few-Shot Retrieval (RAG)	9
2.4	Random-RAG Ablation and LLM-as-Judge Alignment Scoring	9
2.5	Edit-Path Audit: LLM-Direct Edit Generation	10
2.6	Edit-Quality Audit	10
2.7	Golden-Corpus Reasoning Validation	11
3	Prospective Optimization Results for All 42 High-Scoring RIPK1 Candidates	13
4	Template-Inventory Scaling Experiment	20
5	End-to-end Hit Recovery Demonstration on SARS-CoV-2 M^{pro}	22
6	Coverage Across Additional 3D-Generative Baselines	25
7	Systematic Binding-Mode Retention via PLIP	26
8	Failure-Mode Decomposition	28
9	eMolecules Synthesizability Under Enamine Building-Block Constraint	30

1 Methodological Details

1.1 Source-Molecule Generation Procedure

For each of the five 3D-aware generative baselines (LiGAN, AR / 3D-SBDD, Pocket2Mol, DiffSBDD, ResGen), we used the candidate molecules pre-generated and distributed by the GenBench3D benchmark of Baillif *et al.* [1], who generated them across the 94 CrossDocked2020 test pockets with each model’s released checkpoint under its default settings (approximately 75–100 molecules per pocket per model). Total per-model molecule counts after RDKit sanitisation range from 6,285 (LiGAN) to 8,385 (DiffSBDD). Per-pocket counts and per-pocket unsynthesizable fractions are reported in Figure 1. The unsynthesizable fraction varies markedly across pockets and across models (mean / range, per pocket): LiGAN 0.21 / [0.07, 0.32]; AR 0.35 / [0.16, 0.53]; P2M 0.31 / [0.03, 0.64]; DiffSBDD 0.19 / [0.09, 0.31]; ResGen 0.24 / [0.00, 0.61]; a small number of pockets exhibit very high unsynthesizable fractions for P2M and ResGen, reflecting binding sites whose geometry biases the generator toward strained or non-drug-like substructures.

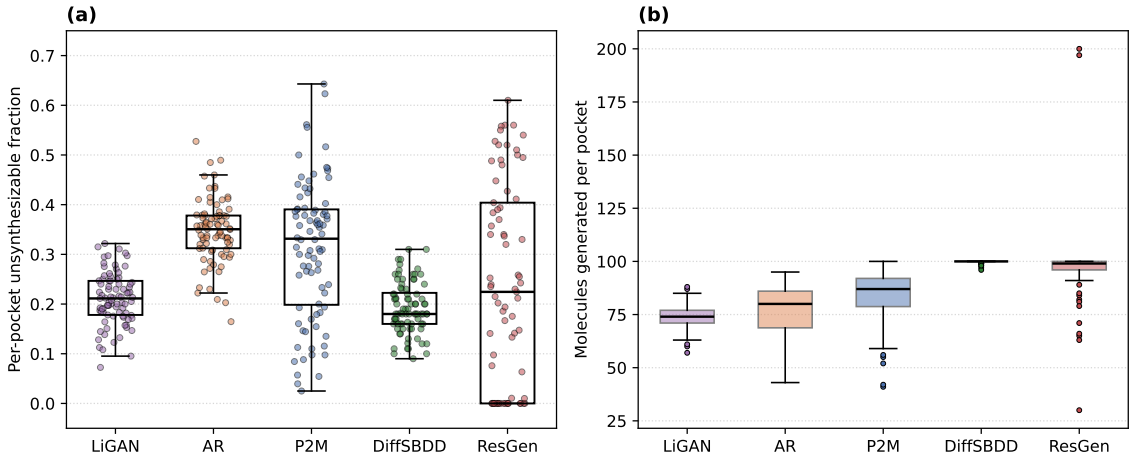


Figure 1: Per-pocket generation distribution across the 94 CrossDocked test pockets for the five 3D-aware generative baselines. (a) Per-pocket fraction of generated molecules classified as unsynthesizable under the SimpRetro+Enamine protocol. (b) Per-pocket count of generated molecules (the per-model total ranges from 6,285 for LiGAN to 8,385 for DiffSBDD).

1.2 Chemistry-Quality Profile of the Source Pools

To characterize the chemistry quality of the unsynthesizable subsets across the five generative baselines, we computed the Ertl & Schuffenhauer synthetic accessibility (SA) score and the rate of compounded chemistry artifacts on the full unsynthesizable pools for P2M and ResGen and on the 10% stratified subsets for LiGAN, AR, and DiffSBDD. Compounded artifacts are defined as molecules containing two or more substructure patterns from the set {cumulene (C=C=C), peroxide (O-O), thiol/S-H, geminal diol, geminal diol/amine, vinyl ether-hydroxyl, and imine-hydroxyl}; this set was chosen because each individual pattern is rarely a synthesis-blocker in isolation but their co-occurrence typically signals a structural artifact of the generator.

Table 1: SA score and compounded-artifact rates per generator (unsynthesizable subsets only). “Any artifact” counts molecules with at least one substructure match; “compounded” counts molecules with two or more distinct patterns. SA shift is computed relative to the P2M baseline.

Source	n	Mean SA	SA shift vs P2M	Any artifact	Compounded
P2M	1,715	3.65	—	12.1%	0.5%
ResGen	1,025	3.51	−0.11	13.0%	0.9%
DiffSBDD	383	5.14	+1.52	22.7%	1.3%
AR	365	5.04	+1.42	17.0%	1.4%
LiGAN	372	4.90	+1.28	47.8%	8.6%

LiGAN, AR, and DiffSBDD show SA distributions shifted right by 1.2–1.5 units versus P2M (the published P2M baseline), confirming that these three baselines produce intrinsically harder-to-synthesize starting points. The compounded-artifact rate of LiGAN (8.6%) is notably higher than the other four baselines (0.5–1.4%), and LiGAN’s any-artifact rate (47.8%) is roughly 4× the P2M rate (12.1%). The full per-generator SA distribution is shown in Supporting Information Figure 2.

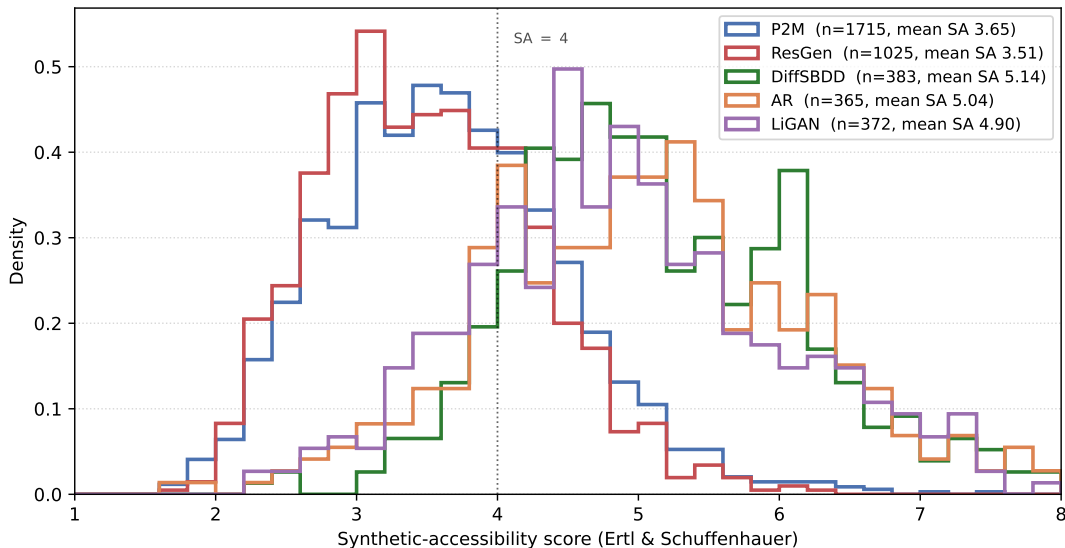


Figure 2: Density of Ertl & Schuffenhauer SA scores across the five 3D-aware generative baselines (unsynthesizable subsets). Dotted vertical line at SA = 4 marks the commonly-used boundary between “easy” and “hard” synthesis. LiGAN, AR, and DiffSBDD distributions are shifted right by 1.2–1.5 units relative to P2M / ResGen.

1.3 Definition of the Editing Action Space

To transform the molecular optimization task into a discrete reasoning problem, we defined a comprehensive action space. The LLM is restricted to outputting a sequence of commands operating on specific atom indices (map numbers). The valid commands and their arguments are defined as follows:

- ["DEL_ATOM", map_num]: Deletes the atom specified by the map number and its incident bonds.
- ["MUTATE_ATOM", map_num, new_symbol]: Changes the atomic element of the target atom (e.g., C to N).
- ["ADD_ATOM", new_id, symbol]: Introduces a new atom. To avoid index conflict, new IDs start from 500.
- ["DEL_BOND", map_num1, map_num2]: Removes the bond between two existing atoms.
- ["ADD_BOND", id1, id2, bond_type]: Creates a bond between atoms. Bond types include "SINGLE", "DOUBLE", "TRIPLE", and "AROMATIC".
- ["CHANGE_BOND", map_num1, map_num2, new_bond_type]: Modifies the order of an existing bond.
- ["SET_CHIRAL", id, tag]: Sets or clears a stereocenter. tag accepts "CW" and "CCW" (set explicit chirality), and "UNSPECIFIED" (clear an existing stereocenter; useful when an edit removes the source’s chiral requirement).
- ["SET_BOND_STEREO", map_num1, map_num2, tag]: Sets or clears double-bond E/Z. tag accepts "E", "Z", and "UNSPECIFIED".
- ["STOP"]: A mandatory token indicating the end of the edit sequence.

The two UNSPECIFIED options let the model symmetrise a previously chiral source. For example, when removing the two methyl substituents of a 2,5-dimethylpiperazine yields an achiral piperazine core (as in the PLK1 case study), clearing the corresponding source stereocentres is part of the chemistry, not a separate metadata step.

1.4 Graph-Matching Edit-Path Extraction (Iterative MCS)

The synthesis-cliff reference corpus is built by computing a sequence of atom-level edits that maps each unsynthesizable source molecule onto a near-neighbour synthesizable target. We use an iterative maximum-common-substructure (MCS) algorithm implemented in `SynCraft-Core/src/utils.py: IterativeMCSExtractor`. The pseudocode below summarises the procedure; the full implementation, including bond-aromaticity matching, atom-map-number bookkeeping, and the fallback relaxation schedule, is open-sourced.

```
def extract_edit_path(src_smi, tgt_smi, action_vocab,
                     max_iters=8, min_mcs_atoms=2):
    """Return ordered list of edits mapping src -> tgt, or None on fail."""
    src = add_atom_maps(canonicalise(src_smi)) # 1-based map ids
    tgt = canonicalise(tgt_smi)
    relax = {"completeRingsOnly": True,
            "ringMatchesRingOnly": True,
            "bondCompare": "CompareOrderExact",
            "atomCompare": "CompareElements"}
    for _ in range(max_iters):
        mcs = rdkit_FMCS([src, tgt], **relax) # maximum common substructure
        if mcs.numAtoms < min_mcs_atoms:
            return None
        anchor = align_via_mcs(src, tgt, mcs) # src_atom -> tgt_atom map
        edits = [] # ordered list
        # 1. atom-side: deletions, additions, element mutations
        for a in src.atoms:
            if a.idx not in anchor:
                edits.append(["DEL_ATOM", a.map])
        for t in tgt.atoms:
            if t.idx not in anchor.values():
                edits.append(["ADD_ATOM", next_new_id(), t.symbol])
        for s, t in anchor.items():
            if src[s].symbol != tgt[t].symbol:
                edits.append(["MUTATE_ATOM", src[s].map, tgt[t].symbol])
        # 2. bond-side
        for b in src.bonds:
            if not bond_anchored(b, anchor):
                edits.append(["DEL_BOND", b.a.map, b.b.map])
            elif b.order != mapped_target_bond(b, anchor).order:
                edits.append(["CHANGE_BOND", b.a.map, b.b.map,
                             mapped_target_bond(b, anchor).order_name])
        for b in tgt.bonds:
            if not bond_anchored(b, invert(anchor)):
                edits.append(["ADD_BOND", id1, id2, b.order_name])
        # 3. stereo
        for s, t in anchor.items():
            edits += stereo_diff(src[s], tgt[t]) # SET_CHIRAL / SET_BOND_STEREO
        edits.append(["STOP"])
        # 4. execute + verify
        rebuilt = apply_edits(src, edits, action_vocab)
        if rebuilt is not None and canonicalise(rebuilt) == tgt:
            return edits
        # else: relax MCS thresholds and retry
        relax = relax_one_level(relax)
    return None
```

Listing 1: Iterative-MCS edit-path extraction.

The relaxation schedule on failure drops `completeRingsOnly`, then `ringMatchesRingOnly`, then loosens `bondCompare` to `CompareAny`; in practice the first or second iteration succeeds for >98% of the cliff pairs in our corpus. The remaining pairs that exhaust all relaxations (typically large scaffold hops) are rejected, contributing a roughly 2% pair-construction yield loss.

1.5 System Prompt

The system prompt establishes the persona of the model and defines the output constraints. The full text used in our experiments is provided below:

```
You are an expert medicinal chemist and a computational chemist. Your primary skill is
analyzing molecular structures to identify parts that are synthetically challenging and
proposing minimal, chemically sound modifications to improve their synthesizability while
preserving key pharmacophoric features.

[TASK]
You will be given a source molecule with atom map numbers. Your tasks are:
1. Reason Step-by-Step: First, analyze the molecule to identify synthetically problematic
   substructures. Then, propose a chemically sound modification.
2. Output Edits: Second, translate your proposed modification into a sequence of machine-
   readable commands in a JSON list.

[OUTPUT FORMAT]
You MUST provide your response in the following strict format. Do not add any text before '[
REASONING]' or after the final '''.

[REASONING]
Your detailed, step-by-step thinking process goes here. Explain what is wrong with the molecule
and why your proposed fix is better.

[EDITS]
'''json
[
  ["COMMAND_1", "ARG_1", ...],
  ["STOP"]
]

[COMMANDS SPECIFICATION]
(The definitions of ADD_ATOM, DEL_ATOM etc. as described in Section "Definition of the Editing
Action Space" above)
```

Listing 2: System Prompt used for SynCraft

1.6 Few-Shot Exemplar Construction

We utilized a Retrieval-Augmented Generation (RAG) approach. For each query molecule, we retrieved the top-k (k=5) most similar examples from our training set based on Morgan Fingerprint similarity. Each exemplar consists of the source SMILES, the ground-truth edit sequence, and a reasoning rationale. An example of a formatted few-shot exemplar is shown below:

```
source_smiles: "[N:1]1([C:2](=[O:3])[CH3:9])[CH2:4][CH2:5][CH2:6][C@H:7]1[CH2:8][NH2:10]"

[REASONING]
The N-acetyl group on the pyrrolidine nitrogen offers limited steric shielding around the
carbonyl and leaves the amide vulnerable to hydrolytic and metabolic cleavage. The
transformation quaternises the acetyl alpha-carbon (atom 9) by adding two methyl
substituents, converting the N-acetyl moiety into a more sterically hindered N-pivaloyl
group. This is a well-established hit-optimisation move that stabilises the amide bond
without disturbing the heterocyclic core.

[EDITS]
'''json
[
  ["ADD_ATOM", 500, "C"],
  ["ADD_ATOM", 501, "C"],
  ["ADD_BOND", 9, 500, "SINGLE"],
  ["ADD_BOND", 9, 501, "SINGLE"],
  ["STOP"]
]'''
```

1.7 Multi-LLM Endpoint Configuration

For the four-LLM comparison reported in Table 1, each language model is invoked through the same SynCraft pipeline (system prompt, action space, 5-shot RAG retrieval) with only the underlying LLM backend substituted. The four backends span roughly a fourteen-month release window (Gemini-2.5-Pro preview release 2025-06-05; GPT-5.2 released 2025-12-11; DeepSeek-V4-Pro/Flash open-weight release 2026-04-24), and the rescue-rate ordering across the ablation reflects this generation gap: the more recently released models are stronger across most thresholds. The LLM-to-LLM performance variation is comparable to or larger in magnitude than the gap between Gemini-based SynCraft and the best projection baseline (Table 1), indicating that reasoning capacity is a first-order factor in the editing paradigm. As stronger reasoning models become available, the SynCraft framework can directly inherit those gains without retraining, whereas projection methods remain constrained by their fixed template inventories. Per-backend reasoning level and decoding settings are summarised below:

- **Gemini-2.5-Pro:** `thinking` enabled at `level=high`; default temperatures (instruction-tuned). `Pass-k = 5`.
- **GPT-5.2:** `reasoning_effort=high`; maximum completion tokens 32,768. `Pass-k = 1` (API-budget constrained).
- **DeepSeek-V4-Pro:** MIT-licensed open weights (<https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro>). `reasoning_effort=high`; `max.tokens` 32,768. `Pass-k = 5`.
- **DeepSeek-V4-Flash:** smaller open-weight variant (<https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash>); same decoding configuration as DSv4-Pro.

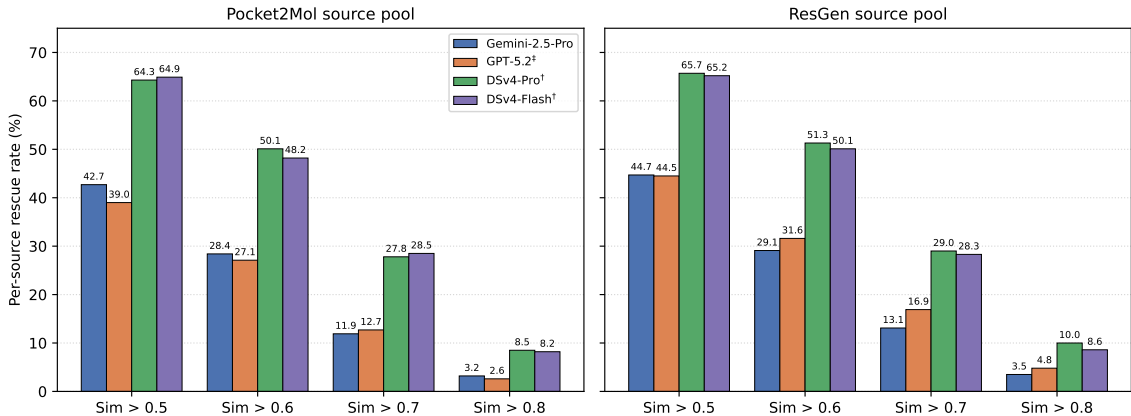


Figure 3: Per-source rescue rate at four Tanimoto-similarity thresholds for the four LLM backends on the Pocket2Mol (left) and ResGen (right) source pools. Numbers correspond to main-text Table 1; `pass-k` is 5 for Gemini-2.5-Pro, DeepSeek-V4-Pro and DeepSeek-V4-Flash, and 1 for GPT-5.2 (API-budget constrained). Sample sizes: $n=1,715$ for Pocket2Mol, $n=1,025$ for ResGen. Each bar represents the fraction of sources successfully rescued at the indicated similarity threshold; no error bars are shown because each LLM $\times$ source-pool $\times$ threshold cell is a single deterministic run (`pass-k` edits are sampled once per source).

1.8 SimpRetro Validation Protocol

All “synthesizable” classifications in this work, whether for source molecules, SynCraft-edited products, or projection-based-method outputs, use the same SimpRetro [2] validation procedure. Specifically:

- **Reaction templates:** extracted from the USPTO database [3] with RDChiral; retained if reaction frequency ≥ 2 in the source corpus.

- **Building blocks:** commercially available Enamine building-block library (the same library used by ChemProjector [4], SynFormer [5], and ReaSyn [6]).
- **Search budget:** 30-minute wall-clock budget per molecule on a single core (Intel Xeon E5-2670), with no hard limit on the number of search-tree steps within that budget.
- **Outcome:** a molecule is classified as “synthesizable” if SimpRetro returns at least one valid synthesis tree before the time budget expires.

This protocol is applied identically across all baselines (P2M, ResGen, LIGAN, AR, DiffSBDD) and all SynCraft variants (multi-LLM comparison, end-to-end skill demo, etc.), ensuring that all reported rescue rates are directly comparable.

1.9 Interaction-Aware Constraints Pipeline

For the structure-based optimization tasks, we implemented an automated pipeline to extract biological constraints.

1. **Docking:** The input SMILES is converted to a 3D conformer and docked into the reference protein structure using AutoDock Vina. We re-dock both the source candidate and every SynCraft-edited product under the same Vina protocol rather than mixing native generator poses (available for the source) with docked poses (the only option for edits); this within-pair symmetry is what makes the source–edit PLIP and Vina comparisons in Section 7 a like-for-like measurement.
2. **Interaction Profiling:** The best-scoring pose is analyzed using the Protein-Ligand Interaction Profiler (PLIP). We detect Hydrogen Bonds, Hydrophobic Interactions, Salt Bridges, Pi-Stacking, Pi-Cation, Halogen Bonds, and Metal Complexes.
3. **Graph Alignment:** A substructure matching algorithm maps the PDB atom serial numbers from the docked pose back to the canonical atom map numbers of the 2D molecular graph. This ensures the constraints correspond correctly to the atoms in the edit action space.
4. **Constraint Injection:** The detected interactions are converted into natural language descriptions and appended to the user prompt.

Because directional contacts (H-bonds, salt bridges, π -stacking, halogen bonds) are geometrically rigid while PLIP’s hydrophobic contacts use a soft 4 Å cut-off and routinely shift between near-isomers, the injected prompt separates the two interaction classes into a *hard* (preserve) block and a *soft* (prefer-where-reasonable) block. The format of the injected constraint prompt is:

```
CRITICAL BIOLOGICAL CONSTRAINTS [PRESERVE; high priority]:
The following directional contacts must be preserved by any proposed edit.
Atom with map number [12] (an N connected to [C:11, C:13]) forms a critical Hydrogen Bond with
ASP156.
Atom with map number [27] (a C of an aromatic ring) participates in pi-stacking with PHE142.

PREFERRED CONTACTS [maintain where reasonable; soft preference]:
The following hydrophobic contacts are detected by PLIP and should be preserved
where chemically consistent with the synthesizability fix, but may be relaxed if
the proposed edit yields a strictly better synthesis route.
Atom with map number [23] (a C atom connected to [C:22]) participates in a hydrophobic contact
with LEU145.
```

The hard/soft distinction reflects standard medicinal-chemistry practice and prevents the LLM from over-constraining the edit space around hydrophobic contacts that are not in fact load-bearing for the binding mode. Empirically, the hard-block residues are recovered far more often than soft-block residues, consistent with the systematic PLIP-retention statistics in Section 7.

2 Ablation Studies

To validate the architectural choices of SynCraft, we conducted a comprehensive ablation study focusing on two critical components: the output modality (Edit Sequence Prediction vs. Direct SMILES Generation) and the inference strategy (Few-shot Retrieval-Augmented Generation vs. Zero-shot).

2.1 Experimental Setup

We evaluated four different configurations on the test sets derived from Pocket2Mol (P2M) and ResGen:

1. **5-shot with edits (SynCraft Standard):** The proposed method using 5 retrieved exemplars to predict edit sequences.
2. **1-shot with edits:** Using only the single nearest neighbor as a demonstration to predict edit sequences.
3. **5-shot direct-SMILES:** Using 5 retrieved exemplars but instructing the LLM to directly generate the full SMILES string of the optimized molecule.
4. **Zero-shot direct-SMILES:** Asking the LLM to optimize the molecule directly without any in-context examples.

The primary metric is the success rate of identifying a synthesizable analog that maintains a Tanimoto similarity above specific thresholds (0.5 to 0.9) relative to the original input.

2.2 Results

The quantitative results on the Pocket2Mol and ResGen datasets are presented in Table 2 and Table 3, respectively.

Table 2: **Ablation Study on Pocket2Mol (P2M) Dataset.** Success rates of finding synthesizable analogs at varying similarity thresholds. The “Total Solved” column indicates the percentage of inputs for which a valid, synthesizable molecule was generated, regardless of similarity. Evaluated at pass-k = 1 to keep the four-condition ablation computationally tractable; the main-text Table 1 headline numbers use pass-k = 5 and therefore exceed the “5-shot with edits” row of this table.

Configuration	Total Solved	Sim > 0.5	Sim > 0.6	Sim > 0.7	Sim > 0.8	Sim > 0.9
5-shot with edits (SynCraft)	61.34%	38.54%	25.71%	10.90%	3.09%	0.47%
1-shot with edits	50.73%	33.24%	21.57%	9.50%	2.92%	0.52%
5-shot direct-SMILES	59.77%	31.66%	19.71%	7.64%	2.10%	0.29%
Zero-shot direct-SMILES	67.35%	23.62%	11.55%	4.31%	1.17%	0.29%

Table 3: **Ablation Study on ResGen Dataset.** Success rates of finding synthesizable analogs at varying similarity thresholds. Evaluated at pass-k = 1, matching Table 2.

Configuration	Total Solved	Sim > 0.5	Sim > 0.6	Sim > 0.7	Sim > 0.8	Sim > 0.9
5-shot with edits (SynCraft)	69.27%	44.68%	29.07%	13.07%	3.51%	0.49%
1-shot with edits	59.32%	38.93%	25.85%	10.63%	2.34%	0.20%
5-shot direct-SMILES	64.20%	34.83%	21.76%	8.39%	2.05%	0.78%
Zero-shot direct-SMILES	72.39%	26.44%	15.22%	5.66%	1.56%	0.39%

2.3 Discussion

2.3.1 Superiority of Edit Prediction over Direct Generation

A key finding from the tables is the distinct behavior between editing and direct generation. While the *Zero-shot direct-SMILES* baseline achieves the highest “Total Solved” rate (67.35% for P2M and 72.39% for ResGen), its performance drops precipitously when a similarity constraint is applied. For instance, on ResGen, the success rate for Zero-shot drops from 72.39% (any molecule) to 26.44% (Sim > 0.5), whereas SynCraft maintains a robust 44.68%.

This phenomenon indicates that without the structural constraints imposed by the edit-based paradigm, LLMs tend to “hallucinate” simple, generic synthesizable molecules that bear little resemblance to the complex input scaffold. By predicting discrete edit operations (e.g., `DEL_ATOM`, `ADD_BOND`), SynCraft forces the optimization to occur *on the graph of the original molecule*. This inherently preserves the majority of the structural features, resulting in significantly higher success rates at medium-to-high similarity thresholds (> 0.6), which is the critical regime for lead optimization.

Comparing the “5-shot with edits” against “5-shot direct-SMILES” further confirms this. Even when provided with the exact same few-shot examples, the edit-based output modality consistently outperforms the SMILES generation modality (e.g., 29.07% vs 21.76% at $\text{Sim} > 0.6$ on ResGen). This validates our hypothesis that LLMs are better reasoned as “chemical editors” rather than “chemical translators” for this task.

2.3.2 Importance of Few-Shot Retrieval (RAG)

The comparison between 5-shot, 1-shot, and Zero-shot configurations highlights the necessity of in-context learning.

- **Contextual Grounding:** The “5-shot” configuration consistently outperforms the “1-shot” configuration, suggesting that providing diverse examples of valid chemical transformations helps the model generalize better to unseen scaffolds.
- **Guidance against Mode Collapse:** The poor performance of the Zero-shot baseline at high similarity thresholds suggests that without retrieving relevant “Synthesis Cliff” pairs, the model lacks the specific knowledge required to perform surgical repairs. Instead, it defaults to generating training-set-like synthesizable molecules, leading to the “structural drift” observed in the data.

In summary, the ablation study provides strong empirical evidence that the combination of **generative editing** and **retrieval-augmented inference** is essential for solving the synthesizability optimization problem effectively.

2.4 Random-RAG Ablation and LLM-as-Judge Alignment Scoring

To test how much of SynCraft’s behaviour comes from the similarity retrieval as opposed to the LLM’s own chemistry knowledge, we ran a controlled random-RAG ablation: the same Gemini-2.5-Pro inference pipeline was re-run on the full P2M and ResGen test pools with exemplars drawn uniformly at random from the reference corpus (matched in count and length; random seed fixed), keeping every other aspect of the prompt and the downstream SimpRetro+Enamine validation identical. Both per-source rescue rate and an independent LLM-as-judge alignment scoring on a 300-query paired subset are summarised in Table 4.

Condition (Gemini-2.5-Pro)	Pass-k	P2M Sim>0.5	ResGen Sim>0.5	Judge alignment (0–5)
Similarity-RAG (5-shot, headline)	5	42.7%	44.7%	2.28
Similarity-RAG (5-shot, matched-passk)	1	38.5%	44.7%	—
Random-RAG (5-shot)	1	22.7%	21.0%	1.76
Zero-shot direct-SMILES (no retrieval)	1	23.6%	26.4%	—

Table 4: Per-source rescue rate at $\text{Sim} > 0.5$ (Gemini-2.5-Pro, SimpRetro+Enamine validation) and mean LLM-as-judge alignment score (300 paired queries, four-axis rubric scoring whether the reasoning chain genuinely uses the retrieved exemplars). The similarity-RAG headline row reproduces the main-text Table 1 number (pass-k = 5); the matched-pass-k row is taken from the ablation Table 2 / 3; random-RAG and zero-shot were evaluated at pass-k = 1.

Two observations. First, at matched pass-k = 1, similarity-RAG clearly outperforms random-RAG (P2M 38.5% vs. 22.7%; ResGen 44.7% vs. 21.0%), and the independent LLM-as-judge alignment score moves in the same direction (Mann–Whitney $U = 53,123$, $p = 3.1 \times 10^{-5}$). The headline pass-k = 5 row (42.7% / 44.7%) is shown for reference. Second, random-RAG performance is essentially at the zero-shot baseline (P2M: 22.7% vs. 23.6%; ResGen: 21.0% vs. 26.4%), indicating that random exemplars do not function as a generic in-context primer on the synthesis-cliff task. The retrieval step therefore contributes specifically when it surfaces structurally similar pairs. We read both components as necessary rather than one being dominant: the LLM provides the chemistry baseline (without it the framework

cannot propose valid edits at all) and similarity retrieval provides the lift that distinguishes the deployed pipeline from a zero-shot prompt. On P2M, retrieval roughly doubles the per-source rescue rate over the zero-shot baseline (22.7% with random exemplars rises to 38.5% at matched pass-k = 1, and 42.7% at the published pass-k = 5) [7]. A chemically unrelated exemplar fails to deliver this lift, suggesting that the teaching signal an exemplar carries is the specific atom-level edit converting an unsynthesizable scaffold into a near-neighbour synthesizable one rather than a generic "edit sequences look like this" template.

2.5 Edit-Path Audit: LLM-Direct Edit Generation

To assess whether SynCraft’s reliance on graph-matching-derived edit ground truth constrains the framework, we sampled $n = 100$ cliff pairs from the corpus and gave DeepSeek-V4-Pro the source SMILES, the target SMILES, and the action space, prompting it to produce an edit sequence directly, bypassing graph matching entirely.

All 100 cases produced parseable JSON. 99/100 reconstructed to a valid product graph; 1/100 failed reconstruction (counted as a failure with similarity 0). 19/100 (19%) of LLM-direct edit sequences exactly match the graph-matching ground-truth edit list; only 1/100 (1%) of canonical SMILES exactly match (stereochemistry differences). The Morgan/Tanimoto similarity of LLM-direct reconstructions to the graph-matching target is mean 0.980 (scoring the 1 failure as 0), median 1.000, with 97/100 reaching similarity ≥ 0.9 . The mean LLM-direct edit length (6.9 operations) is similar to the graph-matching length (6.5).

Interpretation. The graph-matching path is one valid edit sequence connecting source to target, not a unique ground truth. The LLM finds different but chemistry-equivalent edit sequences when given both endpoints, and reaches essentially the same product. This validates the training-corpus design: graph matching serves as a deterministic canonicalization for training stability, not as the only possible correct edit path.

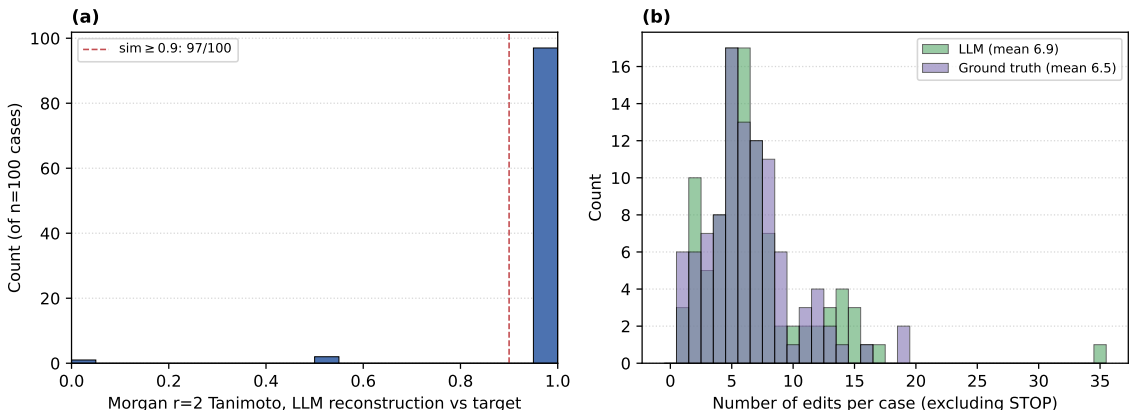


Figure 4: Edit-path audit ($n=100$). (a) Morgan-radius-2 Tanimoto similarity between the LLM-direct edit reconstruction (DeepSeek-V4-Pro given source SMILES, target SMILES, and the SynCraft action vocabulary) and the graph-matching ground-truth target. 97/100 reconstructions reach similarity ≥ 0.9 ; one row failed reconstruction and is scored at similarity 0. (b) LLM edit-count (excluding the terminal STOP token) versus graph-matching ground-truth edit count for the same 100 pairs.

2.6 Edit-Quality Audit

Per-LLM execution success rates. We separately track three orthogonal execution stages for each (source, pass) attempt: (i) the LLM produces a parseable JSON edit list; (ii) the edit list, when applied to the source via `reconstruct_from_sequence`, yields a valence-consistent product graph; (iii) the resulting product passes SimpRetro within the 30-min budget. Stages (i) and (ii) are reported in Table 5; stage (iii) feeds the per-source rescue rates in Table 1 (main text) and the failure-mode decomposition in Section 8.

Table 5: Per-LLM execution success rates for the four-LLM comparison. Computed over the full P2M source pool (1,715 sources) at the indicated pass-k; ResGen-pool numbers are within 1–2 percentage points of these and are not reported separately. JSON parse is the fraction of LLM calls returning a parseable JSON edit list; reconstruction is the fraction of parseable-JSON outputs that yield a valence-consistent product graph after RDKit sanitisation.

LLM	pass-k	JSON parse	Reconstruction
Gemini-2.5-Pro	5	100.0%	66%
GPT-5.2	1	99.6%	60%
DeepSeek-V4-Pro	5	98.3%	63%
DeepSeek-V4-Flash	5	100.0%	64%

Headline observations.

- All four LLMs produce parseable edit-output JSON in $\geq 98\%$ of attempts; the framework does not need format-recovery scaffolding.
- Reconstruction success rate (60–66%) is similar across LLMs; the dominant failure mode at this stage is a proposed bond/atom configuration that fails RDKit’s valence check, not LLM-specific output corruption.
- After SimpRetro validation, the per-source rescue rate at $\text{Sim} > 0.5$ is 39–65% depending on LLM and source pool (Table 1, main text); the failure-mode decomposition by post-retro tier is in Section 8.

Aggregate drug-likeness deltas. On the union of all valid SynCraft-rescued products for the full P2M and ResGen source pools, we compute the best-of-K ($K=5$) chemistry property delta per source between the source molecule and its best-Sim retro-solvable edit. Table 6 summarises the aggregate trend.

Table 6: Aggregate pre/post drug-likeness deltas across the SynCraft rescue corpus. Best-of-K=5 per source; reported as mean of per-source values. SA is the Ertl & Schuffenhauer synthetic accessibility score (lower is better); QED is the Bickerton *et al.* quantitative estimate of drug-likeness (higher is better).

Source pool	Metric	Source (mean)	SynCraft best edit (mean)
Pocket2Mol (1,715 sources)	SA	3.65	3.41 (−0.24)
	QED	0.554	0.583 (+0.029)
ResGen (1,025 sources)	SA	3.51	3.30 (−0.21)
	QED	0.560	0.587 (+0.027)

SA score and QED both move in the favourable direction on both source pools, indicating that SynCraft edits aggregately improve synthesizability without sacrificing drug-likeness. Binding-mode preservation is the orthogonal axis, addressed by the systematic PLIP retention analysis in Section 7.

2.7 Golden-Corpus Reasoning Validation

A separate concern is the quality of the reasoning rationales attached to the 3,332 synthesis-cliff pairs in our reference corpus. These rationales are themselves LLM-generated (post-hoc Gemini-2.5-Pro narrative on extracted graph-matching edit paths) and are subsequently used as in-context exemplars during inference. If the rationales are noisy, the RAG-based inference might inherit that noise.

To validate corpus quality, we deploy a four-axis LLM-as-judge protocol scoring each reasoning chain on: (i) bond-changes-match (does the rationale reference the actual bond changes in the edit?), (ii) liability-exists (does it identify a genuine synthetic liability?), (iii) causal-coherence (is the proposed fix logically connected to the identified liability?), (iv) novel-feature-present (does the edit introduce a chemically distinguishable change?). The judge is a different LLM from the one that generated the rationales (DeepSeek-V4-Pro), reducing self-reference bias.

Results. On the full corpus ($n = 3,032$ chains for which the judge returned a valid four-axis verdict) we report two metrics under permissive (*yes-or-partial*) scoring, where the *partial* tier captures chains that are mostly correct on an axis but elide a sub-detail. Under this rule, per-axis pass rates are 83% (bond-changes-match), 96% (liability-exists), 86% (causal-coherence), and 74% (new-feature-present); **82.5% of chains pass at least three of the four axes** and 68% pass all four. To validate that the judge is discriminating rather than rubber-stamping, we constructed a paired negative-control set ($n = 100$) where the rationale text was shuffled to a non-matching source-target pair. The negative control yields 0% on both pass-all-four and pass- ≥ 3 -of-4, and 0–12% per individual axis (Figure 5), confirming that the judge can distinguish coherent from incoherent rationales.

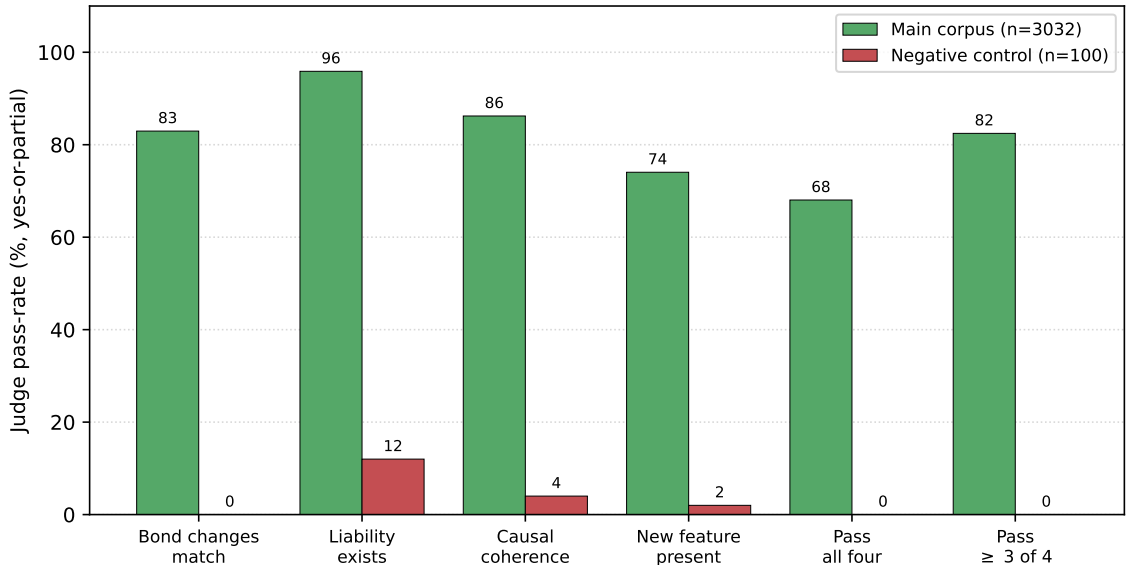


Figure 5: LLM-as-judge alignment on the SynCraft golden reasoning corpus ($n=3,032$) versus a paired negative-control set ($n=100$) where the rationale is shuffled to a non-matching source-target pair. Per-axis pass rate (yes-or-partial scoring) for the four-axis rubric (*bond-changes-match*, *liability-exists*, *causal-coherence*, *new-feature-present*) and the two aggregate criteria (*pass ≥ 3 of 4*, *pass all 4*). The judge clearly discriminates: it rubber-stamps neither the main corpus nor the shuffled control.

Interpretation. The chains that fall below the four-of-four threshold typically omit an explicit statement of one of the rubric elements (most often the *new-feature-present* sub-claim) rather than containing factually incorrect chemistry, consistent with the verdict that only $\sim 18\%$ of chains miss two or more axes even under the strictest reading. We do not filter the corpus to the judge-passed subset for inference, on the grounds that (i) the typical failure is narrative incompleteness rather than chemical error; (ii) RAG averages over $k = 5$ exemplars per query, providing inherent noise robustness; and (iii) the published rescue rates are produced *from this corpus* and thus reflect realistic deployment performance.

3 Prospective Optimization Results for All 42 High-Scoring RIPK1 Candidates

To provide a comprehensive and unbiased assessment of SynCraft’s utility in real-world lead optimization scenarios, we report the complete optimization results for the 42 high-scoring RIPK1 candidates described in the main text. These candidates were originally identified from a generative library targeting the RIPK1 kinase but were discarded due to structural complexity or lack of viable synthetic pathways. SynCraft was tasked with optimizing these molecules using the interaction-aware prompting strategy (PDB ID: 7YDX) to restore synthesizability while preserving predicted binding affinity.

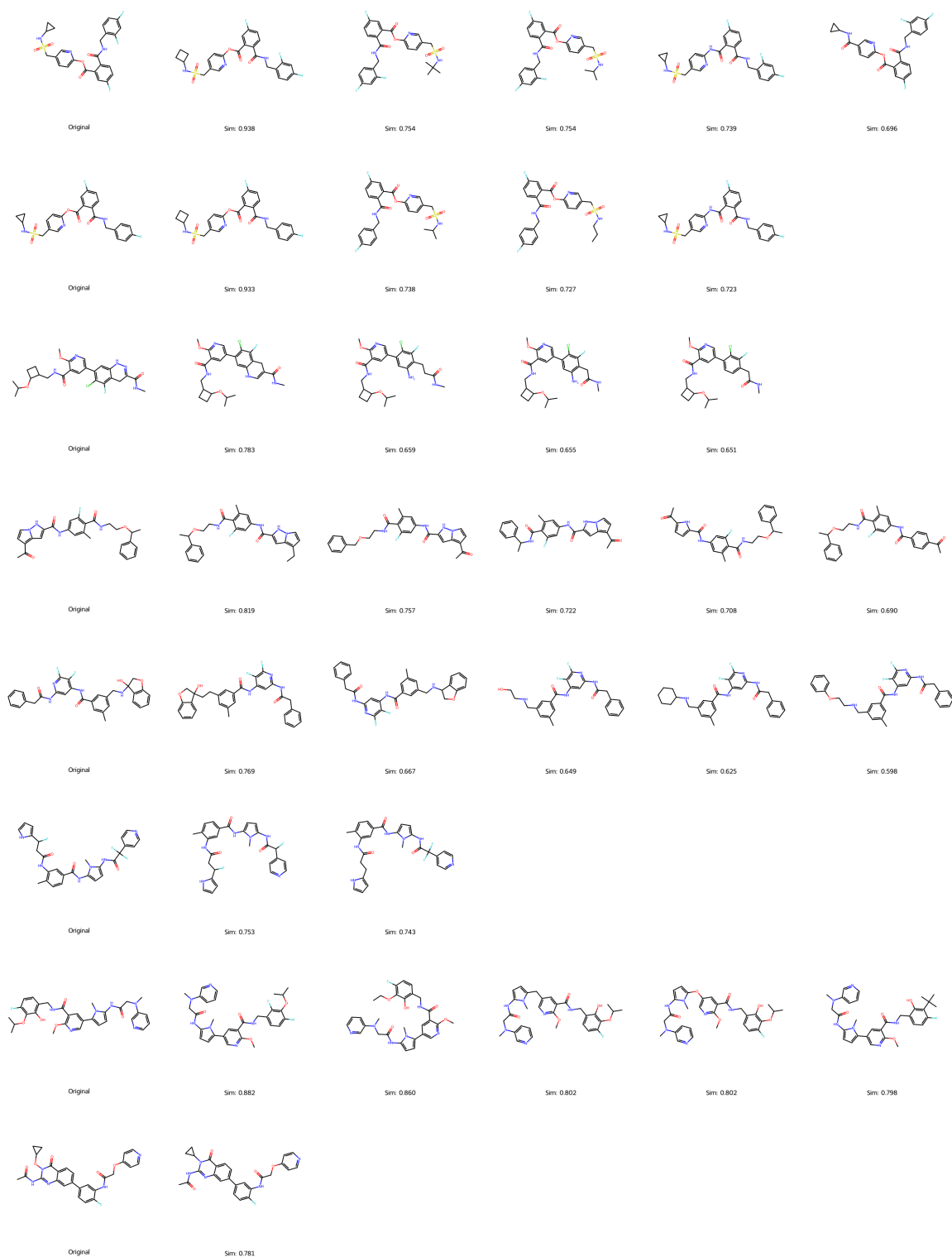


Figure 6: Per-candidate optimization report (1 of 6): RIPK1 sources 1–7 with their SynCraft-edited products, ranked best-of-K by SimpRetro pass status and similarity to source.

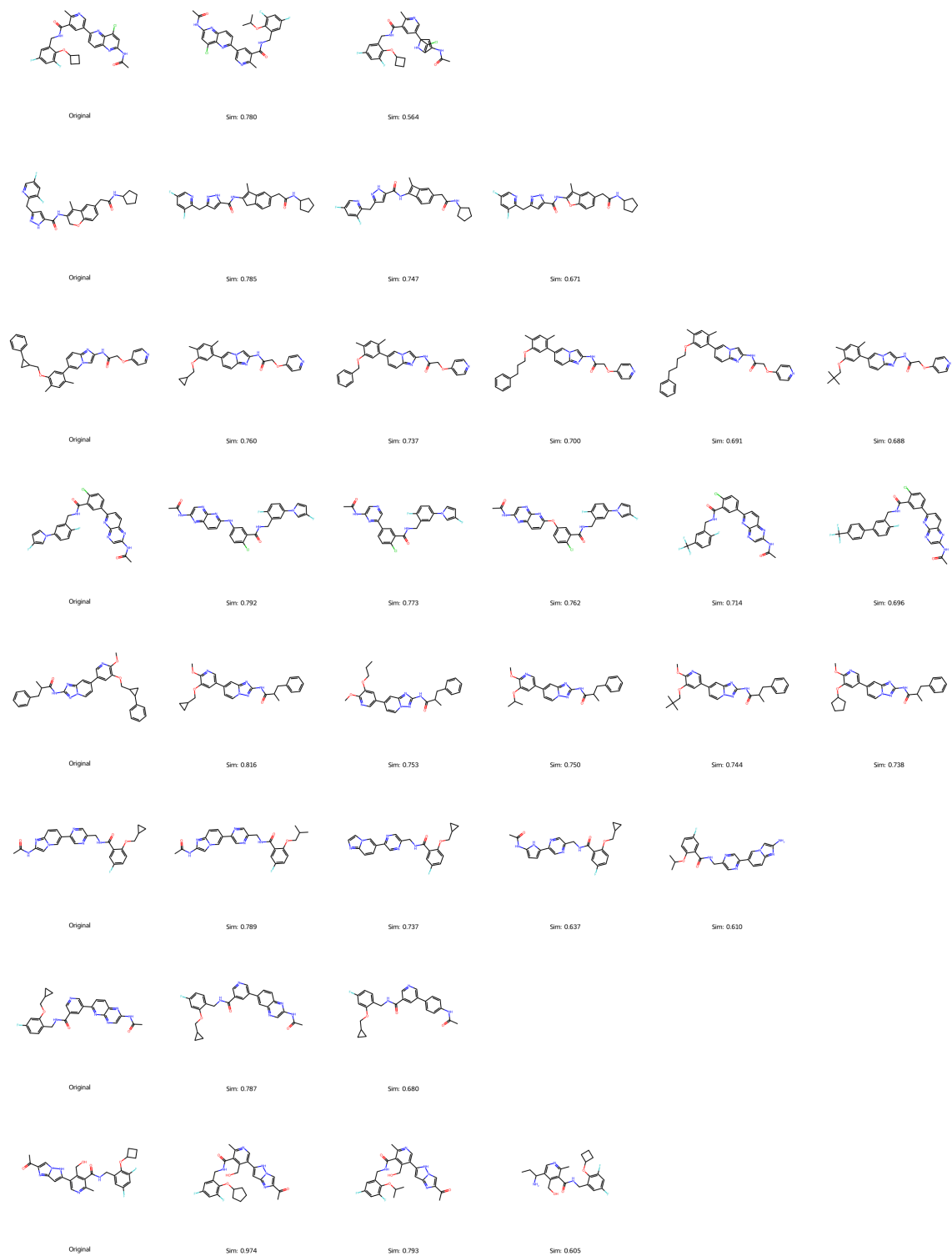


Figure 7: Per-candidate optimization report (2 of 6): RIPK1 sources 8–14.



Figure 8: Per-candidate optimization report (3 of 6): RIPK1 sources 15–21.

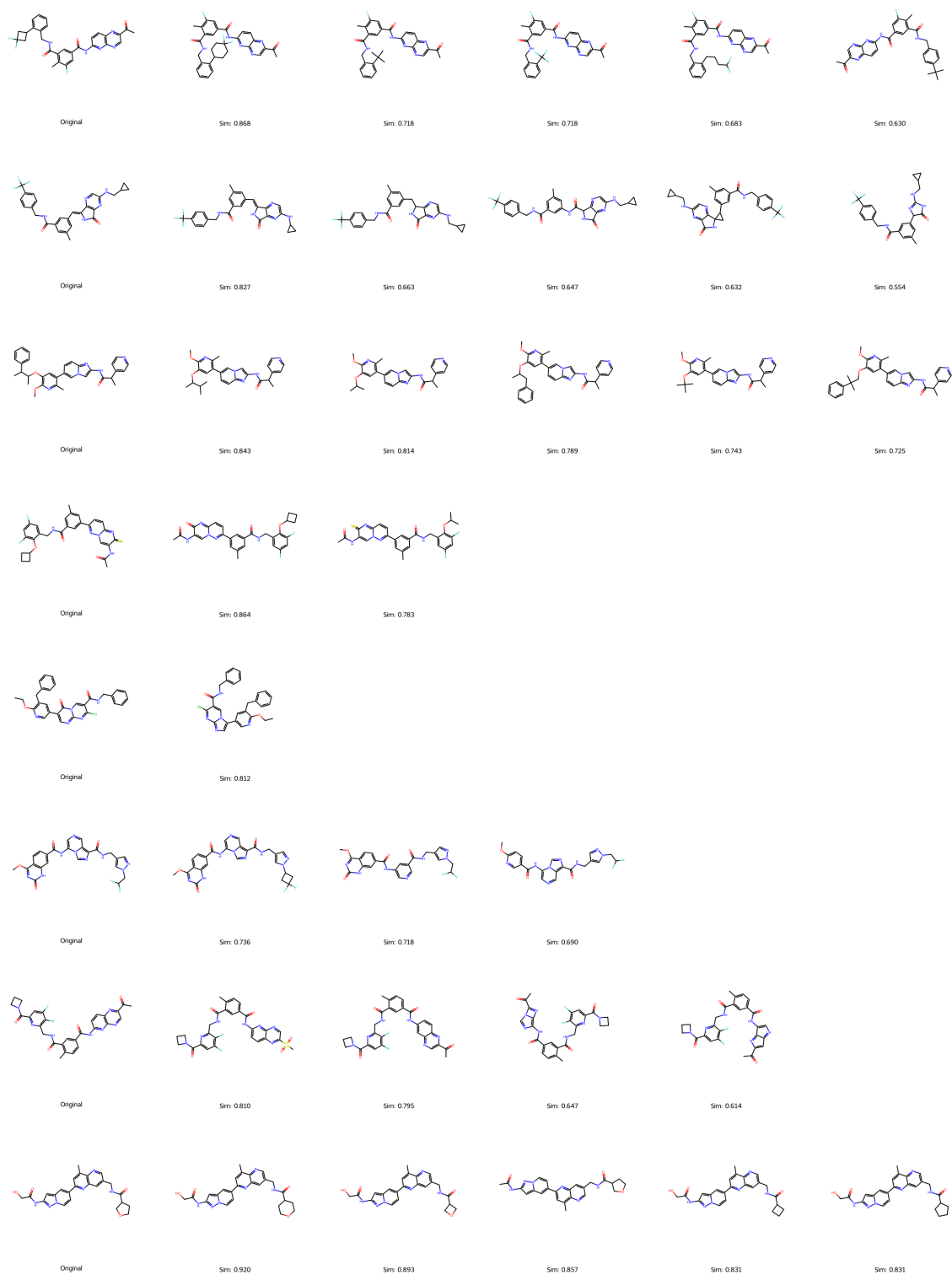


Figure 9: Per-candidate optimization report (4 of 6): RIPK1 sources 22–28.

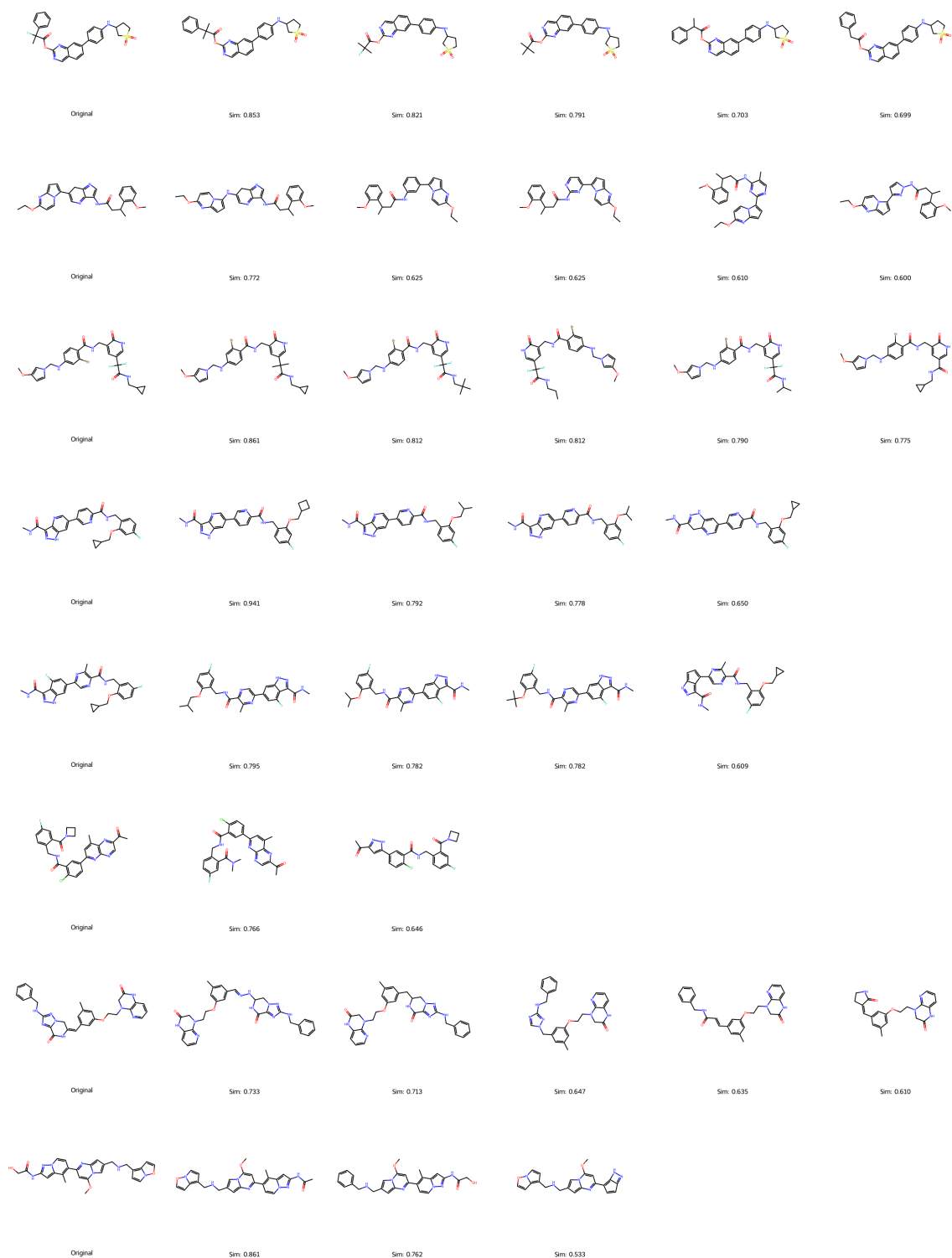


Figure 10: Per-candidate optimization report (5 of 6): RIPK1 sources 29–35.

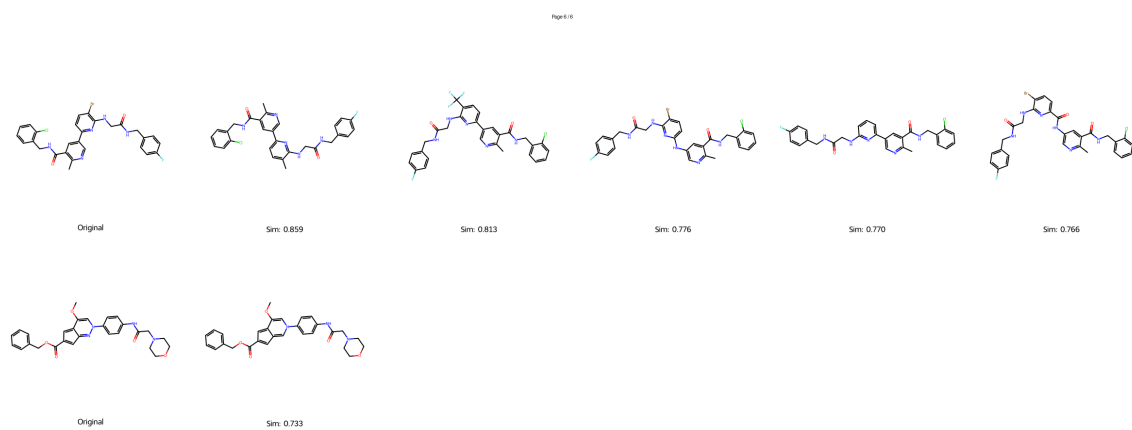


Figure 11: Per-candidate optimization report (6 of 6): RIPK1 sources 36–42.

4 Template-Inventory Scaling Experiment

To test whether the rescue-rate advantage of SynCraft over projection-based methods reflects a difference in available reaction-template inventory rather than a paradigm difference, we retrained SynFormer [5] on the full USPTO 51,720-template superset, an enlarged inventory compared with the curated 115-template library used in the original publication.

Training setup. We retained the published SynFormer architecture and loss formulation. The reaction-classification head was expanded from 120-way (115 templates + 5 reserved indices in the published checkpoint) to 51,725-way (51,720 USPTO templates + the same 5 reserved indices) to cover the full USPTO template superset. Training proceeded for 30,000 steps on four RTX 4090 GPUs (batch size 196, AdamW, learning rate $1e-4$). Validation was computed every 2,000 steps; checkpoints were saved every 2,000 steps with save-top-k=3 by val/loss.

Evaluation. We evaluated the final checkpoint (step 30,000) on the same Pocket2Mol and ResGen unsynthesizable test sets used in Table 1, with REPEAT=10 best-of-K per source molecule. Following the published Table 1 protocol, the metric is the success rate of finding any valid analog at the specified Tanimoto similarity threshold.

Result. The enlarged-template SynFormer-51K achieves the rescue rates shown in Table 7.

Table 7: SynFormer trained on USPTO 51,720 templates, evaluated on the same P2M and ResGen unsynthesizable test sets used in main-text Table 1 (REPEAT=10).

Configuration	Sim > 0.5	Sim > 0.6	Sim > 0.7	Sim > 0.8
P2M-SynFormer (115 templates, published)	30.7%	14.1%	6.2%	2.9%
P2M-SynFormer (51,720 templates, retrained by us)	5.6%	3.0%	1.2%	0.8%
ResGen-SynFormer (115 templates, published)	33.8%	19.9%	8.7%	3.5%
ResGen-SynFormer (51,720 templates, retrained by us)	6.6%	3.3%	1.5%	0.8%

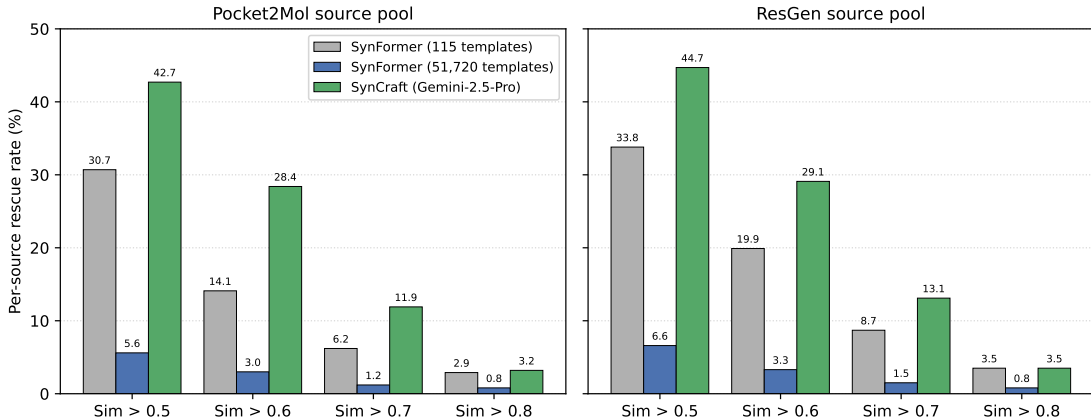


Figure 12: Template-inventory scaling does not close the gap to SynCraft. Per-source rescue rate at increasing Tanimoto similarity thresholds for the published 115-template SynFormer, the same architecture retrained on the full USPTO 51,720-template superset under matched compute, and SynCraft (Gemini-2.5-Pro headline, pass-k=5), on the P2M (left) and ResGen (right) source pools. Numbers correspond to Table 7 and main-text Table 1.

Enlarging the template inventory by the order of magnitude shown above does not improve the projection model’s rescue rate relative to the published 115-template baseline (which uses a curated,

Enamine-compatible subset). Rather than an artifact of insufficient training, several factors plausibly contribute to this outcome:

- **Per-template signal sparsity.** The USPTO 51K-template superset includes long-tail patent-extracted templates that each appear in only tens of reactions in the training corpus; this is two to three orders of magnitude sparser than the per-template signal in the curated 115-template set, which was specifically constructed for high-frequency Enamine-compatible chemistry.
- **Reaction-building-block coverage.** In the matched matrix mapping reaction templates to compatible building blocks, median BB coverage per template is approximately $15\times$ sparser for the 51K set compared with the 115-template set, reducing the likelihood that a chosen template admits a valid retrieval among Enamine commercial inventory.
- **Compounded multi-step error.** At each retrosynthesis step, the model selects among 51,720 reaction templates rather than 115; even modest per-step calibration loss compounds multiplicatively along a tree of depth 3–5.
- **Curated-vs-automatic template quality.** The 115-template baseline was hand-curated for synthetic utility and Enamine compatibility; the remaining 51,605 USPTO templates are dominated by patent one-off conditions, near-duplicates with minor SMARTS perturbations, and reactions with no commercial building-block coverage.

These mechanisms suggest that scaling reaction-template inventory is not a viable route to closing the SynCraft–SynFormer gap. SynCraft’s advantage arises from formulating the rescue task as a local edit on the existing molecular graph rather than as a global reconstruction from atomic building blocks; this formulation more closely matches the locally compositional inductive bias of LLMs trained on natural-language tokens.

5 End-to-end Hit Recovery Demonstration on SARS-CoV-2 M^{pro}

This section provides the detailed protocol and results for the main-text Section “End-to-end Hit Recovery from a 3D-Generative Model”. The demonstration uses the off-the-shelf TargetDiff [8] 3D pocket-aware diffusion model to generate ligand candidates for the SARS-CoV-2 main protease (M^{pro}, PDB ID: 7L12), and applies the SynCraft Agent Skill end-to-end.

Source pool and source quality. We sampled 47 candidate molecules generated by TargetDiff for the M^{pro} 7L12 pocket. Of these, only 1 is directly synthesizable under our SimpRetro protocol (USPTO templates with frequency ≥ 2 , Enamine building blocks, 30-min CPU budget per molecule). The remaining 46 candidates are the targets for SynCraft rescue.

Skill API specification. The SynCraft Agent Skill exposes a single function, `edit_for_synthesizability`, that internally orchestrates four stages: (i) Vina docking of each source against the target receptor, (ii) PLIP interaction profiling of each docked pose, (iii) RAG-based LLM edit prediction (the LLM is given the source SMILES, retrieved exemplars, and either no constraint or the PLIP profile as a structural constraint), (iv) Vina re-docking of edited products. The skill exposes two modes and a *union* default ranking:

- **Rescue-maximizing mode** (formerly “generic edit”): no PLIP constraint; the LLM is free to make any edit that improves synthesizability.
- **Interaction-preserving mode:** PLIP-derived natural-language constraints are added to the LLM’s prompt instructing it to preserve atoms participating in critical interactions.

Per-source ranking is by SA score (lowest first) among edits with Tanimoto similarity > 0.2 to source and Vina loss ≤ 1 kcal/mol.

Per-mode results. We invoked the skill on all 47 actionable M^{pro} candidates with pass-k=10 per mode. Results are summarized in Table 8.

Table 8: Per-mode statistics of SynCraft modes on the 47 actionable TargetDiff M^{pro} candidates (46 requiring rescue; 1 directly synthesizable was excluded from the rate denominators). Under the default ranking that unions the two modes, 32/46 sources (70%) are rescued.

Metric	Rescue-max	Interaction-preserving
Per-source rescue rate (SimpRetro-validated)	63%	37%
Per-product retro solve rate	23%	16%
Mean Δ SA (best edit – source)	−0.75	−0.77
Mean Δ QED	+0.027	+0.122
Ro5 + Veber pass (best edit)	29%	54%
$\geq 50\%$ PLIP retention (best edit)	—	79%

The two modes are complementary: the rescue-maximizing mode produces more synthesizable rescues, while the interaction-preserving mode produces fewer but binding-mode-preserving rescues. The lower per-source rescue rate of the interaction-preserving mode is a direct consequence of the added PLIP constraint: requiring the LLM to keep atoms that participate in key receptor contacts unchanged removes the most aggressive single-edit fixes (whole-scaffold swaps, ring-contraction across an interacting fragment) from the action space, and the remaining edits are correspondingly less effective at clearing the synthesizability bottleneck. The two modes’ retro-validated edit-product sets are entirely disjoint (intersection of zero molecules), but individual source candidates can be rescued by both modes via different edit products; the per-source rescue percentages of the two modes therefore overlap at the source level. The union yields 70% per-source rescue (32 of 46 unsynthesizable sources); counting the one directly-synthesizable source and the 99 unique SynCraft-edited retro-validated products, the rescue pool comprises 100 retrosynthetically validated analogs.

Comparison to source. The agent-skill Vina-distribution analysis on the 47-source M^{Pro} pool is reported in main-text Figure 5: pairing each rescued source to its best retro-validated edit under the default union ranking ($n = 30$ of the 32 union-rescued sources with paired docking data) yields a mean ΔV_{ina} of +0.14 kcal/mol (median +0.17); 60% (18/30) of paired edits fall within ± 0.5 kcal/mol of the source and 93% (28/30) within ± 1.0 kcal/mol, comfortably inside the typical Vina reproducibility envelope.

Representative edit examples. Figure 13 shows three retrosynthetically validated SynCraft edits drawn from the rescued M^{Pro} pool, chosen to span low- to high-complexity edits (3, 4 and 12 structural operations) while covering a representative range of source-to-edit similarities (0.90, 0.63, 0.58). All three picks belong to the rescue-maximising mode under DeepSeek-V4-Pro inference. The first pick (Sim=0.90, 3 ops, $\Delta SA -0.35$) replaces two remote chiral secondary alcohols with simpler methylene linkages, eliminating two stereocentres without disturbing the surrounding scaffold. The second pick (Sim=0.63, 4 ops, $\Delta SA -0.76$) converts a non-aromatic cycloheptatriene ring attached via a hydrazine linker into a flat, Pd-cross-coupling-accessible chloroanilino group. The third pick (Sim=0.58, 12 ops, $\Delta SA -1.29$) demonstrates the model’s capacity for larger-scale restructuring, dissolving a complex multi-heteroatom heterocycle into a simple open-chain glycerol fragment while preserving the benzoic-acid pharmacophore and the pyridinol ether linkage.

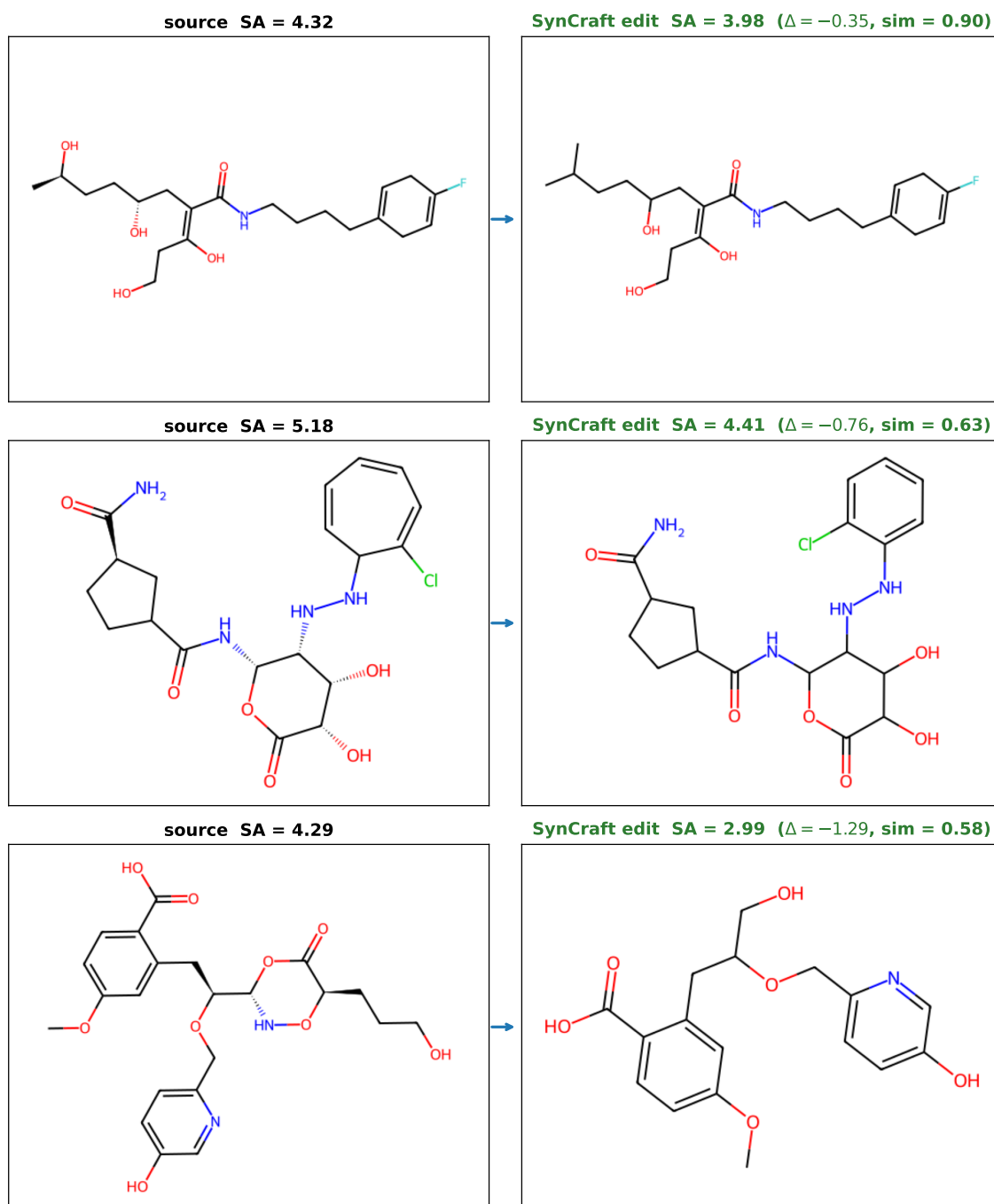


Figure 13: Representative SynCraft-edited rescue products from the M^{Pro} skill demonstration. Three retrosynthetically validated edits chosen to span edit complexity (3, 4 and 12 structural operations) and source-similarity range (0.90, 0.63, 0.58). Each row shows the source candidate (left), the proposed edit operations (annotated arrows), and the retro-validated SynCraft edit (right). Δ SA values: pick 1 -0.35 , pick 2 -0.76 , pick 3 -1.29 .

6 Coverage Across Additional 3D-Generative Baselines

To address whether SynCraft’s advantage generalises beyond P2M and ResGen (the two generative baselines highlighted in main-text Table 1), we ran SynCraft (DeepSeek-V4-Pro, pass-k=1) together with the three projection-based baselines (ChemProjector, SynFormer, ReaSyn) on stratified 10% subsets of the three additional 3D-aware generative baselines from the GenBench3D benchmark: LiGAN, 3D-SBDD (AR), and DiffSBDD. All four methods are evaluated under the same SimpRetro+Enamine protocol used in main-text Table 1. Subsets were stratified by heavy-atom count with seed=42; sizes are 372 / 365 / 383 source molecules for LiGAN / AR / DiffSBDD.

Table 9: Per-source rescue rate at increasing Tanimoto-similarity thresholds (ECFP4 r=2, 4096-bit) on 10% stratified subsets of LiGAN, AR, and DiffSBDD. SynCraft uses DeepSeek-V4-Pro at pass-k=1; projection-based baselines use their default published settings. ChemProjector numbers are restricted from each baseline’s full pool to the 10% subset.

Source pool	Method	Sim>0.5	Sim>0.6	Sim>0.7	Sim>0.8
LiGAN (n=372)	ChemProjector	1.9%	0.8%	0.0%	0.0%
	SynFormer (115 templates)	11.6%	6.2%	3.8%	1.1%
	ReaSyn	14.0%	9.1%	5.9%	2.7%
	SynCraft (DSv4-Pro)	25.5%	17.5%	6.7%	1.3%
AR / 3D-SBDD (n=365)	ChemProjector	10.1%	5.2%	1.9%	0.8%
	SynFormer (115 templates)	7.7%	3.8%	1.9%	0.3%
	ReaSyn	14.5%	6.3%	3.0%	1.6%
	SynCraft (DSv4-Pro)	26.3%	17.8%	8.2%	1.9%
DiffSBDD (n=383)	ChemProjector	6.5%	2.6%	0.5%	0.0%
	SynFormer (115 templates)	6.8%	2.6%	1.6%	0.8%
	ReaSyn	25.6%	10.7%	3.9%	1.6%
	SynCraft (DSv4-Pro)	38.1%	26.1%	14.1%	3.7%

SynCraft (pass-k=1) leads the strongest projection baseline at every Sim>0.5 cell by margins of 11 pp (LiGAN) to 23 pp (DiffSBDD), and is first at 11 of the 12 (source pool, threshold) cells overall. The one exception is LiGAN at Sim>0.8, where ReaSyn (2.7%) edges out SynCraft (1.3%); the four-method success count at this threshold is in single digits, so the ranking is dominated by sampling noise. The ranking among projection methods is non-monotonic across baselines (ChemProjector is competitive on AR but very weak on LiGAN; ReaSyn is the strongest projection method on DiffSBDD; SynFormer is intermediate across the board), whereas SynCraft is the strongest method on every baseline at Sim>0.5–>0.7. Combined with the published main-text Table 1 results for P2M and ResGen, SynCraft is therefore the leading method across all five 3D-generative source distributions.

The pre-screened actionable pool from each baseline differs in chemical character (LiGAN and AR have higher mean SA scores than P2M/ResGen, and contain more chemistry artifacts), and the projection methods correspondingly perform much worse on LiGAN/AR than on the cleaner P2M/ResGen pools. SynCraft’s edit-based paradigm degrades less under this source-quality variation, which we attribute to its ability to identify and locally repair the specific synthesizability bottlenecks in each source rather than searching the full synthesizable chemical space from scratch.

7 Systematic Binding-Mode Retention via PLIP

Beyond the qualitative case studies in main-text Figures 3 and 4, we report a systematic, large-N quantification of binding-mode preservation across two unrelated target families: the RIPK1 kinase domain (Fig. 4 follow-up) and the SARS-CoV-2 M^{Pro} cysteine protease (main-text §End-to-end Hit Recovery + SI Section 5 follow-up).

Retention metric. For each source–edit pose pair, we run PLIP independently on both docked complexes, then represent each interaction profile as the set of (residue identifier, interaction type) tuples (e.g., (*LEU145*, *hydrophobic*), (*HIS41*, *hbond*), ...). Interaction types tracked: hydrogen bond, hydrophobic contact, π -stacking, salt bridge, π -cation, halogen bond, water bridge. We then compute two retention scores per pair:

- **Recall vs source:** $|S \cap E|/|S|$, the fraction of source PLIP interactions that are reproduced in the edited pose.
- **Jaccard:** $|S \cap E|/|S \cup E|$, the symmetric overlap.

where S and E are the source and edit interaction sets respectively. We summarize per-source by taking the best (maximum-recall) edit per source, then report the fraction of sources whose best edit achieves recall ≥ 0.5 , ≥ 0.7 , or $= 1.0$.

Results. The retention statistics across the two targets are summarized in Table 10.

Table 10: Systematic PLIP-based binding-mode retention across two unrelated targets. PLIP-retention rows use the interaction-aware (PLIP-constrained) skill mode (DeepSeek-V4-Pro, pass-k=10). M^{Pro} Vina rows correspond to the default-skill (union) ranking; the per-source pairs are reduced to $n = 30$ because two of the 32 union-rescued M^{Pro} sources did not yield a paired Vina docking score (ligand-preparation failure on the edit side).

Metric	RIPK1 (7YDX)	M ^{Pro} (7L12)
Source–edit pose pairs analyzed	191	260
Number of source candidates	42	47
Mean per-pair recall vs source	0.47	0.42
Median per-pair recall vs source	0.43	0.40
Mean per-pair Jaccard	0.33	0.28
Sources with best-edit recall ≥ 0.5	34/42 (81%)	37/47 (79%)
Sources with best-edit recall ≥ 0.7	21/42 (50%)	19/47 (40%)
Sources with best-edit recall = 1.0	11/42 (26%)	2/47 (4%)
Mean Δ Vina (best edit – source, kcal/mol)	−0.34	+0.14
Sources with best edit Vina $\leq$ source	31/42 (74%)	11/30 (37%)
Sources with best edit Vina within 0.5 kcal/mol	37/42 (88%)	18/30 (60%)

In both targets, the majority of source candidates have at least one SynCraft-edited analog that retains $\geq 50\%$ of the source’s PLIP interaction profile. For RIPK1, the best edit additionally improves binding affinity on average (mean Δ Vina −0.34 kcal/mol); the corresponding M^{Pro} default-skill Vina analysis is reported in main-text Figure 5, where the per-source best retro-validated edit reproduces the source’s binding score to within ± 0.5 kcal/mol for 60% of pairs. The M^{Pro} and RIPK1 targets are from entirely unrelated protein families (cysteine protease vs serine/threonine kinase), and PLIP-retention is consistent across both, supporting the conclusion that binding-mode preservation generalizes rather than being target-specific.

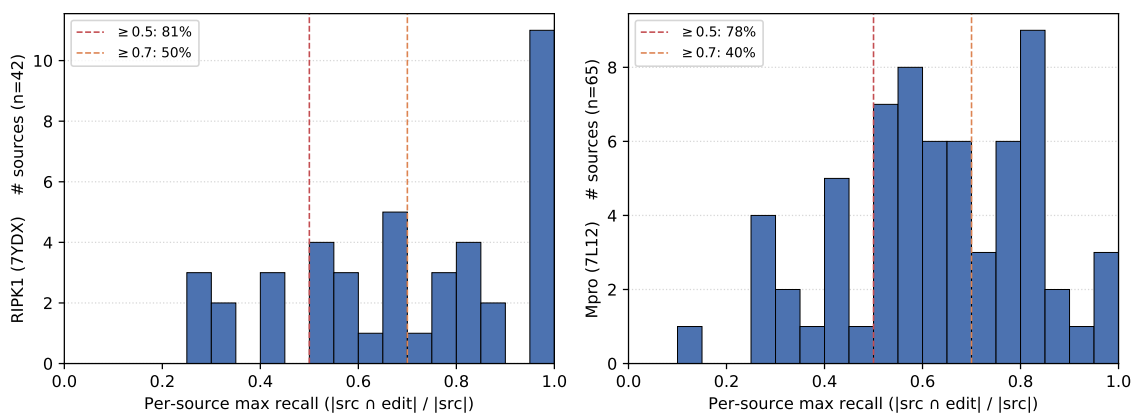


Figure 14: Systematic PLIP-based binding-mode retention. Per-source distributions of the maximum interaction-recall of source PLIP interactions reproduced in any SynCraft-edited product, shown for the RIPK1 kinase (PDB 7YDX; left, 42 sources, 191 source–edit pose pairs) and SARS-CoV-2 M^{pro} (PDB 7L12; right, 65 sources with at least one valid docked source–edit pose pair, 260 pairs). Vertical lines mark the 0.5 and 0.7 recall thresholds. The M^{pro} histogram is drawn over all 65 docked sources, whereas the source-level pass rates in Table 10 are reported on the 47 actionable candidates.

8 Failure-Mode Decomposition

To characterize the failure modes of the SynCraft rescue pipeline, we decompose each source candidate into one of four mutually exclusive outcomes, evaluated across all k inference attempts that source receives (where k is the pass-k of the LLM):

- **Success:** at least one pass produces a valid product that is SimpRetro-solvable AND has Tanimoto similarity (ECFP4, 4096-bit) ≥ 0.5 to source. This is exactly Table 1’s “Sim > 0.5 ” per-source rescue criterion.
- **Sim drift:** at least one pass produces a retro-solvable product, but *every* such product has sim < 0.5 to source; the rescue is structurally too distant.
- **Retro fail:** at least one pass produces a valid (valence-consistent) reconstruction, but no product is SimpRetro-solvable within the 30-minute budget across all passes.
- **Recon fail:** no pass yields a valid reconstructed product (the LLM’s edit list either fails to parse, references an invalid atom-map number, or yields a disconnected/invalid graph).

The per-LLM breakdown across all four backends and both source pools is shown in Figure 15 and Table 11. The four categories form a per-source partition (every source falls into exactly one), so the row sums to 100%; the Success column matches the Sim > 0.5 row of main-text Table 1 for each LLM/pool cell.

Table 11: Per-source failure-mode decomposition for all four LLM backends on the P2M and ResGen source pools (visualised in Figure 15). The four categories form a mutually-exclusive per-source partition (Success = main-text Table 1 Sim > 0.5 cell); each row sums to 100%.

Pool	LLM	Success	Sim drift	Retro fail	Recon fail	
P2M	Gemini-2.5-Pro	42.7%	20.0%	26.3%	11.0%	
P2M	GPT-5.2 [‡]	39.0%	17.8%	38.7%	4.5%	
P2M	DeepSeek-V4-Pro [†]	64.3%	13.9%	21.1%	0.7%	
P2M	DeepSeek-V4-Flash [†]	64.9%	17.2%	17.4%	0.5%	[†] MIT-licensed
ResGen	Gemini-2.5-Pro	44.7%	17.6%	27.2%	10.6%	
ResGen	GPT-5.2 [‡]	44.5%	14.8%	37.0%	3.7%	
ResGen	DeepSeek-V4-Pro [†]	65.7%	12.6%	21.5%	0.2%	
ResGen	DeepSeek-V4-Flash [†]	65.2%	14.5%	19.9%	0.4%	

open-weight, pass-k = 5. [‡] GPT-5.2 evaluated at pass-k = 1. Gemini-2.5-Pro at pass-k = 5 (paper headline protocol).

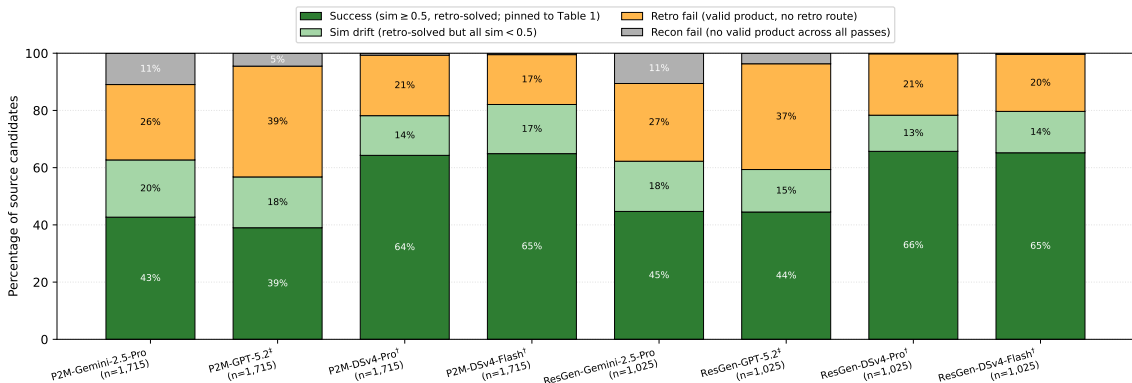


Figure 15: Per-source failure-mode decomposition across the four LLM backends and the two source pools (P2M, ResGen). Each source falls into exactly one of *Success*, *Sim drift*, *Retro fail*, or *Recon fail*; the four bars in each panel sum to 100%. Recon-fail is uniformly small for the DeepSeek backends and GPT-5.2 but noticeably larger for Gemini-2.5-Pro, consistent with the disconnected-fragment behaviour observed in early Gemini-2.5-Flash testing of our pipeline. Sample sizes: $n=1,715$ for Pocket2Mol, $n=1,025$ for ResGen. Each stacked bar decomposes one (LLM $\times$ source-pool) cohort into four mutually exclusive outcomes; no error bars are shown because each source is classified once.

Three structural observations. (i) *Recon-fail is uniformly small for the open-weight DeepSeek backends ($< 1\%$) and small for GPT-5.2 (3–5%), but noticeably larger for Gemini (11%).* The Gemini recon-fail is dominated by edits that reference removed atoms or produce a disconnected fragment; the same behaviour is observed when running the pipeline on the weaker Gemini-2.5-Flash backend. The framework’s downstream sanitisation step explicitly catches these and discards them from the pass-k pool, so they do not contaminate the rescue rate, but they do represent a clear opportunity for LLM-side constraint enforcement. We expect the SynCraft framework to continue benefiting from improvements in the underlying reasoning LLMs without requiring framework-side changes.

(ii) *Retro-fail is the dominant failure mode across all LLMs (17–39% of sources).* The LLM produces a chemistry-reasonable product but no synthesis route is found within the 30-minute SimpRetro budget. This bottleneck is independent of LLM choice and points to retrosynthesis coverage as the next-most-impactful direction for improving headline rescue rates.

(iii) *Sim-drift fractions are similar across LLMs (13–20%),* suggesting that the tendency of an LLM to propose “too distant” rescues when a closer one is not available is a roughly invariant property of the framework rather than an LLM-specific characteristic.

Interpretation: pass-k vs. per-attempt quality. The headline $\text{Sim} > 0.5$ rescue rates in Table 1 differ by up to 22 percentage points across LLMs (e.g., Gemini-2.5-Pro at 42.7% vs. DSv4-Pro at 64.3% on P2M). Two factors contribute. First, pass-k differs across the ablation: DSv4 backends and Gemini were evaluated at $\text{pass-k} = 5$ (consistent with the original SynCraft submission), while GPT-5.2 was evaluated at $\text{pass-k} = 1$ due to API budget. Second, even at matched pass-k, the more recently released open-weight DSv4 backends produce higher per-pass quality. We report the four-LLM comparison neutrally and retain Gemini-2.5-Pro as the headline backend; the multi-LLM evidence supports the claim that SynCraft is framework-agnostic across both proprietary and open-weight LLMs spanning roughly fourteen months of release dates (Section 1.7).

9 eMolecules Synthesizability Under Enamine Building-Block Constraint

To verify that the SimpRetro+Enamine validation protocol used throughout this work does not artificially discount real eMolecules-class compounds, we sampled 4,000 random molecules from the eMolecules database (the same corpus used to construct our synthesis-cliff reference pairs) and ran them through identical SimpRetro validation (USPTO templates with frequency ≥ 2 , Enamine building-block library, 30-minute CPU budget).

Of the 4,000 sampled eMolecules, **3,840 (96.0%)** were successfully solved within the time budget. This confirms that the protocol can recover synthesis routes for the vast majority of commercially-available compounds and that “unsynthesizable” classifications applied to generative-model outputs reflect genuine synthetic infeasibility rather than protocol-induced false negatives.

References

- [1] Benoit Baillif, Jason Cole, Patrick McCabe, and Andreas Bender. Benchmarking structure-based three-dimensional molecular generative models using genbench3d: ligand conformation quality matters. *arXiv preprint arXiv:2407.04424*, 2024.
- [2] Junren Li, Kangjie Lin, Jianfeng Pei, and Luhua Lai. Challenging complexity with simplicity: rethinking the role of single-step models in computer-aided synthesis planning. *Journal of Chemical Information and Modeling*, 64(14):5470–5479, 2024.
- [3] Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, 2012.
- [4] Shitong Luo, Wenhao Gao, Zuofan Wu, Jian Peng, Connor W Coley, and Jianzhu Ma. Projecting molecules into synthesizable chemical spaces. In *International Conference on Machine Learning*, pages 33289–33304. PMLR, 2024.
- [5] Wenhao Gao, Shitong Luo, and Connor W Coley. Generative ai for navigating synthesizable chemical space. *Proceedings of the National Academy of Sciences*, 122(41):e2415665122, 2025.
- [6] Seul Lee, Karsten Kreis, Sriram Prasad Veccham, Meng Liu, Danny Reidenbach, Saeed Paliwal, Weili Nie, and Arash Vahdat. Rethinking molecule synthesizability with chain-of-reaction. *arXiv preprint arXiv:2509.16084*, 2025.
- [7] Yuxin Wang et al. Retrieval-augmented in-context reasoning for synthesis-aware molecular generation. *arXiv preprint arXiv:2505.07027*, 2025.
- [8] Jiaqi Guan, Wesley Wei Qian, Xingang Peng, Yufeng Su, Jian Peng, and Jianzhu Ma. 3D equivariant diffusion for target-aware molecule generation and affinity prediction. In *International Conference on Learning Representations (ICLR)*, 2023.