StructVista: A Benchmark for Molecular Reconstruction and Assembly Reasoning in Scientific Figures

Anonymous authors
Paper under double-blind review

Abstract

Accurate interpretation of molecular structures in scientific figures is fundamental to molecule-centric literature mining and scientific discovery. Existing chemical structure recognition and Markush parsing methods can recover molecules when each molecule is drawn completely or when its components are linked by explicit correspondences. Beyond these settings, many real figures provide the structural components needed to recover the target molecule, but do not explicitly specify how those components map onto the final molecular graph. Correctly reconstructing such molecules requires figure-level integration and structural reasoning. To characterize this transition, we introduce StructVista, a benchmark of 419 expert-verified figures specifying 3,412 compound identifier–structure pairs, organized into three levels ranging from local self-contained recognition, to global instantiation with explicit relations, and finally to context-dependent assembly reasoning. We benchmark seven frontier multimodal large language models and two agent frameworks equipped with chemistry tools. When queried directly, the best-performing model, GPT-6 Sol, achieves only 50.8 F1 overall and exactly recovers 26.5% of figures. Deployed within the Codex framework, the same model improves to 89.2 F1 and 74.9% exact figure recovery. Even so, it exactly recovers only 62.9% of figures requiring context-dependent assembly reasoning, compared with 81.7% of figures with self-contained depictions. These results show that current models still struggle with chemical structure reasoning, particularly when implicit assembly relations must be inferred from figure context. StructVista thus provides a testbed for moving molecular structure interpretation beyond local recognition toward figure-level reconstruction and reasoning.

1 Introduction

Molecular structures reported in figures of scientific literature constitute a rich source of knowledge for biochemical research, supporting applications such as molecule-centric knowledge extraction, structure–function analysis, and scientific discovery. Recovering such information in machine-readable form is an important step and has motivated substantial efforts. Optical chemical structure recognition (OCSR) has made considerable progress in converting molecular depictions into machine-readable representations (Qian et al., 2023; Rajan et al., 2021; Chen et al., 2024; Fang et al., 2025b). More recent work has further extended structure recognition to Markush representations, where shared scaffolds are instantiated using variable substituents (Morin et al., 2025; Strohmeyer et al., 2026; Yan et al., 2025).

Despite this progress, existing methods are designed for two settings, namely figures in which each molecule is drawn completely and Markush representations in which the correspondence between variables and scaffold sites is stated explicitly. Beyond these two settings, we observe a third and qualitatively different regime in real biochemical figures. Here, the figure provides the structural components needed to recover the target molecule; their correspondence to the final molecular graph is not fully specified by explicit labels or standard notation, but can be resolved from context within the same figure. The model needs to determine which region of a reference structure is replaced, which attachment sites of two fragments correspond, which fragment plays a particular structural role, or how partially depicted structures are aligned or fused. In such cases, even correctly recogniz-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

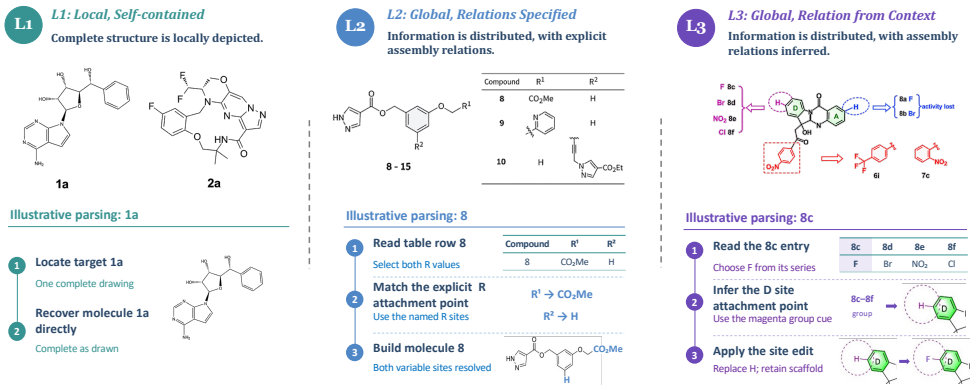


Figure 1: **Three levels of molecular reconstruction in StructVista.** Each panel shows an example figure (top) and the steps required to reconstruct one of its targets (bottom). **L1**: each target is depicted completely within a local region, the setting addressed by OCSR methods. **L2**: structural information is distributed across the figure, but assembly relations are explicitly specified, as in Markush structures whose variables are defined in a table; this setting is addressed by previous Markush resolution pipelines. **L3**: at least one assembly relation must be inferred from figure context, a setting not explicitly addressed by existing methods or benchmarks.

ing all visible molecular components is insufficient to uniquely reconstruct the target structure. The missing information is an *assembly relation* that must itself be resolved from the surrounding figure context. Thus, the problem shifts from retrieving and instantiating distributed structural information to reasoning about how that information should be assembled.

Motivated by this distinction, we introduce StructVista, a three-level benchmark for molecular reconstruction from scientific figures (Figure 1). Rather than categorizing examples by visual complexity alone, StructVista characterizes them according to two factors: whether the information required for reconstruction is locally self-contained or distributed across the figure, and whether the necessary assembly relations are explicitly specified or must be inferred from context. This yields three levels: **L1**, local self-contained reconstruction; **L2**, global instantiation under explicit assembly relations; and **L3**, context-dependent reconstruction requiring at least one assembly relation to be resolved from the full figure.

We construct StructVista from 8,175 articles published in 265 biochemistry journals. After automatic extraction and detection of candidate figures, a domain expert screens 29,124 images and retains only figures whose target structures can all be determined from the figure itself. Reference structures are drafted with an AI-assisted workflow and then independently reviewed by two domain experts, who agree on 96.3% of identifier–structure pairs before adjudication. The resulting benchmark contains 419 figures (180 L1, 134 L2, and 105 L3) specifying 3,412 compound identifier–structure pairs. Given only the figure, a system must recover every compound it specifies, including all members of compound series defined through shared scaffolds and variables.

We evaluate seven frontier multimodal large language models (MLLMs) queried directly, together with two agent frameworks, Codex and Cursor, that can call chemistry tools for structure detection, recognition, and graph validation. Three findings emerge. First, when queried directly, the best-performing MLLM, GPT-6 Sol, achieves only 50.8 F1 and 26.5% exact figure recovery (EFR). Providing additional molecule detection and OCSR outputs raises the best performance to 73.0 F1 and 56.6% EFR. These results suggest that directly reconstructing molecular structures remains challenging for current MLLMs, particularly given their limited ability to recognize chemical structures. Second, agent frameworks equipped with molecule detection and OCSR tools can substantially improve recovery, with Codex GPT-6 Sol reaching 89.2 F1 and 74.9% EFR. This improvement highlights the value of iterative, multi-step processing for complex molecular reconstruction tasks such as StructVista. Third, with overall F1 scores ranging from 4.5 to 89.2 and exact figure recovery rates from 1.0% to 74.9%, StructVista clearly differentiates system capabilities. Together, these

results show that current models and agent frameworks still struggle to reliably solve the molecular reconstruction tasks in StructVista.

In summary, our contributions in this work are threefold:

- We formulate *figure-level molecular reconstruction*, which extends local chemical structure recognition to the integration of non-local structural information and, when necessary, context-dependent assembly reasoning.
- We introduce StructVista, an expert-verified benchmark of 419 figures and 3,412 identifier–structure pairs whose three levels distinguish local self-contained reconstruction, global instantiation under explicit assembly relations, and reconstruction requiring assembly relations to be inferred from figure context.
- We evaluate frontier MLLMs and agent frameworks, revealing persistent challenges in complex molecular reconstruction that stem from limitations in both fundamental chemical structure recognition and reasoning over context-dependent assembly relationships.

2 Related Work

Optical chemical structure recognition. Optical chemical structure recognition (OCSR) converts a depicted molecule into a machine-readable graph. DECIMER formulates the task as image-to-sequence translation with transformer models (Rajan et al., 2021). MolScribe and MolNexTR instead predict atoms and bonds explicitly and incorporate chemical knowledge such as abbreviation expansion and stereochemical interpretation (Qian et al., 2023; Chen et al., 2024), and are commonly evaluated on literature-derived sets such as the ACS benchmark introduced with MolScribe. MolParser extends recognition to in-the-wild depictions through an extended structural representation (Fang et al., 2025b), and MolRecBench-Wild evaluates such recognizers under visual interference and non-standard chemical notation, including icon-based groups (Yang et al., 2026). Across these methods and benchmarks, the input is a single structure crop and the output is the molecule drawn within it. Associating that molecule with a compound identifier, or completing it with information located elsewhere in the figure, lies outside the task definition. The L1 level of StructVista corresponds to this setting; L2 and L3 measure what happens once a crop no longer contains the complete molecule.

Markush structures and figure-level extraction. MarkushGrapher jointly encodes image, text, and layout to recover a Markush template together with its variable definitions, and MarkushGrapher-2 adds dedicated chemical OCR and real-world training data; they are evaluated on the M2S and IP5-M benchmarks, respectively (Morin et al., 2025; Strohmeyer et al., 2026). These systems output a template; enumerating the concrete compounds it specifies is a separate step. Other pipelines perform that step: ChemSchematicResolver and OpenChemIE resolve R-group labels against accompanying tables or text to produce concrete structures (Beard & Cole, 2020; Fan et al., 2024), and BioMiner, together with its BioVista benchmark, evaluates structure extraction linked to bioactivity data at the paper level, including Markush-defined compounds (Yan et al., 2025). This line of work addresses the L2 setting, in which the correspondence between variables and scaffold sites is stated explicitly. To our knowledge, existing evaluations do not isolate cases in which the assembly relation itself—the scope of a replacement, the correspondence between attachment sites, or the atoms shared at a fusion interface—must be inferred from figure context. Table 1 compares StructVista with these structure-extraction benchmarks.

Chemical reasoning benchmarks. ChemBench evaluates broad chemical knowledge and reasoning (Mirza et al., 2025), MolecularIQ focuses on symbolically verifiable operations over molecular graphs (Bartmann et al., 2026), and ReactBench tests topological reasoning over reaction-diagram networks with structured-input controls (Xu et al., 2026). These benchmarks pose questions about given molecules or about relations between molecules, rather than requiring a molecule to be reconstructed from its depiction. StructVista instead targets the reasoning required to determine the intramolecular graph of a single compound when the evidence for it is distributed across a figure.

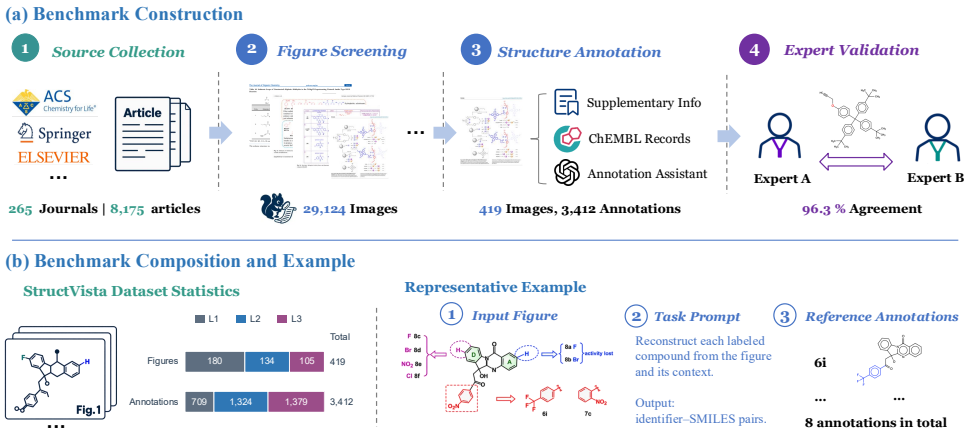


Figure 2: **Construction and composition of StructVista.** (a) The four construction stages, from source collection to independent expert validation. (b) Distribution of figures and annotations (identifier–structure pairs) across levels, and a representative L3 example with its input figure, task prompt, and reference annotations.

3 The StructVista Benchmark

3.1 Task Formulation

Given a scientific figure I, StructVista requires recovering every concrete compound specified by the figure, together with its figure-local identifier. When a figure uses a Markush representation, every compound instance specified by its scaffold and variable definitions must be instantiated into a complete structure; returning a scaffold with unresolved variables is insufficient. Each reconstructed graph must preserve the atom, bond, charge, isotope, and stereochemical information specified by the figure, without inventing unspecified structural details. Formally, the output is

$$\mathcal{Y}(I) = \{(c_i, s_i)\}_{i=1}^{N_I}, \tag{1}$$

where c_i is a figure-local compound identifier and s_i is a SMILES representation of its complete molecular graph G_i. N_I denotes the number of target compounds specified in the figure and is not provided to the system.

The test setting of StructVista is *figure-only*, where all information used for reconstruction must come from the figure itself, and systems may not consult supplementary information, article text, external structure databases, or web search at inference time.

3.2 Reconstruction Levels

Each figure is assigned a reconstruction level according to the structural information and assembly relations required to reconstruct all of its targets. An assembly relation is *explicit* if it is determined by standard chemical notation conventions, such as variable labels defined in a table, numbered attachment points, or abbreviations (Appendix A.1).

- **L1: Local self-contained reconstruction.** Every target can be determined from a self-contained local depiction.
- **L2: Global instantiation under explicit assembly relations.** At least one target requires structural information from multiple regions, but all required structural roles, replacement scopes, and interface correspondences are explicit.
- **L3: Context-driven global assembly reasoning.** At least one target additionally requires resolving an assembly relation that is not explicit. Such a relation may specify a component's structural role, the scope of a replacement, the correspondence between attachment sites, or the shared interface in a fused structure; it must be recoverable from the figure context and must affect the target molecular graph.

Table 1: Comparison with existing structure-extraction benchmarks. **ID**: structures are linked to figure-local compound identifiers. **Inst.**: Markush-defined compounds are instantiated as concrete graphs (○: scaffold–substituent associations only). **Assembly**: cases requiring assembly relations to be inferred from figure context are included and distinguished. **Analysis axes**: dimensions along which results are reported. CSR: ChemSchematicResolver.

Benchmark	Input	Output	ID	Inst.	Assembly	Analysis axes
ACS (Qian et al., 2023)	Structure crop	Graph	–	–	–	–
MolRecBench-Wild (Yang et al., 2026)	Structure crop	CARBON/SMILES	–	–	–	Visual/chemical
CSR eval. (Beard & Cole, 2020)	Diagram	Instances	✓	✓	–	Pipeline stage
IP5-M (Strohmeyer et al., 2026)	Markush crop	Template	–	–	–	Markush type
M2S (Morin et al., 2025)	Scaffold + def.	Template + def.	–	–	–	Markush type
BioVista-Rel. (Yan et al., 2025)	Augmented image	Associations	✓	○	–	Full/Markush
BioVista-E2E (Yan et al., 2025)	Paper	Graph + activity	✓	✓	–	Full/Markush
StructVista	Full figure	All instances	✓	✓	✓	Locality/relation

Because the level reflects the most demanding target, an L3 figure may also contain targets that could be reconstructed locally or through explicit relations. Examples of all three levels, including L3 cases covering different assembly relations, are provided in Appendix A.

3.3 Benchmark Construction

We build StructVista in four stages (Figure 2(a)): source collection, figure screening, structure annotation, and expert validation. The pipeline separates the evidence used to construct labels from the evidence available at evaluation: external records help recover reference structures efficiently, but a figure is retained only if the experts confirm that every target can be determined from the figure alone.

Source collection. We collect 8,175 articles published in 265 journals from major publishers, including ACS, Springer, and Elsevier. Uni-Parser (Fang et al., 2025a) is used to extract figure and table regions from the article PDFs.

Figure screening. MolDetv2 (Fang et al., 2026) identifies the extracted regions that contain molecular structure depictions, yielding 29,124 candidate images. A domain expert reviews every candidate and assigns a preliminary level to each figure in which all in-scope compounds can be reconstructed from the figure itself. Because L3 figures are the least common, we first collect L3 figures from the screened candidates and then randomly sample L1 and L2 figures, yielding 180 L1, 134 L2, and 105 L3 figures.

Structure annotation. For each retained figure, a Codex-based annotation assistant receives the figure together with the source supplementary information and matching ChEMBL records (Zdrazil et al., 2024), and proposes provisional identifier–structure pairs, including the instantiated members of compound series defined through shared scaffolds and variables. These external materials support label construction only and are never provided to evaluated models.

Expert validation. Two domain experts independently review every proposal. They verify identifier assignments, molecular graphs, and coverage of all compounds specified in the figure; correct erroneous structures and associations; add missing targets; confirm or correct the preliminary level of each figure; and discard figures whose targets cannot all be determined from the figure itself. An agreement of 96.3% was achieved on identifier–structure pairs before discussion. They then jointly discussed and resolved all disagreements to produce the final annotations.

3.4 Benchmark Statistics

The resulting benchmark contains 419 figures specifying 3,412 identifier–structure pairs, retained from 29,124 expert-screened images. By level, it comprises 180 L1 figures with 709 pairs, 134 L2 figures with 1,324 pairs, and 105 L3 figures with 1,379 pairs (Figure 2(b)). L2 and L3 figures specify

9.9 and 13.1 pairs on average, compared with 3.9 for L1, because they typically report compound series.

Comparison with existing benchmarks. Table 1 compares StructVista with representative structure-extraction benchmarks; chemical reasoning benchmarks such as ChemBench, MolecularIQ, and ReactBench are excluded because they pose questions about given molecules rather than require reconstructing them. OCSR and Markush benchmarks evaluate individual structure crops or templates without linking them to compound identifiers, while ChemSchematicResolver and BioVista evaluate identifier-linked, instantiated compounds but do not distinguish cases whose assembly must be inferred from figure context. StructVista requires every compound in a complete figure to be instantiated, explicitly targets context-dependent assembly relations, and supports analysis by the locality of structural information and the explicitness of assembly relations.

4 Experimental Setup

Inference frameworks. We evaluate two inference frameworks with the same instruction and output format (Appendix C):

- **Direct.** A multimodal large language model (MLLM) receives the figure in a single request and returns identifier–SMILES pairs without tool use.
- **Agentic.** An agent works in an isolated, offline workspace containing the figure and three chemistry tools: MolDetv2 (Fang et al., 2026) for structure detection, MolParser-Mobile (Fang et al., 2026) for structure recognition, and RDKit (RDKit Developers, 2026) for graph construction and validation. The agent decides how to use these tools and may write additional code; after 10 minutes of wall-clock time, it is instructed to stop and submit its current results.

In both frameworks, each figure is processed in a fresh session without fine-tuning or access to external information.

Evaluated systems. For the direct framework, we evaluate seven frontier MLLMs from six providers: DeepSeek V4.1 Flash (DeepSeek, 2026), GPT-6 Luna (OpenAI, 2026b), GPT-6 Sol (OpenAI, 2026c), Gemini 3.8 Flash (Google, 2026), Qwen 3.8 27B (Qwen Team, 2026), Seed 2.1 Pro (ByteDance Seed, 2026), and MiMo 2.6 Pro (Xiaomi MiMo, 2026). For the agentic framework, we evaluate two agents: Codex (OpenAI, 2026a) with GPT-6 Luna and GPT-6 Sol as backbones, and Cursor with its native model Composer 2.5 (Cursor, 2026). Claude models were not evaluated because provider-side safety filters refused a large fraction of requests involving the chemical structures.

Metrics. We evaluate each system with two complementary metrics: pair-level F1, which measures how many compounds are correctly recovered, and exact figure recovery (EFR), which measures whether a figure is recovered completely. Both rely on the same matching rule: a predicted identifier–SMILES pair is a true positive only if both its identifier and its structure match the same reference pair in $\mathcal{Y}(I)$, with each reference pair matched at most once. Identifiers are matched exactly after trimming leading and trailing whitespace, and structures by RDKit canonical isomeric SMILES (RDKit Developers, 2026) after removing salts and counterions and neutralizing charges, so that isotopes and stereochemistry must still be reproduced exactly. **F1** is computed from true positives, predictions, and references pooled over all figures in each level. **EFR** is the fraction of figures whose predicted set matches the reference set exactly, with no missing or spurious pairs; it reflects practical usability, since a single error requires the whole compound series to be revisited.

5 Experiments

Our experiments address four questions. Section 5.1 measures end-to-end reconstruction from the original figures. Section 5.2 asks whether chemical structure recognition limits performance, by supplying the outputs of a dedicated recognition pipeline to the models. Section 5.3 examines how target multiplicity and molecular recognition difficulty affect molecule-level performance and complete figure recovery. Section 5.4 analyzes the remaining errors.

Table 2: End-to-end reconstruction results (%), grouped by figure-level reconstruction level. Best and second-best results across all systems are shown in **bold** and underlined, respectively.

	L1		L2		L3		Overall	
System	F1	EFR	F1	EFR	F1	EFR	F1	EFR
Direct MLLMs								
DeepSeek V4.1 Flash	14.1	5.0	2.4	0.0	1.9	0.0	4.7	2.1
GPT-6 Luna	23.9	13.3	23.8	11.9	21.1	2.9	22.8	10.3
GPT-6 Sol	50.7	27.8	51.4	28.4	50.2	21.9	50.8	26.5
Gemini 3.8 Flash	45.4	27.8	41.4	19.4	43.1	17.1	42.9	22.4
Qwen 3.8 27B	12.1	3.9	7.6	1.5	14.6	0.0	11.4	2.1
Seed 2.1 Pro	51.8	33.3	36.8	25.4	41.1	10.5	41.7	25.1
MiMo 2.6 Pro	8.1	2.2	2.5	0.0	4.4	0.0	4.5	1.0
Agents with Domain Tools								
Codex (GPT-6 Luna)	71.7	48.3	53.9	28.4	26.2	3.8	47.9	30.8
Codex (GPT-6 Sol)	**90.8**	**81.7**	**90.1**	**75.4**	**87.5**	**62.9**	**89.2**	**74.9**
Cursor (Composer 2.5)	78.7	59.4	52.7	32.1	33.4	3.8	51.0	36.8

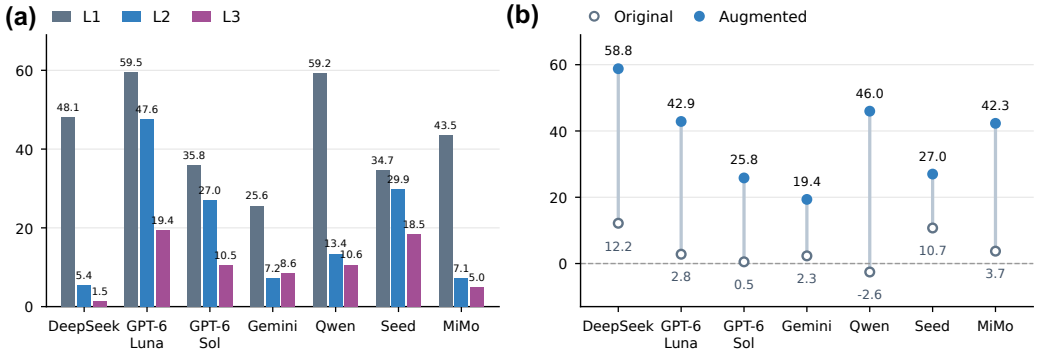


Figure 3: **Effect of OCSR augmentation.** (a) F1 gain on each level: every model gains most on L1 and considerably less on L3. (b) F1 difference between L1 and L3 on the original and augmented inputs: levels that are nearly indistinguishable without augmentation separate once recognition is supplied.

5.1 End-to-End Reconstruction

Table 2 reports pair-level F1 and exact figure recovery (EFR) for seven directly queried MLLMs and three agent configurations, grouped by figure-level reconstruction level.

Direct MLLMs. The best direct model, GPT-6 Sol, achieves 50.8 overall F1 and 26.5% EFR. Performance remains limited even on L1, where every target is locally complete: the highest F1 is 51.8, achieved by Seed 2.1 Pro. For GPT-6 Sol, GPT-6 Luna, and Gemini 3.8 Flash, L1 and L3 F1 differ by less than 3 points. Direct prediction thus remains challenging across levels, with little separation between local and context-dependent reconstruction for these models.

Agent Frameworks. Codex raises overall F1 from 22.8 to 47.9 for GPT-6 Luna and from 50.8 to 89.2 for GPT-6 Sol, but the gains differ across levels. GPT-6 Luna gains 47.8 points on L1 and only 5.1 on L3; Cursor with Composer 2.5 likewise performs much better on L1 than on L3 (78.7 versus 33.4 F1). In contrast, Codex with GPT-6 Sol achieves 90.8 F1 on L1 and 87.5 on L3, although EFR falls from 81.7% to 62.9%. Complete L3 recovery therefore remains less reliable, but its larger number of targets per figure also contributes to the difficulty of this metric.

Together, these results show that tool-equipped agents substantially improve molecular reconstruction, but their gains depend strongly on the backbone, and complete recovery of L3 figures remains challenging. The limited performance of direct MLLMs even on L1 further motivates examining

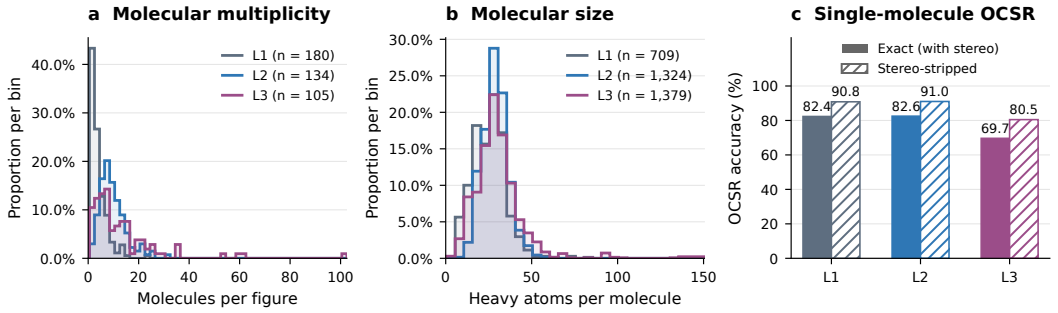


Figure 4: **Figure multiplicity, molecular size, and single-molecule recognition.** (a) Distribution of the number of target compounds per figure. (b) Heavy-atom distributions for molecules in each figure-level group. (c) MolParser-Mobile accuracy on complete RDKit-rendered molecules, with and without stereochemical matching. All groups use figure-level labels.

the role of basic structure recognition. We therefore next provide region-level OCSR predictions to assess how recognition support affects reconstruction across levels.

5.2 OCSR Augmentation Reveals Larger Reconstruction Gaps

The end-to-end results show small L1–L3 gaps for several direct models but uneven gains under agentic workflows. To examine whether recognition support exposes differences between levels, we augment the original figure with region bounding boxes and SMILES predictions obtained using MolDetv2 and MolParser-Mobile, respectively, while retaining the same reconstruction instruction. These predictions provide local structural information without supplying compound identifiers or explicit assembly relations; they may contain errors or unresolved variables. The intervention adds both visual box overlays and textual candidates.

Local reconstruction benefits most. Figure 3(a) shows positive F1 gains on all three levels for all seven evaluated models, with consistently larger gains on L1 than on L3. L1 gains range from 25.6 to 59.5 F1 points, compared with 1.5 to 19.4 on L3. For example, GPT-6 Luna gains 59.5, 47.6, and 19.4 points on L1, L2, and L3, respectively. The L1-minus-L3 gain is positive for all seven models, with pointwise 95% source-paper bootstrap intervals excluding zero (Appendix E). These intervals describe variation across source papers, not repeated model runs.

L1–L3 gaps widen after augmentation. The L1–L3 F1 gap increases for every model (Figure 3(b)), rising from 2.8 to 42.9 points for GPT-6 Luna and from 0.5 to 25.8 points for GPT-6 Sol. The small gaps observed under direct prediction therefore widen when region-level recognition support is provided. The widened gaps suggest that exploiting local structural predictions for figure-level integration and assembly remains challenging for current MLLMs.

5.3 Molecule-Level Accuracy and Complete Figure Recovery

Molecule-level F1 and exact figure recovery capture complementary aspects of reconstruction performance. For Codex with GPT-6 Sol, F1 remains relatively high across L1–L3 (90.8%, 90.1%, and 87.5%), whereas EFR decreases from 81.7% to 75.4% and 62.9%. Thus, recovering most molecules does not necessarily yield a complete and correct figure-level output.

Complete recovery is sensitive to target multiplicity. In the current evaluation, L1, L2, and L3 contain an average of 3.9, 9.9, and 13.1 eligible targets per figure, respectively. Because EFR requires exact recovery of the full target multiset, it is sensitive to omissions and incorrect predictions that molecule-level F1 can partially accommodate. Restricting L1/L3 figures to shared target-count support and reweighting them to a common count distribution yields EFR of 78.3% and 74.5% for Codex with GPT-6 Sol, compared with 81.7% and 62.9% in the full evaluation. This exploratory comparison highlights the contribution of figure composition to measured recovery performance.

Agent workflows extend standalone recognition. To contextualize agent performance, we first examine standalone recognition of the benchmark molecules. The heavy-atom distributions overlap, with a longer upper tail for L3 (Figure 4(b)). A separate diagnostic using MolParser-Mobile on complete RDKit-rendered depictions of 3,412 molecular instances yields exact recognition accuracy of 82.4%, 82.6%, and 69.7% on L1–L3, increasing to 90.8%, 91.0%, and 80.5% when stereochemical distinctions are removed. Across this diagnostic set, exact recognition accuracy is 77.3%, whereas Codex with GPT-6 Sol achieves 88.5% target recall on the figure-level benchmark. Although the two evaluations differ in input format and scoring, this comparison suggests that integrating recognition tools into an iterative agent workflow can overcome some limitations of standalone molecular recognition.

5.4 Error Analysis

Mismatch patterns. Table 3 analyzes GPT-6 Sol under direct, OCSR-augmented, and Codex settings using the same annotations. Across 3,412 targets, these settings leave 1,763, 964, and 391 reference targets unmatched, respectively. Connectivity or bond-order mismatches form the largest category in each setting, decreasing from 1,066 to 587 and 168. On figures labeled L3, augmentation reduces connectivity/bond-order mismatches from 427 to 368 and missing-ID mismatches from 204 to 71, while stereo-only mismatches increase from 59 to 89. Level-specific diagnostics and cases are discussed in Appendix G.

Table 3: Unmatched targets for GPT-6 Sol. Categories describe mismatches, not verified error causes.

Mismatch	Direct	+OCSR	Codex
Connectivity / bond order	1,066	587	168
Stereo only	267	163	122
Missing ID	279	91	61
Structure under another ID	39	61	36
Invalid SMILES only	110	56	0
Tautomer only	2	6	4
Total	1,763	964	391

Valid structures can still misrepresent the figure. All 3,361 scored Codex predictions are parseable, yet 391 reference targets remain unmatched. Inspection of source figures and execution traces reveals errors that are not detected by RDKit validity checks. In a compound table with a shared substituent cell (Figure 8), Codex treats blank entries as hydrogen rather than inheriting the shared methylsulfonyl group, omitting it from four compounds. In a second example (Figure 9), direct prediction, OCSR augmentation, and Codex all omit the explicitly drawn methylene in an N–CH_2–Alk substituent. These examples show how errors in table interpretation and substituent expansion can yield chemically valid molecular graphs that do not match the structures specified by the figure.

6 Conclusion

We introduced StructVista, a benchmark of 419 scientific figures and 3,412 identifier–structure pairs that distinguishes local recognition, explicit instantiation, and context-dependent assembly. Region-level OCSR augmentation yields larger F1 gains on L1 than on L3 across the evaluated MLLMs, widening the L1–L3 performance gap. Complete recovery of L3 figures remains challenging: the strongest system achieves 62.9% EFR. Exact figure recovery varies substantially across systems and is sensitive to target multiplicity. In the exploratory comparison standardized to a common target-count distribution, the L1–L3 gap is small for Codex Sol but remains large for Codex Luna and Cursor. These results do not isolate assembly reasoning from recognition and other system factors. StructVista provides a basis for studying how systems integrate distributed structural evidence to reconstruct complete molecules.

AI USE STATEMENT

AI tools assisted with preliminary data annotation during benchmark construction, as described in Section 3.3. The proposed annotations were subsequently reviewed and corrected by domain experts. AI tools were also used to improve the language and presentation of the manuscript. The authors are responsible for the final annotations and all content of the paper.

ETHICS STATEMENT

The intended use is evaluation of bounded-input molecular structure reconstruction, not autonomous synthesis, clinical advice, or experimental pharmacology. Source figures remain subject to publisher and author rights; image availability in a publication does not establish redistribution permission. The release plan separates original artwork, provenance, newly created annotations, and controlled redraws, and will include only material with an appropriate rights basis.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe benchmark construction, the reconstruction task, scoring rules, inference frameworks, and experimental protocols. The appendix further specifies the annotation schema, convention registry, reference-audit rules, prompts, and result fields. These descriptions document how the benchmark is constructed and how model predictions are evaluated. Experiments involving proprietary models and agents may be affected by stochastic generation and subsequent service updates. Upon acceptance, we will publicly release the benchmark annotations, evaluation code, and experimental configurations to facilitate independent reproduction of our results.

REFERENCES

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. MolecularIQ: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. arXiv:2601.15279, 2026. URL https://arxiv.org/abs/2601.15279.

Edward J. Beard and Jacqueline M. Cole. ChemSchematicResolver: A toolkit to decode 2D chemical diagrams with labels and R-groups into annotated chemical named entities. *Journal of Chemical Information and Modeling*, 60(4):2059–2072, 2020. doi: 10.1021/acs.jcim.0c00042.

ByteDance Seed. Seed2.1, 2026. URL https://seed.bytedance.com/en/seed2_1. Accessed: 2026-09-26.

Yufan Chen, Ching Ting Leung, Yong Huang, Jianwei Sun, Hao Chen, and Hanyu Gao. MolNexTR: A generalized deep learning model for molecular image recognition. *Journal of Cheminformatics*, 2024. doi: 10.1186/s13321-024-00926-w.

Cursor. Introducing Composer 2.5, 2026. URL https://cursor.com/marketing-static/blog/composer-2-5. Accessed: 2026-09-26.

DeepSeek. Introducing DeepSeek-V4.1-Flash: Smarter, Faster, More Efficient, 2026. URL https://www.deepseek.com/en/news/deepseek-v4-1-flash/. Accessed: 2026-09-26.

Vincent Fan, Yujie Qian, Alex Wang, Amber Wang, Connor W. Coley, and Regina Barzilay. OpenChemIE: An information extraction toolkit for chemistry literature. *Journal of Chemical Information and Modeling*, 64(14):5521–5534, 2024.

Xi Fang, Haoyi Tao, Shuwen Yang, Suyang Zhong, Haocheng Lu, Han Lyu, Chaozheng Huang, Xinyu Li, Linfeng Zhang, and Guolin Ke. Uni-parser technical report. *arXiv preprint arXiv:2512.15098*, 2025a.

Xi Fang, Jiankun Wang, Xiaochen Cai, Shangqian Chen, Shuwen Yang, Haoyi Tao, Nan Wang, Lin Yao, Linfeng Zhang, and Guolin Ke. MolParser: End-to-end visual recognition of molecule structures in the wild. arXiv:2411.11098, version 4, 2025b. URL https://arxiv.org/abs/2411.11098.

Xi Fang, Haocheng Lu, Han Lyu, Chengxiang Luo, Linfeng Zhang, and Guolin Ke. Molparser-mobile: Ultrafast ocsr system for large-scale chemical literature mining. *arXiv preprint arXiv:2609.05807*, 2026.

Google. What's New in Gemini 3.8 Flash, 2026. URL https://ai.google.dev/gemini-api/docs/latest-model?hl=en. Accessed: 2026-09-26.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, et al. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17:1027–1034, 2025. doi: 10.1038/s41557-025-01815-x.

Lucas Morin, Valéry Weber, Ahmed Nassar, Gerhard Ingmar Meijer, Luc Van Gool, Yawei Li, and Peter Staar. MarkushGrapher: Joint visual and textual recognition of markush structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14505–14515, 2025. URL https://arxiv.org/abs/2503.16096.

OpenAI. Codex CLI, 2026a. URL https://developers.openai.com/codex/cli/. Accessed: 2026-09-26.

OpenAI. GPT-6 Luna, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-luna. Accessed: 2026-09-26.

OpenAI. GPT-6 Sol, 2026c. URL https://developers.openai.com/api/docs/models/gpt-6-sol. Accessed: 2026-09-26.

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W. Coley, and Regina Barzilay. MolScribe: Robust molecular structure recognition with image-to-graph generation. *Journal of Chemical Information and Modeling*, 2023. doi: 10.1021/acs.jcim.2c01480.

Qwen Team. Qwen3.8-27B, 2026. URL https://huggingface.co/Qwen/Qwen3.8-27B. Accessed: 2026-09-26.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. DECIMER 1.0: Deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13:61, 2021. doi: 10.1186/s13321-021-00538-8.

RDKit Developers. The RDKit book and python API reference. Software documentation, 2026. URL https://www.rdkit.org/docs/RDKit_Book.html. Accessed 15 September 2026.

Tim Strohmeyer, Lucas Morin, Gerhard Ingmar Meijer, Valéry Weber, Ahmed Nassar, and Peter Staar. MarkushGrapher-2: End-to-end multimodal recognition of chemical structures. arXiv:2603.28550, 2026. URL https://arxiv.org/abs/2603.28550.

Xiaomi MiMo. MiMo-V2.6-Pro, 2026. URL https://mimo.mi.com/models/en-US/mimo-v2.6-pro. Accessed: 2026-09-26.

Qiang Xu, Shengyuan Bai, Yu Wang, He Cao, Leqing Chen, Yuanyuan Liu, Bin Feng, Zijing Liu, and Yu Li. ReactBench: A benchmark for topological reasoning in MLLMs on chemical reaction diagrams. arXiv:2604.15994, 2026. URL https://arxiv.org/abs/2604.15994.

Jiaxian Yan, Jintao Zhu, Yuhang Yang, Qi Liu, Kai Zhang, Zaixi Zhang, Xukai Liu, Boyan Zhang, Kaiyuan Gao, Jinchuan Xiao, et al. Biominer: A multi-modal system for automated mining of protein-ligand bioactivity data from literature. *BioRxiv*, pp. 2025–04, 2025.

Haote Yang, Hui Wang, Chen Zhu, Jingchao Wang, Linye Li, Hongbin Lai, Huijie Ao, Yongxuan Lyu, Jiang Wu, Jiaxing Sun, Lua Chen, Yuanyuan Cao, Ruijie Zhang, Shengxin Lu, Lijun Wu, Bin Wang, and Conghui He. MolRecBench-Wild: A real-world benchmark for optical chemical structure recognition. arXiv:2605.05832, 2026. URL https://arxiv.org/abs/2605.05832.

Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen De Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

A Task Scope and Reconstruction Examples

A.1 Reconstruction levels and boundary cases

The three levels distinguish the evidence required for molecular reconstruction, rather than visual density alone. L1 targets have self-contained local depictions; L2 targets require information from multiple regions with explicit correspondences; L3 targets additionally require a figure-specific assembly relation to be resolved from context. Figure labels reflect the most demanding included target. Unless otherwise stated, the reported level-wise statistics group all targets by the label of their source figure; they are not separate measurements of target-level L3 reasoning.

Table 4: Operational distinctions between reconstruction requirements.

Requirement	Information supplied by the depiction	Remaining operation
Local decoding	Complete molecular depiction, conventional atom/bond notation and standard abbreviations	Recover the locally depicted molecular graph
Explicit instantiation	Scaffold sites and substituents linked by named variables or explicit interface labels	Look up the values and instantiate the specified connections
Context-dependent assembly	Components plus figure-specific cues about roles, replacement scope, attachment correspondence or shared atoms	Resolve a necessary relation and construct the resulting graph

A crowded table does not imply L3 if all correspondences are explicit. Blur, missing structural definitions and an unrecognized conventional abbreviation do not establish a reasoning requirement. Conversely, recognizing every component does not necessarily determine which atoms should be joined or replaced. If exchanging symmetric attachment points leaves the molecular graph unchanged, that exchange alone cannot justify an L3 label. For L2, the supplied labels, legend, or standard notation determine the relevant correspondence directly. For L3, identifying that correspondence requires interpreting additional context within the figure. Color or arrows can occur at either level; the deciding factor is whether their meaning directly specifies the graph-changing relation.

A.2 Annotation reliability

Two doctoral researchers, one in computer science and one in cheminformatics, independently reviewed the AI-assisted preliminary identifier–structure annotations and assigned reconstruction-level labels. Structural annotation agreement was 96.3% across 3,412 identifier–structure pairs, reported to one decimal place. The two annotators then jointly discussed disagreements and reached consensus on the final annotations. Level labels revised during this discussion replaced the preliminary labels in the final benchmark annotations. The 96.3% rate measures agreement on identifier–structure pairs.

A.3 Figure-to-output examples

The following examples show source inputs and the complete reference output sets under the current full-structure track. References are serialized as identifier–SMILES pairs; different SMILES strings can represent the same graph under the scoring policy in Appendix D.

L1_003: Local reconstruction. Compound 3 has a complete two-dimensional depiction within a larger figure containing protein views. The surrounding protein renderings do not need to be converted into molecular graphs.

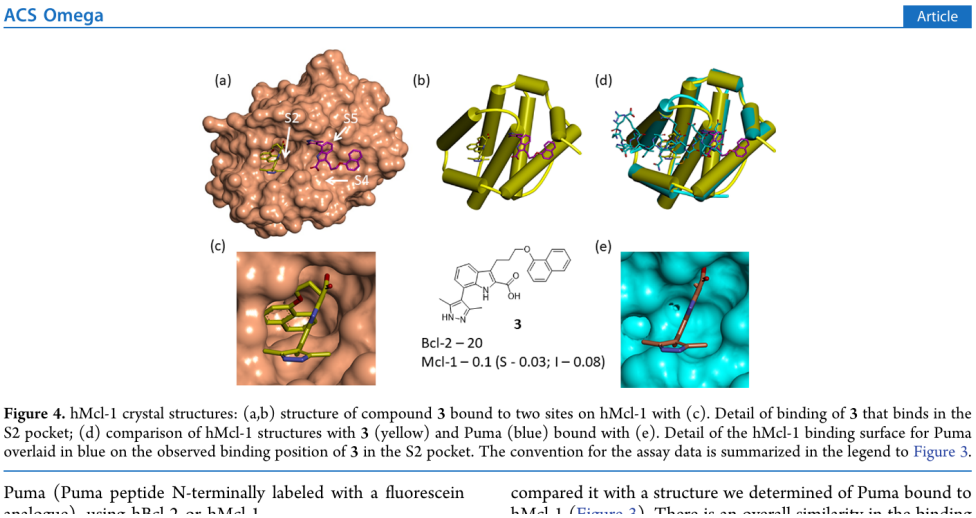


Figure 4. hMcl-1 crystal structures: (a,b) structure of compound **3** bound to two sites on hMcl-1 with (c). Detail of binding of **3** that binds in the S2 pocket; (d) comparison of hMcl-1 structures with **3** (yellow) and Puma (blue) bound with (e). Detail of the hMcl-1 binding surface for Puma overlaid in blue on the observed binding position of **3** in the S2 pocket. The convention for the assay data is summarized in the legend to Figure 3.

Puma (Puma peptide N-terminally labeled with a fluorescein analogue), using hBcl-2 or hMcl-1 compared it with a structure we determined of Puma bound to hMcl-1 (Figure 3). There is an overall similarity in the binding

Figure 5: Input L1_003 (L1). Source DOI: `10.1021/acsomega.9b00611`. The reference pairs are provided in Appendix A.3

```
Identifier;SMILES
3;Cc1n[nH]c(C)c1-c1cccc2c(CCCOc3cccc4ccccc34)c(C(=O)O)[nH]c12
```

L2_032: Explicit scaffold instantiation. The shared scaffold identifies a single R site, and rows 5a and 5b provide substituents with an explicit attachment bond. The reference output contains the two instantiated compounds.

```
Identifier;SMILES
5a;CC#CC(=O)N[C@H]1CCCN(c2c(F)cc(C(N)=O)c3[nH]c(C)c(C)c23)C1
5b;C#CC(=O)N[C@H]1CCCN(c2c(F)cc(C(N)=O)c3[nH]c(C)c(C)c23)C1
```

A011: Modular reconstruction and output scope. Rows 23 and 24 separate the binding module, linker and target ligand. Colored endpoints provide the cues for joining the modules. This L3 example illustrates reconstruction from separately depicted molecular modules and figure-specific attachment cues. Row 25 depicts an aptamer schematically and is not a complete molecular-graph target in the current track.

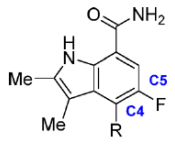
```
Identifier;SMILES
23;Cc1cc(=O)oc2cc(O)cc(OCCCCCN3CCC(C(=O)Nc4ncc(SCc5ncc(C(C)(C)C)o5)s4)CC3)c12
24;Cc1cc(=O)oc2cc(O)cc(OCCCN3CCC(C(=O)Nc4ncc(SCc5ncc(C(C)(C)C)o5)s4)CC3)c12
```

B Dataset Documentation

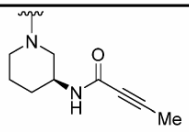
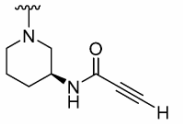
Table 5: Collection size and evaluation populations. Target counts are grouped by figure-level label.

Item	Value
Evaluation figures (L1 / L2 / L3)	180 / 134 / 105; 419 total
Distinct source DOIs in the retained manifest	244
Current full-structure targets (L1 / L2 / L3)	709 / 1,324 / 1,379; 3,412 total
Statistical resampling unit	Source DOI; 244 groups

Table 3. In Vitro Potency of C5-F-Dimethylindole Analogues 5



5

Cmpd	R	BTK IC_{50} (nM)[a]	Ramos IC_{50} (nM)[a]
5a		0.10 ±0.05	7.2 ±2.3
5b		0.07 (n=1)	1.6 (n=1)

[a] IC_{50} values are shown as mean values of at least three determinations unless specified otherwise.

Figure 6: Input L2_032 (L2). Source DOI: 10.1021/acs.jmedchem.9b00167. The reference pairs are provided in Appendix A.3.

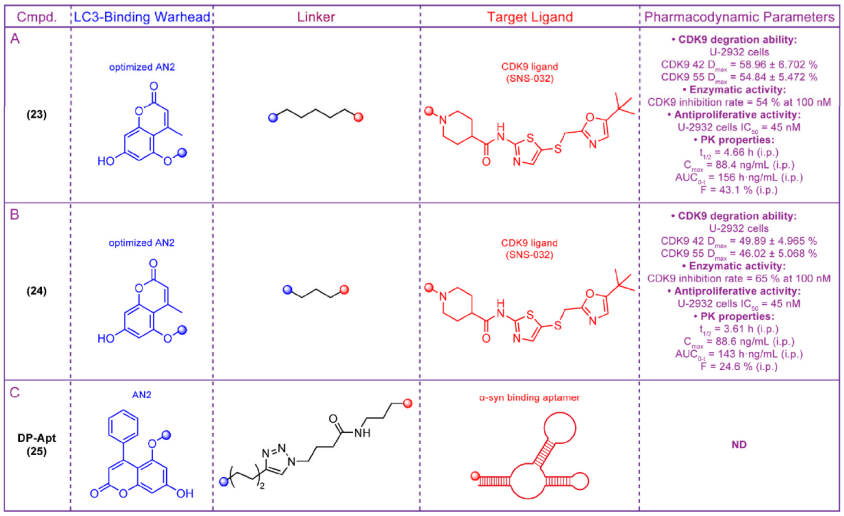


Fig. 9. Structure, biological activity data, and pharmacokinetics properties of ATTEC molecules utilizing AN2 and its derivatives as LC3-binding warheads, compound **23–25.**

associated with neurodegenerative diseases.

directions: (i) small molecule–**AN2** conjugates for the selective degra-

Figure 7: Input A011 (L3). Source DOI: 10.1016/j.ejmech.2026.118585. The reference pairs are provided in Appendix A.3.

Collection versus scoring population. The full-structure evaluation contains 3,412 target instances across 419 figures. Partial or underspecified structures are outside the scored full-structure track. Consequently, exact recovery means matching the reference set, not recovering every partially specified entity visible in the image. Individual annotation instances are counted separately even when they represent the same molecular graph.

Source collection and screening. The source manifest links the 419 evaluation figures to 244 distinct DOIs, image hashes and figure-level labels. For L3 construction, the literature metadata inventory contains 8,175 unique DOIs from successfully parsed articles and recovered early-source records. Among these records, 8,116 have journal metadata, yielding 265 journal-name labels after local alias normalization; 59 lack journal metadata. These journal labels are not an ISSN-verified count of distinct journals. The inventory excludes unprocessed article stock and 694 early-source directories whose DOIs have not been recovered. Across the principal collection rounds, the candidate pools contain 29,124 figure entries. This cumulative candidate count is not a count of unique images independently annotated by both experts. The final L3 evaluation subset comprises 105 figures from 71 papers, all with DOI, title and journal metadata.

Reference annotations. All systems are evaluated against the same final reference set of 3,412 identifier–structure pairs from 419 figures. The reference file is linked to evaluation outputs by a content hash to ensure consistent scoring. Model-output inspection informed annotation verification, and the resulting reference set was applied uniformly to all systems.

Access, provenance and permitted use. The task restricts inference to the supplied figure and task instructions; supplementary information and database records used in annotation are not inference inputs. This restriction does not eliminate possible exposure to published figures during model training. No result partitioned by model knowledge cutoff is reported here. The current source-DOI grouping controls within-paper dependence but is not an assertion of exhaustive cross-paper image or scaffold deduplication.

C Implementation and Input Conditions

C.1 Model configurations

Table 6 gives model identifiers and configurations used in the experiments. Unspecified controls are not reconstructed from model names. For most direct runs, the output-token limit was recorded as provider default rather than a fixed manuscript-wide value; temperature was not explicitly set in the common request builder. Differences in provider defaults, reasoning effort, tool access and execution budget remain part of the comparison.

Table 6: Model identifiers and inference configurations. "Default" indicates no fixed output cap recorded for that run.

System	Model ID	Reasoning control	Output cap
DeepSeek V4.1 Flash	`deepseek-flash`	Default	16,384 (config)
GPT-6 Luna	`gpt-6-luna`	Default	Default
GPT-6 Sol	`gpt-6-sol`	Default	Default
Gemini 3.8 Flash	`gemini-3.8-flash`	Default	Default
Qwen 3.8 27B	`Qwen3.8-27B`	Default	Default
Seed 2.1 Pro	`doubao-seed-2-1-pro`	Default	Default
MiMo 2.6 Pro	`mimo-v2.6-pro`	Disabled	Default
Codex / Luna	`gpt-6-luna`	Medium	600 s/figure
Codex / Sol	`gpt-6-sol`	Medium	600 s/figure
Cursor / Composer	`composer-2.5`	Native agent	600 s/figure

C.2 Shared reconstruction prompt

The shared prompt is reproduced below. Per-image scope instructions, where present, are appended from the benchmark context record. The output format is semicolon-delimited text, not a JSON

array. The prompt preserves depicted salts in the requested output even though the primary scoring policy subsequently normalizes salts and charges.

```
Identify all individually labelled chemical compounds whose structures
    are defined in the supplied image.

For each compound, read its identifier and reconstruct its complete
    molecular structure. When the image uses a shared scaffold,
    substituent table, linker, repeat unit, or other local structural
    definition, combine the relevant information according to the
    depicted attachment points, bond orders, orientation, and constraints
    .

Output complete molecules, not isolated substituents, disconnected
    fragments of a connected molecule, or unresolved Markush variables.
    Expand standard chemical abbreviations where their meaning is
    unambiguous. Do not invent missing structural definitions or unshown
    reaction products. A compound name alone is not a sufficient
    structural definition.

Preserve explicitly depicted stereochemistry. Do not infer absolute
    configuration from an ee value, a star, or a retention time alone.
    For relative stereochemistry, output one representative that
    preserves the specified relationships.

Retain depicted salt components and their multiplicity using dots. Do not
     introduce covalent bonds between noncovalently associated components
    . Follow any explicit image-specific scope or representation
    instructions.

Use the compound identifier, not the table row number or an activity
    value. Output each identifier once. Keep distinct identifiers even if
     their structures are identical. When a compound has multiple
    synonymous labels, output only one visible name or code.

Use only the supplied image and task instructions. Do not search the web
    or consult papers, databases, or supporting information.

Output only:
Identifier;SMILES
One compound per line. No explanation or Markdown.
```

C.3 Direct and OCSR-augmented requests

The common direct request builder sends the encoded image file with the high-detail image hint. It does not resize the file before encoding; any provider-internal image processing is outside this local specification. Direct and augmented conditions retain the shared task instruction and per-image scope context.

In the augmented condition, the supplied image includes detector boxes with local IDs such as B001. MolParser-Mobile candidates are linked to these box IDs through `candidate_smiles` and `markush_annotation` fields. The IDs are not compound identifiers. Candidate strings may contain errors, incomplete graphs or unresolved variables; they are not gold fragments or target answers. The augmentation does not explicitly supply the inter-region assembly relations. It jointly changes the visual input by overlaying boxes and the textual input by appending candidates, so the measured effect is that of this combined augmentation.

The recorded augmentation instruction is:

```
The annotated image contains detector box IDs B001, B002, etc. These are
    NOT compound identifiers. The following raw MolParser-Mobile OCSR
    candidates are linked to those boxes. They may be incorrect,
```

```
    incomplete, or contain Markush variables. Use the original image to
    correct and assemble molecules; output the original compound
    identifiers, not the box IDs. No candidate is ground truth.
```

It is followed by the candidate list and the instruction:

```
Also identify structures outside the detector boxes. Output only original
     compound identifiers and complete SMILES.
```

C.4 Agent execution and attempts

Agents operate in per-image workspaces and can use detection, recognition, and local graph-processing tools. Each agent run allows 600 seconds per image, at most eight detector/recognizer calls and two health calls. These are specialist service-call limits, not a limit of eight total shell commands or eight reasoning steps. Final outputs are collected as identifier–SMILES lines; optional notes are retained for diagnostics.

Transport failures and timeouts may trigger recovery attempts. Scores are computed from the final saved output for each figure in the evaluated run, with completion status recorded separately. Missing outputs retain their reference targets in the recall denominator; a timed-out run with a partial saved output can still contribute pair-level matches but cannot pass the completed-response EFR check. Wall-clock budgets and reasoning effort differ between direct and agentic settings, precluding a tool-only causal interpretation.

D Scoring Rules and Statistical Analysis

D.1 Identifier and graph matching

The evaluator matches predictions and references within the same figure using maximum one-to-one assignment. Identifier aliases explicitly recorded for a reference are accepted; aliases do not add targets to the denominator. Identifiers are trimmed before exact alias comparison.

The primary graph comparison parses molecules with RDKit, removes exact disconnected salt/counterion fragments from a versioned list, applies feasible charge neutralization, and compares canonical isomeric SMILES. Stereochemistry and isotope distinctions remain except for explicitly registered reference-bound relative-stereochemistry alternatives. Tautomers are not globally canonicalized for the main score. The salt-removal fallback never turns a salt-only structure into an empty graph. This normalization defines an operational equivalence class; it does not assess literal preservation of every depicted charge or counterion.

D.2 Prediction filtering and metrics

For a group of figures, let C be the number of matched pairs, P the number of retained predictions and G the number of reference targets. Micro-precision, recall and F1 are C/P, C/G and $2C/(P+G)$, respectively; precision is reported as zero when there are no retained predictions. EFR requires both an exact multiset match of the retained prediction pairs and recorded evidence of a completed response. It is computed over the 419 evaluation figures.

A strict prediction-line sensitivity score retains malformed nonempty output lines in P. This is distinct from the main parseable-prediction score and is sensitive to verbose or repetitive malformed responses. For example, GPT-6 Sol has 50.8 main F1 and 49.87 strict F1 on the same reference set. The difference results from including malformed or unparseable prediction rows in the strict prediction denominator. All Codex Sol scored predictions are parseable, so its two overall F1 values coincide.

D.3 Uncertainty

We use 2,000 percentile bootstrap replicates over 244 normalized source DOIs, with seed 20260926. Each sampled DOI contributes all its figures and targets. Identical sampling multiplicities are used

for original and augmented inputs and for cross-level contrasts. F1 is recomputed from pooled counts in each replicate. The intervals are pointwise and conditional on the retained runs; they neither account for repeated-inference stochasticity nor apply a multiple-comparison correction. DOI grouping captures within-paper dependence but not every possible cross-paper shared scaffold or reproduced figure.

E Complete Results and Scoring Sensitivity

E.1 Precision, recall, F1 and exact figure recovery

Table 7: Complete overall results (%). Main P/R/F1 use parseable predictions; EFR uses 419 figures.

System	Setting	Precision	Recall	F1	EFR
DeepSeek V4.1 Flash	Direct	5.2	4.3	4.7	2.1
GPT-6 Luna	Direct	27.3	19.6	22.8	10.3
GPT-6 Sol	Direct	53.5	48.3	50.8	26.5
Gemini 3.8 Flash	Direct	44.8	41.2	42.9	22.4
Qwen 3.8 27B	Direct	14.8	9.2	11.4	2.1
Seed 2.1 Pro	Direct	43.0	40.5	41.7	25.1
MiMo 2.6 Pro	Direct	5.2	3.9	4.5	1.0
Codex (GPT-6 Luna)	Agent	58.2	40.7	47.9	30.8
Codex (GPT-6 Sol)	Agent	89.9	88.5	89.2	74.9
Cursor (Composer 2.5)	Agent	55.1	47.5	51.0	36.8
DeepSeek V4.1 Flash	Augmented	17.8	16.1	16.9	18.4
GPT-6 Luna	Augmented	65.6	59.2	62.2	50.1
GPT-6 Sol	Augmented	74.3	71.7	73.0	56.6
Gemini 3.8 Flash	Augmented	56.1	53.1	54.6	30.1
Qwen 3.8 27B	Augmented	39.3	30.2	34.2	22.4
Seed 2.1 Pro	Augmented	68.8	67.3	68.1	51.8
MiMo 2.6 Pro	Augmented	22.0	17.0	19.2	11.0

Table 8: OCSR-augmented F1 (%) by figure-level label.

Model	L1	L2	L3	Overall
DeepSeek V4.1 Flash	62.2	7.8	3.4	16.9
GPT-6 Luna	83.3	71.4	40.5	62.2
GPT-6 Sol	86.5	78.4	60.7	73.0
Gemini 3.8 Flash	71.1	48.6	51.7	54.6
Qwen 3.8 27B	71.3	21.0	25.3	34.2
Seed 2.1 Pro	86.6	66.7	59.6	68.1
MiMo 2.6 Pro	51.7	9.6	9.3	19.2

E.2 Paired augmentation gains

Table 9: Paired F1 gains and their pointwise 95% source-paper bootstrap intervals (percentage points). The final column is the L1 gain minus the L3 gain.

Model	L1	L2	L3	L1 minus L3
DeepSeek V4.1 Flash	48.1 [39.4, 57.0]	5.4 [1.2, 10.4]	1.5 [-1.5, 4.7]	46.6 [37.4, 55.9]
GPT-6 Luna	59.5 [51.8, 66.9]	47.6 [37.6, 56.9]	19.4 [10.2, 27.3]	40.1 [28.7, 51.6]
GPT-6 Sol	35.8 [27.9, 43.0]	27.0 [17.1, 36.7]	10.5 [-0.5, 22.4]	25.3 [11.3, 38.4]
Gemini 3.8 Flash	25.6 [17.1, 34.1]	7.2 [0.4, 14.4]	8.6 [-1.8, 17.8]	17.0 [4.4, 30.6]
Qwen 3.8 27B	59.2 [51.7, 66.0]	13.4 [6.8, 19.9]	10.6 [4.3, 18.9]	48.5 [36.9, 57.9]
Seed 2.1 Pro	34.7 [25.9, 43.3]	29.9 [19.2, 40.8]	18.5 [8.7, 27.9]	16.3 [3.0, 29.3]
MiMo 2.6 Pro	43.5 [35.6, 51.2]	7.1 [1.4, 13.7]	5.0 [0.5, 9.2]	38.6 [29.7, 47.5]

All seven point estimates favor a larger L1 gain, and their pointwise intervals under the main policy exclude zero. L3 gain intervals include zero for some models. These results concern the combined box-and-candidate input augmentation, not an intervention providing verified-correct components.

E.3 Strict prediction-line sensitivity

Table 10: Strict prediction-line sensitivity. F1 values are percentages; differential gains and intervals are percentage points.

Model	Direct parseable F1	Direct strict F1	Strict L1 minus L3 gain
DeepSeek V4.1 Flash	4.7	4.4	42.5 [32.4, 52.4]
GPT-6 Luna	22.8	21.3	40.8 [29.6, 52.4]
GPT-6 Sol	50.8	49.9	26.1 [11.9, 39.3]
Gemini 3.8 Flash	42.9	6.9	32.6 [-4.5, 72.0]
Qwen 3.8 27B	11.4	9.0	49.4 [39.8, 57.9]
Seed 2.1 Pro	41.7	40.5	17.6 [4.2, 30.6]
MiMo 2.6 Pro	4.5	1.6	42.1 [30.5, 49.9]

All differential-gain point estimates remain positive under strict prediction-line scoring, but the Gemini interval includes zero. Gemini's overall direct F1 changes from 42.9 to 6.9 because many malformed lines enter the strict prediction denominator. The two scoring policies capture different aspects of output quality.

F Figure Multiplicity and Single-Molecule Recognition

F.1 Target-count stratification and standardization

This analysis uses the final evaluation set of 419 figures and 3,412 target instances. We report EFR in the fixed bins 1–3, 4–6, 7–10 and 11 or more targets. As an exploratory adjustment, we restrict L1/L3 comparisons to exact counts represented in both levels and weight the within-count EFR by their pooled count distribution on this support. The common counts are 1–11, 13, 14, 16 and 24, retaining 180 L1 and 77 of 105 L3 figures. This selection excludes 28 L3 figures outside common support.

For Codex Sol, unweighted EFR on common support is 81.7% for L1 and 70.1% for L3. Standardization yields 78.3% and 74.5%, compared with the unadjusted full-sample rates of 81.7% and 62.9%. Both support selection and weighting affect the comparison. The analysis does not control molecular complexity or image quality; sparse exact-count cells also limit interpretation. It is not a causal decomposition of the original EFR gap.

Table 11: EFR by target count, shown as exact / eligible figures.

System	Level	1–3	4–6	7–10	11+
Codex (GPT-6 Luna)	L1	56/105	19/44	10/22	2/9
Codex (GPT-6 Luna)	L2	5/8	8/30	18/48	7/48
Codex (GPT-6 Luna)	L3	2/15	1/23	0/21	1/46
Codex (GPT-6 Sol)	L1	91/105	34/44	16/22	6/9
Codex (GPT-6 Sol)	L2	6/8	21/30	42/48	32/48
Codex (GPT-6 Sol)	L3	13/15	15/23	15/21	23/46
Cursor (Composer 2.5)	L1	68/105	23/44	12/22	4/9
Cursor (Composer 2.5)	L2	7/8	9/30	15/48	12/48
Cursor (Composer 2.5)	L3	2/15	1/23	0/21	1/46

F.2 Rendered-molecule diagnostic

All 3,412 annotation instances in the rendering manifest are rendered independently, without cross-figure structure deduplication. Rendering uses RDKit 2026.03.6, CoordGen coordinates, a Cairo

Table 12: Exploratory EFR standardized to a common exact-target-count distribution (%). All systems use the same 180 L1 and 77 L3 figures.

System	L1	L3	L1–L3 gap
Codex (GPT-6 Luna)	44.8	6.8	37.9
Codex (GPT-6 Sol)	78.3	74.5	3.9
Cursor (Composer 2.5)	56.5	6.8	49.6

flexible canvas, fixed bond length 32, a monochrome atom palette, padding 0.1 and an additional 20-pixel white border. MolParser-Mobile receives only complete rendered molecule images, without identifiers, reference SMILES, detector outputs or source-figure context. Greedy decoding uses maximum length 256, with at most 16 independent crops per request and shuffled input order (seed 20260924).

Exact canonical comparison preserves fragments, charges, isotopes, tautomer identity and stereochemistry. A secondary comparison removes stereochemistry only. Invalid predictions remain incorrect; the experiment records no API failures or truncated candidates. This scoring differs from the main track's desalt/neutralization policy. The diagnostic measures MolParser-Mobile performance on rendered complete molecules, not MolParser-Mobile candidate quality on the original multi-region figures.

Table 13: Rendered-molecule accuracy with source-paper bootstrap intervals.

Figure group	Instances	Exact (%)	Stereo-stripped (%)
L1	709	82.4 [77.8, 86.8]	90.8 [87.3, 94.0]
L2	1,324	82.6 [75.9, 88.3]	91.0 [86.5, 94.8]
L3	1,379	69.7 [59.6, 77.7]	80.5 [71.3, 87.8]

Table 14: Descriptive properties of the final evaluation set. Counts refer to target instances rather than unique molecular graphs.

Group	Instances	Mean targets/figure	Median targets	Mean heavy atoms	Median atoms
L1	709	3.94	3.0	25.19	26.0
L2	1,324	9.88	9.0	28.85	29.0
L3	1,379	13.13	8.0	30.52	29.0

The L3 recognition deficit persists after removing stereochemistry, but this result does not determine how much of the original figure-level difference is caused by perception versus assembly. The rendered-molecule diagnostic uses the final reference set of 3,412 target instances from 419 figures. All rendered inputs, reference structures, and reported recognition statistics are aligned with this version.

G Mismatch Diagnostics and Failure Cases

G.1 Automatic diagnostic categories

The analysis partitions unmatched reference targets, not all predicted rows. If there is no candidate under an accepted identifier alias, the evaluator checks whether the reference structure occurs under another identifier; otherwise the target is missing. If same-identifier candidates exist but all are unparseable, the target is assigned to invalid SMILES. Remaining cases are tested for stereo-only and tautomer-only equivalence. Other discrepancies enter the residual connectivity/bond-order category, which can include different elements, atom counts or local recognition mistakes; it is not an atom-mapped attribution to an assembly operation.

These categories are mutually exclusive for the current reference-target accounting. A wrong prediction may create both a false negative and a false positive, so those counts must not be added as

the number of erroneous molecules. The presence of the correct structure under another identifier is a diagnostic signal rather than conclusive evidence of an identifier mistake, because some distinct identifiers legitimately refer to the same structure.

Table 15: Unmatched targets for GPT-6 Sol under direct prediction, OCSR augmentation, and Codex, grouped by figure-level label. Conn. denotes connectivity or bond-order mismatches.

Setting	Level	Targets	Unmatched	Conn.	Stereo	Missing ID	Other ID	Invalid	Tautomer
original	L1	709	362	186	88	41	18	29	0
original	L2	1324	661	453	120	34	2	51	1
original	L3	1379	740	427	59	204	19	30	1
augmented	L1	709	94	41	19	4	30	0	0
augmented	L2	1324	307	178	55	16	14	38	6
augmented	L3	1379	563	368	89	71	17	18	0
agent	L1	709	62	28	23	1	10	0	0
agent	L2	1324	140	56	52	22	6	0	4
agent	L3	1379	189	84	47	38	20	0	0

G.2 Paired target outcomes

For GPT-6 Sol, augmentation recovers 1,041 targets missed under direct prediction but loses 242 targets previously recovered. Codex recovers 1,396 targets missed under direct prediction and loses 24. These are paired changes in matching outcomes, not a causal attribution to particular tool calls or a claim that each newly unmatched target reflects a confirmed model mistake.

G.3 Operational failures

The Codex GPT-6 Sol run contains five timeouts. Four produce no output, affecting 35 reference targets. The fifth retains 52 matches among 54 targets. Saved partial predictions contribute to pair-level scoring, while timed-out responses do not satisfy the completion requirement for EFR.

G.4 Representative cases

The following cases were selected for source-image and execution-trace inspection. They demonstrate possible failure mechanisms but are not a random sample from which their frequencies can be estimated.

Merged-cell inheritance: L2_004. The table shares a methylsulfonyl substituent across several rows. Codex interprets blank entries for 7p–7s as hydrogen rather than carrying over the shared value. The resulting graphs omit the substituent while remaining parseable and passing the recorded RDKit checks. This is a table-relation failure within an L2 figure, not evidence that the failure is exclusive to L3.

Explicit linker omission: A026. For target 2s, the local template contains N–CH_2–Alk with Alk = nPr. The explicitly drawn methylene must remain in the expanded structure. All three settings instead produce an N-propyl substituent, losing one carbon. The Codex trace includes stereochemical checks, illustrating that those checks do not establish correctness of the linker length.

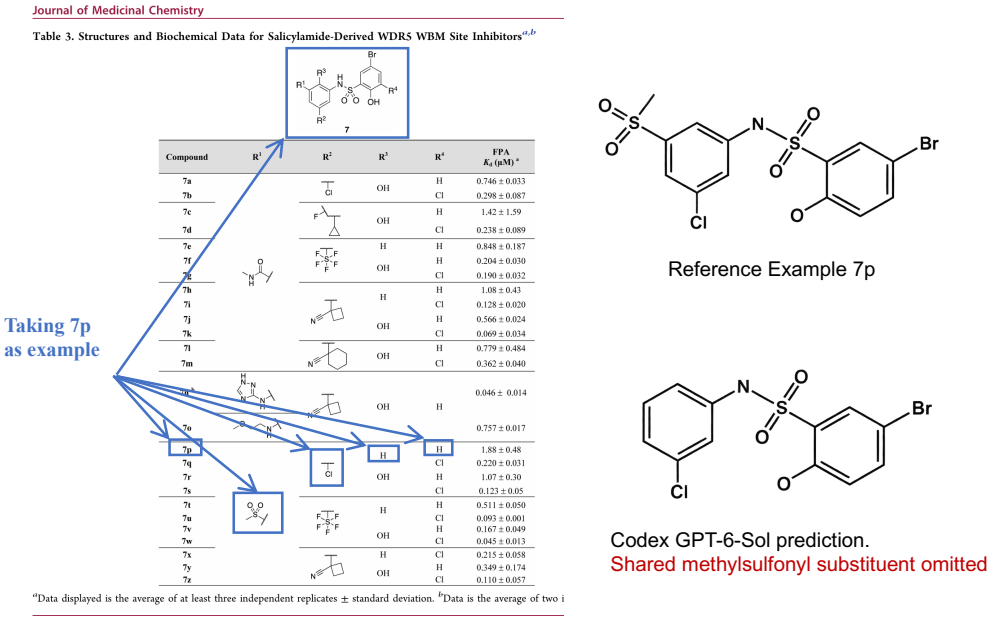


Figure 8: Shared-substituent omission in L2_004, illustrated by compound 7p. The reference structure incorporates the methylsulfonyl group specified by the shared R^1 cell, whereas the Codex Sol prediction omits this substituent.

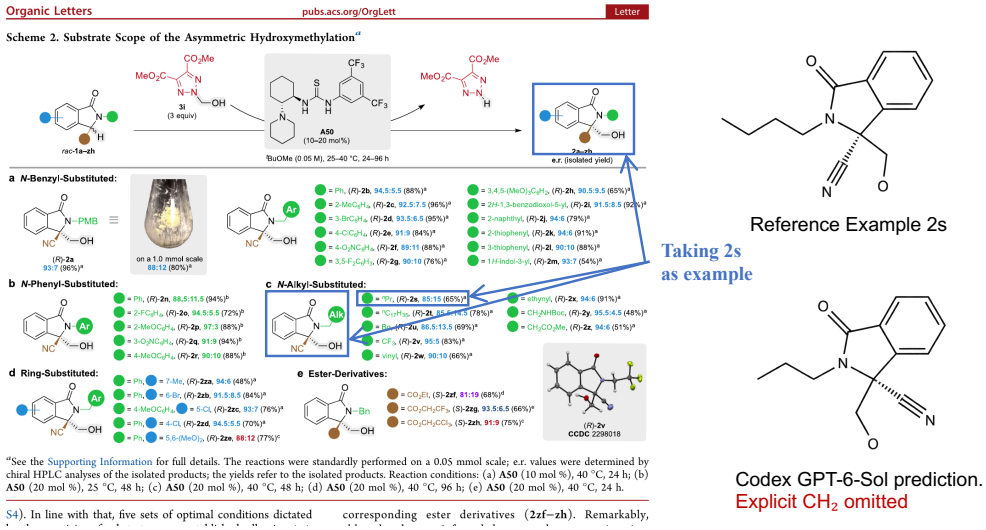


Figure 9: Explicit methylene omission in A026, illustrated by compound 2s. The template specifies N–CH_2–Alk with Alk = nPr, yielding an N-butyl substituent. The Codex Sol prediction instead contains an N-propyl substituent, omitting the explicitly depicted methylene.