RxnExBench: Benchmarking Vision-Language Models for Structured Chemical Reaction Extraction from Scientific Literature

Anonymous ACL submission

Abstract

Screen Tables and *Substrate Scopes* are information-dense figures widely used in synthetic chemistry literature, yet automated extraction of structured reaction data from them remains challenging. We introduce **RxnExBench**, a multimodal benchmark for fine-grained reaction-level evaluation. Built from high-impact chemistry journals, it contains 3,565 expert-annotated reactions and 2,361 molecular structures. The dataset adopts a unified schema with 15 reaction roles and 26 core fields, supporting downstream AI-for-Chemistry tasks such as retrosynthetic planning, yield prediction, and reaction condition recommendation. RxnExBench further provides a reasoning-driven difficulty taxonomy, a modular annotation pipeline, and the Oracle-MolRec protocol, which disentangles molecular recognition from higher-level reasoning. Evaluation of 9 state-of-the-art VLMs shows that performance degrades substantially as schema complexity and reasoning demands increase. Under the Oracle-MolRec setting, results further reveal that accurate molecular recognition alone is insufficient, with multimodal integration and chemistry-aware reasoning remaining the primary bottlenecks. We release all data and code to facilitate future research in chemistry-aware multimodal understanding and AI-for-Chemistry dataset construction.

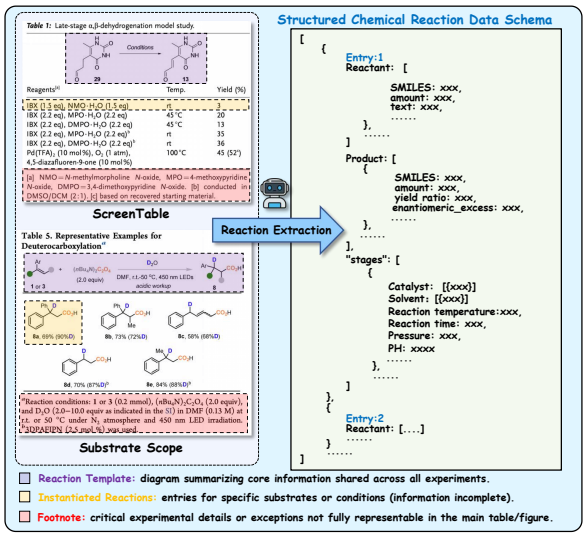


Figure 1: *Screen Table* and *Substrate Scope* are information-rich figure types commonly used in chemical papers to present chemical reactions. A typical *Screen Table* or *Substrate Scope* usually contains: a "reaction template" presented as an image (purple box), "instantiated" chemical reactions in a tabular or substrate array format (yellow box), and a footnote for important supplementary information that cannot be fully displayed in the figure (red box). The task of **RxnExBench** is to extract structured information for all chemical reactions from these figures according to the defined schema.

1 Introduction

Recent years have witnessed significant progress in applying artificial intelligence to tasks such as chemical reaction prediction (Joung et al., 2025), reaction condition optimization (Wang et al., 2024), and retrosynthetic planning (Xiong et al., 2025; Deng et al., 2025; Xu et al., 2025). However, a critical bottleneck in current AI-for-Chemistry research remains the scarcity of large-scale, high-quality chemical reaction datasets (Ding et al., 2025; Zhong et al., 2024). While the vast chemical literature contains abundant reaction information, it is predominantly presented in unstructured formats, such as complex figures and tables, making it unsuitable for direct use in training machine learning models. Consequently, automatically extracting structured reaction data from scientific papers has become an important and actively pursued research direction.

Existing efforts have largely focused on textual reaction information extraction(Guo et al., 2021; Zhong et al., 2023b), molecular structure recognition(Qian et al., 2023b; Wang et al., 2025), and reaction scheme parsing(Qian et al., 2023a;

Table 1: Comparison of related benchmarks and our **RxnExBench**. StructRxnEx indicates that the benchmark's task is structured chemical reaction information extraction; R-group Resolved means that for all instantiated chemical reactions, the R-groups in all molecular structure diagrams have been resolved; "CLS figures" means that the benchmark contains the "complex-line-substrate" figures (see Figure 3 for details).

Benchmark	Figure Types	StructRxnEx	R-group Resolved	CLS Figures	Difficulty Levels	# Schema Fields	#Rxn	Avg Tokens /Fig	Avg Rxn /Fig
ChemTable	Screen Table, Substrate Scope	×	×	–	–	–	–	–	–
MERMES-T24	Substrate Scope	×	×	–	–	–	–	–	–
ChemEAGLE	Substrate Scope	✓	✓	×	Level 3	11	1120	1,179.91	9.21
MerMaid	ScreenTable	×	✓	×	Level 1,2	15	1343	2,707.64	12.32
RxnExBench	**ScreenTable, Substrate Scope**	✓	✓	✓	**Level 1,2,3**	**26**	**3565**	**7,205.28**	**13.40**

Chen et al., 2025b; Song et al., 2025). Yet, these highly informative but complex graphical formats—specifically *Screen Tables* and *Substrate Scopes*—remain largely underexplored, despite being ubiquitous in synthetic chemistry papers.

As illustrated in Figure 1, *Screen Tables* compare closely related reactions under varying conditions to identify optimal parameters, while *Substrate Scopes* present reaction outcomes across diverse reactants under fixed conditions to demonstrate reaction generality. Both formats encode dense experimental knowledge through compact notations, domain-specific abbreviations, and embedded molecular diagrams. Consequently, they represent highly valuable yet notoriously challenging sources for automated extraction. Accurate parsing therefore requires not only visual understanding and OCR capabilities, but also chemistry-aware symbolic interpretation and cross-modal reasoning.

The recent rise of Vision-Language Models (VLMs), such as Gemini(Google DeepMind, 2026), Qwen-VL(Bai et al., 2025b), and InternVL(Chen et al., 2024), has shown great promise in scientific document understanding, with notable advances in layout analysis(Wei et al., 2024; Niu et al., 2025), molecular graph recognition(Wang et al., 2025), reaction diagram parsing(Song et al., 2025), and property extraction(Yan et al., 2025). Nevertheless, the performance of VLMs in comprehending and structuring reactions from *Screen Tables* and *Substrate Scopes* has not been systematically evaluated. As shown in Table 1, existing benchmarks exhibit clear limitations: MERMES(Leong et al., 2024) focuses solely on condition extraction without modeling full reactions; ChemTable(Zhou et al., 2025) addresses table structure but ignores reaction semantics; while ChemEAGLE(Chen et al., 2025c) and MerMaid(Leong et al., 2025) suffer from narrow chart coverage and oversimplified evaluation protocols, ultimately failing to reflect the true complexity of real-world chemical literature.

To address this critical gap, we introduce **RxnExBench**, a comprehensive benchmark designed to evaluate VLMs on structured reaction extraction from *Screen Tables* and *Substrate Scopes*. Shifting the paradigm from figure-level evaluation to fine-grained, reaction-level understanding, RxnExBench adopts a unified schema featuring 15 reaction roles and 26 core fields, explicitly designed to support downstream AI-for-Chemistry tasks. Furthermore, we categorize the extraction process into three reasoning-driven difficulty levels, providing a granular diagnosis of model capabilities. Compared to existing benchmarks, RxnExBench introduces substantially greater complexity, multimodal integration, and reasoning diversity. Our main contributions are summarized as follows:

1. **A systematic framework for chemistry-aware extraction and reasoning evaluation.** To tackle the significant challenge of extracting structured chemical reactions from *Screen Tables* and *Substrate Scopes*, we formalize this end-to-end task and design a comprehensive 26-field schema aligned with major AI-for-Chemistry applications. Recognizing that this extraction demands holistic multimodal reasoning beyond basic chart parsing, we further introduce a three-tier classification (local reading, global integration, and R-group-aware spatial reasoning) grounded in chemistry, which empirically correlates with VLM error rates.

2. **A modular high-quality annotation pipeline.** To support the complex demands of this task, we develop a multi-stage labeling workflow (template annotation, instantiation, Ketcher-based molecule drawing, RDKit canonicalization). Executed by 40 chemistry-trained annotators with three rounds of

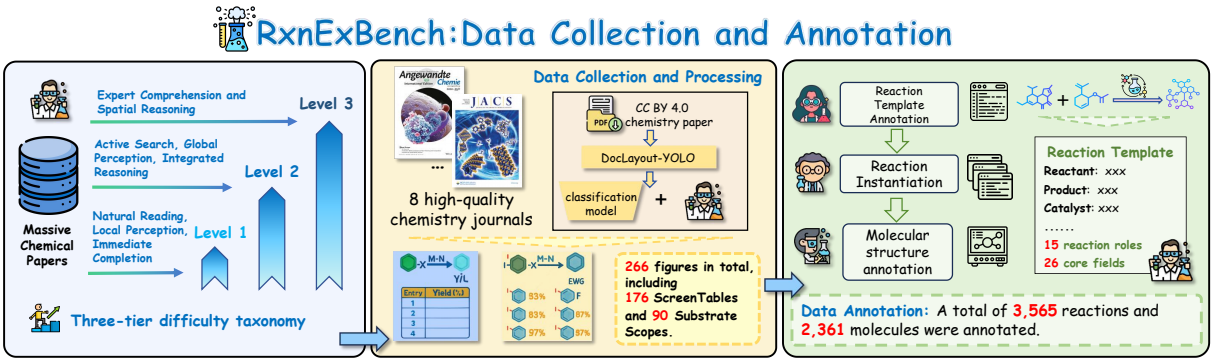


Figure 2: Data collection and annotation of RxnExBench

cross-validation, this pipeline yields a robust dataset of 3,565 reactions, 2,361 molecular structures and 53,659 annotated fields.

3. **Empirical study and the Oracle-MolRec protocol.** We benchmark 9 state-of-the-art VLMs and design the Oracle-MolRec inference protocol to explicitly decouple fundamental molecular recognition from higher-level reasoning. Our evaluation shows that, while VLMs handle basic chart parsing, they struggle with holistic chemistry-grounded integration as schema demands grow; moreover, Oracle-MolRec reveals that even with perfect molecular recognition, higher-level reasoning remains the dominant bottleneck for fully automated extraction.

2 Related Work

Large-scale structured chemical reaction datasets extracted from the scientific literature (Schneider et al., 2016a,b) have significantly advanced AI-for-Chemistry research. Recent efforts in automated extraction have predominantly focused on *text-based* approaches, including deep learning methods (Guo et al., 2021) and, more recently, large language model (LLM)-driven techniques (Zhong et al., 2023b). However, a substantial portion of critical reaction information in chemistry papers is conveyed through *images*, motivating growing interest in *image-based* reaction extraction. Several works have targeted general reaction diagram parsing (Qian et al., 2023a; Chen et al., 2025b; Song et al., 2025), yet *Screen Tables* and *Substrate Scopes*, which are two highly informative and challenging chart types, remain underexplored despite their ubiquity in synthetic chemistry literature.

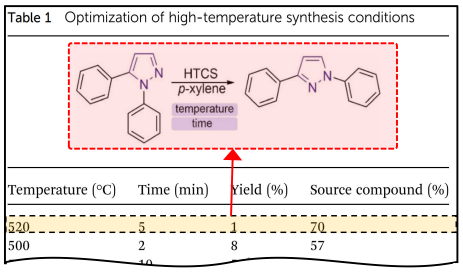
As summarized in Table 1, existing datasets exhibit notable limitations. **MERMES** (Leong et al., 2024) constructs the MERMES-T24 dataset from 65 papers and evaluates five VLMs on reaction diagram parsing and cross-modal reasoning. However, its annotations focus only on reaction conditions and a single substrate, making it suitable for partial element extraction but inadequate for assessing complete reaction reconstruction. **MerMaid** (Leong et al., 2025) covers broader reaction scenarios and introduces SMILES fields to evaluate molecular-to-SMILES conversion. It specifically targets R-group-dependent reasoning in *Screen Tables*, but does not address *Substrate Scopes* or multi-condition comparison. **ChemEAGLE** (Chen et al., 2025c) evaluates VLMs on *Substrate Scope* understanding, yet considers only a limited set of reaction roles, falling short in assessing holistic, multi-modal integration into full reaction records. **ChemTable** (Zhou et al., 2025) includes both *Substrate Scopes* and *Screen Tables*, but primarily measures table-structure comprehension rather than structured chemical reaction extraction.

Table 2: Statistics of reaction figures and counts.

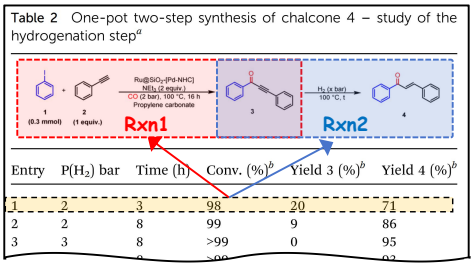
Type	#Figures	#Min Rxn	#Max Rxn	#Avg Rxn	#Total Rxn
ScreenTable	176	3	29	11.53	2,029
SubstrateScope	90	3	44	17.07	1,536
Level 1	60	4	19	9.85	591
Level 2	90	3	29	12.12	1,091
Level 3	116	3	44	16.23	1,883
Total	**266**	**3**	**44**	**13.40**	**3,565**

3 RxnExBench

We present **RxnExBench**, a multimodal benchmark designed to comprehensively and rigorously evaluate Vision-Language Models (VLMs) on their ability to understand, interpret, and extract *structured chemical reaction information* from two critical yet understudied chart types in synthetic chemistry literature: *Screen Tables* and *Substrate Scopes*.



(a) Simple-Line-Substate figure



(b) Complex-Line-Substate figure

Figure 3: In *Screen Tables*, each row typically corresponds to one instantiated reaction; in *Substrate Scopes*, each substrate likewise. Figures where all rows/substrates meet this condition are called "simple-line-substrate" (SLS) figures, as shown in (a). However, some rows or substrates may contain multiple reactions; these are termed "complex-line-substrate" (CLS) figures, as shown in (b).

3.1 Structured Reaction Schema

Inspired by prior works (Zhong et al., 2023a; CAS, 2025), we define a structured schema for representing chemical reactions (Figure 1). Each reaction consists of three top-level components: `reactants`, `products`, and `stages`. The `reactants` and `products` store detailed information for each molecular participant, while each `stage` represents a reaction step with associated conditions such as `reagent`, `catalyst`, and `solvent`. The schema covers all essential reaction information, including SMILES strings, reactant amounts, and product yields (Appendix A). This design is both semantically complete and practically useful, supporting downstream AI-for-Chemistry tasks such as product prediction (Joung et al., 2025), yield estimation, condition recommendation (Wang et al., 2024), and retrosynthetic planning (Xiong et al., 2025; Deng et al., 2025; Xu et al., 2025). Accordingly, RxnExBench evaluates whether VLMs can accurately extract complete schema-compliant reaction records from real-world *Screen Table* and *Substrate Scope* figures.

3.2 Difficulty-Aware Image Collection and Curation

As illustrated in Figure 1, *Screen Tables* and *Substrate Scopes* are two prevalent yet distinct formats for presenting reaction data in chemical publications (see Appendix A for detailed descriptions). In practice, they exhibit substantial heterogeneity in visual layout, information density, and required inference complexity.

Based on systematic analysis of real-world figures, we propose a three-tier difficulty taxonomy along two orthogonal axes—*visual scope* (local vs. global perception) and *reasoning depth* (literal extraction vs. chemistry-aware inference). Each level is defined by concrete, observable operations a reader must perform:

- **Level 1: Natural Reading, Local Perception, Immediate Completion.** Extracting structured chemical reaction information from Level 1 figures follows standard human reading order: first parse the reaction template and populate known information into the schema fields, then sequentially process each row (for *Screen Table*) or substrate entry (for *Substrate Scope*) to update fields such as specific reactants, catalysts, temperature, or time—thereby instantiating complete chemical reactions, as shown in Figure 3(a). This requires basic document understanding, OCR, molecular diagram recognition, and foundational chemical knowledge.

- **Level 2: Active Search, Global Perception, Integrated Reasoning.** For Level 2 figures, linear reading is insufficient. Models must actively search and integrate information across distant regions—for example, when key conditions are specified only in footnotes, as illustrated by the purple arrow in Stage 2 of Figure 4. Another challenge arises when a row or substrate corresponds to multiple reactions. We define figures with one reaction per row/substrate as "simple-line-substrate" (SLS) figures, and those containing multiple reactions as "complex-line-substrate" (CLS) figures (Figure 3(b)). These CLS figures belong to Level 2 or higher, requiring cross-region awareness and contextual integration.

- **Level 3: Expert Comprehension and Spatial Reasoning.** Building on Level 2, Level 3 requires domain-specific chemical reasoning.

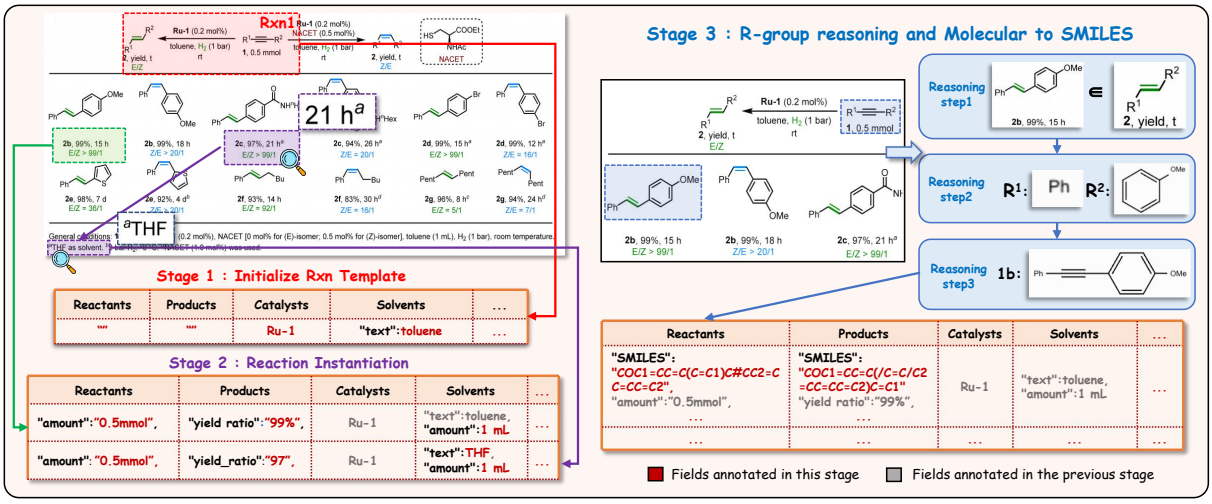


Figure 4: Three stages of structured chemical reaction extraction from *Screen Tables* and *Substrate Scopes*, describing the processes of human expert reading, VLM understanding, and manual data annotation.

As shown in the Stage 3 of Figure 4, in *Substrate Scopes* (and complex *Screen Tables*), the model must infer variable R-groups from substrate structures and propagate these substitutions back to reactants to reconstruct full molecular identities: a non-trivial spatial and symbolic inference task.

To construct a representative dataset, we collected ∼5,000 CC-4.0-licensed papers from 8 high-impact chemistry journals (list in Appendix B). After PDF-to-image conversion and chart region extraction using DocLayout-YOLO, we obtained ∼370K candidate figures. A custom classifier then identified 2,269 *Substrate Scope* and 2,413 *Screen Table* images as our initial pool.

From this large-scale collection, chemistry experts performed multi-round manual screening, preliminary annotation, and difficulty assessment to select the most representative and reasoning-intensive examples. The final benchmark contains 90 *Substrate Scope* and 176 *Screen Table* figures with coverage across all three difficulty levels (Table 2).

3.3 Data Annotation Pipeline

Fully annotating a single *Screen Table* or *Substrate Scope* still requires an average of 230 and 339 minutes[1] respectively—reflecting both the deep chemical knowledge required and the visual parsing burden, and motivating our decomposition into focused, independently quality-controlled sub-tasks. We therefore design a modular multi-stage annotation pipeline (Figure 2) that decomposes the workflow into three primary stages and several focused sub-tasks with independent quality control (details in Appendix C.2), improving both annotation efficiency and reliability:

1. **Reaction Template Annotation**: Annotate all directly observable information in the generic reaction template (e.g., fixed reactants, base conditions). Fields requiring instance-specific context, such as R-group identities, yields, temperatures, or SMILES are left blank.

2. **Template Instantiation**: For each substrate (*Substrate Scope*) or table row (*Screen Table*), instantiate the abstract template into a concrete reaction by filling in context-dependent fields (e.g., specific catalyst, solvent, yield). SMILES strings are still deferred.

3. **Molecular Structure (SMILES) Annotation**: Chemists use the open-source tool `Ketcher` to reconstruct complete molecular structures from diagrams and substrate context, addressing challenges posed by complex long-form SMILES and implicit R-group reasoning unsupported by existing diagram-to-SMILES models. The resulting `molfiles` are then automatically converted into standardized SMILES using `RDKit`.

We employed 40 chemistry-trained annotators with three rounds of cross-validation. In total, we invested 1,183 person-hours and produced 3,565 annotated reactions and 2,361 canonical SMILES strings. For more annotation details, please refer to Appendix C.

[1]Detailed statistics in Appendix C.1

Table 3: F1 scores of VLMs on RxnExBench (precision/recall in Appendix G). Results are shown for the full dataset and three difficulty levels under three evaluation criteria. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	All Levels (266 Figures)			Level 1 (60 Figures)			Level 2 (90 Figures)			Level 3 (116 Figures)		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.52%	1.08%	0.94%	5.60%	4.24%	4.24%	2.40%	1.20%	0.74%	1.62%	0.00%	0.00%
Gemini 3.1 Pro	19.47%	**9.57%**	**8.86%**	**40.42%**	**20.56%**	**19.86%**	15.46%	10.22%	**8.25%**	15.24%	**5.77%**	**5.77%**
GPT-5.5	**23.16%**	9.16%	6.81%	33.50%	12.79%	12.63%	**28.13%**	**13.93%**	6.41%	**16.91%**	5.18%	5.18%
GPT-4o	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.17%	0.00%	0.00%
Doubao Seed 2.0 Pro	19.81%	6.59%	5.35%	33.56%	11.36%	10.68%	25.70%	7.14%	4.55%	11.80%	4.71%	4.10%
GLM-5V Turbo	5.08%	1.92%	1.80%	15.48%	6.80%	6.29%	2.77%	1.29%	1.20%	2.90%	0.60%	0.60%
Intern-S1	2.88%	0.06%	0.06%	4.59%	0.00%	0.00%	6.90%	0.19%	0.19%	0.00%	0.00%	0.00%
Llama 4 Maverick	2.27%	0.86%	0.86%	4.58%	2.38%	2.38%	2.83%	0.10%	0.10%	0.99%	0.80%	0.80%
Qwen3-VL-235B	1.68%	0.26%	0.06%	5.62%	1.53%	0.34%	2.23%	0.00%	0.00%	0.00%	0.00%	0.00%

3.4 Evaluation Task Design

We formulate the task as end-to-end extraction: given a *Screen Table* or *Substrate Scope* image, the VLM must output *all* reactions in the image as a single JSON object compliant with the schema (prompt template in Appendix E).

We evaluate using *reaction-level* precision, recall and F1. For each image, predicted and ground-truth reaction lists are matched, and three criteria determine whether a prediction is a true positive (TP):

- **Criterion 1** (for retrosynthesis, etc.): match only on `reactants`, `products`, and `reagents`.
- **Criterion 2** (for yield prediction): adds `catalyst` and `solvent` to Criterion 1.
- **Criterion 3** (for condition recommendation, etc.): the strictest standard, requiring exact match on `reactants`, `products`, `reagents`, `catalyst`, `ligand`, `solvent`, `pressure`, `reaction time`, `reaction temperature`, and `pH`.

A prediction is counted as TP *only if all fields specified by the criterion match exactly*; otherwise, it is FP or FN. Full evaluation protocol is detailed in Appendix F.

4 Experiments

4.1 Model Selection and Experimental Setup

We conducted an in-depth evaluation of state-of-the-art Vision-Language Models (VLMs), including both open-source models and commercial APIs: Claude Opus 4.6 (Anthropic, 2026), Gemini 3.1 Pro (Google DeepMind, 2026), GPT-5.5 (OpenAI, 2026), GPT-4o (Hurst et al., 2024), Doubao Seed 2.0 Pro (Seed, 2026), GLM-5V Turbo (Hong et al., 2026), Intern-S1 (Bai et al., 2025a), Llama 4 Maverick (Llama Team, 2025), and Qwen3-VL-235B (Leong, 2025).

While we primarily focus on end-to-end VLM extraction, we also evaluated two pipeline-based tools (MerMaid (Leong et al., 2025) and ChemEAGLE (Chen et al., 2025c)); results are in Appendix H.

4.2 Overall Performance and Preliminary Findings

The F1 scores of all models across different difficulty levels and evaluation criteria on RxnExBench are presented in Table 3, with precision and recall metrics detailed in Appendix G. Our analysis reveals three key findings:

First, complete and accurate extraction of structured chemical reaction information from *Screen Tables* and *Substrate Scopes* remains highly challenging for state-of-the-art VLMs. Under Criterion 1—the most practically relevant setting for retrosynthesis-style downstream tasks—GPT-5.5 leads at 23.16% F1, with Doubao Seed 2.0 Pro (19.81%) and Gemini 3.1 Pro (19.47%) forming a clear second tier. The remaining six models cluster below 6%. Under the stress-test Criterion 3, which requires exact matching across ten fields, even the strongest model reaches only 8.86% F1. Together, these results highlight substantial room for improvement in current VLMs for structured chemical information extraction. Human expert performance on the same task is reported in Appendix C.5.

Second, extraction difficulty increases signif-

icantly when downstream tasks require more fields from the *Schema*. As shown in Figure 5, when progressing from Criterion 1 to Criterion 3, all VLMs exhibit a consistent and significant performance decline, demonstrating that field complexity is a critical factor affecting extraction accuracy.

Third, VLM performance varies substantially across difficulty levels. Figure 5 illustrates the average performance distribution of all 9 models across the three difficulty levels, revealing a clear stepwise performance degradation from Level 1 to Level 3. For instance, Gemini 3.1 Pro achieves only 15.24% F1 on Level 3 data under Criterion 1, significantly lower than that on Level 1 data under the same criterion (40.42%).

4.3 Reaction Count as a Proxy for Basic Understanding Ability

Beyond aggregate F1, we introduce *reaction-count consistency* as a complementary diagnostic that targets a more fundamental ability: whether a VLM correctly perceives *how many* reactions a figure contains, independently of whether each individual reaction is extracted correctly. Because every downstream field is bound to a specific reaction, miscounting reactions propagates errors into the entire output; accurate reaction counting is therefore a prerequisite for any structured extraction and a clean probe of basic chart parsing, decoupled from chemistry-aware reasoning.

As defined in Section 3.2, RxnExBench figures are labeled as SLS or CLS . We analyze the two groups separately. For SLS figures, VLMs can generally match the ground truth reaction count, indicating basic chart parsing ability. However, for CLS figures, performance drops sharply: even the best model (Doubao 2.0) achieves only 60% consistency with ground truth, with most VLMs significantly undercounting reactions, demonstrating that multi-reaction rows require broader context awareness beyond natural reading order.

4.4 Error Analysis

To systematically analyze VLM failure modes, we categorize errors into three non-exclusive types: **Field Missing** (required fields omitted), **SMILES Error** (incorrect SMILES for molecules drawn as diagrams), and **Content Error** (incorrect content for non-SMILES fields).

We performed systematic error analysis for all VLMs on RxnExBench (detailed results in Appendix G). SMILES-related errors dominate the

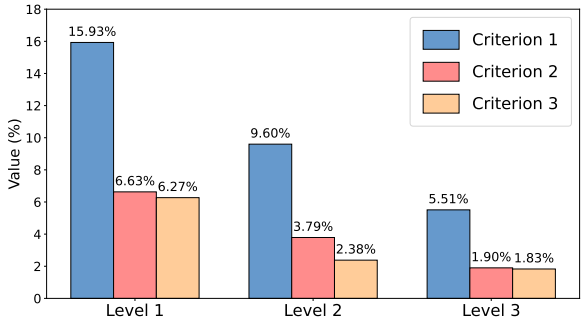


Figure 5: Average F1 scores of all VLMs across RxnExBench difficulty levels and evaluation criteria.

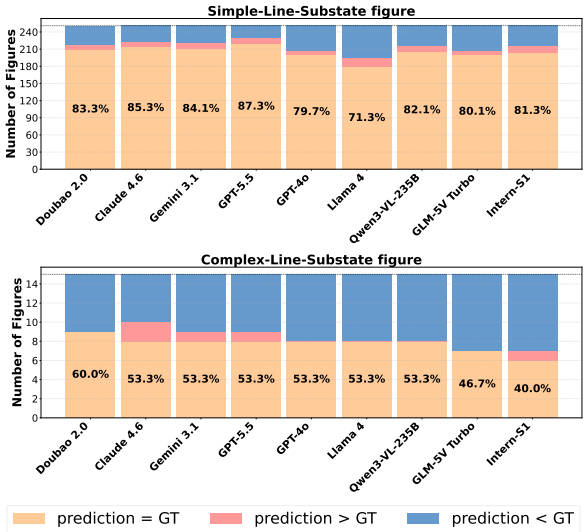


Figure 6: Comparison of reaction counts in figure between VLM predictions and ground truth, grouped by simple-line-substrate (SLS) and complex-line-substrate (CLS) figures. While VLMs achieve accurate reaction counts on SLS figures, they exhibit substantial omission errors on CLS figures.

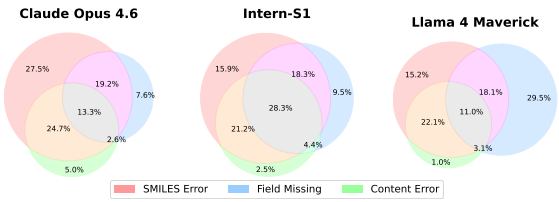


Figure 7: Error type distributions for three representative models.

failure modes: 75.51% of extraction errors involve incorrect molecular structures, highlighting molecular recognition as a major bottleneck in automated chemical reaction extraction.

However, the distribution of error types varies substantially across models (Figure 7). For Claude Opus 4.6, 84.7% of erroneous reactions involve SMILES errors, with 27.5% arising exclusively from molecular recognition, indicating that molecular structure recognition is its primary fail-

Table 4: VLM performance (F1 scores) under Oracle-MolRec inference setting on the **RxnExBench-MolIdt** subset. (Red numbers in parentheses) show improvement over the standard inference setting.

Models	Criterion 1	Criterion 2	Criterion 3
Claude Opus 4.6	30.04%(↑27.52%)	17.20%(↑16.12%)	14.78%(↑13.84%)
Gemini 3.1 Pro	36.70%(↑17.23%)	**18.25%**(↑8.68%)	**15.88%**(↑7.02%)
GPT-5.5	**42.67%**(↑19.51%)	16.83%(↑7.67%)	15.29%(↑8.48%)
GPT-4o	25.25%(↑25.16%)	5.32%(↑5.32%)	4.57%(↑4.57%)
Doubao Seed 2.0 Pro	36.73%(↑16.92%)	11.94%(↑5.35%)	10.41%(↑5.06%)
GLM-5V Turbo	31.95%(↑26.87%)	11.99%(↑10.07%)	11.11%(↑9.31%)
Intern-S1	20.77%(↑17.89%)	7.41%(↑7.35%)	5.87%(↑5.81%)
Llama 4 Maverick	20.34%(↑18.07%)	9.08%(↑8.22%)	8.25%(↑7.39%)
Qwen3-VL-235B	26.90%(↑25.22%)	7.24%(↑6.98%)	5.94%(↑5.88%)

ure mode. In contrast, Intern-S1 exhibits strong overlap among all three error categories, suggesting simultaneous weaknesses in visual perception, chemical reasoning, and molecular recognition. Llama 4 Maverick shows a different pattern: Field Missing errors form the largest independent component (29.5%), largely because the model often fails to identify complete reactions (Figure 6).

Beyond these model-specific patterns, we observe two consistent trends. First, models that perform well under Criterion 1 but degrade sharply under Criterion 3 (e.g., Gemini 3.1 Pro and GPT-5.5) frequently identify reactants and products correctly while missing or merging condition fields, indicating that condition extraction becomes a major challenge under stricter evaluation. Second, all models systematically undercount reactions on CLS figures (Section 4.3), revealing weaknesses in layout-aware reasoning when a single row corresponds to multiple reactions. Together, these findings identify molecular recognition, condition extraction, and layout reasoning as three coupled bottlenecks for current VLMs.

4.5 Disentangling Molecular Recognition from Higher-Level Reasoning

Since molecular recognition (OCSR) is a well-established task with many specialized models, integrating external OCSR systems into the extraction pipeline could alleviate a major bottleneck. To estimate the upper bound of this approach, we introduce the Oracle-MolRec protocol, which provides ground-truth SMILES for every directly drawn molecule together with references linking each SMILES to its corresponding structure in the image, thereby simulating perfect molecular recognition. We instantiate three variants: Oracle-MolRec-IDT, which pairs molecular identifiers with SMILES (default setting); Oracle-MolRec-BBOX-IDX, which annotates molecules with indexed bounding boxes and provides index–SMILES mappings; and Oracle-MolRec-BBOX-CORD, which supplies BBOX-coordinate–SMILES pairs. Prompt templates are provided in Appendix E.

Table 4 compares Oracle-MolRec-IDT against standard inference. Performance improves substantially across all criteria, confirming the importance of accurate molecular structures for reaction extraction. Under Criterion 1, the best F1 score increases from 23.16% to 42.67%; however, performance under the stricter Criteria 2 and 3 remains limited, reaching only 18.25% and 15.88%, respectively. Notably, under Criterion 3, 4 of the 9 evaluated VLMs still achieve below 10% F1 even with perfect molecular recognition.

These results show that parsing *Screen Tables* and *Substrate Scopes* requires far more than localized molecular recognition. Beyond accurate OCSR, successful extraction depends on holistic multimodal perception, long-range integration of templates, instances, and footnotes, and chemistry-aware reasoning. In this sense, Oracle-MolRec serves not only as an oracle baseline, but also as a probe for disentangling molecular recognition from higher-level reasoning. Similar gain and plateau patterns are consistently observed under both Oracle-MolRec-BBOX-IDX and Oracle-MolRec-BBOX-CORD; detailed comparisons are provided in Appendix G.

5 Conclusion

We introduce RxnExBench, the first benchmark that jointly covers *Screen Tables* and *Substrate Scopes* with full reaction-level schema, R-group resolution, CLS figures, and a reasoning-driven difficulty taxonomy. The benchmark contains 3,565 expert-annotated reactions and 2,361 molecular structures across three reasoning-driven difficulty levels under a 26-field reaction schema.

Experiments on 9 state-of-the-art VLMs show that current models still struggle with this task, particularly in multimodal integration and chemistry-aware reasoning. Oracle-MolRec results further reveal that perfect molecular recognition alone is insufficient for accurate reaction extraction. We hope RxnExBench will support future research in automated reaction extraction and AI-for-Chemistry.

Limitations

Although RxnExBench provides a challenging and high-quality benchmark for evaluating Vision-Language Models on structured chemical reaction extraction, the current dataset is still primarily focused on English-language organic synthesis literature and adopts relatively strict reaction-level matching criteria. Therefore, the reported results should be viewed as a snapshot of the current capabilities of general-purpose VLMs rather than the upper bound of task performance. In the future, RxnExBench can be further expanded to cover broader chemistry sub-disciplines and literature sources, while also enabling the exploration of supervised fine-tuning, tool-augmented approaches, and related strategies to further advance this field.

References

Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. 2018. Predicting reaction performance in c–n cross-coupling using machine learning. *Science*, 360(6385):186–190.

Anthropic. 2026. Claude opus 4.6 system card. System card. Accessed 2026-05-24.

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, and 1 others. 2025a. Intern-s1: A scientific multimodal foundation model. *arXiv preprint arXiv:2508.15763*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

CAS. 2025. Cas scifinder discovery platform. https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder. Accessed: 2025-02-14.

Guikun Chen, Xu Zhang, Xiaolin Hu, Yong Liu, Yi Yang, and Wenguan Wang. 2025a. Chemical knowledge-informed framework for privacy-aware retrosynthesis learning. *Nature Communications*, 16(1):8389.

Yufan Chen, Ching Ting Leung, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. 2025b. Towards large-scale chemical reaction image parsing via a multimodal large language model. *arXiv preprint arXiv:2503.08156*.

Yufan Chen, Ching Ting Leung, Bowen Yu, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. 2025c. A multi-agent system enables versatile information extraction from the chemical literature. *arXiv preprint arXiv:2507.20230*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. 2017. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–443.

Yafeng Deng, Xinda Zhao, Hanyu Sun, Yu Chen, Xiaorui Wang, Xi Xue, Liangning Li, Jianfei Song, Chang-Yu Hsieh, Tingjun Hou, and 1 others. 2025. Rsgpt: a generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints. *Nature Communications*, 16(1):7012.

Hu Ding, Pengxiang Hua, and Zhen Huang. 2025. Survey on recent progress of ai for chemistry: Methods, applications, and opportunities. *arXiv preprint arXiv:2502.17456*.

Liliana C Gallegos, Guilian Luchini, Peter C St. John, Seonah Kim, and Robert S Paton. 2021. Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties. *Accounts of Chemical Research*, 54(4):827–836.

Google DeepMind. 2026. Gemini 3.1 pro - model card. Model card. Accessed 2026-05-24.

Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. 2021. Automated chemical reaction extraction from scientific literature. *Journal of chemical information and modeling*, 62(9):2035–2045.

Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, and 1 others. 2026. Glm-5v-turbo: Toward a native foundation model for multimodal agents.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Joonyoung F Joung, Mun Hong Fong, Nicholas Casetti, Jordan P Liles, Ne S Dassanayake, and Connor W Coley. 2025. Electron flow matching for generative reaction mechanism prediction. *Nature*, pages 1–9.

Brent A Koscher, Richard B Canty, Matthew A McDonald, Kevin P Greenman, Charles J McGill, Camille L Bilodeau, Wengong Jin, Haoyang Wu, Florence H Vermeire, Brooke Jin, and 1 others. 2023. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science*, 382(6677):eadi1407.

Greg Landrum. 2006. Rdkit: Open-source cheminformatics.

Shi Xuan Leong, Sergio Pablo-García, Brandon Wong, and Alán Aspuru-Guzik. 2025. Mermaid: Universal multimodal mining of chemical reactions from pdfs using vision-language models. *Matter*, 8(12).

Shi Xuan Leong, Sergio Pablo-García, Zijian Zhang, and Alán Aspuru-Guzik. 2024. Automated electrosynthesis reaction mining with multimodal large language models (mllms). *Chemical Science*, 15(43):17881–17891.

SX Leong. 2025. Automated electrosynthesis reaction mining with multimodal large language models (mllms). https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef&from=research.latest-advancements-list. Accessed: 2025-11-14.

Llama Team. 2025. Llama: Open foundation and application models. https://www.llama.com/. Accessed: 2025-11-14.

Babak Mahjour, Rui Zhang, Yuning Shen, Andrew McGrath, Ruheng Zhao, Osama G Mohamed, Yingfu Lin, Zirong Zhang, James L Douthwaite, Ashootosh Tripathi, and 1 others. 2023. Rapid planning and analysis of high-throughput experiment arrays for reaction discovery. *Nature Communications*, 14(1):3924.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, and 1 others. 2025. Mineru2. 5: A decoupled vision-language model for efficient high-resolution document parsing. *arXiv preprint arXiv:2509.22186*.

OpenAI. 2026. Introducing gpt-5.5. OpenAI. Accessed 2026-05-24.

Daniel Probst, Philippe Schwaller, and Jean-Louis Reymond. 2022. Reaction classification and yield prediction using the differential reaction fingerprint drfp. *Digital discovery*, 1(2):91–97.

Yujie Qian, Jiang Guo, Zhengkai Tu, Connor W Coley, and Regina Barzilay. 2023a. Rxnscribe: a sequence generation model for reaction diagram parsing. *Journal of chemical information and modeling*, 63(13):4030–4041.

Yujie Qian, Jiang Guo, Zhengkai Tu, Zhening Li, Connor W Coley, and Regina Barzilay. 2023b. Molscribe: robust molecular structure recognition with image-to-graph generation. *Journal of chemical information and modeling*, 63(7):1925–1934.

Nadine Schneider, Daniel M Lowe, Roger A Sayle, Michael A Tarselli, and Gregory A Landrum. 2016a. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *Journal of medicinal chemistry*, 59(9):4385–4402.

Nadine Schneider, Nikolaus Stiefl, and Gregory A Landrum. 2016b. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of chemical information and modeling*, 56(12):2336–2346.

Bytedance Seed. 2026. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. Technical report, Technical report, Technical report, Bytedance, 2025. URL https://lf3-static . . . .

Jiahe Song, Chuang Wang, Bowen Jiang, Yinfan Wang, Hao Zheng, Xingjian Wei, Chengjin Liu, Junyuan Gao, Yubin Wang, Lijun Wu, and 1 others. 2025. Rxncaption: Reformulating reaction diagram parsing as visual prompt guided captioning. *arXiv preprint arXiv:2511.02384*.

Jason Y Wang, Jason M Stevens, Stavros K Kariofillis, Mai-Jan Tom, Dung L Golden, Jun Li, Jose E Tabora, Marvin Parasram, Benjamin J Shields, David N Primer, and 1 others. 2024. Identifying general reaction conditions by bandit optimization. *Nature*, 626(8001):1025–1033.

Jingchao Wang, Haote Yang, Jiang Wu, Yifan He, Xingjian Wei, Yinfan Wang, Chengjin Liu, Lingli Ge, Lijun Wu, Bin Wang, and 1 others. 2025. Gtr-cot: Graph traversal as visual chain of thought for molecular structure recognition. *arXiv preprint arXiv:2506.07553*.

Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, and 1 others. 2024. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*.

Jiacheng Xiong, Wei Zhang, Yinquan Wang, Jiatao Huang, Yuqi Shi, Mingyan Xu, Manjia Li, Zunyun Fu, Xiangtai Kong, Yitian Wang, and 1 others. 2025. Bridging chemistry and artificial intelligence by a reaction description language. *Nature Machine Intelligence*, pages 1–12.

Li-Cheng Xu, Miao-Jiong Tang, Junyi An, Fenglei Cao, and Yuan Qi. 2025. A unified pre-trained deep learning framework for cross-task reaction performance prediction and synthesis planning. *Nature Machine Intelligence*, pages 1–11.

Jiaxian Yan, Jintao Zhu, Yuhang Yang, Qi Liu, Kai Zhang, Zaixi Zhang, Xukai Liu, Boyan Zhang, Kaiyuan Gao, Jinchuan Xiao, and 1 others. 2025. Biominer: A multi-modal system for automated mining of protein-ligand bioactivity data from literature. *bioRxiv*, pages 2025–04.

Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, Jinfeng Xiao, Minhao Jiang, Vivian Hu, Xuan Wang, Heng Ji, Martin Burke, Huimin Zhao, and Jiawei Han. 2023a. Reaction miner: An integrated system for chemical reaction extraction from textual data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*,

pages 389–402, Singapore. Association for Computational Linguistics.

Ming Zhong, Siru Ouyang, Yizhu Jiao, Priyanka Kargupta, Leo Luo, Yanzhen Shen, Bobby Zhou, Xianrui Zhong, Xuan Liu, Hongxiang Li, and 1 others. 2023b. Reaction miner: An integrated system for chemical reaction extraction from textual data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 389–402.

Zipeng Zhong, Jie Song, Zunlei Feng, Tiantao Liu, Lingxiang Jia, Shaolun Yao, Tingjun Hou, and Mingli Song. 2024. Recent advances in deep learning for retrosynthesis. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 14(1):e1694.

Yitong Zhou, Mingyue Cheng, Qingyang Mao, Yucong Luo, Qi Liu, Yupeng Li, Xiaohan Zhang, Deguang Liu, Xin Li, and Enhong Chen. 2025. Benchmarking multimodal llms on recognition and understanding over chemical tables. *arXiv preprint arXiv:2506.11375*.

A Chemical Reaction Related Concepts and Reaction Schema

A.1 Introduction of Screen Tables and Substrate Scopes

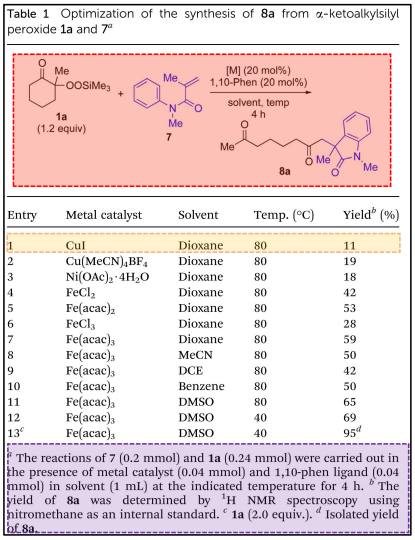


Table 1 Optimization of the synthesis of **8a** from α-ketoalkylsilyl peroxide **1a** and **7**[a]

Entry	Metal catalyst	Solvent	Temp. (°C)	Yield[b] (%)
1	CuI	Dioxane	80	11
2	$Cu(MeCN)_4BF_4$	Dioxane	80	19
3	$Ni(OAc)_2 \cdot 4H_2O$	Dioxane	80	18
4	$FeCl_2$	Dioxane	80	42
5	$Fe(acac)_2$	Dioxane	80	53
6	$FeCl_3$	Dioxane	80	28
7	$Fe(acac)_3$	Dioxane	80	59
8	$Fe(acac)_3$	MeCN	80	50
9	$Fe(acac)_3$	DCE	80	42
10	$Fe(acac)_3$	Benzene	80	50
11	$Fe(acac)_3$	DMSO	80	65
12	$Fe(acac)_3$	DMSO	40	69
13[c]	$Fe(acac)_3$	DMSO	40	95[d]

[a] The reactions of **7** (0.2 mmol) and **1a** (0.24 mmol) were carried out in the presence of metal catalyst (0.04 mmol) and 1,10-phen ligand (0.04 mmol) in solvent (1 mL) at the indicated temperature for 4 h. [b] The yield of **8a** was determined by ^1H NMR spectroscopy using nitromethane as an internal standard. [c] **1a** (2.0 equiv.). [d] Isolated yield of **8a**.

Figure 8: Example of Screen Table.

As shown in Figure 8, *Screen Table* is a common form of table in chemical papers, used to systematically compare the performance of a set of highly similar chemical reactions under strictly controlled variables (such as different catalysts, solvents, temperatures, reaction times, etc.), thereby identifying the optimal reaction conditions.

A typical *Screen Table* usually includes a "reaction template" presented as an image (as indicated by the red box in Figure 8), which summarizes the core information shared by the entire set of reactions. This includes reactants and products (typically represented as molecular structure diagrams) and some reaction conditions (often annotated as text). In the template, certain conditions are explicitly specified (such as the reagent "1,10-Phen" labeled in Figure 8), while others are represented by abstract symbols (e.g., "solvent/temp" above the reaction arrow).

Below the template, a structured table is usually attached, where each row (i.e., an *entry*) corresponds to the reaction results under one or more specific experimental conditions. However, the information in a single row is often incomplete and must be combined with the reaction template above to reconstruct a complete reaction description. Common scenarios include: instantiating abstract conditions from the template (e.g., replacing "solvent/temp" with specific compound names), or making localized modifications to a condition in the template (e.g., explicitly setting the reaction temperature to "80 °C" in row 1).

Additionally, some *Screen Tables* include footnotes at the bottom of the table (as shown in the purple box in Figure 8) to supplement experimental details or special notes that cannot be fully expressed in the table.

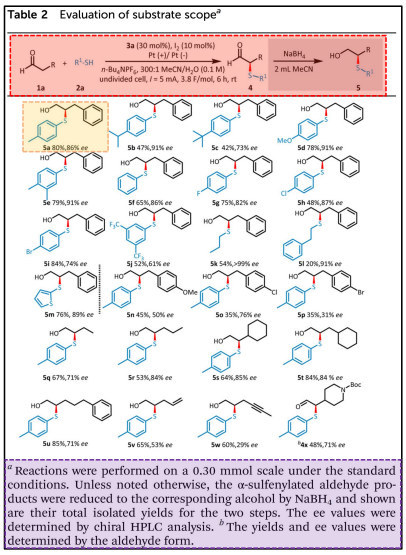


Table 2 Evaluation of substrate scope[a]

[a] Reactions were performed on a 0.30 mmol scale under the standard conditions. Unless noted otherwise, the α-sulfenylated aldehyde products were reduced to the corresponding alcohol by $NaBH_4$ and shown are their total isolated yields for the two steps. The ee values were determined by chiral HPLC analysis. [b] The yields and ee values were determined by the aldehyde form.

Figure 9: Example of Substrate Scope.

Similar to *Screen Table*, *Substrate Scope* (as shown in Figure 9) also includes a reaction tem-

plate in the form of an image. The key difference is that in *Substrate Scope*, the reaction conditions are typically fixed and explicitly defined, while the molecular structures of the reactants and products contain one or more variable substituents (such as R and R^1 in Figure 9; other common forms include R′ or specific symbols).

Below the reaction template, molecular structure diagrams of specific "substrates" and their corresponding reaction results (such as yield, selectivity, etc.) are neatly arranged. These substrate structures (e.g., **5a–5w** in Figure 9) essentially represent "instantiations" of the product in the template (such as molecule **5**); they clarify the exact form of the product generated when the substrate participates in the reaction by specifying the specific chemical structure of the R-group.

Similar to *Screen Tables*, *Substrate Scopes* often include footnotes below the substrate array to supplement critical experimental details that are difficult to present directly in the diagram (as shown in the purple box in Figure 9) .

A.2 Structured Chemical Reaction Schema

We define a structured schema for representing chemical reactions. Table 5 presents the detailed specifications of the Structured Reaction Schema. Since some reaction roles or fields may contain multiple instances, we have introduced the corresponding "list" data type in addition to the "string" type. The data types are indicated in parentheses. Additionally, the fields within each reaction role are organized in a "dictionary" format. Specific annotation examples are shown in Appendix D.

B Data Collection

Our data come from 8 high-impact chemistry journals. The list of journals is shown in Table 6.

Table 7: Average Per-image Annotation Statistics

Figure Type	Screen Table	Substrate Scope
Reactions	11.53	17.07
Annotated Fields	141.81	248.17
Molecular structures	36.17	43.67
Atoms	441.65	584.00
Chemical bonds	456.76	605.00

C Data annotation

C.1 Statistics Related to Data Annotation

The RxnExBench dataset comprises 266 images and 3,565 chemical reactions in total. During the annotation of reaction templates, 53,659 fields and 10,296 molecular structures were annotated. The 10,296 figure counts molecule *occurrences*—a single molecule may be referenced by multiple reactions; after RDKit canonicalization and deduplication, these correspond to the 2,361 unique canonical SMILES reported in the main text. Collectively, these structures contain 130,290 atoms and 134,840 chemical bonds. The average statistics per image are summarized in Table 7.

Table 5: Structured chemical reaction schema of RxnExbench.

Reaction roles		Fine grained fields
reactants (list)		SMILES (str) text (str) amount (str) conversion rate (str)
products (list)		SMILES (str) text (str) yield ratio (str) isolated yield (str) enantiomeric excess (str) regioisomeric ratio (str) enantiomeric ratio (str) alkene configuration (str) diastereomeric ratio (str)
stage (list)	reagents (list)	SMILES (str) text (str) amount (str)
	catalysts (list)	SMILES (str) text (str) amount (str)
	solvents (list)	text (str) ratio (str) amount (str)
	ligand (list)	SMILES (str) text (str) amount (str)
	pressure (str)	-
	reaction time (str)	-
	reaction temperature (str)	-
	light condition (str)	-
	atmosphere (str)	-
	pH (str)	-
	stir speed (str)	-
	vacuum condition (str)	-
	scale (str)	-

C.2 Detailed Data Annotation Process

Our annotation pipeline consists of three main stages, as outlined in the main text: **Reaction Template Annotation**, **Template Instantiation**, and **Molecular Structure (SMILES) Annotation**. To provide necessary context and enable collaborative work, two preparatory steps are integrated into this workflow: **Molecular Structure Localization** pre-

Table 6: Journal List

Journal Name
Angewandte Chemie International Edition
Chemical Science
Journal of the American Chemical Society
ACS Catalysis
Green Chemistry
ACS Central Science
Nature Chemistry
Accounts of Chemical Research

cedes the Reaction Template Annotation, and the **Template Instantiation** stage is itself divided into two sub-steps: **Pre-annotation of Instantiation References** and **Instantiation Execution**.

Molecular Structure Localization. In this initial step, annotators are required to draw bounding boxes around all molecular structures depicted as diagrams (non-string form) within the image. This localization serves as a critical reference for subsequent stages: these diagram-based structures must be specifically annotated later using the `Ketcher` tool to generate their corresponding SMILES strings. The bounding boxes allow us to visually guide annotators to the correct molecules in later steps and to organize the molecular annotation content using the unique indices of these boxes.

Reaction Template Annotation. As described in the main text, annotators delineate the area of a generic reaction template using a polygon on the annotation platform. The initial annotation is performed based solely on the information contained within this polygon. Consistent with the main text, fields requiring instance-specific context, such as R-group identities, yields, temperatures, or SMILES, are left blank at this stage.

Pre-annotation of Instantiation References. A single image often contains one reaction template that must be instantiated multiple times. For instance, each row in a *Screen Table* or each substrate molecule in a *Substrate Scope* figure typically corresponds to one concrete reaction instance. Since our annotation process is collaborative, with multiple annotators working from a shared pool of images, different instantiation tasks for the same image might be handled by different individuals. To provide context, this sub-step requires annotators to first register all information that easily identifies each reaction instance across the image. This includes details like row numbers in a *Screen Table* or the bounding box indices of substrate molecules in a *Substrate Scope*. This registry of references is crucial for the subsequent execution of instantiation tasks.

Instantiation Execution. After the reference information for all instances within an image is established in the previous sub-step, a distinct instantiation task is created for each entry in this registry. Annotators, provided with a specific instantiation context (e.g., a row number, a substrate box index), proceed to convert the abstract template from the **Reaction Template Annotation** into a concrete reaction. They fill in the context-dependent fields (e.g., specific catalyst, solvent, yield). For molecular SMILES, only the ID of the corresponding molecular bounding box is recorded at this point; the actual SMILES string is added later.

Molecular Structure (SMILES) Annotation. In this step, annotators use the `Ketcher` tool to annotate specified molecular structures one at a time. The molecule to be annotated is highlighted using its pre-defined bounding box. For molecules within reaction templates that contain undefined R-groups—whose identities are determined only during template instantiation—the relevant instantiation reference information (from the pre-annotation sub-step) is also provided. This ensures the annotator knows exactly which reaction instance (e.g., which row or substrate) the R-group corresponds to, allowing for the correct structure to be drawn.

Finally, the `mol_file` information for each molecule, obtained from the **Molecular Structure (SMILES) Annotation** step, is programmatically converted into a standardized SMILES string using `RDKit`(Landrum, 2006). Based on the associated molecular box ID and the instantiation reference information, these SMILES strings are automatically inserted into their corresponding slots within the instantiated reaction templates from the **Instantiation Execution** step. This process yields the final, complete ground truth (GT) data.

C.3 Data Annotation Platform

Each annotation step was conducted on specifically designed web-based platforms. These platforms enabled the efficient coordination of annotator team for parallel annotation, while also facilitating multiple rounds of internal quality checks and flexible data processing. Example interface screenshots for each annotation step are as follows: Figure 10

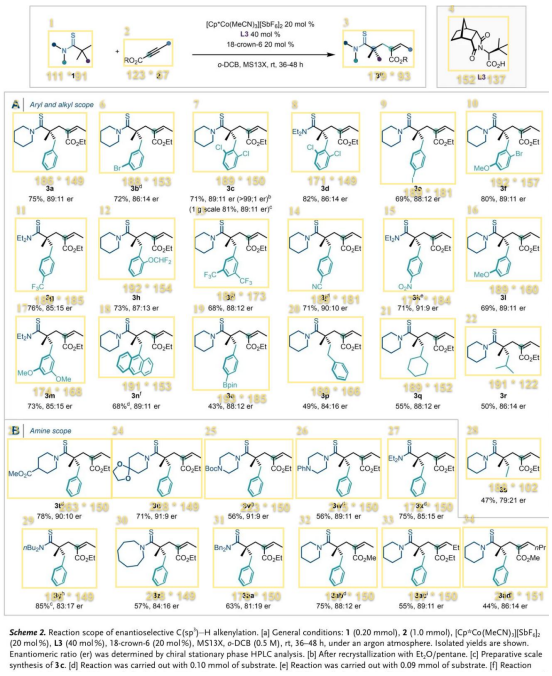


Figure 10: Screenshot for Molecular Structure Localization

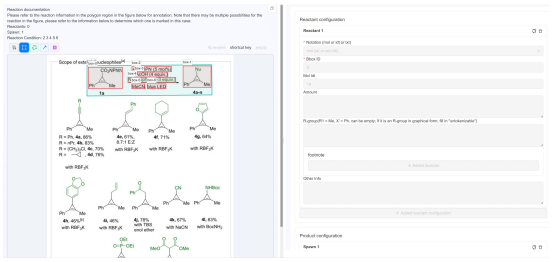


Figure 11: Screenshot for Reaction Template Annotation

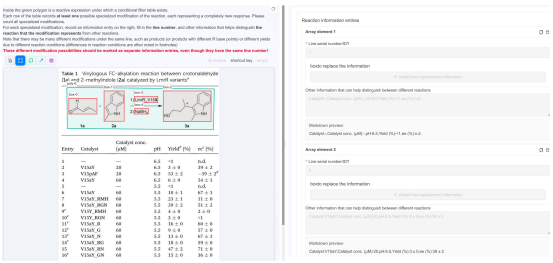


Figure 12: Screenshot for Pre-annotation of Instantiation References

shows the Molecular Structure Localization interface, Figure 11 illustrates the Reaction Template Annotation interface, Figure 12 presents the Pre-annotation of Instantiation References interface, Figure 13 displays the Instantiation Execution interface, and Figure 14 depicts the Molecular Structure (SMILES) Annotation interface.

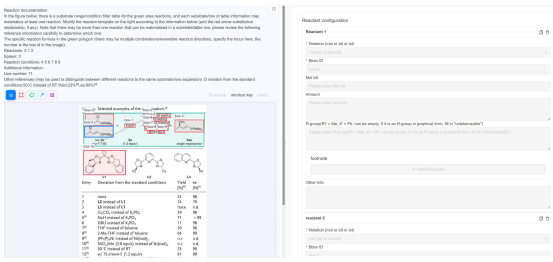


Figure 13: Screenshot for Instantiation Execution

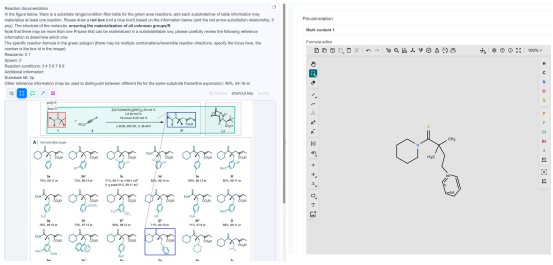


Figure 14: Screenshot for Molecular Structure (SMILES) Annotation

C.4 Data Annotation Team

We recruited a total of 40 professional chemical annotators to participate in our chemical reaction annotation task. All annotators possess backgrounds in chemistry-related disciplines, including 9 master's degree candidates and 31 bachelor's degree holders. Their academic specialties cover multiple chemistry-related fields such as Chemistry, Applied Chemistry, Materials Chemistry, Chemical Engineering and Technology, and Organic Chemistry. Prior to formally commencing the annotation work, all annotators passed rigorous chemical knowledge assessments and comprehensive annotation protocol training. This ensures their solid foundation in chemistry and consistent adherence to annotation standards, thereby guaranteeing the professionalism and reliability of the annotated data. During the annotation process, samples were periodically cross-checked by different annotators, and disagreements were resolved through discussion and expert review. In addition, the modular design of the pipeline—which separates template annotation, instantiation, and molecular structure annotation—helps reduce ambiguity and improve consistency across different annotation stages.

All annotators were compensated under written agreements at rates that meet or exceed the prevailing wage standards for skilled scientific annotation work in their respective regions, in compliance with applicable local labor regulations.Before participating, annotators were informed in writing of the

research purpose, the structure of the annotation task, the planned public release of the resulting benchmark under a CC-BY-4.0 license, and their right to withdraw at any time; all annotators provided informed consent.

C.5 Human Performance Upper Bound

To contextualize VLM results and quantify the inherent difficulty of RxnExBench, we conducted a dedicated human-performance study. Three chemistry-trained experts who did *not* participate in the original annotation pipeline were asked to perform the end-to-end extraction task on figures sampled across all three difficulty levels, following exactly the same instructions and schema given to the VLMs. Their predictions were scored against the released ground truth using the identical evaluation pipeline (Appendix F). The averaged F1 scores are reported in Table 8.

Three observations emerge:

(i) Even trained chemists do not reach 100%. Overall F1 reaches 91.33%, 85.27%, and 81.66% under Criteria 1, 2, and 3 respectively. The residual gap reflects the genuine ambiguity of chemical literature: footnote conventions, abbreviation expansions, stereochemistry annotation, and tautomer choice sometimes admit more than one defensible interpretation. These numbers therefore serve as an empirical *performance ceiling* for RxnExBench rather than a theoretical upper bound of 100%.

(ii) Human accuracy degrades monotonically with difficulty level. Under Criterion 1, expert F1 drops from 98.98% on Level 1 to 93.49% on Level 2 and 87.68% on Level 3, mirroring the same Level 1 > Level 2 > Level 3 ordering observed for every VLM (Table 3). This empirically validates that the proposed reasoning-driven taxonomy reflects actual cognitive load rather than surface heterogeneity.

(iii) Human accuracy also degrades monotonically with stricter criteria. On Level 1 figures alone, F1 falls from 98.98% (Criterion 1) to 94.92% (Criterion 3) as the number of required-match fields grows from three to ten. This confirms that the field-level extension from Criterion 1 to Criterion 3 introduces genuine, additive difficulty, and is not a metric artifact.

The VLM–human gap is large and consistent. The best VLM in our evaluation (GPT-5.5) reaches 23.16% F1 under Criterion 1, leaving a gap of 68.17 absolute points to human performance. Under the strictest Criterion 3, the gap widens to 72.80 points (8.86% vs. 81.66%). This sizable and consistent margin indicates that fully automated structured reaction extraction from *Screen Tables* and *Substrate Scopes* remains an open problem, and that RxnExBench captures challenges that current frontier VLMs are far from solving.

Table 8: Human expert performance (F1, %) on RxnExBench, broken down by difficulty level and evaluation criterion. Numbers are averaged over three independent experts. The overall human F1 of 91.33% under Criterion 1 serves as an empirical performance ceiling for this benchmark.

Subset	Criterion 1	Criterion 2	Criterion 3
Level 1	98.98%	96.79%	94.92%
Level 2	93.49%	86.25%	81.76%
Level 3	87.68%	81.09%	77.43%
All	**91.33%**	**85.27%**	**81.66%**

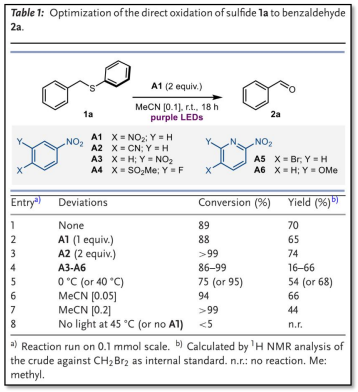


Table 1: Optimization of the direct oxidation of sulfide **1a** to benzaldehyde **2a**.

Entry[a]	Deviations	Conversion (%)	Yield (%)[b]
1	None	89	70
2	**A1** (1 equiv.)	88	65
3	**A2** (2 equiv.)	>99	74
4	**A3-A6**	86–99	16–66
5	0 °C (or 40 °C)	75 (or 95)	54 (or 68)
6	MeCN [0.05]	94	66
7	MeCN [0.2]	>99	44
8	No light at 45 °C (or no **A1**)	<5	n.r.

[a] Reaction run on 0.1 mmol scale. [b] Calculated by ^{1}H NMR analysis of the crude against CH_2Br_2 as internal standard. n.r.: no reaction. Me: methyl.

Figure 15: Example of *Screen Table*. The annotation results for this *Screen Table* are shown in Figures 27 and 28.

D Examples of RxnExBench Annotation Result

This section presents the detailed annotation results of RxnExBench. Figure 15 shows an example of the *Screen Table*, while Figure 27 and Figure 28 provide examples of its corresponding detailed reaction annotations.

Figure 16 shows an example of the *Substrate Scope*, while Figures 29 and 30 provide examples of its corresponding detailed reaction annotations.

E Inference Prompt

This section presents all inference prompts used in **RxnExBench**. Figures 19 and 20 show the standard inference prompts for *Screen Tables* and *Sub-*

Table 9: Criterion details of RxnExBench.

Criterion	Downstream Task Category	Reaction Roles	Evaluation Description
1	Retrosynthesis Analysis(Chen et al., 2025a); Main Product Prediction(Coley et al., 2017); Reaction Classification(Probst et al., 2022)	reactant, product, reagent	Focuses on the correctness of the primary reaction roles. (SMILES or text field indicating the identity of the substance)
2	Yield Prediction(Ahneman et al., 2018); Selectivity prediction(Gallegos et al., 2021)	reactant, product, reagent, catalyst, solvent	Focuses on the correctness of reaction roles and fine-grained fields. (SMILES or text indicating the identity of the substance, and fields such as amount and yield ratio that describe quantitative reaction information.)
3	Reaction Condition Recommendation; Reaction Kinetics Modeling; HTE Screening(Koscher et al., 2023; Mahjour et al., 2023)	reactant, product, reagent, catalyst, ligand, solvent, pressure, reaction_time, reaction_temperature, pH	Focuses on the correctness of reaction roles and fine-grained fields.(SMILES or text indicating the identity of the substance, and fields such as amount and yield ratio that describe quantitative reaction information.)

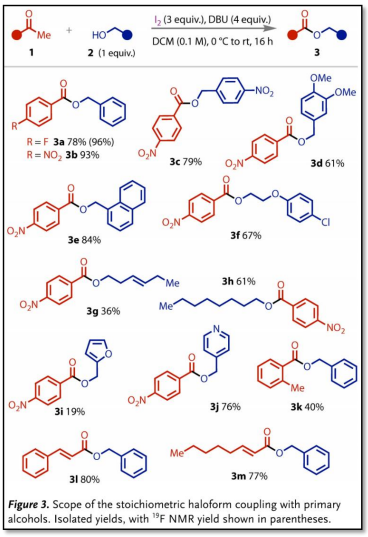


Figure 16: Example of *Substrate Scope*. The annotation results for this *Substrate Scope* are shown in Figures 29 and 30.

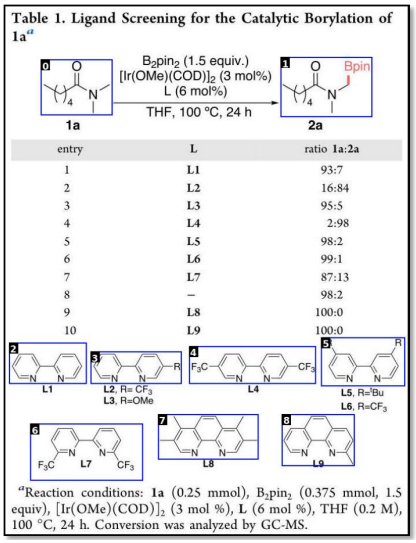


Figure 17: Example of the input image of Oracle-MolRec-BBOX-IDX inference. The ground-truth bounding box of each molecule is drawn on the original image as visual prompts and a unique index number is labeled at the top-left corner. The corresponding prompt for this image is shown in Figure 23.

strate Scopes, respectively. Figures 21 and 22 illustrate the inference prompts adopted in the Oracle-MolRec inference setting for *Screen Tables* and *Substrate Scopes*. Figures 23 and 24 display the inference prompts used in the Oracle-MolRec-BBOX-IDX setting for *Screen Tables* and *Substrate Scopes*. Finally, Figures 25 and 26 present the inference prompts for the Oracle-MolRec-BBOX-CORD setting for *Screen Tables* and *Substrate Scopes*. For detailed experimental setup, please refer to Appendix G.

F Criterion and evaluation details of RxnExBench

As shown in Table 9, we design a three-level evaluation criterion based on the requirements of downstream AI-for-Chemistry tasks. **Criterion 1**, targeting tasks such as single-step retrosynthesis, checks only the reactant, product, and reagent fields. **Criterion 2**, used for tasks such as yield prediction, evaluates all fields in Criterion 1 and additionally checks catalyst and solvent information. **Criterion 3** is the most stringent: designed for tasks such as reaction-condition recommendation, it requires verifying a total of ten fields. Under all three criteria, a prediction is counted as a true positive (TP) only when it exactly matches the ground-truth annotation across all required fields.

For SMILES fields, we use `RDKit` to canonicalize all molecular representations before compari-

Table 10: Recall scores of VLMs on RxnExBench. Results are shown for the full dataset and three difficulty levels under three evaluation criteria. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	All Levels (266 Figures)			Level 1 (60 Figures)			Level 2 (90 Figures)			Level 3 (116 Figures)		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.50%	1.07%	0.93%	5.58%	4.23%	4.23%	2.38%	1.19%	0.73%	1.59%	0.00%	0.00%
Gemini 3.1 Pro	18.99%	**9.34%**	**8.64%**	**39.26%**	**19.97%**	**19.29%**	15.12%	9.99%	**8.07%**	14.87%	**5.63%**	**5.63%**
GPT-5.5	**22.97%**	9.09%	6.76%	33.67%	12.86%	12.69%	**28.14%**	**13.93%**	6.42%	**16.62%**	5.10%	5.10%
GPT-4o	0.08%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.16%	0.00%	0.00%
Doubao Seed 2.0 Pro	19.30%	6.42%	5.22%	33.50%	11.34%	10.66%	25.39%	7.06%	4.49%	11.31%	4.51%	3.93%
GLM-5V Turbo	4.74%	1.80%	1.68%	15.40%	6.77%	6.26%	2.75%	1.28%	1.19%	2.55%	0.53%	0.53%
Intern-S1	2.83%	0.06%	0.06%	4.57%	0.00%	0.00%	6.78%	0.18%	0.18%	0.00%	0.00%	0.00%
Llama 4 Maverick	1.99%	0.76%	0.76%	4.57%	2.37%	2.37%	2.66%	0.09%	0.09%	0.80%	0.64%	0.64%
Qwen3-VL-235B	1.60%	0.25%	0.06%	5.58%	1.52%	0.34%	2.20%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 11: Precision scores of VLMs on RxnExBench. Results are shown for the full dataset and three difficulty levels under three evaluation criteria. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	All Levels (266 Figures)			Level 1 (60 Figures)			Level 2 (90 Figures)			Level 3 (116 Figures)		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.55%	1.09%	0.95%	5.61%	4.25%	4.25%	2.42%	1.21%	0.74%	1.64%	0.00%	0.00%
Gemini 3.1 Pro	19.96%	**9.82%**	**9.08%**	**41.65%**	**21.18%**	**20.47%**	15.82%	10.45%	**8.44%**	15.63%	**5.92%**	**5.92%**
GPT-5.5	**23.34%**	9.23%	6.87%	33.33%	12.73%	12.56%	**28.11%**	**13.92%**	6.41%	**17.20%**	5.27%	5.27%
GPT-4o	0.09%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.17%	0.00%	0.00%
Doubao Seed 2.0 Pro	20.34%	6.77%	5.50%	33.62%	11.38%	10.70%	26.01%	7.23%	4.60%	12.33%	4.92%	4.28%
GLM-5V Turbo	5.47%	2.07%	1.94%	15.56%	6.84%	6.32%	2.79%	1.30%	1.21%	3.36%	0.70%	0.70%
Intern-S1	2.93%	0.06%	0.06%	4.62%	0.00%	0.00%	7.01%	0.19%	0.19%	0.00%	0.00%	0.00%
Llama 4 Maverick	2.65%	1.01%	1.01%	4.60%	2.39%	2.39%	3.03%	0.10%	0.10%	1.32%	1.06%	1.06%
Qwen3-VL-235B	1.76%	0.28%	0.06%	5.65%	1.54%	0.34%	2.27%	0.00%	0.00%	0.00%	0.00%	0.00%

son. Two molecules are considered equivalent only if their canonicalized molecular graphs are identical, implemented by computing Tanimoto similarity on RDKit fingerprints with a strict threshold of 1.0. For other string-based fields (e.g., textual descriptions and reaction conditions), we apply a normalized string-matching procedure to improve robustness against superficial formatting variations. Specifically, strings are lowercased and normalized by removing redundant whitespace and standardizing common symbols and delimiters (e.g., commas, semicolons, and temperature/unit formatting). Matching is then performed on the normalized strings while preserving the original semantic content. This strategy tolerates minor formatting inconsistencies without relaxing the underlying field semantics.

The Tanimoto similarity is used to measure the similarity between the fingerprints of two molecules, and is defined as

$$T(A, B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

where A and B are the binary fingerprints of the two molecules, and $|A \cap B|$ denotes the number of bits that are 1 in both fingerprints.

Tanimoto @1.0 indicates that two molecules are considered chemically equivalent only if their fingerprints match exactly, i.e., the Tanimoto similarity equals 1.0.

G Detailed experimental results

This subsection presents the detailed experimental results of RxnExBench. Table 10 shows the Recall scores of VLMs on RxnExBench, while Table 11 presents the Precision scores of VLMs on RxnExBench. The results are displayed for the full dataset and three difficulty levels under three evaluation criteria. Figure 18 presents the error type distributions of all models.

As described in Section 4.5, the Oracle-MolRec protocol comprises three variants—Oracle-MolRec-IDT, Oracle-MolRec-BBOX-IDX, and Oracle-MolRec-BBOX-CORD—differing in how each molecule is referenced in the VLM prompt. An example input image with the bounding-box visual prompts used in Oracle-MolRec-BBOX-

Table 12: The Recall scores of each VLM under the four inference settings: Standard Inference, Oracle-MolRec-IDT, Oracle-MolRec-BBOX-CORD, and Oracle-MolRec-BBOX-IDX, on the **RxnExBench-MolIDT** subset. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	Standard Inference			Oracle-MolRec-IDT			Oracle-MolRec-BBOX-CORD			Oracle-MolRec-BBOX-IDX		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.50%	1.07%	0.93%	29.63%	16.96%	14.58%	28.83%	14.77%	12.31%	29.98%	15.13%	12.95%
Gemini 3.1 Pro	18.99%	**9.34%**	**8.64%**	35.15%	**17.47%**	**15.21%**	31.85%	**16.32%**	13.98%	34.99%	**19.18%**	**16.16%**
GPT-5.5	**22.97%**	9.09%	6.76%	**41.78%**	16.48%	14.97%	**45.04%**	16.12%	**14.26%**	**45.75%**	14.97%	12.67%
GPT-4o	0.08%	0.00%	0.00%	24.15%	5.08%	4.37%	22.16%	5.36%	4.25%	21.37%	4.01%	2.94%
Doubao Seed 2.0 Pro	19.30%	6.42%	5.22%	35.31%	11.48%	10.01%	35.19%	11.32%	10.44%	36.30%	10.56%	9.93%
GLM-5V Turbo	4.74%	1.80%	1.68%	31.53%	11.83%	10.96%	30.30%	10.64%	9.13%	33.36%	9.21%	7.86%
Intern-S1	2.83%	0.06%	0.06%	18.71%	6.67%	5.28%	20.77%	6.55%	6.04%	19.86%	5.68%	4.96%
Llama 4 Maverick	1.99%	0.76%	0.76%	17.51%	7.82%	7.11%	14.42%	5.36%	4.57%	19.06%	8.18%	6.63%
Qwen3-VL-235B	1.60%	0.25%	0.06%	25.54%	6.87%	5.64%	23.23%	6.16%	5.08%	25.34%	7.70%	6.24%

Table 13: Results of Mermaid on RxnExBench

Criterion	Recall	Precision	F1-score
Criterion1	3.99%	4.68%	4.31%
Criterion2	0.59%	0.69%	0.64%
Criterion3	0.59%	0.69%	0.64%

Table 14: Results of ChemEAGLE on RxnExBench

Criterion	Recall	Precision	F1-score
Criterion1	0.31%	0.43%	0.36%
Criterion2	0.00%	0.00%	0.00%
Criterion3	0.00%	0.00%	0.00%

IDX is shown in Figure 17. The Recall, Precision, and F1 score results in Tables 12, 15, and 16 show that all three Oracle-MolRec variants significantly outperform standard inference, demonstrating the substantial value of accurate molecular structural information for reaction extraction. This also indicates that enhancing the parsing capability of molecular structures (SMILES) is crucial for VLMs in chemical reaction extraction.

H Evaluation of MerMaid and ChemEAGLE

We also evaluated two popular pipeline-based reaction extraction tools, MerMaid and ChemEAGLE. Since the reaction-role fields supported by both tools are fewer than those evaluated in RxnExBench, we only assess the intersection of the reaction roles for each tool.

H.1 Evaluation of MerMaid

Table 13 presents the performance of MerMaid on RxnExBench. For MerMaid, its extraction capability is limited to reactions in *Screen Tables*, so our evaluation uses only the **RxnExBench-ScreenTable** subset. To align MerMaid's output fields with our ground truth, we take the intersection of both field sets and normalize synonymous fields into a unified schema. Specifically: **Criterion 1** evaluates the SMILES or text content for reactant, product, and reagent roles. **Criterion 2** extends Criterion 1 by additionally evaluating solvent, as well as quantitative reaction fields such as amount and yield ratio. **Criterion 3** further extends Criterion 2 by including reaction time and atmosphere.

H.2 Evaluation of ChemEAGLE

Table 14 presents the performance of ChemEAGLE on RxnExBench. We follow the same procedure to align ChemEAGLE's output fields with our ground truth by taking the intersection and unifying synonymous fields. Specifically: **Criterion 1** evaluates the SMILES or text content for reactant, product, and reagent roles. **Criterion 2** extends Criterion 1 by additionally evaluating solvent and catalyst, as well as quantitative reaction fields such as yield ratio. **Criterion 3** further extends Criterion 2 by including ligand, reaction time, and reaction temperature.

From the results, both MerMaid and ChemEAGLE perform poorly on RxnExBench. Our analysis of their failure cases shows that these pipeline-based methods struggle to recognize complex molecular structures. For molecules containing R-group placeholders, MerMaid often fails to correctly instantiate and substitute the R-groups, while ChemEAGLE is unable to perform the necessary reasoning to infer the reactant SMILES from the

Table 15: The Precision scores of each VLM under the four inference settings: Standard Inference, Oracle-MolRec-IDT, Oracle-MolRec-BBOX-CORD, and Oracle-MolRec-BBOX-IDX, on the **RxnExBench-MolIDT** subset. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	Standard Inference			Oracle-MolRec-IDT			Oracle-MolRec-BBOX-CORD			Oracle-MolRec-BBOX-IDX		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.55%	1.09%	0.95%	30.47%	17.44%	14.99%	29.50%	15.12%	12.60%	30.80%	15.54%	13.30%
Gemini 3.1 Pro	19.96%	**9.82%**	**9.08%**	38.39%	**19.09%**	**16.62%**	35.98%	**18.44%**	**15.79%**	38.83%	**21.29%**	**17.94%**
GPT-5.5	**23.34%**	9.23%	6.87%	**43.60%**	17.20%	15.62%	**45.89%**	16.43%	14.53%	**46.30%**	15.15%	12.82%
GPT-4o	0.09%	0.00%	0.00%	26.46%	5.57%	4.79%	24.37%	5.90%	4.67%	23.41%	4.40%	3.22%
Doubao Seed 2.0 Pro	20.34%	6.77%	5.50%	38.27%	12.44%	10.85%	37.88%	12.18%	11.24%	38.36%	11.16%	10.49%
GLM-5V Turbo	5.47%	2.07%	1.94%	32.38%	12.15%	11.26%	31.08%	10.92%	9.37%	34.30%	9.47%	8.08%
Intern-S1	2.93%	0.06%	0.06%	23.35%	8.33%	6.59%	22.00%	6.94%	6.39%	21.84%	6.25%	5.46%
Llama 4 Maverick	2.65%	1.01%	1.01%	24.24%	10.83%	9.84%	20.10%	7.48%	6.37%	27.03%	11.60%	9.40%
Qwen3-VL-235B	1.76%	0.28%	0.06%	28.41%	7.64%	6.27%	25.98%	6.88%	5.68%	28.84%	8.77%	7.10%

Table 16: The F1 scores of each VLM under the four inference settings: Standard Inference, Oracle-MolRec-IDT, Oracle-MolRec-BBOX-CORD, and Oracle-MolRec-BBOX-IDX, on the **RxnExBench-MolIDT** subset. **C** refers to **Criterion**. **Boldfaced** numbers denote the best performance, while underlined numbers represent the runner-up.

Models	Standard Inference			Oracle-MolRec-IDT			Oracle-MolRec-BBOX-CORD			Oracle-MolRec-BBOX-IDX		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
Claude Opus 4.6	2.52%	1.08%	0.94%	30.04%	17.20%	14.78%	29.16%	14.94%	12.45%	30.39%	15.34%	13.12%
Gemini 3.1 Pro	19.47%	**9.57%**	**8.86%**	36.70%	**18.25%**	**15.88%**	33.79%	**17.32%**	**14.83%**	36.81%	**20.18%**	**17.00%**
GPT-5.5	**23.16%**	9.16%	6.81%	**42.67%**	16.83%	15.29%	**45.46%**	16.28%	14.39%	**46.02%**	15.06%	12.74%
GPT-4o	0.09%	0.00%	0.00%	25.25%	5.32%	4.57%	23.21%	5.62%	4.45%	22.34%	4.19%	3.07%
Doubao Seed 2.0 Pro	19.81%	6.59%	5.35%	36.73%	11.94%	10.41%	36.48%	11.74%	10.83%	37.30%	10.85%	10.20%
GLM-5V Turbo	5.08%	1.92%	1.80%	31.95%	11.99%	11.11%	30.69%	10.78%	9.25%	33.82%	9.34%	7.97%
Intern-S1	2.88%	0.06%	0.06%	20.77%	7.41%	5.87%	21.37%	6.74%	6.21%	20.80%	5.95%	5.20%
Llama 4 Maverick	2.27%	0.86%	0.86%	20.34%	9.08%	8.25%	16.79%	6.24%	5.32%	22.36%	9.59%	7.78%
Qwen3-VL-235B	1.68%	0.26%	0.06%	26.90%	7.24%	5.94%	24.53%	6.50%	5.37%	26.98%	8.20%	6.64%

complex product structures. In addition, both tools exhibit weak performance in handling multi-stage or multi-reaction scenarios, where they fail to accurately parse and analyze multiple reaction stages or parallel reactions.

Figure 31 and Figure 32 respectively show the results of ChemEAGLE and MerMaid on their original benchmarks and on RxnExBench (using reactant and product SMILES prediction as an example). As shown, both pipeline-based tools can produce valid reactant and product SMILES for simple images, but on the more complex images in RxnExBench, neither model is able to generate accurate SMILES.

RxnExBench provides a more comprehensive and challenging benchmark for chemical reaction extraction, and we believe it will advance research in chemical literature understanding and the development of large-scale AI4Chemistry datasets.

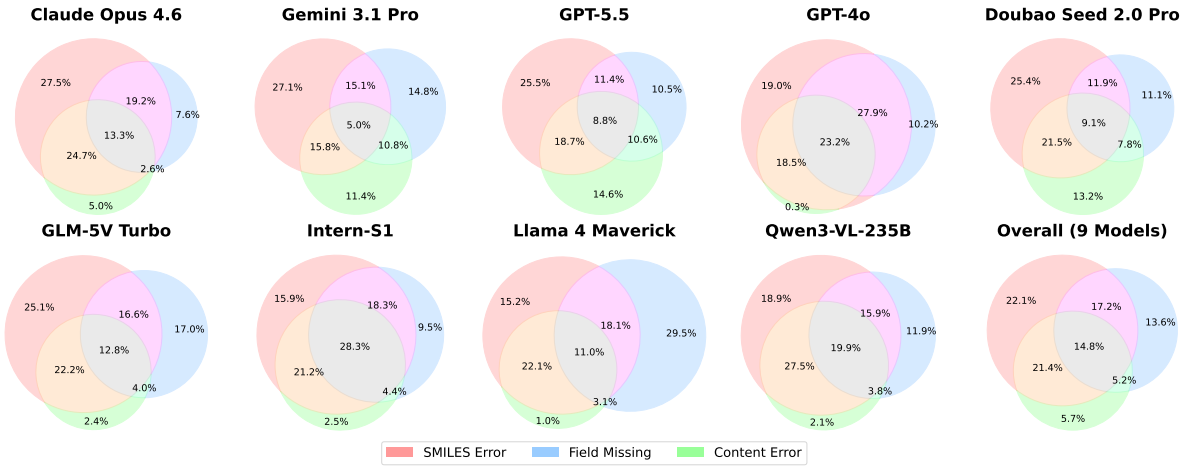


Figure 18: Venn diagrams showing error type distribution for each model. The diagrams illustrate the distribution and overlap of three error types: SMILES Error, Field Missing, and Content Error. Percentages indicate the proportion of each error category among all errors.

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a ScreenTable figure.
Your task is to parse the entire ScreenTable figure row by row. For each row entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one row entry.

Entity & text extraction rules.
- entry:
  - The row index must be placed in "entry" field. The reaction data for that row must be placed in "reactions".
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount:
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-stage conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse screenTable row by row in order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse the ScreenTable row by row in order. The row index must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
            "SMILES": ""
            "text":"",
            "yield_ratio":"",
            "isolated_yield":"",
            "enantiomeric_excess":"",
            "regioisomeric_ratio": "",
            "enantiomeric_ratio":"",
            "alkene_configuration":"",
            "diastereomeric_ratio":"",
          }
        ],
        "stages": [
          {
            "reagents": [
              {
                "SMILES":"",
                "text": "",
                "amount": ""
              }
            ],
            "catalysts": [
              {
                "SMILES":"",
                "text":"",
                "amount":""
              }
            ],
            "solvents": [
              {
                "text":"",
                "ratio":"",
                "amount":""
              }
            ],
            "ligand": [
              {
                "SMILES":"",
                "text":"",
                "amount":""
              }
            ],
            "pressure":"",
            "reaction_time":"",
            "reaction_temperature":"",
            "light_condition":"",
            "atmosphere":"",
            "PH":"",
            "stir_speed": "",
            "vacuum_condition":"",
            "scale":""
          }
        ]
      }
    ]
  },
  {
    "entry": "2",
    "reactions": [...]
  }
]
```

Figure 19: Standard inference prompt of *Screen Table*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a SubstrateScope figure.
Your job is to parse the entire SubstrateScope figure Substrate by Substrate. For each substrate entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one substrate entry.

Entity & text extraction rules.
- entry:
  - The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
  - The "identifier" must be the label used for that substrate (e.g., "1a", "2c", or an explicit label adjacent to that substrate in boldface).
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-step conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1a",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
            "SMILES": ""
            "text":"",
            "yield_ratio":"",
            "isolated_yield":"",
            "enantiomeric_excess":"",
            "regioisomeric_ratio": "",
            "enantiomeric_ratio":"",
            "alkene_configuration":"",
            "diastereomeric_ratio":"",
          }
        ],
        "stages": [
          {
            "reagents": [
              {
                "SMILES":"",
                "text": "",
                "amount": ""
              }
            ],
            "catalysts": [
              {
                "SMILES":"",
                "text":"",
                "amount":""
              }
            ],
            "solvents": [
              {
                "text":"",
                "ratio":"",
                "amount":""
              }
            ],
            "ligand": [
              {
                "SMILES":"",
                "text":"",
                "amount":""
              }
            ],
            "pressure":"",
            "reaction_time":"",
            "reaction_temperature":"",
            "light_condition":"",
            "atmosphere":"",
            "PH":"",
            "stir_speed": "",
            "vacuum_condition":"",
            "scale":""
          }
        ]
      }
    ]
  },
  {
    "entry": "2b",
    "reactions": [...]
  }
]
```

Figure 20: Standard inference prompt of *Substrate Scope*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a ScreenTable figure.
Your task is to parse the entire ScreenTable figure row by row. For each row entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one row entry.

Entity & text extraction rules.
- entry:
  - The row index must be placed in "entry" field. The reaction data for that row must be placed in "reactions".
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - At the end of the instructions, we provide a mapping table of all molecules and their SMILES. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount:
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-stage conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse screenTable row by row in order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse the ScreenTable row by row in order. The row index must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "1a": "C(C)C(=O)N(C)C",
    "2a": "CCC(=O)N(C)CB1OC(C)(C)C(C)(C)O1",
    "L1": "C1C=NC(C2C=CC=CN=2)=CC=1",
    "L3": "COC1=CC=C(C2=CC=CC=N2)N=C1",
    "L4":
"FC(F)(F)C1=CC=C(C2=CC=C(C(F)(F)F)C=N2)N=C
1",
    "L6":
"FC(F)(F)C1=CC=NC(C2=CC(C(F)(F)F)=CC=N2)=C
1",
    "L7":
"FC(F)(F)C1=CC=CC(C2=CC=CC(C(F)(F)F)=N2)=N
1",
    "L8":
"C1(C)C=NC2C3C(C=CC=2C=1C)=C(C)C(C)=CN=3
",
    "L9":
"C1C=NC2C3=NC(C)=CC=C3C=CC=2C=1"
  }
```

Figure 21: Oracle-MolRec-IDT inference prompt of *Screen Table*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a SubstrateScope figure.
Your job is to parse the entire SubstrateScope figure Substrate by Substrate. For each substrate entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one substrate entry.

Entity & text extraction rules.
- entry:
  - The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
  - The "identifier" must be the label used for that substrate (e.g., "1a", "2c", or an explicit label adjacent to that substrate in boldface).
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - At the end of the instructions, we provide a mapping table of all molecules and their SMILES. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-step conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1a",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2b",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "2a": "C1C=CC=C(CCCN=C=O)C=1",
    "2b": "COC1=CC=C(CCCCN=C=O)C=C1",
    "2c": "ClCCCCN=C=O",
    "2d": "CCOC(CCCN=C=O)=O",
    "2e": "CC(CCN=C=O)C",
    "2f": "C1OCCC(CN=C=O)C1.O",
    "2g": "C=CCCN=C=O",
    "2h": "CC(=O)OCCCCN=C=O",
    "2i": "C(C)CCCCCCN=C=O",
    "2j": "COC(=O)CCCCN=C=O",
    "2k": "C(#C)CCCCN=C=O",
    "2l": "C1=CN=C(Cl)C(C(OCCCN=C=O)=O)=C1",
    "2m": "C1OC(C(CCCN=C=O)=O)=CC=1",
    "2n": "C/C(/C)=C\\CCN=C=O",
    "2o": "CCC(N=C=O)C",
    "2p": "C1CCC(N=C=O)CC1",
    "2q": "C1CC(N=C=O)CC1",
    "2r": "CC(OC(N1CC(N=C=O)C1)=O)(C)C",
    "2s": "C1C2CC3(CC1CC(C3)C2)N=C=O",
    "2t":
"N1(CCCN=C=O)C2C=CC=CC=2C2C=CC=CC1=2",
    "2u":
"C1C=C2C(=O)N(CCN=C=O)C(=O)C2=CC=1",
    "2v": "C=CCCCCCCCCCN=C=O",
    "2w": "CCCCC(CN=C=O)CC",
    "2x": "C1C(CN=C=O)CCC1",
    "2y": "C1CC(N=C=O)CC1",
    "2z": "CC(OCCCCN=C=O)=O",
    "2aa": "CCOC(=O)CCCN=C=O",
    "2bb": "[Si](C(C)(C)C)(C)(C)OCCCN=C=O",
    "2cc": "FCCCCN=C=O",
    "2dd": "C(Cl)CCCCCN=C=O",
    "2ee": "C1C=C(CCN=C=O)C=CC=1F",
    "2ff": "C1C=C(CCN=C=O)C=CC=1Cl",
    "2gg": "C12OCCC1=CC(CCN=C=O)=CC=2",
    "2hh": "C(#C)CCN=C=O",
      ... ,
  }
```

Figure 22: Oracle-MolRec-IDT inference prompt of *Substrate Scope*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a ScreenTable figure.
Your task is to parse the entire ScreenTable figure row by row. For each row entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one row entry.

Entity & text extraction rules.
- entry:
  - The row index must be placed in "entry" field. The reaction data for that row must be placed in "reactions".
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - Each molecule in the image is enclosed by a blue-lined bounding box and uniquely identified by a white index number (idx) placed at the top-left corner of the box.
  - At the end of the instructions, we provide a mapping table of all molecules idx and their SMILES. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount:
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-stage conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse screenTable row by row in order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse the ScreenTable row by row in order. The row index must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "0": "C(C)C(=O)N(C)C",
    "1": "CCC(=O)N(C)CB1OC(C)(C)C(C)(C)O1",
    "2": "C1C=NC(C2C=CC=CN=2)=CC=1",
    "3": "COC1=CC=C(C2=CC=CC=N2)N=C1",
    "4":
"FC(F)(F)C1=CC=C(C2=CC=C(C(F)(F)F)C=N2)N=C
1",
    "5":
"FC(F)(F)C1=CC=NC(C2=CC(C(F)(F)F)=CC=N2)=C
1",
    "6":
"FC(F)(F)C1=CC=CC(C2=CC=CC(C(F)(F)F)=N2)=N
1",
    "7":
"C1(C)C=NC2C3C(C=CC=2C=1C)=C(C)C(C)=CN=3
",
    "8": "C1C=NC2C3=NC(C)=CC=C3C=CC=2C=1"
  }
```

Figure 23: Oracle-MolRec-BBOX-IDX inference prompt of *Screen Table*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a SubstrateScope figure.
Your job is to parse the entire SubstrateScope figure Substrate by Substrate. For each substrate entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one substrate entry.

Entity & text extraction rules.
- entry:
  - The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
  - The "identifier" must be the label used for that substrate (e.g., "1a", "2c", or an explicit label adjacent to that substrate in boldface).
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - Each molecule in the image is enclosed by a blue-lined bounding box and uniquely identified by a white index number (idx) placed at the top-left corner of the box.
  - At the end of the instructions, we provide a mapping table of all molecules idx and their SMILES. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-step conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1a",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2b",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "0": "C1C=CC=C(CCCN=C=O)C=1",
    "1": "COC1=CC=C(CCCCN=C=O)C=C1",
    "2": "ClCCCCN=C=O",
    "3": "CCOC(CCCN=C=O)=O",
    "4": "CC(CCN=C=O)C",
    "5": "C1OCCC(CN=C=O)C1.O",
    "6": "C=CCCN=C=O",
    "7": "CC(=O)OCCCCN=C=O",
    "8": "C(C)CCCCCCN=C=O",
    "9": "COC(=O)CCCCN=C=O",
    "10": "C(#C)CCCCN=C=O",
    "11": "C1=CN=C(Cl)C(C(OCCCN=C=O)=O)=C1",
    "12": "C1OC(C(CCCN=C=O)=O)=CC=1",
    "13": "C/C(/C)=C\\CCN=C=O",
    "14": "CCC(N=C=O)C",
    "15": "C1CCC(N=C=O)CC1",
    "16": "C1CC(N=C=O)CC1",
    "17": "CC(OC(N1CC(N=C=O)C1)=O)(C)C",
    "18": "C1C2CC3(CC1CC(C3)C2)N=C=O",
    "19":
"N1(CCCN=C=O)C2C=CC=CC=2C2C=CC=CC1=2",
    "20":
"C1C=C2C(=O)N(CCN=C=O)C(=O)C2=CC=1",
    "21": "C=CCCCCCCCCCN=C=O",
    "22": "CCCCC(CN=C=O)CC",
    "23": "C1C(CN=C=O)CCC1",
    "24": "C1CC(N=C=O)CC1",
    "25": "CC(OCCCCN=C=O)=O",
    "26": "CCOC(=O)CCCN=C=O",
    "27": "[Si](C(C)(C)C)(C)(C)OCCCN=C=O",
    "28": "FCCCCN=C=O",
    "29": "C(Cl)CCCCCN=C=O",
    "30": "C1C=C(CCN=C=O)C=CC=1F",
    "31": "C1C=C(CCN=C=O)C=CC=1Cl",
    "32": "C12OCCC1=CC(CCN=C=O)=CC=2",
    "33": "C(#C)CCN=C=O",
      ... ,
  }
```

Figure 24: Oracle-MolRec-BBOX-IDX inference prompt of *Substrate Scope*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a ScreenTable figure.
Your task is to parse the entire ScreenTable figure row by row. For each row entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one row entry.

Entity & text extraction rules.
- entry:
  - The row index must be placed in "entry" field. The reaction data for that row must be placed in "reactions".
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - At the end of the instructions, we provide a mapping table of the absolute position coordinates of each molecule in the image with its corresponding SMILES representation. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount:
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-stage conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse screenTable row by row in order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse the ScreenTable row by row in order. The row index must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "[61, 90, 183, 185]": "C(C)C(=O)N(C)C",
    "[470, 86, 626, 188]":
"CCC(=O)N(C)CB1OC(C)(C)C(C)(C)O1",
    "[7, 573, 139, 643]":
"C1C=NC(C2C=CC=CN=2)=CC=1",
    "[146, 588, 301, 643]":
"COC1=CC=C(C2=CC=CC=N2)N=C1",
    "[314, 579, 530, 643]":
"FC(F)(F)C1=CC=C(C2=CC=C(C(F)(F)F)C=N2)N=C
1",
    "[539, 564, 682, 647]":
"FC(F)(F)C1=CC=NC(C2=CC(C(F)(F)F)=CC=N2)=C
1",
    "[43, 711, 235, 808]":
"FC(F)(F)C1=CC=CC(C2=CC=CC(C(F)(F)F)=N2)=N
1",
    "[261, 695, 432, 772]":
"C1(C)C=NC2C3C(C=CC=2C=1C)=C(C)C(C)=CN=3
",
    "[472, 695, 596, 783]":
"C1C=NC2C3=NC(C)=CC=C3C=CC=2C=1"
  }
```

Figure 25: Oracle-MolRec-BBOX-CORD inference prompt of *Screen Table*

You are a chemistry expert who excels at parsing chemical reactions. You will receive an image that is a SubstrateScope figure.
Your job is to parse the entire SubstrateScope figure Substrate by Substrate. For each substrate entry, instantiate the reaction template with the entry's information to produce one or more concrete reactions. Output a strict JSON array, where each element corresponds to one substrate entry.

Entity & text extraction rules.
- entry:
  - The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
  - The "identifier" must be the label used for that substrate (e.g., "1a", "2c", or an explicit label adjacent to that substrate in boldface).
- Molecules:
  - If the scheme shows a molecular structure, convert it to SMILES and put the string in the "SMILES" field; leave "text": "".
  - If the scheme shows a molecular formula/abbreviation/text (e.g., "Et3N", "NBS", "PhI(OAc)2", "R–Br"), put that string in "text" and leave "SMILES": "".
  - Output atom-level SMILES only. Do not use abbreviations or pseudo-atoms such as Me, Et, iPr, tBu, Ph, Bn, Boc, OMe, OAc, OCO2Me, etc.
  - At the end of the instructions, we provide a mapping table of the absolute position coordinates of each molecule in the image with its corresponding SMILES representation. Please refer to this table carefully whenever you need to convert molecular structures to SMILES.
- Conditions:
  - Record temperature/time/pressure/atmosphere/light exactly as written (e.g., "rt", "25 °C", "16 h", "O2", "N2", "hv", "Blue LED") and other relevant conditions exactly as written.
  - If a field is not present, leave it as "".
- Selectivity / ratios:
  - yield_ratio: copy as shown (e.g., "78%").
  - enantiomeric_excess: e.g., "94%".
  - enantiomeric_ratio: e.g., "95:5".
  - diastereomeric_ratio: e.g., "10:1".
  - regioisomeric_ratio: e.g., "20:1".
  - alkene_configuration: e.g., "E", "Z", "E/Z 9:1".
- amount
  - Fill in all amount-related information for the object except for the reaction scale, such as mL, mol%, g, equiv, ratio, etc. Separate multiple entries with semicolons if several types of amount information are present.
- Footnote
  - Footnotes provide additional critical information about specific conditions, or essential factors in the reaction. If the image contains any footnotes, please ensure that incorporate these footnotes into reaction analysis.
- Multiple reactions:
  - If there are multiple reactions in the reaction template, add multiple reaction objects in the corresponding list.
- Multi-step conditions:
  - Use the "stages" list. Stage order must match the scheme order.
  - Put stage-specific conditions inside that stage only.
- Order:
  - Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The JSON array should follow this order.

{JSON schema.}

Strict output rules
1. Output valid JSON only. No extra text, comments, or trailing commas.
2. Use exactly the fields above. For any missing item, set the field to "".
3. If the image contains multiple reaction schemes or different reactions, output them as a list of JSON objects.
4. For molecular structures, fill "SMILES" and leave "text": "". For molecular formula/abbreviation/text, fill "text" and leave "SMILES": "".
5. If the image contains any footnotes that are crucial for accurate parsing, treat them as supplemental reaction information and apply them to each reaction they refer to.
6. Parse substrate left-to-right, top-to-bottom or the figure's natural reading order. The identifier must be placed in "entry". The reaction data for that row must be placed in "reactions".
7. Please do not add extra text information to the fields.
8. Please carefully review the mapping table below. Use it when you need to convert molecular structures to SMILES.

{Molecules-SMILES MAP}

Your answer is (JSON only):

JSON schema.

```
[
  {
    "entry": "1a",
    "reactions": [
      {
        "reactants": [
          {
            "SMILES": "",
            "text": "",
            "amount": "",
            "conversion_rate":""
          }
        ],
        "products": [
          {
                ...
          }
        ],
        "stages": [
                ...
        ]
      }
    ]
  },
  {
    "entry": "2b",
    "reactions": [...]
  }
]
```

Molecules-SMILES MAP.

```
{
    "[22, 158, 249, 249]":
"C1C=CC=C(CCCN=C=O)C=1",
    "[261, 138, 546, 246]":
"COC1=CC=C(CCCCN=C=O)C=C1",
    "[558, 143, 767, 244]":
"ClCCCCN=C=O",
    "[775, 138, 1011, 239]":
"CCOC(CCCN=C=O)=O",
    "[1026, 155, 1197, 233]":
"CC(CCN=C=O)C",
    "[1195, 128, 1368, 254]":
"C1OCCC(CN=C=O)C1.O",
    "[22, 273, 195, 336]":
"C=CCCN=C=O",
    "[216, 271, 461, 357]":
"CC(=O)OCCCCN=C=O",
    "[478, 273, 744, 339]":
"C(C)CCCCCCN=C=O",
    "[786, 260, 1045, 343]":
"COC(=O)CCCCN=C=O",
    "[1069, 264, 1323, 337]":
"C(#C)CCCCN=C=O",
    "[45, 364, 317, 480]":
"C1=CN=C(Cl)C(C(OCCCN=C=O)=O)=C1",
    "[352, 367, 588, 472]":
"C1OC(C(CCCN=C=O)=O)=CC=1",
    "[593, 367, 777, 490]":
"C/C(/C)=C\\CCN=C=O",
    "[815, 362, 953, 440]":
"CCC(N=C=O)C",
    "[981, 372, 1157, 455]":
"C1CCC(N=C=O)CC1",
    "[1172, 382, 1358, 465]":
"C1CC(N=C=O)CC1",
    "[135, 490, 420, 598]":
"CC(OC(N1CC(N=C=O)C1)=O)(C)C",
    ... ,
  }
```

Figure 26: Oracle-MolRec-BBOX-CORD inference prompt of *Substrate Scope*

```
"instance_info": [
    {
      "entry": "1",
      "reactions": [
        {
          "reactants": [
            {
              "text": "",
              "amount": "",
              "SMILES":
              "C1=CC=CC(CSC2=CC=CC=C2)=C1",
              "conversion_rate": ""
            }
          ],
          "products": [
            {
              "text": "",
              "yield_ratio": "70%",
              "enantiomeric_excess": "",
              "regioisomeric_ratio": "",
              "enantiomeric_ratio": "",
              "alkene_configuration": "",
              "diastereomeric_ratio": "",
              "SMILES": "C1=CC=CC(C=O)=C1",
              "isolated_yield": ""
            }
          ],
          "stages": [
            {
              "reagents": [
                {
                  "amount": "2 equiv.",
                  "SMILES": "",
                  "text": "A1"
                }
              ],
              "catalysts": [],
              "solvents": [
                {
                  "text": "MeCN",
                  "amount": "0.1",
                  "ratio": ""
                }
              ],
              "ligand": [],
              "pressure": "",
              "reaction_time": "18 h",
              "reaction_temperature": "r.t.",
              "light_condition": "purple LEDs",
              "atmosphere": "",
              "description": "",
              "PH": "",
              "speed": "",
              "vacuum_condition": "",
              "cooling_heating_condition": "",
              "current_conditions": "",
              "scale": "0.1 mmol",
              "stir_speed": ""
            }
          ]
        }
      ]
    },
    {
      "entry": "2",
      "reactions": [
        {
          "reactants": [
            {
              "text": "",
              "amount": "",
              "SMILES":
              "C1=CC=CC(CSC2=CC=CC=C2)=C1",
              "conversion_rate": ""
            }
          ],
          "products": [
            {
              "text": "",
              "yield_ratio": "65%",
              "enantiomeric_excess": "",
              "regioisomeric_ratio": "",
              "enantiomeric_ratio": "",
              "alkene_configuration": "",
              "diastereomeric_ratio": "",
              "SMILES": "C1=CC=CC(C=O)=C1",
              "isolated_yield": ""
            }
          ],
          "stages": [
            {
              "reagents": [
                {
                  "amount": "1 equiv.",
                  "SMILES": "O=[N+]([O-
])C1=CC=C([N+](=O)[O-])C=C1",
                  "text": ""
                }
              ],
              "catalysts": [],
              "solvents": [
                {
                  "text": "MeCN",
                  "amount": "0.1",
                  "ratio": ""
                }
              ],
              "pressure": "",
              "reaction_time": "18 h",
              "reaction_temperature": "r.t.",
              "light_condition": "purple LEDs",
              "atmosphere": "",
              "description": "",
              "PH": "",
              "speed": "",
              "vacuum_condition": "",
              "cooling_heating_condition": "",
              "current_conditions": "",
              "scale": "0.1 mmol",
              "ligand": [],
              "stir_speed": ""
            }
          ]
        }
      ]
    },
    {.....},
```


Figure 27: Annotation result of Screen Table.

```
{
  "entry": "8",
  "reactions": [
    {
      "reactants": [
        {
          "text": "",
          "amount": "",
          "SMILES":
          "C1=CC=CC(CSC2=CC=CC=C2)=C1",
          "conversion_rate": ""
        }
      ],
      "products": [
        {
          "text": "",
          "yield_ratio": "n.r.",
          "enantiomeric_excess": "",
          "regioisomeric_ratio": "",
          "enantiomeric_ratio": "",
          "alkene_configuration": "",
          "diastereomeric_ratio": "",
          "SMILES": "C1=CC=CC(C=O)=C1",
          "isolated_yield": ""
        }
      ],
      "stages": [
        {
          "reagents": [
            {
              "amount": "2 equiv.",
              "SMILES": "",
              "text": "A1"
            }
          ],
          "catalysts": [],
          "solvents": [
            {
              "text": "MeCN",
              "amount": "0.1",
              "ratio": ""
            }
          ],
          "pressure": "",
          "reaction_time": "18 h",
          "reaction_temperature": "45 °C",
          "atmosphere": "",
          "description": "",
          "PH": "",
          "speed": "",
          "vacuum_condition": "",
          "cooling_heating_condition": "",
          "current_conditions": "",
          "scale": "0.1 mmol",
          "ligand": [],
          "light_condition": "",
          "stir_speed": ""
        }
      ]
    },
    {
      "reactants": [
        {
          "text": "",
          "amount": "",
          "SMILES":
          "C1=CC=CC(CSC2=CC=CC=C2)=C1",
          "conversion_rate": ""
        }
      ],
      "products": [
        {
          "text": "",
          "yield_ratio": "n.r.",
          "enantiomeric_excess": "",
          "regioisomeric_ratio": "",
          "enantiomeric_ratio": "",
          "alkene_configuration": "",
          "diastereomeric_ratio": "",
          "SMILES": "C1=CC=CC(C=O)=C1",
          "isolated_yield": ""
        }
      ],
      "stages": [
        {
          "reagents": [],
          "catalysts": [],
          "solvents": [
            {
              "text": "MeCN",
              "amount": "0.1",
              "ratio": ""
            }
          ],
          "ligand": [],
          "pressure": "",
          "reaction_time": "18 h",
          "reaction_temperature": "r.t.",
          "light_condition": "purple LEDs",
          "atmosphere": "",
          "description": "",
          "PH": "",
          "speed": "",
          "vacuum_condition": "",
          "cooling_heating_condition": "",
          "current_conditions": "",
          "scale": "0.1 mmol",
          "stir_speed": ""
        }
      ]
    }
  ]
}
```


Figure 28: Annotation result of Screen Table.

```
"instance_info": [
    {
      "entry": "3a",
      "reactions": [
        {
          "reactants": [
            {
              "text": "",
              "amount": "",
              "SMILES": "CC(=O)C1=CC=C(F)C=C1",
              "conversion_rate": ""
            },
            {
              "amount": "1 equiv.",
              "SMILES": "OCC1=CC=CC=C1",
              "text": "",
              "conversion_rate": ""
            }
          ],
          "products": [
            {
              "text": "",
              "yield_ratio": "96%",
              "enantiomeric_excess": "",
              "regioisomeric_ratio": "",
              "enantiomeric_ratio": "",
              "alkene_configuration": "",
              "diastereomeric_ratio": "",
              "isolated_yield": "78%",
              "SMILES":
                "C1(=CC=C(F)C=C1)C(=O)OCC1=CC=CC=C1
"
            }
          ],
          "stages": [
            {
              "reagents": [
                {
                  "text": "I₂",
                  "amount": "3 equiv.",
                  "SMILES": ""
                },
                {
                  "text": "DBU",
                  "amount": "4 equiv.",
                  "SMILES": ""
                }
              ],
              "catalysts": [],
              "solvents": [
                {
                  "text": "DCM",
                  "amount": "0.1 M",
                  "ratio": ""
                }
              ],
              "ligand": [],
              "pressure": "",
              "reaction_time": "16 h",
              "reaction_temperature": "0 °C to rt",
              "light_condition": "",
              "atmosphere": "",
              "description": "",
              "PH": "",
              "speed": "",
              "vacuum_condition": "",
              "cooling_heating_condition": "",
              "current_conditions": "",
              "stir_speed": "",
              "scale": ""
            }
          ]
        }
      ]
    },
```

```
    {
      "entry": "3b",
      "reactions": [
        {
          "reactants": [
            {
              "text": "",
              "amount": "",
              "SMILES": "CC(=O)C1=CC=C([N+](=O)[O-
])C=C1",
              "conversion_rate": ""
            },
            {
              "amount": "1 equiv.",
              "SMILES": "OCC1=CC=CC=C1",
              "text": "",
              "conversion_rate": ""
            }
          ],
          "products": [
            {
              "text": "",
              "yield_ratio": "",
              "enantiomeric_excess": "",
              "regioisomeric_ratio": "",
              "enantiomeric_ratio": "",
              "alkene_configuration": "",
              "diastereomeric_ratio": "",
              "isolated_yield": "93%",
              "SMILES":
"O=C(OCC1=CC=CC=C1)C1=CC=C([N+](=O)[O-])C=C1"
            }
          ],
          "stages": [
            {
              "reagents": [
                {
                  "text": "I₂",
                  "amount": "3 equiv.",
                  "SMILES": ""
                },
                {
                  "text": "DBU",
                  "amount": "4 equiv.",
                  "SMILES": ""
                }
              ],
              "catalysts": [],
              "solvents": [
                {
                  "text": "DCM",
                  "amount": "0.1 M",
                  "ratio": ""
                }
              ],
              "ligand": [],
              "pressure": "",
              "reaction_time": "16 h",
              "reaction_temperature": "0 °C to rt",
              "light_condition": "",
              "atmosphere": "",
              "description": "",
              "PH": "",
              "speed": "",
              "vacuum_condition": "",
              "cooling_heating_condition": "",
              "current_conditions": "",
              "stir_speed": "",
              "scale": ""
            }
          ]
        }
      ]
    },
```

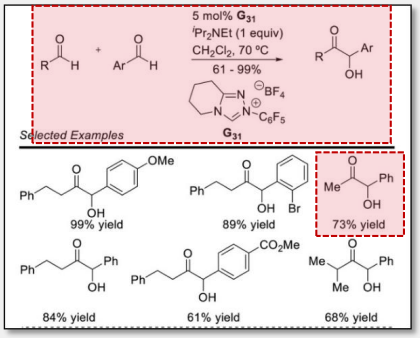

Figure 29: Annotation result of Substrate Scope.

```
{
  "entry": "3c",
  "reactions": [
    {
      "reactants": [
        {
          "text": "",
          "amount": "",
          "SMILES":
            "CC(=O)C1=CC=C([N+](=O)[O-])C=C1",
          "conversion_rate": ""
        },
        {
          "amount": "1 equiv.",
          "SMILES": "O=[N+]([O-])C1=CC=C(CO)C=C1",
          "text": "",
          "conversion_rate": ""
        }
      ],
      "products": [
        {
          "text": "",
          "yield_ratio": "",
          "enantiomeric_excess": "",
          "regioisomeric_ratio": "",
          "enantiomeric_ratio": "",
          "alkene_configuration": "",
          "diastereomeric_ratio": "",
          "isolated_yield": "79%",
          "SMILES": "O=C(OCC1=CC=C([N+](=O)[O-
])C=C1)C1=CC=C([N+](=O)[O-])C=C1"
        }
      ],
      "stages": [
        {
          "reagents": [
            {
              "text": "I₂",
              "amount": "3 equiv.",
              "SMILES": ""
            },
            {
              "text": "DBU",
              "amount": "4 equiv.",
              "SMILES": ""
            }
          ],
          "catalysts": [],
          "solvents": [
            {
              "text": "DCM",
              "amount": "0.1 M",
              "ratio": ""
            }
          ],
          "ligand": [],
          "pressure": "",
          "reaction_time": "16 h",
          "reaction_temperature": "0 °C to rt",
          "light_condition": "",
          "atmosphere": "",
          "description": "",
          "PH": "",
          "speed": "",
          "vacuum_condition": "",
          "cooling_heating_condition": "",
          "current_conditions": "",
          "stir_speed": "",
          "scale": ""
        }
      ]
    }
  ]
},
```

```
{
  "entry": "3d",
  "reactions": [
    {
      "reactants": [
        {
          "text": "",
          "amount": "",
          "SMILES": "CC(=O)C1=CC=C([N+](=O)[O-
])C=C1",
          "conversion_rate": ""
        },
        {
          "amount": "1 equiv.",
          "SMILES": "COC1=CC=C(CO)C=C1OC",
          "text": "",
          "conversion_rate": ""
        }
      ],
      "products": [
        {
          "text": "",
          "yield_ratio": "",
          "enantiomeric_excess": "",
          "regioisomeric_ratio": "",
          "enantiomeric_ratio": "",
          "alkene_configuration": "",
          "diastereomeric_ratio": "",
          "isolated_yield": "61%",
          "SMILES":
"COC1=CC=C(COC(=O)C2=CC=C([N+](=O)[O-])C=C2)C=C1OC"
        }
      ],
      "stages": [
        {
          "reagents": [
            {
              "text": "I₂",
              "amount": "3 equiv.",
              "SMILES": ""
            },
            {
              "text": "DBU",
              "amount": "4 equiv.",
              "SMILES": ""
            }
          ],
          "catalysts": [],
          "solvents": [
            {
              "text": "DCM",
              "amount": "0.1 M",
              "ratio": ""
            }
          ],
          "ligand": [],
          "pressure": "",
          "reaction_time": "16 h",
          "reaction_temperature": "0 °C to rt",
          "light_condition": "",
          "atmosphere": "",
          "description": "",
          "PH": "",
          "speed": "",
          "vacuum_condition": "",
          "cooling_heating_condition": "",
          "current_conditions": "",
          "stir_speed": "",
          "scale": ""
        }
      ]
    }
  ]
}
...
]
```


Figure 30: Annotation result of Substrate Scope.

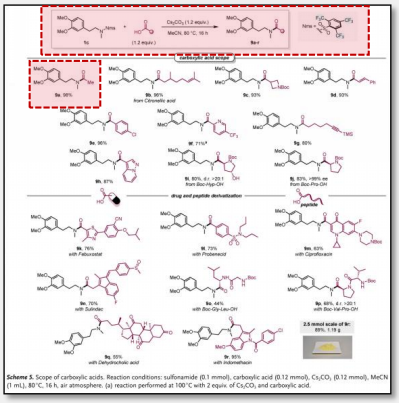
Substrate Scope Example of ChemEAGLE



ChemEAGLE result

Reactants:
SMILES: CC=O ✓
SMILES: O=Cc1ccccc1 ✓
Products:
SMILES: CC(=O)C(O)c1ccccc1 ✓

Substrate Scope Example of RxnExBench



ChemEAGLE result

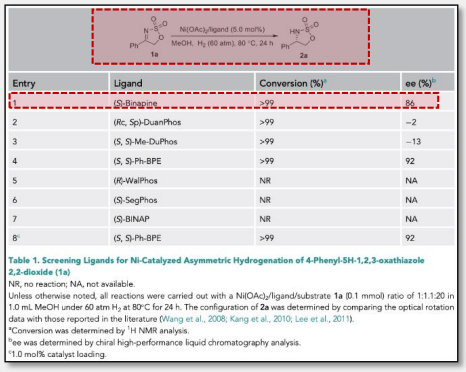
Reactants:
SMILES: *N(C)CCc1ccc(OC)c(OC)c1 ✗
SMILES: O=C(O)O ✗
Products:
SMILES: COc1ccc(CCN(C)C(C)=O)cc1OC ✗

GT

Reactants:
SMILES:COC1=CC=C(CCN(C)S(=O)(=O)C2=C(C(F)(F)F)C=C(C(F)(F)F)C=C2C(F)(F)F)C=C1OC
SMILES: CC(=O)O
Products:
SMILES: COC1=CC=C(CCN(C)C(C)=O)C=C1OC

Figure 31: Example inference results of ChemEAGLE on ChemEAGLE dataset and on RxnExBench.

Screen Table Example of Mermaid Dataset



Entry	Ligand	Conversion (%)[a]	ee (%)[b]
1	(S)-Binapine	>99	84
2	(Rc, Sp)-DuanPhos	>99	−2
3	(S, S)-Me-DuPhos	>99	−13
4	(S, S)-Ph-BPE	>99	92
5	(R)-WalPhos	NR	NA
6	(S)-SegPhos	NR	NA
7	(S)-BINAP	NR	NA
8[c]	(S, S)-Ph-BPE	>99	92

Table 1. Screening Ligands for Ni-Catalyzed Asymmetric Hydrogenation of 4-Phenyl-5H-1,2,3-oxathiazole 2,2-dioxide (1a)

NR, no reaction; NA, not available.

Unless otherwise noted, all reactions were carried out with a $Ni(OAc)_2$/ligand/substrate **1a** (0.1 mmol) ratio of 1:1.1:20 in 1.0 mL MeOH under 60 atm H_2 at 80°C for 24 h. The configuration of **2a** was determined by comparing the optical rotation data with those reported in the literature (Wang et al., 2008; Kang et al., 2010; Lee et al., 2011).

[a]Conversion was determined by ^{1}H NMR analysis.

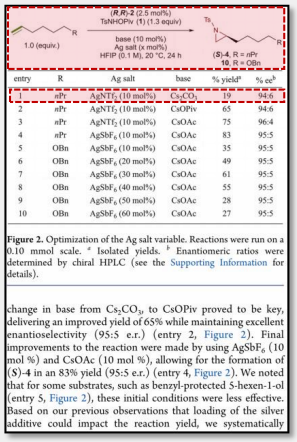
[b]ee was determined by chiral high-performance liquid chromatography analysis.

[c]1.0 mol% catalyst loading.

Mermaid result

Reactants:
SMILES: O=S1(=O)N=C(c2ccccc2)CO1 √
Products:
SMILES: O=S1(=O)N[C@@H](c2ccccc2)CO1 √

Screen Table Example of RxnExBench (with R-group)



entry	R	Ag salt	base	% yield[a]	% ee[b]
1	nPr	$AgNTf_2$ (10 mol%)	Cs_2CO_3	19	94:6
2	nPr	$AgNTf_2$ (10 mol%)	CsOPiv	65	94:6
3	nPr	$AgNTf_2$ (10 mol%)	CsOAc	75	96:4
4	nPr	$AgSbF_6$ (10 mol%)	CsOAc	83	95:5
5	OBn	$AgSbF_6$ (10 mol%)	CsOAc	35	95:5
6	OBn	$AgSbF_6$ (20 mol%)	CsOAc	49	95:5
7	OBn	$AgSbF_6$ (30 mol%)	CsOAc	61	95:5
8	OBn	$AgSbF_6$ (40 mol%)	CsOAc	55	95:5
9	OBn	$AgSbF_6$ (50 mol%)	CsOAc	28	95:5
10	OBn	$AgSbF_6$ (60 mol%)	CsOAc	27	95:5

Figure 2. Optimization of the Ag salt variable. Reactions were run on a 0.10 mmol scale. [a] Isolated yields. [b] Enantiomeric ratios were determined by chiral HPLC (see the Supporting Information for details).

change in base from Cs_2CO_3, to CsOPiv proved to be key, delivering an improved yield of 65% while maintaining excellent enantioselectivity (95:5 e.r.) (entry 2, Figure 2). Final improvements to the reaction were made by using $AgSbF_6$ (10 mol %) and CsOAc (10 mol %), allowing for the formation of (*S*)-**4** in an 83% yield (95:5 e.r.) (entry 4, Figure 2). We noted that for some substrates, such as benzyl-protected 5-hexen-1-ol (entry 5, Figure 2), these initial conditions were less effective. Based on our previous observations that loading of the silver additive could impact the reaction yield, we systematically

Mermaid result

Reactants:
SMILES: *CCCCC=C ×
Products:
SMILES: *CCCC[C@H]1CN1S(=O)(=O)c1ccc(C)cc1 ×

GT

Reactants:
SMILES: C=CCCCCCCC
Products:
SMILES:
CCCCCCC[C@H]1CN1S(=O)(=O)C1=CC=C(C)C=C1

Figure 32: Example inference results of MerMaid on MerMaid dataset and on RxnExBench.