AUTOCHEM: RECONSTRUCTING CHEMICAL REACTIONS FROM DISTRIBUTED MULTIMODAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Chemical AI systems require detailed experiment-level reaction records, yet such data remain difficult to recover from scientific literature. Information about a single experiment is often distributed across reaction figures, Supporting Information (SI), and shared procedures, connected by local compound labels and procedural references. We formulate this challenge as *reaction reconstruction* and introduce **AutoChem**, a fully automated, multimodal multi-agent framework that converts raw papers and SI into structured reaction records. AutoChem explicitly resolves compound identities and general-procedure inheritance, aligns textual and visual evidence, and preserves experiment-specific conditions, outcomes, and provenance. On 22 expert-annotated papers covering 14 reaction-information fields, AutoChem achieves **79.56% micro-F1**, outperforming direct extraction with the same GPT-5-mini backbone by **32.26 points**. Ablations highlight procedure inheritance as a key contributor to extraction accuracy. The reconstructed records yield a reaction-centric knowledge graph with **11,138 nodes** and **39,075 edges**, together with **344** controlled-variable optimization questions: 179 for condition optimization and 165 for substrate optimization. Without retrieval, none of the evaluated LLM configurations significantly exceeds its random baseline after multiple-testing correction. Cross-paper KG retrieval improves Top-1 accuracy across three models, with the largest numerical gain observed for GPT-5-mini (from 45.9% to 50.3%), suggesting downstream utility of the automatically reconstructed knowledge. Code, benchmark, and demo are available at `https://anonymous.4open.science/r/AutoChem`.

1 INTRODUCTION

AI systems are increasingly used for synthesis planning and laboratory automation (Jablonka et al., 2024; M. Bran et al., 2024; Boiko et al., 2023). These applications require detailed experiment-level reaction records specifying molecular participants, procedures, conditions, and outcomes. Although scientific literature provides rich experimental evidence, well-structured, machine-readable records remain scarce. Recent LLM-based extraction studies show promise in recovering scientific data (Dagdelen et al., 2024; Polak & Morgan, 2024), but converting complete chemical papers into comprehensive reaction records remains challenging.

The challenge arises from both information fragmentation and multimodal reporting (Figure 1). A single experiment may span a main-text scheme, Supporting Information (SI), tables, and shared procedures: molecular structures appear in figures, while quantities, conditions, and outcomes are distributed across text and annotations. Local compound labels and procedure references connect these sources. Recovering a complete reaction therefore requires linking complementary evidence, beyond recognizing entities or parsing individual schemes. We call this task ***reaction reconstruction***: assembling distributed textual and visual evidence into a coherent experiment-level record.

Existing methods have advanced chemical information extraction, but reconstructing complete experiments remains challenging when their evidence is distributed across document regions and modalities. Image-based systems parse reaction diagrams and molecular graphics (Qian et al., 2023; Wilary & Cole, 2023; Chen et al., 2025b), while text-oriented approaches extract entities, roles, and conditions using supervised models, weak supervision, or adapted LLMs (Guo et al., 2022; Zhong et al., 2023; Zhang et al., 2024; Kumari & Garg, 2025; Vangala et al., 2024). Document-level and

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

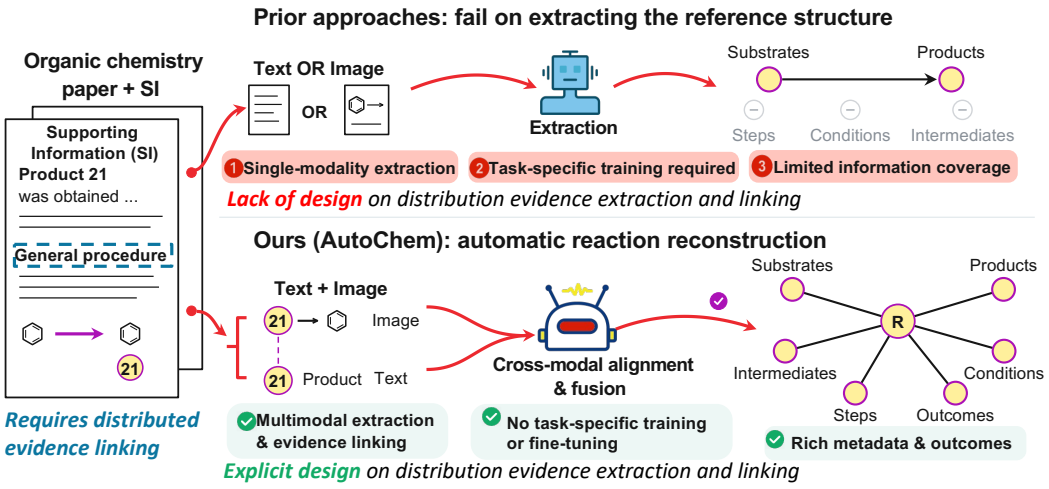


Figure 1: **From fragmented evidence to experiment-level reaction records.** Information about a single experiment is often distributed across reaction figures, Supporting Information (SI), and shared general procedures, connected by local compound labels and procedure references. Extracting these sources independently can leave molecular identities, conditions, and outcomes disconnected or incomplete. AutoChem links textual and visual evidence through compound-identity resolution and procedure inheritance to reconstruct coherent experiment-level reaction records without task-specific training or fine-tuning.

multimodal systems, including OpenChemIE, MERMaid, ChemEAGLE, ChemMiner, and ReactionSeek, further incorporate cross-modal associations, document context, or entity resolution (Fan et al., 2024; Leong et al., 2024; 2025; Chen et al., 2025c;a; Li et al., 2026). Even when multiple modalities are processed, complete reconstruction requires linking their evidence to the same experiment, including resolving local compound identifiers and recovering information inherited from shared procedures. Missing these dependencies can leave participants, conditions, or outcomes disconnected or incomplete, even when individual pieces of evidence are extracted correctly. This motivates explicit resolution of entity and procedure references during experiment-level reconstruction.

We introduce **AutoChem**, a fully automated, multimodal multi-agent framework for reaction reconstruction from raw papers and SI. AutoChem explicitly resolves compound labels and general-procedure inheritance, recovers structures and experimental information from reaction figures, and aligns the two modalities into standardized records. This design connects information omitted from individual entries with the evidence needed to complete each experiment. The records are organized into a reaction-centric knowledge graph (KG), with explicit reaction nodes preserving participants, conditions, outcomes, and provenance. The framework uses GPT-5-mini without task-specific training or fine-tuning.

On 22 expert-annotated papers covering 14 reaction-information fields, AutoChem achieves 79.56% micro-F1, exceeding direct extraction with the same backbone by 32.26 points. Ablations identify procedure inheritance as the largest contributor among the evaluated components. The reconstructed KG supports 344 controlled-variable questions on condition and substrate optimization. Without retrieval, no evaluated LLM configuration significantly exceeds its random baseline after multiple-testing correction. Cross-paper KG retrieval yields numerical Top-1 gains across three models, including an increase from 45.9% to 50.3% for GPT-5-mini, suggesting downstream utility of the reconstructed knowledge. Such an automated source of literature-grounded evidence could also support future self-improving chemical AI systems (Lu et al., 2024; Zhang et al., 2025; Novikov et al., 2025).

Our contributions are threefold:

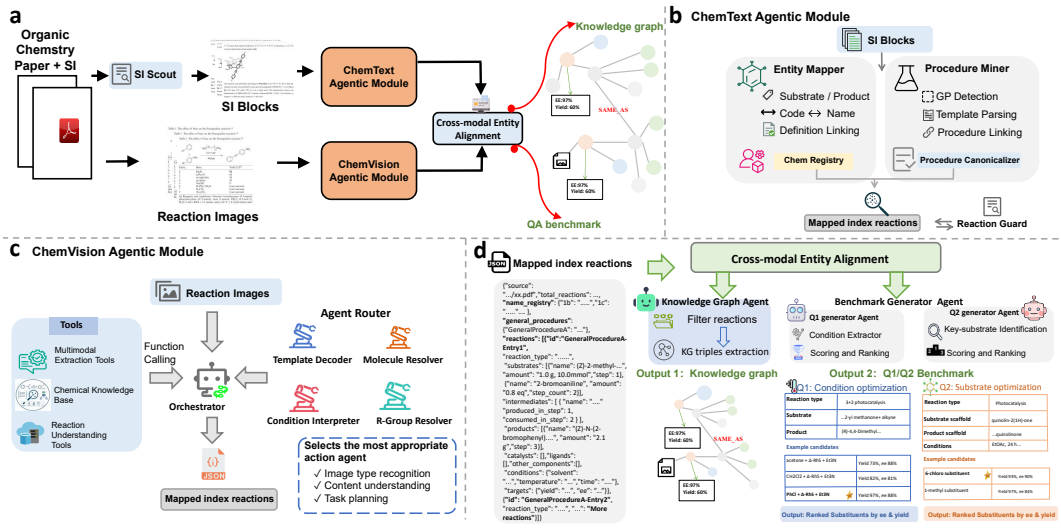


Figure 2: **Overview of AutoChem.** (a) Overall literature-to-knowledge workflow from chemical papers and SI to mapped index reactions, a reaction-centric knowledge graph, and Q1/Q2 optimization benchmarks. (b) The ChemText Agentic Module reconstructs textual reaction information through SI Scout, Entity Mapper, Procedure Miner, Procedure Canonicalizer, and Reaction Guard. (c) The ChemVision Agentic Module dynamically routes reaction images to specialized agents for template decoding, molecule resolution, condition interpretation, and R-group resolution. (d) Cross-modal Entity Alignment consolidates textual and visual reaction records, which are then processed by the Knowledge Graph Agent and Benchmark Generator Agent.

- **Reaction reconstruction and AutoChem.** We formulate experiment-level reconstruction from distributed chemical evidence and introduce a training-free framework that integrates entity resolution, procedure inheritance, and cross-modal alignment.
- **Evaluation and analysis.** We demonstrate substantial gains over direct LLM extraction and identify the contribution of procedure inheritance through controlled ablations.
- **Literature-grounded knowledge resources.** We automatically construct a reaction-centric KG and optimization benchmark, and assess their downstream utility through cross-paper retrieval experiments.

2 Method

2.1 Problem Formulation and Overview

Given chemical papers and their Supporting Information (SI), AutoChem reconstructs experiment-level reaction records from distributed textual and visual evidence. Each record contains molecular participants, procedures, conditions, quantitative outcomes, and literature provenance. This unit of representation preserves differences between experiments that share a substrate–product pair but use different conditions or report different outcomes. Let $\mathcal{X}$ denote the literature collection and $\mathcal{R} = \{r_i\}_{i=1}^{N}$ the reconstructed mapped index reactions. The framework produces a reaction-centric knowledge graph G and condition- and substrate-optimization task sets:

$$\mathcal{X} \to \mathcal{R} \to \{G, \mathcal{Q}_{\text{cond}}, \mathcal{Q}_{\text{sub}}\}.$$

As shown in Figure 2, ChemText and ChemVision recover complementary evidence, which cross-modal alignment consolidates before graph and benchmark construction. The reaction schema and detailed agent procedures appear in Appendix A.2.1 and Appendix A.2, respectively.

2.2 CHEMTEXT AGENTIC MODULE

Individual SI entries often identify compounds by local labels and refer to shared general procedures instead of repeating experimental details. ChemText explicitly resolves both dependencies before assembling reaction records. The **SI Scout** first identifies reaction-relevant pages and regions, including compound definitions, experimental descriptions, product entries, and general procedures. Restricting downstream processing to these regions limits unnecessary computation while retaining the context needed for reconstruction. The **Entity Mapper** identifies substrates and products and links document-local identifiers, such as "1a" or "21", to their chemical names and definitions. These mappings form a document-level **Chem Registry**,

$$\mathcal{D}_{\text{chem}} = \{c_j \mapsto e_j\}_{j=1}^{N_c},$$

where c_j denotes a local compound identifier and e_j its normalized chemical entity. This registry resolves abbreviated references to compounds whose full identities are introduced elsewhere in the document.

The **Procedure Miner** detects the boundaries and identifiers of general procedures, extracts reusable procedural templates, and links experimental entries to their referenced procedures. For example, an entry stating "following General Procedure A" may omit the catalyst, solvent, or temperature; linking it to that procedure recovers the shared experimental context. The **Procedure Canonicalizer** normalizes the recovered descriptions into a procedure registry, $\mathcal{D}_{\text{GP}} = \{g_k \mapsto p_k\}_{k=1}^{N_g}$, where g_k denotes a canonical procedure identifier and p_k its normalized description. Canonicalization provides a consistent representation when equivalent procedures appear in different textual forms, supporting their reuse across individual experiments.

ChemText combines the Chem Registry, linked procedures, and entry-specific evidence into textual mapped index reactions, $\mathcal{R}^{\text{text}} = \{r_i^{\text{text}}\}$. Each record contains the participants, linked procedural information, reaction steps, conditions, quantities, and outcomes recoverable from text. The **Reaction Guard** checks reaction coverage, consistency between entities and procedures, and duplicated entries. These checks address errors arising from the assembly of distributed evidence before records are passed to cross-modal alignment. Further details appear in Appendix A.2.2.

2.3 CHEMVISION AGENTIC MODULE

Reaction figures provide molecular structures and experimental information that may be missing or ambiguous in text. Their layouts also differ: some depict complete molecules, whereas others encode multiple variants through shared scaffolds and R-group definitions. ChemVision therefore selects extraction operations according to the available visual evidence. An **Orchestrator** interprets the image and determines the required operations, while an **Agent Router** invokes specialized agents based on the image type, extracted evidence, and current reconstruction requirements. The module has access to multimodal extraction tools, chemical knowledge bases, and reaction-understanding tools to support these operations.

The **Template Decoder** parses generic reaction templates and explicit R-group definitions, identifying shared scaffolds, variable positions, and their correspondence to individual reaction instances. The **Molecule Resolver** recognizes molecular structures and assigns their reaction roles, linking structures to explicit substituent assignments when available. The **R-Group Resolver** reconstructs complete molecular variants from the scaffolds and substituent definitions. This distinction is important for substrate-scope figures: recognizing a generic scaffold alone does not identify the individual compounds represented by different R-group assignments. Template decoding and variant resolution together connect the shared visual representation to experiment-specific molecular identities.

The **Condition Interpreter** recovers experimental information embedded in figures, including reagents, solvents, temperatures, reaction times, yields, and stereochemical outcomes when reported. The specialized agents produce visual mapped index reactions, $\mathcal{R}^{\text{visual}} = \{r_i^{\text{visual}}\}$, associating molecular participants with the available conditions and outcomes. These records retain evidence recoverable from reaction graphics and complement the procedure-rich textual records. Cross-modal entity alignment subsequently consolidates the two pathways into unified experiment-level records. Further agent and execution details appear in Appendix A.2.3.

2.4 CROSS-MODAL ENTITY ALIGNMENT

Textual and visual pathways provide complementary but partially overlapping descriptions of the same experiment. For example, a compound may appear as a local identifier in SI and as a molecular drawing in a reaction figure. Treating these observations independently would split one entity across modalities. AutoChem aligns them using document-local identifiers, molecular information, reaction context, and surrounding experimental evidence to determine whether they refer to the same chemical entity.

After entity alignment, the associated textual and visual evidence is combined into unified mapped index reactions. This separates the resolution of chemical identity from the assembly of experiment-level information: aligned entities connect molecular structures to the procedures, conditions, quantities, and outcomes recovered by the two pathways. Each record retains literature provenance and is serialized using a structured JSON schema. The serialization provides a common machine-readable interface for the Knowledge Graph Agent and Benchmark Generator Agent. Further alignment and serialization details appear in Appendix A.2.4.

2.5 REACTION-CENTRIC KNOWLEDGE GRAPH CONSTRUCTION

The **Knowledge Graph Agent** filters valid mapped index reactions and extracts structured relations to construct a heterogeneous graph $G = (V, E, A)$, where V denotes nodes, E denotes typed edges, and A contains node- and reaction-level attributes. The node types include Reaction, Substrate, Product, Catalyst, Condition, Scaffold, Substituent, and Paper. Each reported experiment is represented as an explicit **Reaction** node linked to its molecular participants, experimental conditions, structural information, and literature provenance.

Quantitative outcomes, including yield and enantiomeric excess, are stored as attributes of the corresponding reaction instance. This preserves the association between an outcome and the experiment that produced it, allowing reactions with the same substrate–product pair but different conditions or results to remain distinguishable. Paper nodes provide links to the source literature.

Substrates are further decomposed into shared scaffolds and substituents. This organization connects reactions with a common molecular core but different substitution patterns and provides the structural basis for the Q2 comparison groups defined below. The complete graph schema and interactive examples appear in Appendix A.2.5 and Appendix A.2.6.

2.6 OPTIMIZATION BENCHMARK CONSTRUCTION

Comparing arbitrary literature reactions would confound changes in substrates with changes in conditions. The **Benchmark Generator Agent** therefore forms controlled-variable groups. For condition optimization (Q1), it fixes the reaction type, substrate, and product while allowing conditions to vary:

$$\kappa_{\mathrm{cond}}(r) = \big(\mathrm{Type}(r), S(r), P(r)\big).$$

For substrate optimization (Q2), it fixes reaction type, substrate and product scaffolds, and conditions while comparing substituent variants:

$$\kappa_{\mathrm{sub}}(r) = \big(\mathrm{Type}(r), \mathrm{Scaffold}(S(r)), \mathrm{Scaffold}(P(r)), \mathrm{Cond}(r)\big).$$

Within each question, candidates are ranked by reported experimental outcomes. Questions with both yield and enantiomeric excess (ee) use $s(r) = 0.5\,\mathrm{yield}(r) + 0.5\,\mathrm{ee}(r)$; yield-only questions use $s(r) = \mathrm{yield}(r)$. Both outcomes are on a percentage scale, and equal weights define a fixed trade-off without prioritizing either outcome. Identical scores receive the same rank, with all top-ranked ties accepted as gold answers. This scoring rule defines the benchmark objective; construction details appear in Appendix A.2.7.

3 EXPERIMENTS

We evaluate full-document reaction reconstruction, LLM ranking on the resulting optimization benchmark, and the contribution of individual reconstruction components.

Table 1: Full-document reaction-extraction performance over 14 information fields.

Model	Correct Predictions	Total Predictions	Precision (%)	Recall (%)	F1 (%)
Base Models					
DeepSeek-V4-Flash	3,671	7,304	50.26	22.25	30.84
Gemini-3.8-Flash	7,537	12,612	59.76	45.85	51.89
GPT-5-mini	6,734	12,062	55.83	41.03	47.30
Llama 4 Maverick	3,948	9,733	40.56	23.98	30.14
Agentic System Methods					
OpenChemIE	355	1,427	24.88	2.15	3.95
Ours					
AutoChem	12,894	15,870	**81.29**	**78.11**	**79.56**

3.1 Experimental Settings

Evaluation data. The extraction corpus contains 22 chemistry papers with their SI and expert annotations over 14 fields covering participants, conditions, quantities, and outcomes. It spans diverse reaction families, including cross-coupling, photocatalytic transformations, and Ni/photoredox dual catalysis. The optimization benchmark contains 344 questions and 1,018 candidate options: 179 condition-optimization questions (Q1) and 165 substrate-optimization questions (Q2). Corpus details appear in Appendix A.3.

Models and protocol. AutoChem uses GPT-5-mini (Singh et al., 2025) as its backbone without task-specific training or fine-tuning. For reaction extraction, we compare against four direct LLM baselines: GPT-5-mini, DeepSeek-V4-Flash (Xu et al., 2026), Gemini-3.8-Flash (Team et al., 2023), and Llama 4 Maverick (Meta, 2025), together with the chemical literature extraction toolkit OpenChemIE (Fan et al., 2024). The standalone GPT-5-mini baseline provides a same-backbone comparison to assess the benefit of the reconstruction workflow. For reaction optimization, we evaluate the same four base models alongside GPT-4o (Hurst et al., 2024) and GPT-5. GPT-5-mini is evaluated with both minimal and high reasoning effort, yielding seven configurations in total. All configurations receive identical questions and candidate ordering and are evaluated against the same gold answers. In the no-KG setting, models have no access to source reaction IDs, gold labels, experimental outcomes, or the knowledge graph. We also report question-specific uniform-random baselines. Detailed model configurations and implementation settings appear in Appendix A.3.

Metrics and statistical tests. Extraction is measured by field-instance micro-averaged precision, recall, and F1 against expert annotations; field-level analysis also reports document-level F1 distributions. Annotation, reaction alignment, and scoring rules appear in Appendix A.3.1. Benchmark metrics are Top-1 accuracy and MRR on all 344 questions, and Top-3 accuracy on the 98 questions with more than two options. Random baselines account for each question's candidate count and tied gold answers. Extraction ablations use paired document-level Wilcoxon tests with Benjamini–Hochberg correction. Benchmark comparisons use Poisson–binomial tests against chance, Cochran's Q across configurations, and pairwise McNemar tests, with Holm correction for multiple comparisons as detailed in Appendix A.3.

3.2 Reaction Extraction Performance

AutoChem improves full-document reaction reconstruction. Table 1 reports 81.29% precision, 78.11% recall, and 79.56% F1 for AutoChem, exceeding the strongest direct LLM baseline, Gemini-3.8-Flash, by 27.67 F1 points. OpenChemIE obtains 3.95% F1 under this full-document, 14-field evaluation. The same-backbone comparison yields a 32.26-point gain over standalone GPT-5-mini (47.30% F1), supporting the benefit of the reconstruction workflow without changing the underlying LLM. Gains in both precision and recall indicate more complete recovery with fewer errors relative to the number of extracted field instances.

Table 2: Performance on the reaction-optimization benchmark. Overall Top-1 and MRR are evaluated on all 344 questions, while Top-3 is evaluated on the 98 questions containing more than two candidate options. Q1 and Q2 report task-specific Top-1 accuracy for condition optimization and substrate optimization, respectively. Random denotes the question-specific uniform-random baseline.

	Overall			Task-specific Top-1	
Model	**Top-1 (%)**	**Top-3 (%)**	**MRR**	**Q1 (%)**	**Q2 (%)**
Random baseline	46.4	73.9	0.699	52.3	40.1
GPT-5-mini (minimal)	45.9	72.4	0.695	52.0	39.4
GPT-5-mini (high)	**51.7**	71.4	**0.726**	56.4	**46.7**
GPT-4o	48.8	73.5	0.714	55.3	41.8
GPT-5	45.1	**77.6**	0.697	52.0	37.6
DeepSeek-V4-Flash	48.0	73.5	0.712	54.7	40.6
Gemini-3.8-Flash	50.6	76.5	0.722	**57.0**	43.6
Llama 4 Maverick	48.8	68.4	0.708	**57.0**	40.0

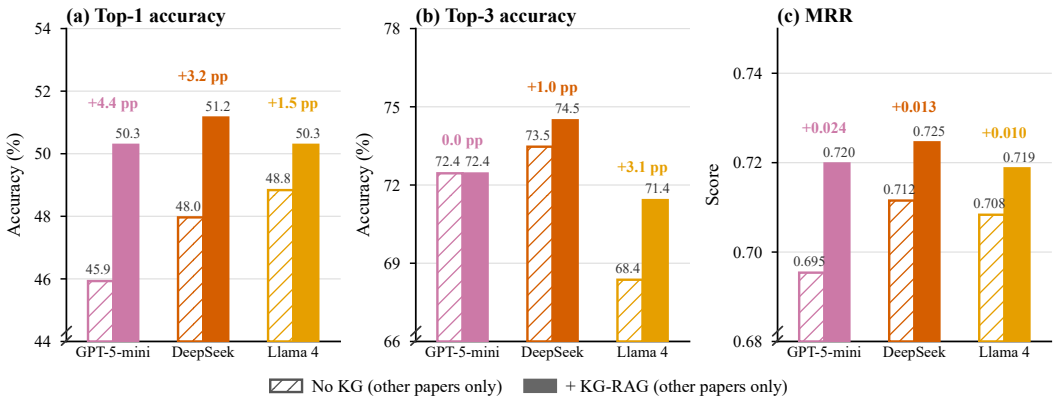


Figure 3: Ranking performance with and without cross-paper KG retrieval. Top-1 and MRR are evaluated on all 344 questions; Top-3 is evaluated on the 98 questions with more than two candidate options. Hollow bars denote the no-KG baseline and filled bars denote KG-RAG.

3.3 Reaction Optimization Benchmark Evaluation

Table 2 reports overall ranking performance and task-specific Top-1 accuracy for the seven model configurations.

No model significantly exceeds chance after correction. GPT-5-mini (high) achieves the highest numerical Top-1 accuracy (51.7%) and MRR (0.726), while GPT-5 leads on Top-3 (77.6%). The random Top-1 baseline is 46.4%, reflecting the prevalence of two-option questions (246 of 344) and tied gold answers. GPT-5-mini (high) exceeds this baseline by 5.3 points, but its Poisson–binomial result ($p = 0.0367$) does not remain significant after Holm correction across seven configurations ($p_{\text{Holm}} = 0.2571$). Cochran's Q likewise finds no significant overall difference ($Q = 7.874$, $p = 0.247$), and none of the 21 pairwise comparisons remains significant after correction. These results do not establish a reliable ordering among the tested configurations.

Cross-paper knowledge retrieval. We evaluate whether the reaction knowledge automatically reconstructed by AutoChem provides useful context for optimization questions derived from other papers. This assesses the downstream utility of the reconstructed records beyond extraction accuracy. For each of the 344 questions, we exclude all KG records originating from its source paper and mask outcome fields in the retrieved evidence, including yield, ee, er, and dr, limiting direct access to question-specific answers. We compare three LLMs with and without retrieval, providing the same retrieved evidence to all models to examine whether cross-paper reaction context improves candidate ranking.

Table 3: Full-document extraction performance and ablation effects on 22 papers. Full reports the equal-weight mean of run-level micro-averaged F1 scores (%). w/o Scout removes the SI Scout, w/o Proc. removes the Procedure Miner, and w/o Mapper removes the Entity Mapper. Delta columns report percentage-point changes relative to Full. A superscript * indicates a statistically significant difference from Full under a paired document-level Wilcoxon signed-rank test after Benjamini–Hochberg correction ($q < 0.05$).

Field	Full	Δ w/o Scout	Δ w/o Proc.	Δ w/o Mapper
Substrate	83.52	-1.56 ± 2.33	$-15.14^{*} \pm 2.14$	-9.33 ± 7.42
Product	86.61	-2.77 ± 1.55	-4.58 ± 1.29	$-8.58^{*} \pm 2.74$
Intermediate	65.79	$+4.99 \pm 3.75$	-31.09 ± 25.73	-4.90 ± 2.55
Catalyst	84.31	$+1.13 \pm 2.37$	$-36.76^{*} \pm 11.44$	-3.05 ± 6.14
Ligand	76.91	-0.10 ± 0.36	$-28.17^{*} \pm 7.26$	-5.43 ± 6.55
Solvent	81.41	$+0.12 \pm 1.18$	$-29.03^{*} \pm 7.87$	-2.77 ± 4.01
Other comp.	77.05	$+0.20 \pm 4.82$	$-27.93^{*} \pm 10.51$	-2.12 ± 4.50
Reaction cond.	45.34	$+3.39 \pm 1.45$	$-21.38^{*} \pm 3.83$	-0.02 ± 2.16
Yield	90.52	-0.89 ± 1.43	-3.47 ± 1.86	$-3.86^{*} \pm 6.22$
ee	87.32	-1.44 ± 3.11	-1.44 ± 3.11	-2.55 ± 4.72
Substrate amt.	81.24	-0.65 ± 1.37	$-20.15^{*} \pm 5.70$	-9.23 ± 8.41
Catalyst amt.	84.79	$+0.14 \pm 0.88$	$-37.67^{*} \pm 11.94$	-4.31 ± 6.61
Ligand amt.	76.84	$+0.07 \pm 0.35$	$-42.04^{*} \pm 1.78$	-4.86 ± 4.87
Solvent amt.	73.31	$+0.13 \pm 1.73$	$-27.40^{*} \pm 9.13$	-2.46 ± 3.37
Model tokens (M)	9.179 ± 0.641	$+0.347 \pm 0.430$	-2.276 ± 0.346	-1.693 ± 0.858
Wall-clock (h)	3.84 ± 0.20	$+0.23 \pm 0.47$	-0.80 ± 0.19	$+0.19 \pm 0.61$

KG retrieval improves Top-1 across three models. Figure 3 shows Top-1 gains from 45.9% to 50.3% for GPT-5-mini ($+4.4$ points; 15 additional correct answers), from 48.0% to 51.2% for DeepSeek-V4-Flash ($+3.2$), and from 48.8% to 50.3% for Llama 4 Maverick ($+1.5$). MRR also increases for all three, from 0.695 to 0.720, 0.712 to 0.725, and 0.708 to 0.719, respectively. Top-3 changes are smaller: GPT-5-mini remains at 72.4%, while DeepSeek and Llama improve from 73.5% to 74.5% and 68.4% to 71.4%. These numerical gains are consistent with cross-paper evidence helping models rank the correct candidate first; they do not establish statistical significance for the retrieval effect.

3.4 Ablation Study

We remove the SI Scout, Procedure Miner, and Entity Mapper individually to examine region filtering, procedure inheritance, and entity resolution. Table 3 reports field-level F1 changes and resource costs relative to the full system; significance is assessed after Benjamini–Hochberg correction ($q < 0.05$).

Procedure inheritance drives the largest extraction gains. Removing the Procedure Miner significantly reduces F1 in 10 of 14 fields, with the largest losses in ligand amount (-42.04 points), catalyst amount (-37.67), and catalyst (-36.76). Solvent, ligand, other components, and reaction conditions also decline substantially. This pattern supports explicit procedure inheritance: these details are often specified once in a general procedure and omitted from individual entries.

Entity mapping supports identity-sensitive fields. Removing the Entity Mapper lowers substrate, product, and substrate-amount F1 by 9.33, 8.58, and 9.23 points, respectively; the product decrease is significant after correction. Yield F1 also decreases significantly (-3.86 points). The participant-level losses are consistent with the need to resolve local compound identifiers before combining evidence across document regions.

Modules exhibit distinct accuracy–efficiency trade-offs. Removing the SI Scout increases mean token use by 0.347M and runtime by 0.23 h, with mixed, nonsignificant field-level F1 changes. This is consistent with its role in filtering unnecessary processing, although the resource differences have appreciable variability. Removing the Procedure Miner saves 2.276M tokens and 0.80 h at a substantial accuracy cost. Removing the Entity Mapper saves 1.693M tokens, while its mean

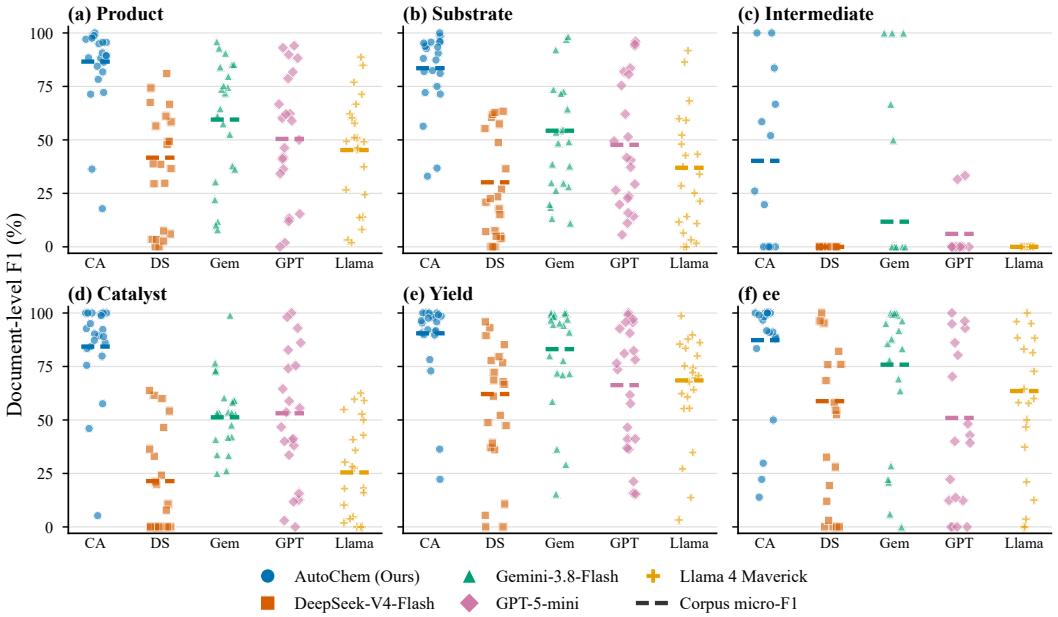


Figure 4: Field-level reaction-extraction performance on six representative fields across the 22-document corpus. Panels show product, substrate, intermediate, catalyst, yield, and enantiomeric excess (ee). Markers denote document-level F1 scores for individual papers, while horizontal dashed lines indicate corpus-level micro-F1 for each model. Results for the remaining eight fields are provided in Appendix A.3.2.

runtime increases by 0.19 h with considerable variability. Thus, its primary observed benefit is extraction accuracy, rather than a demonstrated runtime reduction.

3.5 Field-level Extraction Analysis

Extraction gains span multiple information fields. Figure 4 shows AutoChem's advantage across participants, catalysts, and outcomes, rather than in a single field. Product, substrate, catalyst, yield, and ee F1 range from 83.52% to 90.52%, whereas intermediate F1 is lower at 65.79%. Intermediate extraction also has variable ablation effects: removing the Procedure Miner changes F1 by -31.09 ± 25.73 points (Table 3). Reaction conditions remain particularly challenging (45.34% F1), and document-level scores vary across papers. Together with the procedure-inheritance ablation, these results point to integrating distributed experimental context as a remaining challenge. The other eight fields appear in Appendix A.3.2.

4 Conclusion

We introduced AutoChem, a fully automated framework for reconstructing experiment-level reaction knowledge from chemical papers and SI. By explicitly resolving compound identities, recovering inherited procedures, and integrating textual and visual evidence, AutoChem achieves 79.56% micro-F1 on 22 expert-annotated papers, exceeding its standalone GPT-5-mini backbone by 32.26 points. Ablations highlight the importance of procedure inheritance. The resulting reaction-centric knowledge graph and 344-question optimization benchmark connect reconstruction quality with downstream evaluation, while numerical gains from cross-paper retrieval suggest that the reconstructed records provide useful context for reaction ranking. These findings support automated literature reconstruction as a source of experimentally grounded knowledge for chemical AI. Future work will broaden reaction coverage, improve condition reconstruction, and evaluate the role of these resources in iterative model-improvement workflows.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting, editing, and language refinement, as well as code development. The research ideas were developed independently by the authors, and the literature survey was conducted manually without generative AI assistance. The authors manually reviewed the codebase, including AI-assisted code. The use of LLMs within AutoChem is described in the Method section. We take responsibility for the final manuscript, implementation, and reported results.

ETHICS STATEMENT

AutoChem is a research framework for reconstructing chemical reaction knowledge from scientific literature and evaluating reaction-optimization reasoning. Automatically extracted records and model-generated rankings may contain errors or omissions and should not be treated as validated experimental instructions. Before informing laboratory work, these outputs should be checked against the original literature by qualified researchers. Any experimental application requires independent assessment of chemical feasibility and safety and adherence to applicable laboratory protocols.

REPRODUCIBILITY STATEMENT

To support reproducibility, we release the AutoChem implementation, the constructed knowledge graph, and the source code for the interactive demo at `https://anonymous.4open.science/r/AutoChem`. The framework is described in Section 2, and detailed model configurations, implementation settings, and evaluation protocols are provided in Appendix A.3.

REFERENCES

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Andres M Bran, Zlatko Jončev, and Philippe Schwaller. Knowledge graph extraction from total synthesis documents. In *Proceedings of the 1st Workshop on Language+ Molecules (L+ M 2024)*, pp. 74–84, 2024.

Kexin Chen, Yuyang Du, Junyou Li, Hanqun Cao, Menghao Guo, Xilin Dang, Lanqing Li, Jiezhong Qiu, Guangyong Chen, and Pheng Ann Heng. Chemminer: A large language model agent system for chemical literature data mining. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 7645–7653. IEEE, 2025a.

Yufan Chen, Ching Ting Leung, Jianwei Sun, Yong Huang, Linyan Li, Hao Chen, and Hanyu Gao. Towards large-scale chemical reaction image parsing via a multimodal large language model. *Chemical Science*, 16(45):21464–21474, 2025b.

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

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418, 2024.

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

Jiang Guo, A Santiago Ibanez-Lopez, Hanyu Gao, Victor Quach, Connor W Coley, Klavs F Jensen, and Regina Barzilay. Automated chemical reaction extraction from scientific literature. *Journal of chemical information and modeling*, 62(9):2035–2045, 2022.

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

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169, 2024.

Madhavi Kumari and Prabha Garg. Enhanced chemical reaction mining through combined model adaptation and prompt engineering. *Journal of Chemical Information and Modeling*, 65(19): 10348–10363, 2025.

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

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

Jiawei Li, Minzhou Li, Qi Yang, and Sanzhong Luo. Reactionseek: Llm-powered literature data mining and knowledge discovery in organic synthesis. *Nature Communications*, 17(1):3356, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

AI Meta. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 2025.

Alexander Novikov et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Maciej P Polak and Dane Morgan. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569, 2024.

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

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Sarveswara Rao Vangala, Sowmya Ramaswamy Krishnan, Navneet Bung, Dhandapani Nandagopal, Gomathi Ramasamy, Satyam Kumar, Sridharan Sankaran, Rajgopal Srinivasan, and Arijit Roy. Suitability of large language models for extraction of high-quality chemical reaction dataset from patent literature. *Journal of Cheminformatics*, 16(1):131, 2024.

Damian M Wilary and Jacqueline M Cole. Reactiondataextractor 2.0: a deep learning approach for data extraction from chemical reaction schemes. *Journal of Chemical Information and Modeling*, 63(19):6053, 2023.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents. *arXiv preprint arXiv:2505.22954*, 2025.

Wei Zhang, Qinggong Wang, Xiangtai Kong, Jiacheng Xiong, Shengkun Ni, Duanhua Cao, Buying Niu, Mingan Chen, Yameng Li, Runze Zhang, et al. Fine-tuning large language models for chemical text mining. *Chemical science*, 15(27):10600–10611, 2024.

Ming Zhong, Siru Ouyang, Minhao Jiang, Vivian Hu, Yizhu Jiao, Xuan Wang, and Jiawei Han. Reactie: Enhancing chemical reaction extraction with weak supervision. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12120–12130, 2023.

A Appendix

A.1 Related Work

Reaction Extraction from Chemical Literature

Reaction extraction from chemical literature has been studied across visual, textual, and document-level settings. Image-based methods such as RxnScribe and ReactionDataExtractor recover machine-readable reactions from graphical reaction schemes (Qian et al., 2023; Wilary & Cole, 2023), while text-oriented approaches such as ChemRxnExtractor and ReactIE extract reaction entities, roles, conditions, and outcomes from chemical descriptions using task-specific trained models (Guo et al., 2022; Zhong et al., 2023). More recent work adapts LLMs to chemical text mining through fine-tuning or model adaptation (Zhang et al., 2024; Kumari & Garg, 2025), and document-level systems such as OpenChemIE, MERMaid, and ReactionSeek further integrate information across text, tables, and reaction figures (Fan et al., 2024; Leong et al., 2025; Li et al., 2026).

Despite this progress, complete reaction reconstruction from raw literature remains challenging because information for a single experiment may be distributed across main-text figures, Supporting Information, general procedures, local compound identifiers, and cross-referenced document regions. Existing systems primarily focus on extracting and standardizing information from individual modalities or predefined reaction fields. In contrast, AutoChem jointly resolves textual and visual evidence across document regions and reconstructs coherent reaction records without task-specific training, fine-tuning, or manually pre-segmented reaction procedures.

Chemical Reaction Resources and Benchmarks

Literature-mining systems increasingly transform extracted chemical information into structured resources for downstream AI applications. ReactionSeek, for example, standardizes multimodal reaction information from the Organic Syntheses collection into machine-readable reaction records and uses the resulting dataset for predictive modeling, natural-language retrieval, and large-scale trend analysis (Li et al., 2026; Bran et al., 2024). Such standardized records provide a useful representation of individual experiments, but do not explicitly organize the relationships among reactions, molecular participants, structural variations, conditions, outcomes, and literature provenance.

AutoChem extends literature-derived reaction data in two directions. First, it organizes reconstructed experiments into a *reaction-centric heterogeneous knowledge graph*, where reactions serve as explicit nodes connecting molecular and experimental information. Second, rather than using extracted reactions only as a reference dataset, it reorganizes them into controlled-variable optimization benchmarks. Condition optimization (Q1) compares alternative reaction conditions for comparable transformations, while substrate optimization (Q2) compares structurally related substrates under controlled settings. This allows the same literature-derived reaction records to support both structured knowledge organization and systematic evaluation of reaction optimization reasoning.

A.2 Additional Method Details

A.2.1 Reaction Schema and Formal Definition

Given a collection of chemical papers and their Supporting Information, we denote the input document collection as

$$\mathcal{X} = \{x_1, x_2, \ldots, x_n\},$$

where each document may contain experimental descriptions, general procedures, reaction schemes, molecular structures, tables, and cross-references between different document regions.

Rather than representing a reaction only as a substrate–product pair, we model each experimentally reported reaction as

$$r_i = (S_i, P_i, I_i, A_i, \Gamma_i, H_i, O_i, M_i),$$

where S_i, P_i, and I_i denote substrates, products, and intermediates, respectively. A_i denotes auxiliary reaction components such as catalysts, ligands, and reagents; Γ_i represents reaction conditions; H_i contains experimental procedures and step-level information; O_i contains quantitative outcomes such as yield and enantiomeric excess (ee); and M_i stores document metadata and literature provenance.

The literature-processing stage produces a collection of mapped and standardized reaction records,

$$\mathcal{R} = \{r_i\}_{i=1}^{N},$$

which we refer to as *mapped index reactions*. The two extraction pathways initially produce

$$\mathcal{R}^{\text{text}} = \{r_i^{\text{text}}\} \qquad \text{and} \qquad \mathcal{R}^{\text{visual}} = \{r_i^{\text{visual}}\},$$

which are subsequently consolidated through cross-modal entity alignment.

Based on the reconstructed reaction collection $\mathcal{R}$, AutoChem constructs a reaction-centric heterogeneous knowledge graph

$$G = (V, E, A),$$

and two families of controlled-variable reaction-optimization tasks,

$$\mathcal{Q} = \mathcal{Q}_{\text{cond}} \cup \mathcal{Q}_{\text{sub}},$$

where $\mathcal{Q}_{\text{cond}}$ corresponds to Q1 condition optimization and $\mathcal{Q}_{\text{sub}}$ corresponds to Q2 substrate optimization. The complete processing flow is therefore

$$\mathcal{X} \to \left\{\mathcal{R}^{\text{text}}, \mathcal{R}^{\text{visual}}\right\} \to \mathcal{R} \to \{G, \mathcal{Q}\}.$$

A.2.2 ChemText Agentic Module Details

Chemical SI documents rarely describe all information associated with a reaction in a single contiguous span. Compound identities may be introduced in one section, general experimental procedures defined elsewhere, and individual reaction entries may refer back to these procedures through abbreviated identifiers. The ChemText Agentic Module therefore coordinates multiple components to recover reaction information distributed across the document.

SI Scout. The **SI Scout** identifies pages and document regions likely to contain reaction-relevant information, including compound definitions, experimental descriptions, product entries, and general procedures. Restricting downstream processing to relevant regions reduces unnecessary document processing while retaining the context required for complete reaction reconstruction.

Entity Mapper and Chem Registry. The **Entity Mapper** identifies substrates and products and resolves different textual expressions referring to the same chemical entity. A common pattern in SI documents is that a compound is initially introduced using its full chemical name and later referenced only by a document-local identifier such as "1a" or "21". The Entity Mapper therefore performs substrate/product identification, code–name mapping, and definition linking.

The resulting mappings are stored in a document-level **Chem Registry**,

$$\mathcal{D}_{\text{chem}} = \{c_j \mapsto e_j\}_{j=1}^{N_c},$$

where c_j denotes a document-local compound identifier and e_j denotes the corresponding normalized chemical entity. This registry provides a consistent reference system for resolving chemical identities across different document regions.

Procedure Miner and Procedure Canonicalizer. The **Procedure Miner** identifies general procedures (GPs), detects their boundaries and identifiers, parses reusable procedural templates, and links individual reaction entries to the corresponding procedures. This linking step is important because reaction-specific entries frequently omit experimental details that have already been specified in a shared general procedure.

Because equivalent procedures may occur under different textual forms, the **Procedure Canonicalizer** normalizes the recovered descriptions into a consistent procedure registry,

$$\mathcal{D}_{\text{GP}} = \{g_k \mapsto p_k\}_{k=1}^{N_g},$$

where g_k denotes a canonical procedure identifier and p_k denotes its normalized experimental description.

Mapped index reactions and Reaction Guard. Information from the Chem Registry, canonicalized general procedures, and reaction-specific descriptions is combined into textual mapped index reactions,

$$\mathcal{R}^{\text{text}} = \{r_i^{\text{text}}\}.$$

Each record contains the reaction participants, linked procedural information, reaction steps, experimental conditions, quantities, and reported outcomes recoverable from textual evidence.

The resulting records are subsequently examined by the **Reaction Guard**, which checks reaction coverage, consistency between chemical entities and procedural information, and duplicated reaction entries. Validated textual records are then passed to the cross-modal alignment stage.

A.2.3 ChemVision Agentic Module Details

Reaction schemes frequently contain structural and experimental information that cannot be recovered reliably from text alone. Graphical representations also vary across publications: some define explicit R-groups, some contain multiple substrate or product variants, and others directly report reaction conditions or experimental outcomes. AutoChem therefore employs a dynamically routed ChemVision Agentic Module rather than a single fixed visual extraction pipeline.

Tools, Orchestrator, and Agent Router. The ChemVision Agentic Module has access to multimodal extraction tools, chemical knowledge bases, and reaction-understanding tools. Given a reaction image, the **Orchestrator** first interprets its content and determines which extraction operations are required. The **Agent Router** then dynamically invokes specialized agents according to the image type, extracted evidence, and current reconstruction requirements.

Template Decoder. The **Template Decoder** parses generic reaction templates and explicit R-group definitions. It identifies shared molecular scaffolds, variable positions, and the correspondence between generic structures and individual reaction instances.

Molecule Resolver. The **Molecule Resolver** recognizes molecular structures appearing in reaction images and associates them with their corresponding reaction roles. When explicit substituent definitions are available, the resolved molecular structures are linked to the corresponding variable assignments.

Condition Interpreter. The **Condition Interpreter** recovers experimental conditions and quantitative outcomes encoded in reaction figures, including reagents, solvents, temperatures, reaction times, yields, and stereochemical outcomes when available.

R-Group Resolver. The **R-Group Resolver** reconstructs complete molecular variants from shared molecular scaffolds and explicit or implicit substituent definitions. This step is particularly important when individual substrates or products are represented through R-group substitutions rather than independently redrawn molecular structures.

Together, the specialized visual agents generate

$$\mathcal{R}^{\text{visual}} = \{r_i^{\text{visual}}\},$$

which provide complementary structural and experimental information to the textual mapped index reactions.

A.2.4 Cross-modal Entity Alignment Details

Textual and visual pathways provide complementary but partially overlapping descriptions of the same experiment. For example, a compound may appear only as a local identifier such as "21" in the SI text while its complete molecular structure is shown in a reaction image. Treating these observations independently would fragment a single chemical entity across modalities.

AutoChem therefore performs **Cross-modal Entity Alignment** to link textual compound references with corresponding molecular structures in reaction figures. Let e_i^{text} and e_j^{visual} denote textual and visual observations of chemical entities, respectively. Alignment considers document-local identifiers, molecular information, reaction context, and surrounding experimental evidence to determine whether two observations correspond to the same entity. A matched pair is consolidated as

$$e_k = \text{Align}\left(e_i^{\text{text}}, e_j^{\text{visual}}\right).$$

After entity alignment, complementary textual and visual reaction evidence is combined into a mapped index reaction,

$$r_i = \text{Fuse}\left(r_i^{\text{text}}, r_i^{\text{visual}}\right).$$

The resulting collection

$$\mathcal{R} = \{r_i\}_{i=1}^{N}$$

contains standardized substrates, products, intermediates, auxiliary components, experimental procedures, step-level information, reaction conditions, quantitative outcomes, and literature provenance.

Each mapped index reaction is serialized using a structured JSON schema. The JSON object serves as a machine-readable serialization of the reaction-level record and provides a common interface for the downstream Knowledge Graph Agent and Benchmark Generator Agent.

A.2.5 Reaction-Centric Knowledge Graph Schema

Given the mapped index reaction collection $\mathcal{R}$, the **Knowledge Graph Agent** first filters valid reaction records and then extracts structured graph relations to construct a reaction-centric heterogeneous knowledge graph,

$$G = (V, E, A),$$

where V denotes the node set, E denotes the typed edge set, and A contains node-level and reaction-level attributes.

The node set is defined as

$$V = V_R \cup V_S \cup V_P \cup V_C \cup V_{\text{Cond}} \cup V_{\text{Scaf}} \cup V_{\text{Sub}} \cup V_{\text{Paper}},$$

where the subsets correspond to Reaction, Substrate, Product, Catalyst, Condition, Scaffold, Substituent, and Paper nodes, respectively.

Let $\mathcal{R}_{\text{rel}}$ denote the set of relation types. The edge set satisfies

$$E \subseteq V \times \mathcal{R}_{\text{rel}} \times V.$$

Reaction-centric modeling. Rather than representing a reaction as a single edge from substrate to product, the Knowledge Graph Agent instantiates each experimentally reported reaction as an explicit **Reaction** node. Molecular participants, experimental conditions, structural information, and literature provenance are connected to this reaction node. This organization preserves the context of each individual experiment while allowing related reactions to be compared directly.

Quantitative experimental outcomes are represented as reaction-level attributes,

$$A(r) = \{\text{yield}, \text{ee}, \text{dr}, \text{temperature}, \text{time}, \ldots\}.$$

In particular, outcomes such as yield and enantiomeric excess are attached to the corresponding reaction instance rather than represented as generic relations disconnected from experimental context.

Scaffold–substituent decomposition. To explicitly represent structural variation across substrate families, each substrate is further decomposed into a molecular scaffold and its substituents,

$$S \rightarrow \big(\text{Scaffold}(S), \text{Substituent}(S)\big).$$

This decomposition makes reactions sharing a common molecular core but differing in substitution patterns directly comparable. It also provides the structural basis for constructing Q2 substrate-optimization tasks.

A.2.6 Interactive Knowledge Graph Visualization

To qualitatively inspect the reaction-centric knowledge graph constructed by AutoChem, we provide an interactive visualization interface that exposes the resulting graph at complementary levels of granularity. Figure 5 presents three representative views.

At the graph level, Overview mode jointly visualizes heterogeneous node and relation types and supports search and filtering by node type, relation type, and source modality (Figure 5, panel (a)). The example graph contains Reaction, Substrate, Product, Catalyst, Ligand, Solvent, Condition, and Paper nodes, providing a global view of how experimental reactions are connected to their chemical and literature context.

The interface also supports provenance-centered inspection. Selecting a Paper node highlights the reactions reported in that source and exposes document-level metadata and source modality (Figure 5, panel (b)). Typed provenance relations such as REPORTED_IN therefore allow individual reconstructed reactions to be traced back to their source literature.

At the reaction level, Focus mode organizes reactions by source paper and displays the local neighborhood of a selected Reaction node (Figure 5, panel (c)). The selected reaction is connected through typed relations to its substrates, products, solvents, experimental conditions, and source paper. The

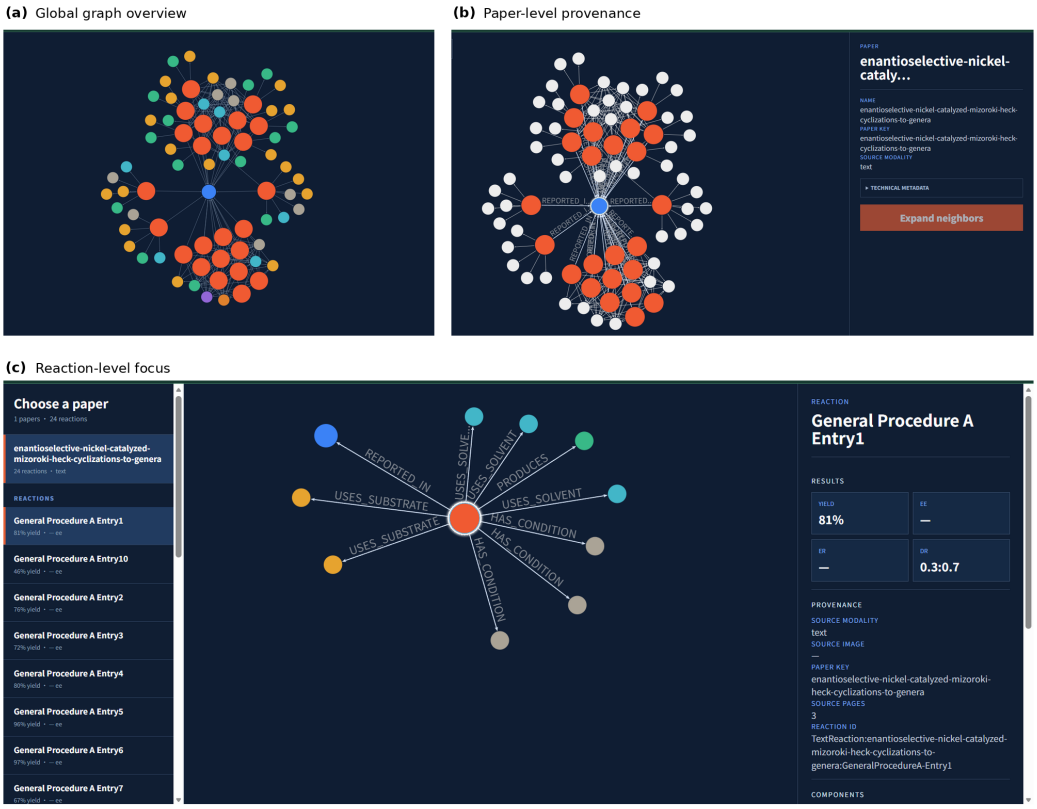


Figure 5: Representative screenshots of the interactive reaction-centric knowledge graph constructed by AutoChem. (a) Graph-wide Overview mode visualizes the heterogeneous structure of the reconstructed knowledge graph. (b) Paper-level inspection exposes document metadata and provenance relations linking reconstructed reactions to their literature source. (c) Reaction-level Focus mode displays the local neighborhood of an individual Reaction node together with typed relations to substrates, products, solvents, conditions, and the source paper; the accompanying evidence panel additionally reports experimental outcomes and provenance metadata.

accompanying evidence panel further exposes reaction-level outcomes and provenance metadata, including reported yield and stereochemical outcomes, source modality, source pages, paper identifier, and reaction identifier.

Together, these views illustrate the reaction-centric representation described in Appendix A.2.5, in which each experimentally reported reaction serves as an explicit hub connecting molecular participants, experimental context, quantitative outcomes, and literature provenance.

A.2.7 Optimization Benchmark Construction Details

The reconstructed reaction collection contains experiments reported under heterogeneous substrates, catalysts, and reaction conditions. Directly comparing arbitrary reactions would therefore introduce multiple uncontrolled variables. The **Benchmark Generator Agent** first reorganizes mapped index reactions into controlled-variable comparison groups and then constructs Q1 and Q2 benchmark instances.

Q1 Generator Agent: Condition Optimization. The **Q1 Generator Agent** constructs condition-optimization tasks in two stages. A **Condition Extractor** identifies reactions representing comparable transformations and extracts their experimental conditions. A subsequent **Scoring and Ranking** stage compares the experimentally reported outcomes of candidate conditions.

For condition optimization, reactions are grouped by the same underlying transformation while experimental conditions are allowed to vary. The condition-comparison key is

$$\kappa_{\text{cond}}(r) = \big(\text{Type}(r), S(r), P(r)\big),$$

where $\text{Type}(r)$ denotes reaction type and $S(r)$ and $P(r)$ denote the substrate and product configurations.

For a comparison key k, the corresponding reaction group is

$$\mathcal{R}_{\text{cond}}(k) = \{r \in \mathcal{R} \mid \kappa_{\text{cond}}(r) = k\}.$$

Reactions in the same group therefore represent comparable transformations while differing primarily in experimental conditions. These groups are converted into Q1 instances whose candidate conditions are ranked according to the scoring rule defined below.

Q2 Generator Agent: Substrate Optimization. The **Q2 Generator Agent** constructs substrate-optimization tasks. A **Key-substrate Identification** stage identifies structurally related substrates and their scaffold–substituent relationships, after which **Scoring and Ranking** compares their experimentally reported performance under controlled reaction settings.

Using the scaffold–substituent decomposition introduced above, the substrate-comparison key is

$$\kappa_{\text{sub}}(r) = \big(\text{Type}(r), \text{Scaffold}(S(r)), \text{Scaffold}(P(r)), \text{Cond}(r)\big).$$

The corresponding comparison group is

$$\mathcal{R}_{\text{sub}}(k) = \{r \in \mathcal{R} \mid \kappa_{\text{sub}}(r) = k\}.$$

Reactions in the same group share a comparable transformation context and experimental conditions while differing in their substrate substituents. These groups are converted into Q2 instances comparing structurally related substrate variants.

Scoring and ranking. Both Q1 and Q2 use experimentally reported outcomes to derive benchmark supervision. For reactions reporting both yield and enantiomeric excess (ee), we denote the outcome vector as

$$\mathbf{o}(r) = \big(\text{yield}(r), \text{ee}(r)\big).$$

Because yield and ee are both reported on the same percentage scale, we combine them using an equal-weight arithmetic score:

$$s(r) = 0.5 \cdot \text{yield}(r) + 0.5 \cdot \text{ee}(r).$$

Thus, neither yield nor ee is treated as lexicographically dominant. When the two outcomes favor different candidates, their ordering is determined by the composite score $s(r)$. For example, a reaction with 97% yield and 84% ee receives

$$s(r) = 0.5 \times 97 + 0.5 \times 84 = 90.5,$$

whereas a reaction with 93% yield and 90% ee receives

$$s(r) = 0.5 \times 93 + 0.5 \times 90 = 91.5,$$

and is therefore ranked higher.

For reactions for which ee is not reported, ranking is based on yield alone:

$$s(r) = \text{yield}(r).$$

Within each benchmark instance q, all candidates are ranked in descending order of $s(r)$. Let $\mathcal{R}_q$ denote the candidate reaction set for question q. The ground-truth ordering is therefore defined as

$$\pi_q = \text{Rank}_{\downarrow}\left(\{s(r) \mid r \in \mathcal{R}_q\}\right),$$

where $\text{Rank}_{\downarrow}$ denotes ranking from the highest to the lowest score. Candidates with identical scores are assigned the same rank and are treated as tied gold candidates when they occupy the top rank. The same scoring and tie-handling rules are applied to both Q1 and Q2.

The final benchmark is

$$\mathcal{Q} = \mathcal{Q}_{\text{cond}} \cup \mathcal{Q}_{\text{sub}},$$

where $\mathcal{Q}_{\text{cond}}$ contains Q1 condition-optimization questions and $\mathcal{Q}_{\text{sub}}$ contains Q2 substrate-optimization questions. Through this controlled-variable reorganization and deterministic scoring procedure, heterogeneous experimental records are converted into standardized evaluation units while preserving experimentally reported outcomes as ground-truth supervision.

A.3 Detailed Experimental Settings

Reaction-extraction evaluation. The extraction corpus consists of 22 chemistry papers and their associated Supporting Information. To evaluate robustness across heterogeneous synthetic literature, the corpus covers multiple reaction families, including photocatalytic transformations, Suzuki coupling, Buchwald–Hartwig amination, photoredox Minisci alkylation, Ni/photoredox dual catalysis, Sonogashira coupling, and radical decarboxylative C–C coupling. These reaction families differ substantially in catalyst systems, reagents, reaction conditions, substrate classes, and reporting conventions, providing a heterogeneous test bed for full-document reaction reconstruction.

Reference annotations cover 14 reaction-information fields: substrate, product, intermediate, catalyst, ligand, solvent, other components, reaction conditions, yield, enantiomeric excess (ee), substrate amount, catalyst amount, ligand amount, and solvent amount. All extraction systems are evaluated against the same reference annotations and field definitions.

A.3.1 Expert Annotation and Field-level Scoring Protocol

Gold-standard initialization and review interface. To provide a consistent interface for expert verification, the initial reaction records were deterministically converted from the unified knowledge-graph (KG) edge table into one spreadsheet row per `reaction_id`. The paper identifier was taken from `pdf_name`, and the local suffix of `reaction_id` was combined with the sorted `source_pages` to form the reaction-location field. Edges labeled `USES_SUBSTRATE`, `PRODUCES`, `USES_CATALYST`, `USES_LIGAND`, and `USES_SOLVENT` populated their corresponding entity columns. `USES_INTERMEDIATE` and `PRODUCES_INTERMEDIATE` populated the intermediate column, with produced intermediates preferred when both forms were present. `USES_OTHER_COMPONENT` and the legacy `USES_ADDITIVE`/`USES_REAGENT` relations were folded into the other-components column. Non-solvent facts from `HAS_CONDITION` were consolidated into a single reaction-condition column, while legacy solvent facts were assigned to the solvent column to avoid duplication.

Entity amounts stored on KG edges were rendered as `name (amount)`. Step indices were retained using `step1:`, `step2:`, and analogous prefixes, and duplicate values within a reaction were removed while preserving source order. Yield, dr, ee, and er were taken from the first nonempty outcome value associated with the grouped reaction. Scaffold, substituent, name-resolution, and provenance-only edges were not expanded into review columns. Cross-modal aggregate edges marked `text+image` were skipped because the underlying text and image records were already represented separately.

Table 4: Schema of the 19-column expert-review spreadsheet, grouped by function.

Spreadsheet group and fields	Review representation	Role in evaluation
Identity and provenance: Paper; reaction location	`pdf_name`; local reaction identifier plus sorted PDF pages	Used for document grouping and reaction-row alignment; not scored as chemical fields.
Chemical entities: Substrate; product; intermediate; catalyst; ligand; solvent; other components	Relation-specific entity names with optional amounts and retained reaction step labels	Seven entity fields are scored. Substrate, catalyst, ligand, and solvent amounts are parsed and scored separately.
Conditions: Reaction condition	Temperature, time, atmosphere, light source, wavelength, and other non-solvent condition facts consolidated by reaction step	Scored as one canonical condition bundle per reaction.
Reported outcomes: Yield; dr; ee; er; E:Z; rr	KG outcome attributes, supplemented or corrected during expert review when supported by the source literature	Yield and ee are scored scalar fields. dr, er, E:Z, and rr are retained as auxiliary outcomes and provenance.
Expert-review metadata: Direct extraction error; procedure/order error; source-text note	Blank in the machine export and completed during source verification	Used for annotation auditing and error analysis.

The resulting reaction-summary CSV served only as an initialization and review interface rather than defining the final gold reaction set. During annotation, the chemistry-domain expert independently re-read the original article and Supporting Information, verified the machine-initialized rows against the source literature, corrected incomplete or incorrect entries, and added reaction rows that were missing from the initial export. The final gold standard was therefore determined from the source literature rather than from the machine-generated spreadsheet. The review schema is summarized in Table 4.

Expert annotation rubric. The final gold standard contains 1,163 reaction rows from 22 papers. A chemistry-domain expert checked every reaction against the original article and Supporting Information, including general procedures, reaction schemes, tables, and experiment-specific descriptions. The expert verified both the machine-initialized reaction rows and the source documents themselves, allowing reactions omitted from the initial export to be added to the final gold set.

A correct cell was left unchanged. For an incorrect or incomplete cell, the expert appended a structured revision marker containing the corrected value in the same cell and recorded the corresponding error type in the audit columns. When multiple revision markers occurred, the final revision was selected during gold normalization.

Explicit statements that a field was not reported, was not applicable, or had no separately added component were normalized to an empty gold value rather than retained as literal chemical entities. The direct-extraction-error column records omissions, incorrect entities or values, role assignments, and general-procedure mismatches. The procedure/order-error column records missing or incorrect addition, premixing, degassing, transfer, irradiation, and work-up operations. The source-text-note column is reserved for inconsistencies in the paper or SI itself, such as incompatible names, units, or quantities. The annotation protocol relied on expert source verification rather than independent duplicate annotation; accordingly, inter-annotator agreement is not reported.

Reaction alignment. Prediction and gold rows were first aligned within the same paper using reaction-location and PDF-page evidence together with bidirectionally unique product, substrate, and outcome fingerprints. A nonunique candidate was not force-matched. Unmatched predicted rows contributed false-positive units for their nonempty fields, whereas unmatched gold rows contributed false-negative units. Field-level scoring was then applied to aligned reaction pairs according to Table 5.

Metric aggregation. The 14 primary scoring fields are substrate, product, intermediate, catalyst, ligand, solvent, other components, reaction conditions, yield, ee, substrate amount, catalyst amount, ligand amount, and solvent amount. Counts are micro-aggregated across evaluation units as

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN}, \qquad F_1 = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision} + \text{Recall}}.$$

Table 5: Field-level matching and error-counting rules applied after reaction-row alignment.

Evaluation unit	Match / TP rule	FP/FN rule
Chemical entity	A predicted entity is uniquely equivalent to one gold entity.	Each unmatched predicted entity contributes one FP; each unmatched gold entity contributes one FN.
Entity amount	The parent entities match and their normalized quantity tuples are equal.	Different nonempty quantities contribute one FP and one FN; a one-sided quantity contributes the corresponding FP or FN.
Yield or ee	Both nonempty scalar values are equal after deterministic normalization.	Different nonempty values contribute one FP and one FN; a one-sided value contributes the corresponding FP or FN.
Reaction condition	The canonical condition bundles match after deterministic normalization.	A non-matching pair contributes one FP and one FN; prediction-only and gold-only bundles contribute an FP and FN, respectively.
Procedure/order audit	The condition is correct and the expert procedure/order-error field is empty.	A condition mismatch, a nonempty procedure/order-error field, or both contribute one FP and one FN in the corresponding audit.

For configurations with repeated runs, the main tables report the equal-weight mean of run-level micro-averaged scores.

Reaction-optimization benchmark. The benchmark contains 344 questions with 1,018 candidate options, including 179 condition-optimization questions (Q1) and 165 substrate-optimization questions (Q2). Q1 compares alternative reaction conditions while controlling the underlying transformation, whereas Q2 compares structurally related substrate variants under controlled reaction settings. Among the 344 questions, 230 contain both yield and ee information and 114 use yield only. For questions containing both outcomes, candidates are ranked using the equal-weight score $0.5 \times \mathrm{yield} + 0.5 \times \mathrm{ee}$; for yield-only questions, candidates are ranked directly by yield. Candidates receiving the same score are treated as tied gold candidates. A total of 312 questions are constructed from text-derived reaction records and 32 from image-derived records.

Benchmark model configurations. We evaluate seven LLM configurations: GPT-5-mini with minimal reasoning effort, GPT-5-mini with high reasoning effort, GPT-4o, GPT-5, DeepSeek-V4-Flash, Gemini-3.8-Flash-Nothinking, and Llama-4-Maverick-17B-128E-Instruct. All configurations are evaluated on the same 344 benchmark questions with identical candidate options, prompt version, option ordering, and gold definitions.

GPT-5-mini (minimal) and GPT-5 use minimal reasoning effort, whereas GPT-5-mini (high) uses high reasoning effort. GPT-4o is evaluated without an explicit reasoning-effort parameter. DeepSeek-V4-Flash is evaluated with thinking disabled, while Gemini-3.8-Flash-Nothinking and Llama-4-Maverick-17B-128E-Instruct are evaluated using their corresponding provider configurations without an OpenAI-style reasoning-effort parameter. The models observe only the benchmark questions and candidate options and do not have access to source reaction IDs, gold labels, experimental outcomes, or the underlying knowledge graph.

Benchmark metrics and random baseline. Top-1 is counted as correct when the highest-ranked candidate belongs to the gold set. Top-3 measures whether at least one gold candidate occurs among the first three predictions and is evaluated only on the 98 questions containing more than two candidates. MRR is computed using the reciprocal rank of the first gold candidate.

Because candidate-set sizes and the number of valid gold answers vary across questions, we compute a question-specific uniform-random baseline rather than assuming a fixed random accuracy. Aggregated over the complete benchmark, this procedure yields random baselines of 46.4% for Top-1, 73.9% for Top-3, and 0.699 for MRR. The relatively high Top-1 baseline primarily reflects the fact that 246 of the 344 benchmark questions contain only two candidate options.

Statistical testing. For extraction ablations, we compute paired document-level differences between each ablated configuration and the Full system and apply a two-sided Wilcoxon signed-rank test. Since 14 fields are tested for each ablation, we apply Benjamini–Hochberg correction and use $q < 0.05$ as the significance criterion.

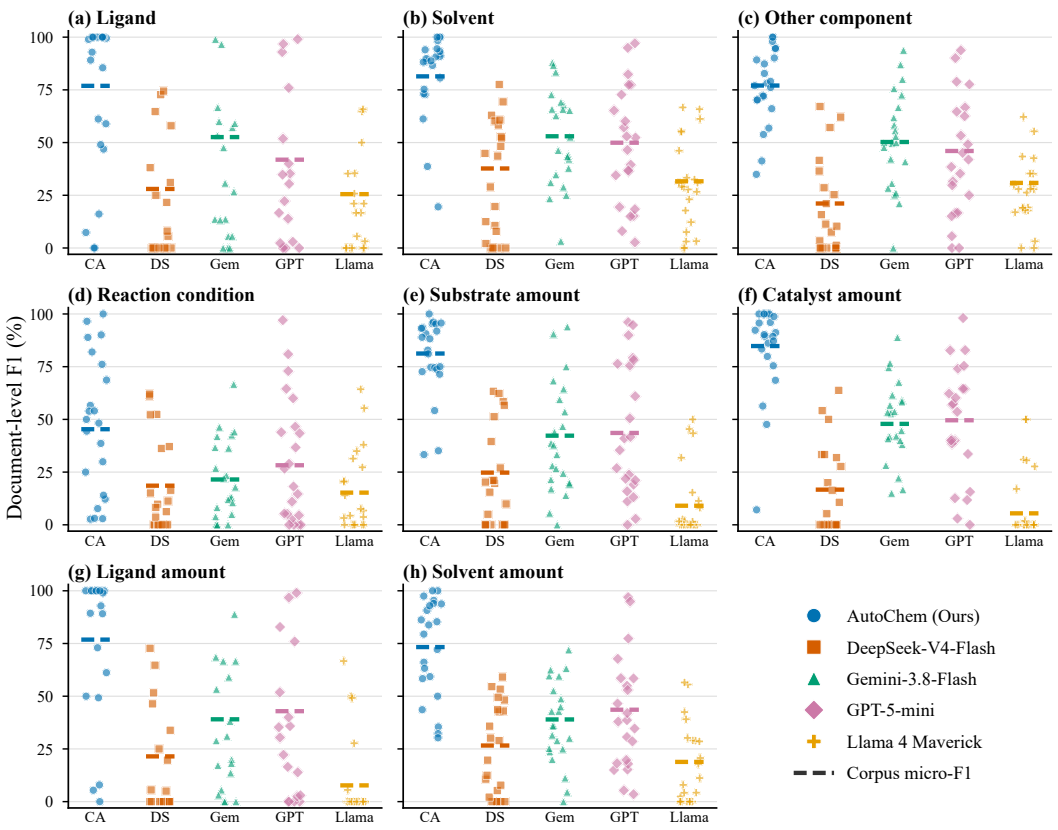


Figure 6: Additional field-level reaction-extraction performance across the 22-document corpus. Panels show ligand, solvent, other component, reaction condition, substrate amount, catalyst amount, ligand amount, and solvent amount. Markers denote document-level F1 scores for individual papers, while horizontal dashed lines indicate corpus-level micro-F1 for each model.

For benchmark evaluation, deviation from the question-specific random Top-1 baseline is assessed separately for each of the seven model configurations using a two-sided Poisson–binomial test. The resulting seven p-values are adjusted using Holm correction.

Differences in Top-1 accuracy among the seven model configurations are assessed using Cochran's Q test over the fully paired set of 344 questions. All 21 pairwise model comparisons are subsequently evaluated using exact McNemar tests with Holm correction across the complete comparison family. All hypothesis tests are two-sided.

A.3.2 Additional Field-Level Extraction Results

Figure 6 reports field-level extraction performance for the remaining eight reaction-information fields. These fields cover additional experimental components, reaction conditions, and component quantities and complement the representative results reported in Figure 4 of the main paper.

The additional fields exhibit the same broad pattern observed for the representative fields in the main paper. AutoChem generally achieves higher corpus-level F1 than direct LLM extraction across experimental components and component quantities, although substantial document-level variation remains. In particular, ligand, solvent, and other-component extraction benefit from reconstructing information that may be inherited from shared experimental procedures rather than repeated in each reaction-specific description.

Reaction conditions remain substantially more challenging than most participant and quantity fields. Their extraction frequently requires integrating heterogeneous information such as temperature, time, atmosphere, concentration, and other procedural details distributed across reaction-specific

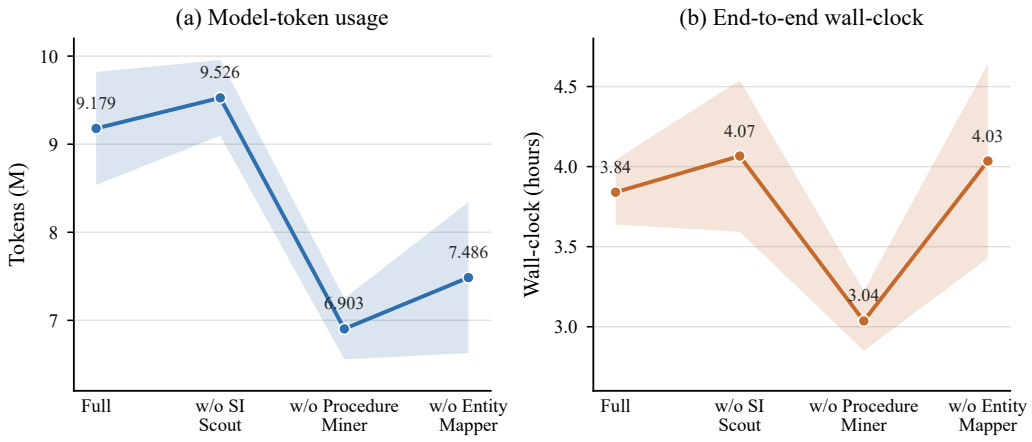


Figure 7: Descriptive model-token usage and end-to-end wall-clock time for the 22-document corpus. Solid lines and markers indicate means, and shaded bands indicate $\pm$ one sample SD across available resource results.

descriptions and general procedures. The four component-quantity fields likewise exhibit noticeable cross-document variation, reflecting differences in whether quantities are stated locally or inherited from shared procedural context. These results are consistent with the ablation analysis in Section 3.4, where removing the Procedure Miner produces particularly large degradations for experimental components and their associated quantities.

A.4 Efficiency Analysis

We further analyze the computational cost of the three ablated components. Figure 7 reports model-token consumption and end-to-end wall-clock time over the 22-document extraction corpus.

The SI Scout improves computational efficiency while largely preserving extraction accuracy. As shown in Figure 7, removing the SI Scout increases model-token consumption from 9.179M to 9.526M and wall-clock time from 3.84 h to 4.07 h. This corresponds to increases of approximately 3.8% in token usage and 6.0% in runtime. Together with the small change in overall micro-F1 from 79.56 to 79.38, these results indicate that the SI Scout primarily reduces unnecessary document processing without materially affecting extraction quality.

The Procedure Miner incurs substantial computational cost but provides the largest accuracy benefit. Removing the Procedure Miner reduces token consumption from 9.179M to 6.903M and wall-clock time from 3.84 h to 3.04 h, corresponding to reductions of approximately 24.8% and 20.8%, respectively. However, overall micro-F1 decreases sharply from 79.56 to 59.15. Thus, although procedure mining requires additional computation, this cost is accompanied by a substantial improvement in recovering experimental information distributed through general procedures and cross-referenced reaction descriptions.

The Entity Mapper primarily supports chemical-identity resolution rather than runtime reduction. Removing the Entity Mapper lowers token consumption from 9.179M to 7.486M, a reduction of approximately 18.4%, but slightly increases wall-clock time from 3.84 h to 4.03 h. Together with the decrease in overall micro-F1 from 79.56 to 76.70, this pattern indicates that the Entity Mapper contributes to reliable chemical-entity resolution while its computational effect is not simply a reduction in end-to-end runtime.

Overall, the three components exhibit complementary accuracy–efficiency trade-offs: the SI Scout reduces both token usage and wall-clock time while largely preserving extraction accuracy, the Procedure Miner trades additional computation for substantial extraction gains, and the Entity Mapper primarily supports consistent chemical-identity resolution across distributed document regions.