Knowledge Graph Extraction from Total Synthesis Documents

Andres M Bran[1,2] **, Zlatko Joncev**[1] **,**
Philippe Schwaller[1,2]**,**
[1] Laboratory of Artificial Chemical Intelligence (LIAC), Lausanne, Switzerland,
[2] National Centre of Competence in Research (NCCR) Catalysis, Lausanne, Switzerland

Correspondence: andres.marulandabran@epfl.ch

Abstract

Knowledge graphs (KGs) have emerged as a powerful tool for organizing and integrating complex information, making it a suitable format for scientific knowledge. However, translating scientific knowledge into KGs is challenging as a wide variety of styles and elements to present data and ideas is used. Although efforts for KG extraction (KGE) from scientific documents exist, evaluation remains challenging and field-dependent; and existing benchmarks do not focuse on scientific information. Furthermore, establishing a general benchmark for this task is challenging as not all scientific knowledge has a ground-truth KG representation, making any benchmark prone to ambiguity. Here we propose Graph of Organic Synthesis Benchmark (GOSyBench), a benchmark for KG extraction from scientific documents in chemistry, that leverages the native KG-like structure of synthetic routes in organic chemistry. We develop KG-extraction algorithms based on LLMs (GPT-4, Claude, Mistral) and VLMs (GPT-4o), the best of which reaches 73% recovery accuracy and 59% precision, leaving a lot of room for improvement. We expect GOSyBench can serve as a valuable resource for evaluating and advancing KGE methods in the scientific domain, ultimately facilitating better organization, integration, and discovery of scientific knowledge.

Knowledge graphs (KGs) have emerged as a powerful tool for representing and organizing complex information, enabling efficient storage, retrieval, and analysis of data across various domains (Hogan et al., 2021). The extraction of knowledge graphs from unstructured data sources, such as text documents, has gained significant attention in recent years due to its potential to unlock valuable insights and facilitate knowledge discovery. KGs have also recently been used in Retrieval-Augmented Generation (RAG) pipelines (Abu-Rasheed et al., 2024), as a strategy to ground text generation from large language models (LLMs) with domain-specific facts, thus improving performance across tasks (Khattab et al., 2023; Khattab and Zaharia, 2020).

0.1 Extraction of Knowledge Graphs

The field of Knowledge Graph Extraction (KGE) has witnessed substantial progress, with numerous approaches being developed to automatically construct KGs from textual data. These methods range from rule-based systems to machine learning-based techniques, and more recently, LLM-driven extraction (Meyer et al., 2023; Shu et al., 2024). Several benchmarks have been proposed to evaluate the performance of KGE systems, from open-domain ones like Open Graph Benchmark (Hu et al., 2020) and Text2KGBbench (Mihindukulasooriya et al., 2023), to more field specific ones like PharmaKG for biomedical data mining (Zheng et al., 2020). These benchmarks focus on evaluating algorithms on the extraction of specific facts from short sentences or paragraphs, while extraction from complete documents, and specially scientific ones, remains largely untested.

Scientific literature contains a wealth of knowledge that can be represented in KGs, the extraction of which would enable more efficient knowledge integration and facilitate discovery. Excellent efforts have been made to extract specific types of scientific information, such as entities and relations in chemical literature (Lowe and Sayle, 2013; Swain and Cole, 2016; Mavračić et al., 2021). While these advances have enabled the extraction of influential reaction datasets (Lowe, 2012), they are tailored to patents, which have a more standardized format and contain less scientific details as journal papers do. Moreover, these methods focus on extracting single reactions or short sequences, mostly ignoring the underlying network of objects and concepts originally expressed in the texts.

The lack of benchmarks specifically designed for

Proceedings of the 1st Workshop on Language + Molecules (L+M 2024), pages 75–85
August 15, 2024 ©2024 Association for Computational Linguistics

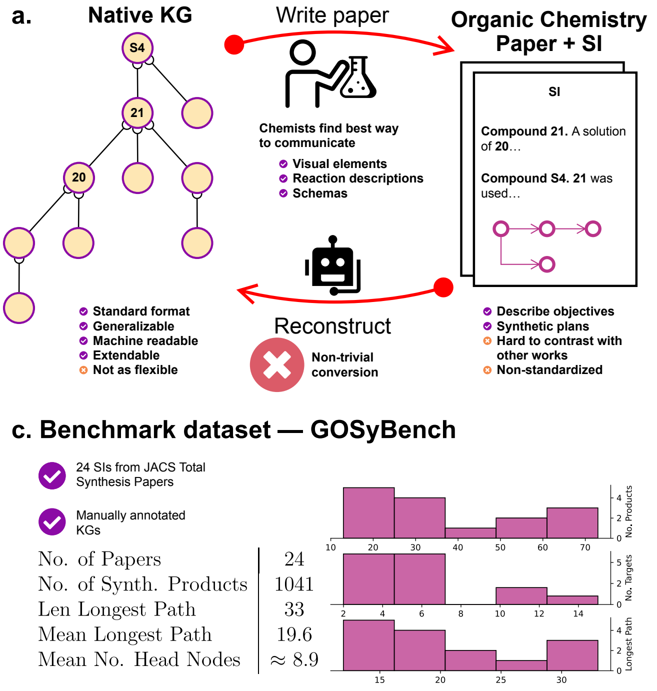

a.
Native KG
S4
21
20
Write paper
Chemists find best way to communicate
Visual elements
Reaction descriptions
Schemas
Organic Chemistry Paper + SI
SI
Compound 21. A solution of 20…
Compound S4. 21 was used…
Standard format
Generalizable
Machine readable
Extendable
Not as flexible
Reconstruct
Non-trivial conversion
Describe objectives
Synthetic plans
Hard to contrast with other works
Non-standardized
b. Evaluation methodology
Ground Truth KG
Extracted KG
$P_S(G)$ — set of all linear paths in G
$S(G,G') = \frac{1}{|G|}\sum_{p\in P_S(G)}\sum_{p'\in P_S(G')} 1_{p=p'}$
$S(G_{Extracted}, G_{GT})$ path recovery accuracy
$S(G_{GT}, G_{Extracted})$ path recovery precision
c. Benchmark dataset — GOSyBench
24 SIs from JACS Total Synthesis Papers
Manually annotated KGs
No. of Papers 24
No. of Synth. Products 1041
Len Longest Path 33
Mean Longest Path 19.6
Mean No. Head Nodes ≈ 8.9
No. Products
No. Targets
Longest Path
d. KG Extraction algorithms
PDF
(optional) select relevant sections
vision-based parsing
GPT-4o
text-based parsing
Isolated reaction procedures
Compound 21. A solution of 20…
Structured data extraction — LLM
{ 'key': '21', 'role': 'product', },
{ 'key': '20', 'role': 'reactant', },
Graph wiring
KG Backbone
Iterative refining — LLM
Systematic and accurate KG representation of paper

Figure 1: **Example Knowledge Graph and evaluation strategy. a.** Shows the data representation used for the task, where each node Si in the directed graph represents abstractly a substance, and each edge V (i → j) expresses that substance Sj is used in a reaction that has substance Si as a product. The goal of the KG is to accurately represent the information presented in the paper. **b.** Evaluation methodology followed in this work. **c.** Summary statistics of the resulting dataset. These highlight aspects critical to graph complexity, like number of substances (nodes), maximum path length, number of head nodes (indegree(Si) = 0), among others. **d.** Algorithm developed for KGE.

evaluating KGE in science poses a challenge, as the diverse nature of scientific knowledge and the absence of ground-truth KGs make it difficult to establish a standardized evaluation framework. The heterogeneity of scientific literature, with its wide range of domains, writing styles, and presentation formats, further complicates the development of a comprehensive benchmark.

0.2 KGs in Organic Chemistry

A knowledge graph is defined generally as a graph of data, intended to convey knowledge. Here, nodes represent entities of interest and edges represent relations between these entities (Hogan et al., 2021). As such, synthetic sequences in Organic Chemistry are susceptible of being represented under such a structure.

Research in synthetic organic chemistry (OC) focuses very generally on the synthesis of organic compounds through a suitable sequence of reactions. Under this conception, substances are *concepts* that are connected through reactions as *relationships*. Each substance may serve as product or reactant for a multitude of different reactions, leading to the natural definition of networks of chemical reactions. This has previously been studied under different models with different levels of depth (Fialkowski et al., 2005). This bare abstraction defines the backbone of a KG, and is this native KG-like structure makes OC an ideal domain for exploring KGE techniques.

But reactions –defined as an experimentally executed transformation that leads from one substance to another– are not the only type of relationships that may exist between substances. In research works in OC, substances are synthesized not only because they will be directly used as building blocks for the synthetic targets, but some are synthesized also to serve as model systems for more complex and valuable structures, some are synthesized but paths need to be abandoned due to unsuccessful reactions, and sometimes even substances are synthesized to facilitate structural elucidation of their precursors. Indeed, many more relationships are built on top of the reaction-graph backbone, that are of interest for organic chemists: these go beyond to inform about strategic aspects of synthesis and multi-level chemistry-driven decision processes.

This work focuses mainly on the extraction of the main backbone from research papers. These are typically given in papers' Supporting Information (SI) files, and contain detailed descriptions of synthetic routes and experimental procedures. These documents exhibit a wide variety of representations, designs, and conventions, making it challenging to extract consistent and comprehensive KGs, see Appendix A for examples. Despite the heterogeneity in the representation of OC knowledge, the underlying structure remains the same: a network of chemical reactions and synthetic plans. This property allows for the definition of a ground-truth KG, making OC a suitable domain for developing and evaluating KGE methods in science.

In this paper, we propose **GOSyBench**, a benchmark for KGE from scientific documents in the domain of organic chemistry. By leveraging the native KG-like structure of synthetic routes, we aim to provide a standardized evaluation framework for assessing the performance of KGE algorithms in extracting scientific knowledge. Our KG ontology defines substances as *entities*, with *reference_key* and *substance_name* as properties, that are connected by reactions as relationships. Furthermore, we develop novel KGE algorithms based on LLMs, and conduct extensive experiments and ablation studies to validate their effectiveness using our proposed benchmark.

1 Methods

1.1 Guidance / structured output generation

Despite their usefulness in various domains, one of the limitations of LLMs is their incapacity to generate consistent and controllable outputs that fit use-case specific guidelines. Recent research has focused in steering LLM generation through the enforcement of grammars in the resulting generations (Rebedea et al., 2023; Khattab et al., 2023). This not only helps steer models towards non-harmful outcomes, but also enables tool usage in agent-like scenarios (Boiko et al., 2023; Bran et al., 2024) and facilitates parsing of the results and integration in existing software (Liu, 2024).

1.2 Benchmark dataset curation

The dataset curation pipeline used involved a combination of automated knowledge extraction and expert human labeling. Initially, 24 Supplementary Information files (SIs) on total synthesis were manually selected from the Journal of the American Chemical Society (JACS), with the format and content of their SI used as a criterion. The SIs were selected such that the obtained sample represents a wide variety of text formatting, varying use of visual elements, order and location of relevant sections, among others, see Appendix A for examples.

The SIs were then processed using the KGE method presented in Section 1.3, resulting in a collection of 24 knowledge graphs, where each contains an approximation to the complete network of chemical reactions expressed in the SI. The process then continued with manual curation, which generally involved node relabeling, node creation/removal, and edge creation/removal. The resulting objects are directed graphs, with individual substances as nodes, and reactions as edges. Some statistics of the dataset are described in Figure 1, which highlights the size and overall complexity of the KGs being extracted.

1.3 KGE method

The Knowledge Graph Extraction method developed for this work has several steps, as shown in Figure 1d. Initially, the SI PDF is pre-processed to select the relevant sections describing the reaction procedures, as explained in more detail in Appendix B. This aims to lower the amount of text that needs to be processed in the steps following, and prevents errors by erroneous addition of spurious nodes to the graph. The PDF is then processed into text and split into single text segments describing chemical reactions. Two methods were tested for this: one based in rule-based text parsing from PDF, and one based in Vision-Language Models (VLMs), namely the recent GPT-4o by OpenAI. The latter method was implemented in view of the variability of representations and interleaved use of visual elements observed in SIs, as shown in Appendix A.

Resulting *reaction blocks* are then each processed individually by an LLM-powered generation pipeline, that detects and extracts all the substances declared in the input reaction. Each of these substances is represented as a structured object containing three main properties: *reference_key*, *substance_name*, and *role_in_reaction*. Each collection of substances is converted into a *reaction_unit*, a structured object resembling a node in a tree, where the head node is the product of the reaction and the children are all the substances with a role different than *product*.

Finally, a graph is constructed by connecting all the different *reaction_unit* objects, using each substance's *reference_key* as the node label.

The reported benchmark was used to perform ablations on 3 of the design choices for the algorithm, namely to test the effect of SI preprocessing to select relevant sections, the use of rule-based or vision-based PDF parsing, and the choice of LLM used for structured object generation. The results are shown in Figure 2.

1.4 PDF Parsing methods

Two parsing methods have been tested in this work. One is a simple, rule-based algorithm that is based on general observations from the structure of SIs in organic chemistry papers, while the other is fully driven by a Vision-Language Model (VLM), which aims to recover information by directly processing documents as humans would read it, without loss of visual elements.

1.4.1 Rule-based — Text

This approach consists of parsing the input PDF file using the PyMuPDF package (noa), which yields the complete text from the PDF, including titles and paragraphs, but also formatting details such as bold letters. Unfortunately it also includes spurious formatting details like page numbers and side notes from journals. Using this information, the text is split using "long sequences of bold letters" as a splitting criteria, which leads to a list of text segments. The idea behind this parsing is that most authors state products in bold font with the name of the product (*IUPAC*, or simply a reference name), followed by a reference key, and then proceed with the description of the reaction procedure in normal font (see Appendix A). This pattern is somewhat consistent and in some cases leads to very nicely parsed documents.

1.4.2 Image-based — Vision

The effectiveness of the rule-based method above is endangered by the variety of formats and representation styles that authors decide to use in their papers, as shown in the Appendix A. Understanding of these documents is heavily dependent on the reader's ability to interpret the visuals and contrast them and connect them with the text, thus the purely rule-based method falls short in some cases.

Leveraging the recent advances in VLM research, we propose directly using one such model for this task. In particular, we use the recently released GPT-4o, one of the most powerful end-to-end Large Multimodal Models (LMMs) from OpenAI.

The pipeline starts with the conversion of the input PDF into a suitable format, and for this we simply convert each page from the PDF into a png image using the pdf2image package (Belval, 2024). The images are then processed into overlapping batches of images, each batch in a single VLM call. This process ensures that the VLM sees a more global structure of the paper and thus has better context to give an appropriate response.

The VLM is then queried with all the images from a batch and a prompt with instructions (see Appendix C). The expected output of this is a summary of the relevant information for *each* reaction the VLM can identify in the image context; each reaction separated by a given separator token.

1.5 Evaluation metrics

A wealth of methods exist to compare graphs, each suitable for certain sets of use cases (Thompson et al., 2022; Shimada et al., 2016; Hartle et al., 2020). These include direct comparison of the node or edge sets, subgraph matching, spectral analysis, and the use of graph kernels, among others. In this work, we take an approach based on subgraph matching, that aims to capture the similarities relevant to synthetic routes in organic chemistry.

Appealing to the specific structure of the types of graphs used in this work, namely directed graphs with mostly a tree-like structure, we use 3 metrics based on the ratio of paths shared between the compared graphs, as shown in Equation 1.

$$S(G, G') = \frac{1}{|P_S(G)|} \sum_{p \in P_S(G)} \sum_{p' \in P_S(G')} 1_{p=p'} \tag{1}$$

Where $P_S(G)$ defines the set of all the linear paths p in G, and the $1_p = p'$ operator is defined as 1 if the condition $p = p'$ is met, 0 otherwise. The key difference between the methods used here is the definition of the equivalence operator $=$, which can take multiple forms depending on the property of interest. In particular, two options are defined: exact match and preservation of partial order. Exact match directly compares the two paths based on the exact sequence of nodes defined by each. This method thus directly measures to what extent the exact KG is reconstructed from documents.

The second method aims to capture a more nuanced structure in the retrieved KGs, through a slightly less strict comparison metric based on ordered sets. In this method, two paths are considered equivalent if the order relationships defined by each path are preserved in the other. Take for example the following two paths

$$p_0 = 6 \rightarrow S2 \rightarrow 7$$
$$p_1 = 6 \rightarrow 7$$

Where p_0 defines the order $6 \succ S2 \succ 7$. In this example, $p_0 \neq p_1$ under exact match, however they are under the PO equivalence as the order relationship $6 \succ 7$ exists in both paths. Such a less strict definition is particularly relevant in our case as it is typical in SIs to describe the formation of an intermediate and continue using it "without further purification". In these cases, the complete sequence p_0 with intermediate $S2$ may be reduced by the extraction models to p_1, which is not necessarily incorrect however missing some information.

A last method is used, which uses exact match as equivalence operator, but both G and G' are preprocessed to remove the leaves (nodes with $outdegree(n) = 0$), thus only comparing the backbone of the synthetic tree without considering reagents. Figure 1b shows such removed nodes in yellow, and the nodes belonging to the backbone in red.

2 Results

The proposed benchmark was used to perform ablations on 3 of the components of the KGE algorithm described in Section 1.3. Namely, we assess the effect of SI preprocessing (Appendix B), the parsing of PDFs using a rule-based approach, or directly through Vision-Language Models (VLMs), and the choice of LLM for parsing of reaction descriptions into formatted reaction units. In addition, we evaluate the performance of multiple LLMs from different providers on the latter task across multiple metrics using a more specific benchmark, aimed at selecting suitable LLMs for this task, without the need to execute the whole extraction pipeline.

2.1 KGE Benchmark

The aim of these experiments is to determine the effectivity of a given system at extracting a KG in the required format, not only at assessing the capabilities of LLMs, hence 2 binary variables are ablated that deal with document preprocessing, parsing and chunking. The latter is the LLM used, however here we have restricted ourselves to only testing models provided by OpenAI, mainly due to rate limit constraints from the other providers.

In Figure 2 we display the per-paper performance for each variation of the system in Figure 1d, across six metrics, all different forms of accuracy (left column=) and precision (right column) of synthetic path recovery. The upper row shows the results on exact path reconstruction, middle row a more relaxed version of this based on comparing the orders defined by each path, and bottom row compares the pruned graphs, assessing the similarity between the tree backbones; see Section 1.5 for details.

For each comparison method, $S(G_{EX}, G_{GT})$ measures the system's ability to reconstruct Ground Truth paths — highly important for organic chemistry as it defines the specific sequence of reactions, while $S(G_{GT}, G_{Extracted})$ measures the precision or "purity" of the resulting graphs, thus also accounting for erroneous introduction of nodes or edges in the extraction process.

The results show that the overall performance varies widely as a function of the paper, which is to be expected given the high variability in styles and formats used in these documents (see Appendix A). A systematic difference is found between the 2 models tested, with a clear advantage for GPT-4-turbo, the most advanced model, especially on reconstruction accuracy. The gap is nevertheless reduced in reconstruction precision which, as will be shown in the next section, can be attributed to the smaller model being better at detecting wrong inputs, thus introducing less noise into the extracted KG.

Interestingly, comparing the pruned graphs demonstrates GPT-3.5's poor performance on precision, with most values below 0.1, however the corresponding accuracy is relatively high, even surpassing GPT-4 based methods on the same metric. Such results imply that smaller models perform poorly in general conditions, however the information recovered by these is typically valid. More advanced models seem not to have a strong filter and generate valid structured outputs despite noisy filters, which in turns generate accurate but noisy KGs. These observations will be further elaborated in the following section.

From the results presented here it seems that using vision models like GPT-4o (columns in Figure

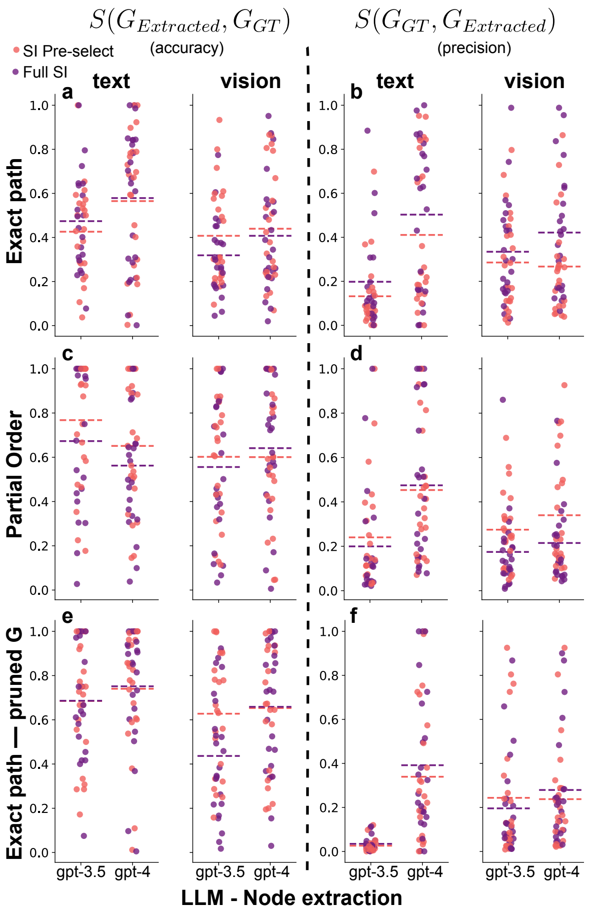


Figure 2: **Results from Knowledge Graph Extraction benchmark.** System performance on GOSyBench for multiple system ablations. The two main columns show accuracy (left) and precision (right). Each sub column shows the result for PDF parsing methods text-based (left) and vision-based (right). Rows present different metrics used for graph comparison, and the color distinguishes between SI pre-processing methods.

2), or preprocessing the document before to select the most relevant parts of the SI (colors in Figure) do not improve the system's performance. Vision only helps slightly improve the accuracy of the system when a smaller model is used, however such system still underperforms relative to the larger GPT-4.

A more in-depth exploration of the results is needed to determine how to best leverage vision models for this task.

2.2 LLM Performance across tasks

To assess the effect of the choice of LLM in the KGE method developed in this work, another benchmark with a narrower scope was produced. The benchmark aims to assess LLM's abilities to recover specific information from reaction description text samples. This involved the creation of 3 smaller datasets, each designed to test the models at specific tasks, namely ability to recognize and retrieve the correct product and reactant sets, ability to produce empty responses whenever a non-reaction text is given, and the ability to correctly retrieve the *reference_key* of substances.

All of these are elements of utmost importance for the algorithm's success at reconstructing a paper's KG, as failure to correctly perform these contaminates the resulting KG with spurious nodes and edges, and leads to the loss of real nodes and edges.

For the sake of completeness and ease of implementation, we have tested LLMs from 3 API providers, namely OpenAI, Anthropic and Mistral. Moreover, the models tested span a wide range of sizes and scores on standard benchmarks. As shown in Figure 3, the top-performing model in terms of product and reactants retrieval accuracy is *gpt-4-turbo*, on of the most advanced models as shown by benchmarks, in terms of reasoning capabilities. Nevertheless, other models, some smaller and far cheaper, perform almost on-par with gpt-4 on this metric (mistral small and medium, mixtral 8x7b, all claude models).

Surprisingly, the "smarter" models do not perform as good on other tasks, particularly "Wrong inp" and "Key exact". Smaller, less poweful models, like *mistral-small*, *mixtral-8x7b* and *gpt-3.5-turbo* do better in rejecting wrong inputs than their more advanced counterparts despite their less developed reasoning capabilities. An important observation is that, when given a non-reaction text, smaller models give an error as they fail to find the requested information and fail to produce an answer in the requested format, thus being caught as exception during model validation. In counterpart, larger models tend to give a response, despite the input text not containing the desired information, typically through hallucinations.

In spite of these observations, the ablations in Section 2.1 have been performed only with OpenAI models as we had higher rate limits, allowing us to perform multiple experiments concurrently.

3 Conclusions

We have proposed a novel benchmark for knowledge graph extraction in science from full papers. We exploit the native KG-like structure of synthetic

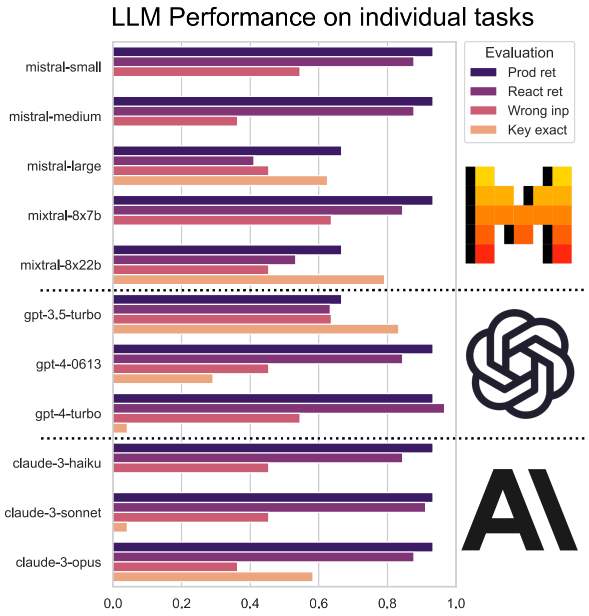


Figure 3: **Capability-specific benchmark for LLMs.** The performance of multiple LLM across multiple scales and providers is shown. Models are evaluated on 4 metrics: **Prod ret** evaluates the accuracy of retrieving the correct product name from an input paragraph (which involves separating product name from its reference key), **React ret** evaluates the same, for retrieval of reactants used in the described reaction, **Wrong inp** assesses how good the models are at rejecting inputs that do not describe a chemical reaction, and **Key exact** evaluates the ability of models to output the exact reference key for products.

organic chemistry and propose a benchmark with 24 manually curated papers. This benchmark is continuously growing to incorporate more high quality samples of challenging papers. We developed an LLM-based algorithm for KGE and evaluate each individual part using a small, handcrafted benchmark to test the capabilities of LLMs for each specific task, and find that advanced models have better recall of input context, however smaller models are advantageous to detect text that can not be identified as a reaction, thus not contaminating the generated KG with spurious nodes. Finally, we perform ablations on our algorithm and show that the usage of Language-Vision Models (LVMs) does not directly improve the system's performance, despite having empirical reasons to believe so. Overall, there is still a lot of room for improvement as our algorithms reach a maximum of 73% average in accuracy, and 59.7% in precision. More work needs to go into desiging and optimizing algorithms for this task, however we believe the release of GOSyBench sets the field into the right direction by providing a challenging, diverse and high-quality dataset for benchmarking.

4 Future work and outlook

The efforts presented here deal with the extraction and evaluation of the reaction networks from chemistry papers, which is only the backbone structure of a much richer KG for organic chemistry. However as discussed in Section 0.2, additional relationship types between substances are implicitly reported in papers, such as failed reactions and abandoned synthetic plans, use of substances as model systems, among others. All these are important details that describe not only a successful route to a target substance, but encode also the difficulties, lessons, and other valuable insights that are reported in chemistry papers. From early experiments, we have found that extracting such new connections is possible with LLMs thanks to their summarizing and reasoning capabilities. Achieving such a milestone has the potential to unlock promising advances in reaction search and chemical knowledge retrieval in general.

In addition to this, the currently presented ontology can further be enhanced with additional substance properties reported in papers. Starting with extraction of the SMILES strings for each molecule (Mavračić et al., 2021; Rajan et al., 2021, 2023), along with yields, scalability, and analytical results, the resulting KGs can continuously be populated with more substance-specific details to better represent the knowledge in papers.

Additionally, papers report multiple visualizations that display different views, or highlight different aspects of the molecules and reactions in question. The interplay between text and image modalities is strong in papers, and leveraging VLMs will be an essential step towards better KGE in chemistry, as has been shown in this work.

References

pymupdf/PyMuPDF: PyMuPDF is a high performance Python library for data extraction, analysis, conversion & manipulation of PDF (and other) documents.

Hasan Abu-Rasheed, Christian Weber, and Madjid Fathi. 2024. Knowledge graphs as context sources for llm-based explanations of learning recommendations. *ArXiv*, abs/2403.03008.

Edouard Belval. 2024. Belval/pdf2image. Original-date: 2017-05-28T19:00:59Z.

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

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

Marcin Fialkowski, Kyle J. M. Bishop, Victor A. Chubukov, Christopher J. Campbell, and Bartosz A. Grzybowski. 2005. Architecture and Evolution of Organic Chemistry. *Angewandte Chemie International Edition*, 44(44):7263–7269. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200502272.

Harrison Hartle, Brennan Klein, Stefan McCabe, Alexander Daniels, Guillaume St-Onge, Charles Murphy, and Laurent Hébert-Dufresne. 2020. Network comparison and the within-ensemble graph distance. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2243):20190744.

Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Computing Surveys*, 54(4):71:1–71:37.

Weihua Hu, Matthias Fey, M. Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and J. Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. *ArXiv*.

O. Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *ArXiv*, abs/2310.03714.

O. Khattab and Matei A. Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Jason Liu. 2024. jxnl/instructor. Original-date: 2023-06-14T10:42:23Z.

Daniel M. Lowe and Roger A. Sayle. 2013. Leadmine : A grammar and dictionary driven approach to chemical entity recognition.

Daniel Mark Lowe. 2012. *Extraction of chemical structures and reactions from the literature*. Ph.D. thesis, University of Cambridge.

Juraj Mavračić, Callum J. Court, Taketomo Isazawa, Stephen R. Elliott, and Jacqueline M. Cole. 2021. ChemDataExtractor 2.0: Autopopulated Ontologies for Materials Science. *Journal of Chemical Information and Modeling*, 61(9):4280–4289. Publisher: American Chemical Society.

Lars Meyer, Claus Stadler, Johannes Frey, Norman Radtke, Kurt Junghanns, Roy Meissner, Gordian Dziwis, Kirill Bulert, and Michael Martin. 2023. Llm-assisted knowledge graph engineering: Experiments with chatgpt. *ArXiv*, abs/2307.06917.

Nandana Mihindukulasooriya, Sanju Mishra Tiwari, Carlos F. Enguix, and Kusum Lata. 2023. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. *ArXiv*, abs/2308.02357.

Luc Patiny and Guillaume Godin. 2023. Automatic extraction of FAIR data from publications using LLM.

Kohulan Rajan, Henning Otto Brinkhaus, M. Isabel Agea, Achim Zielesny, and Christoph Steinbeck. 2023. DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. *Nature Communications*, 14(1):5045. Publisher: Nature Publishing Group.

Kohulan Rajan, Achim Zielesny, and Christoph Steinbeck. 2021. DECIMER 1.0: deep learning for chemical image recognition using transformers. *Journal of Cheminformatics*, 13(1):61.

Traian Rebedea, Razvan Laurentiu Dinu, Makesh Narsimhan Sreedhar, Christopher Parisien, and Jonathan Cohen. 2023. Nemo guardrails: A toolkit for controllable and safe llm applications with programmable rails. In *Conference on Empirical Methods in Natural Language Processing*.

Yutaka Shimada, Yoshito Hirata, Tohru Ikeguchi, and Kazuyuki Aihara. 2016. Graph distance for complex networks. *Scientific Reports*, 6(1):34944.

Dong Shu, Tianle Chen, Mingyu Jin, Yiting Zhang, Chong Zhang, Mengnan Du, and Yongfeng Zhang. 2024. Knowledge graph large language model (kg-llm) for link prediction. *ArXiv*, abs/2403.07311.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904. Publisher: American Chemical Society.

Rylee Thompson, Boris Knyazev, Elahe Ghalebi, Jungtaek Kim, and Graham W. Taylor. 2022. On Evaluation Metrics for Graph Generative Models. *arXiv preprint*. ArXiv:2201.09871 [cs].

Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and

Zhangming Niu. 2020. Pharmkg: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics*.

A Supplementary Information Files

A typical practice in organic chemistry publishing is having Supplementary Information files (SIs) where all information regarding experimental procedures, analytical results, and sometimes computational and theoretical predictions, are reported. In these documents, which all share a general *underlying* structure, reactions are described with references to other substances in the same document, with a notation shared between the SI and the main manuscript. Hence, a numeration scheme exists for the substances in each paper that can be followed to find the experimental procedure for the preparation of any compound synthesized as part of the research work. Despite of this homogeneity, large differences are noticeable, as is evident from figures 4, 5 and 6.

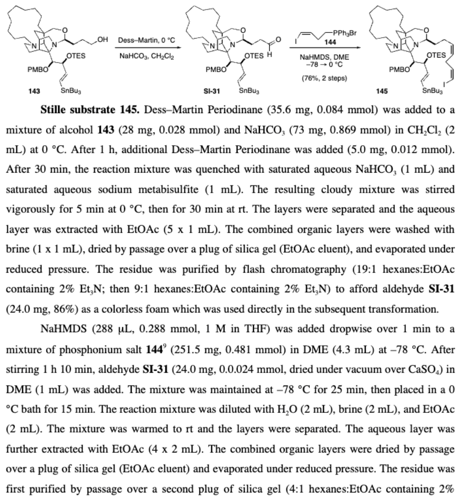
As these examples show, representations and formats are far from standardized. The SI displayed in Figure 4 shows a common format: compound name and reference in bold, accompanied by the molecular structure of the product substance, and followed by the reaction procedure. Notice however that a subsequent reaction is described directly in the same paragraph, without *announcing* the next product.

Figure 5 shows an SI with a heavier use of visual elements, where colored marbles are used to reference individual steps in a short reaction sequence. The marbles are then used throughout to refer to specific intermediates, with no reference in text to the products' reference keys. Lastly, Figure 6 shows another example where the product is not directly announced in the text, but rather a new reaction procedure is presented after a graphical depiction of the reaction in question, making it impossible for a text parser to grasp this information.

B SI Preprocessing

SIs in chemistry research papers contain many sections, however the one of interest for this work is the part on Experimental Methods. For our purposes, it may make sense to extract the most relevant parts of the document and process only that, however no naming convention or guidelines exist for this, making it difficult to identify and isolate the specific sections.

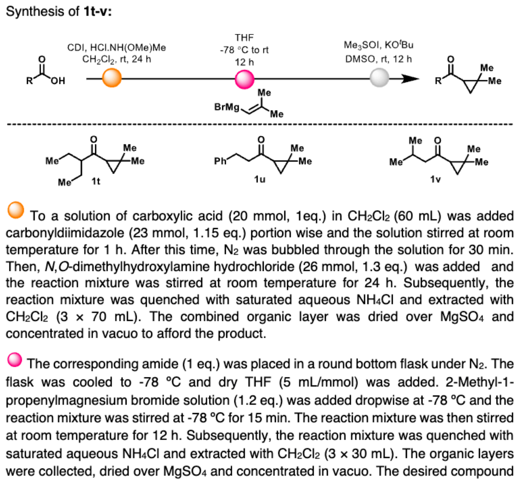
To address this, we develop a simple rule-based method to identify the relevant sections, partially inspired by Patiny and Godin (2023). For this, we rely on the observation that reaction descriptions



Stille substrate 145. Dess–Martin Periodinane (35.6 mg, 0.084 mmol) was added to a mixture of alcohol **143** (28 mg, 0.028 mmol) and $NaHCO_3$ (73 mg, 0.869 mmol) in CH_2Cl_2 (2 mL) at 0 °C. After 1 h, additional Dess–Martin Periodinane was added (5.0 mg, 0.012 mmol). After 30 min, the reaction mixture was quenched with saturated aqueous $NaHCO_3$ (1 mL) and saturated aqueous sodium metabisulfite (1 mL). The resulting cloudy mixture was stirred vigorously for 5 min at 0 °C, then for 30 min at rt. The layers were separated and the aqueous layer was extracted with EtOAc (5 x 1 mL). The combined organic layers were washed with brine (1 x 1 mL), dried by passage over a plug of silica gel (EtOAc eluent), and evaporated under reduced pressure. The residue was purified by flash chromatography (19:1 hexanes:EtOAc containing 2% Et_3N; then 9:1 hexanes:EtOAc containing 2% Et_3N) to afford aldehyde **SI-31** (24.0 mg, 86%) as a colorless foam which was used directly in the subsequent transformation.

NaHMDS (288 µL, 0.288 mmol, 1 M in THF) was added dropwise over 1 min to a mixture of phosphonium salt **144**[9] (251.5 mg, 0.481 mmol) in DME (4.3 mL) at –78 °C. After stirring 1 h 10 min, aldehyde **SI-31** (24.0 mg, 0.0.024 mmol, dried under vacuum over $CaSO_4$) in DME (1 mL) was added. The mixture was maintained at –78 °C for 25 min, then placed in a 0 °C bath for 15 min. The reaction mixture was diluted with H_2O (2 mL), brine (2 mL), and EtOAc (2 mL). The mixture was warmed to rt and the layers were separated. The aqueous layer was further extracted with EtOAc (4 x 2 mL). The combined organic layers were dried by passage over a plug of silica gel (EtOAc eluent) and evaporated under reduced pressure. The residue was first purified by passage over a second plug of silica gel (4:1 hexanes:EtOAc containing 2%

Figure 4: Example of an SI. Taken from https://pubs.acs.org/doi/10.1021/ja074300t. This example shows



Synthesis of **1t-v:**

● To a solution of carboxylic acid (20 mmol, 1eq.) in CH_2Cl_2 (60 mL) was added carbonyldiimidazole (23 mmol, 1.15 eq.) portion wise and the solution stirred at room temperature for 1 h. After this time, N_2 was bubbled through the solution for 30 min. Then, *N,O*-dimethylhydroxylamine hydrochloride (26 mmol, 1.3 eq.) was added and the reaction mixture was stirred at room temperature for 24 h. Subsequently, the reaction mixture was quenched with saturated aqueous NH_4Cl and extracted with CH_2Cl_2 (3 × 70 mL). The combined organic layer was dried over $MgSO_4$ and concentrated in vacuo to afford the product.

● The corresponding amide (1 eq.) was placed in a round bottom flask under N_2. The flask was cooled to -78 °C and dry THF (5 mL/mmol) was added. 2-Methyl-1-propenylmagnesium bromide solution (1.2 eq.) was added dropwise at -78 °C and the reaction mixture was stirred at -78 °C for 15 min. The reaction mixture was then stirred at room temperature for 12 h. Subsequently, the reaction mixture was quenched with saturated aqueous NH_4Cl and extracted with CH_2Cl_2 (3 × 30 mL). The organic layers were collected, dried over $MgSO_4$ and concentrated in vacuo. The desired compound

Figure 5: Example of an SI. Taken from https://pubs.acs.org/doi/10.1021/jacs.1c01356. This example shows

typically follow the pattern "reaction setup → workup → analytics", as the example below. As can be seen, the analytics section has a higher ratio of certain special and numeric characters relative to other parts of the text.

Example of a typical synthesis paragraph obtained from an SI file:

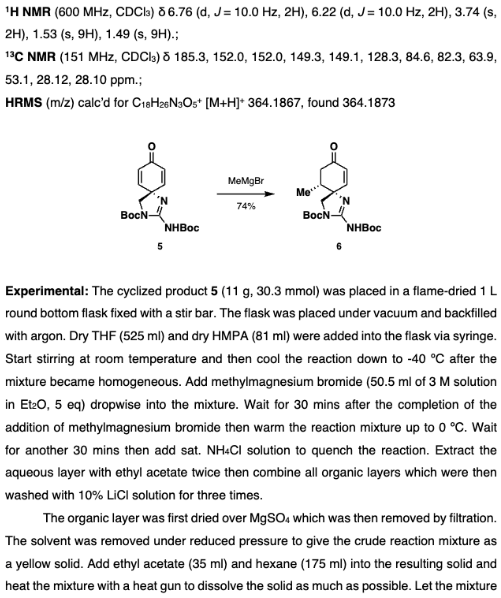
> To a solution of alkene **5** (266 mg, 0.92 mmol, 1.0 equiv.) in DCM (30 mL) was bubbled ozone (40% in air) at -78 °C until the starting material disappeared (TLC analysis, about 1 min), and the mixture was purged with air at -78 °C followed by addition of PPh3 (250 mg, 0.95 mmol, 1.0 equiv.). The mixture was warmed up to room temperature slowly, and stirred at the same temperature for 12 h. After removal of the solvent, the residue was purified by a flash column chromatography on silica gel (hexane/EtOAc = 5 : 1 to 3 : 1) to give compound 6 as a colorless oil (173 mg, 65%), which is an inconsequential 1.05: 1 mixture.
> Rf = 0.25 (hexane/EtOAc = 8:1, PMA); $[\alpha]$21
> D = - 4.44 (c 1.31, CHCl3); 1H NMR (400 MHz, CDCl3) δ 9.77 – 9.70 (m, 1.69H, overlap), 2.63 – 2.48 (m, 2.21H, overlap), 2.42 – 2.18 (m, 9.27H, overlap), 2.18 – 2.06 (m, 3.58H, overlap), 2.00 – 1.82 (m, 5.93H, overlap), 1.82 – 1.72 (m, 4.72H, overlap), 1.71 – 1.60 (m, 3.95H, overlap), 1.58 – 1.49

To leverage this, we split the complete document into sentences, and then calculate the ratio of special characters to normal letters for each. Plotting the values of these ratio with the line index in the x-axis, patterns like those in Figure 7 are apparent. An alogrithm is also applied for smoothing and performing selection by selecting the longest region with a prominent signal as the "relevant" SI. We find that this strategy generally leads to an accurate selection of the relevant parts.

C Vision-Language Models

The following prompt was used as a template to pass the images to GPT-4o for the vision-based parsing method exposed in Figure 1.

> These are some pages from the SI of an organic chemistry paper. Describe all the reactions shown there, if any. Separate each reaction with {SEPARATOR}, describe products and reactants for each reaction. Ignore all characterization data. Consider work-up and purification as part of the same reaction. Use the following format to represent the products and main reactants: {SUBSTANCE_FORMAT}. Do not rewrite



^{1}H NMR (600 MHz, $CDCl_3$) δ 6.76 (d, *J* = 10.0 Hz, 2H), 6.22 (d, *J* = 10.0 Hz, 2H), 3.74 (s, 2H), 1.53 (s, 9H), 1.49 (s, 9H).;

^{13}C NMR (151 MHz, $CDCl_3$) δ 185.3, 152.0, 152.0, 149.3, 149.1, 128.3, 84.6, 82.3, 63.9, 53.1, 28.12, 28.10 ppm.;

HRMS (m/z) calc'd for $C_{18}H_{26}N_3O_5^+$ [M+H]$^+$ 364.1867, found 364.1873

O
BocN
N
NHBoc
5
MeMgBr
74%
O
Me
BocN
N
NHBoc
6

Experimental: The cyclized product **5** (11 g, 30.3 mmol) was placed in a flame-dried 1 L round bottom flask fixed with a stir bar. The flask was placed under vacuum and backfilled with argon. Dry THF (525 ml) and dry HMPA (81 ml) were added into the flask via syringe. Start stirring at room temperature and then cool the reaction down to -40 °C after the mixture became homogeneous. Add methylmagnesium bromide (50.5 ml of 3 M solution in Et_2O, 5 eq) dropwise into the mixture. Wait for 30 mins after the completion of the addition of methylmagnesium bromide then warm the reaction mixture up to 0 °C. Wait for another 30 mins then add sat. NH_4Cl solution to quench the reaction. Extract the aqueous layer with ethyl acetate twice then combine all organic layers which were then washed with 10% LiCl solution for three times.

The organic layer was first dried over $MgSO_4$ which was then removed by filtration. The solvent was removed under reduced pressure to give the crude reaction mixture as a yellow solid. Add ethyl acetate (35 ml) and hexane (175 ml) into the resulting solid and heat the mixture with a heat gun to dissolve the solid as much as possible. Let the mixture

Figure 6: Example of an SI. Taken from https://pubs.acs.org/doi/10.1021/jacs.3c01991. This example shows

the reaction procedures, just describe the substances involved.

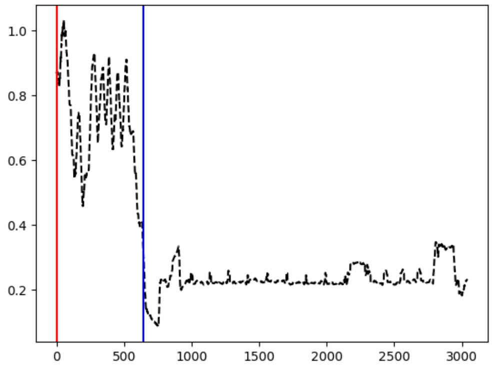


Figure 7: SIs were processed like this. Based on the frequency of special characters etc. Based on the observation that, most commonly, text-summaries of analytical data are given after the end of each reaction, giving a distinctive signal to each line in the document, producing more or less a spectrum that can then be analysed and processed.