Large Language Models Ensemble for Biochemical Properties Detection in Scientific Articles

Marcos Paulo Silva Gôlo[1,*,†], Jose Gilberto Barbosa de Medeiros Junior[1,†], Gabriele Souza Vilas Boas[1,†], Fábio Manoel França Lobato[1,2,†], Diego Furtado Silva[1] and Ricardo Marcondes Marcacini[1]

[1]*Institute of Mathematical and Computer Sciences, University of São Paulo, Brazil*
[2]*Engineering and Geoscience Institute, Federal University of Western Pará, Brazil*

Abstract

The discovery of new drugs from natural products is a task with a significant impact on both society and industry. However, it presents substantial challenges when conducted manually. With the growing volume of scientific data, machine learning–based approaches have become essential for extracting relevant biochemical information from specialized texts, particularly in chemistry-related articles. Recent advances have explored knowledge graph representations and graph embedding techniques in the NatUKE benchmark. In this work, we propose LLM-based strategies for the automatic extraction of biochemical properties: (i) the use of General Models with Prompt-Engineering in a zero-shot scenario; (ii) few-shot prompting with open-source models; (iii) supervised fine-tuning of an open-source LLM; and (iv) an ensemble of proprietary LLMs in zero-shot mode. Our analysis includes comparisons between single-task and multi-task scenarios. Among all strategies, the ensemble of proprietary LLMs in a zero-shot scenario outperformed the other strategies, including the graph embedding–based methods.

Keywords

Chemistry Text Mining, Knowledge Graphs, NatUKE Benchmark, LLM Fine-tuning, Few-shot Learning

1. Introduction

The discovery of new drugs is a cornerstone of medical advancement and a crucial effort to address complex diseases. Natural products is one of the most promising sources for this intent, whose diverse chemical properties have served as the basis for the development of bioactive compounds [1]. The systematic exploration of these substances offers not only substantial benefits to society, through more effective treatments, but also strategic opportunities for pharmaceutical companies pursuing innovation [2]. However, identifying relevant biochemical properties in compounds derived from natural products remains a challenging task, particularly when performed manually by domain experts [3]. This gap highlights the need for computational

BiKE'25: Second International Biochemical Knowledge Extraction Challenge, June 1-2, 2025, co-located with Extended Semantic Web Conference (ESWC), Portoroz, Slovenia.

*Corresponding author.

†These authors contributed equally.

marcosgolo@usp.br (M. P. S. Gôlo); gilberto.barbosa@usp.br (J. G. B. d. M. Junior); gabrielevilasboas@usp.br (G. S. V. Boas); fabio.lobato@ufopa.edu.br (F. M. F. Lobato); diegofsilva@icmc.usp.br (D. F. Silva); ricardo.marcacini@usp.br (R. M. Marcacini)

approaches to facilitate and accelerate the discovery process [3].

For computational approaches to effectively contribute to the extraction of biochemical properties, the availability of high-quality data is essential. In this context, scientific papers in the field of chemistry serve as a valuable source of information, providing detailed descriptions of compounds, experimental methodologies, and results related to their biochemical properties. With advancements in Natural Language Processing (NLP), it has become possible to automatically explore these texts to identify patterns and infer relevant characteristics [4]. Several benchmarks have been proposed to evaluate model performance in such tasks, with NatUKE standing out as one of the main datasets focused on chemical text analysis, providing specific annotations for biochemical properties. The NatUKE dataset comprises hundreds of manually annotated scientific articles aimed at predicting five key properties of chemical compounds [5].

Several recent studies have explored knowledge graph–based representations for extracting biochemical properties from scientific papers, particularly within the context of the NatUKE benchmark. Graph embedding methods such as DeepWalk [6], Node2Vec [7], Metapath2Vec [8], and more recently, EPHEN [5], have been applied to capture structural relationships among compounds, properties, and chemical concepts. Further advancements have been proposed by [9], who employed breadth-first search (BFS) strategies to improve the semantic coverage of graph paths; by [10], who used GPT-3.5 to preprocess the text of PDFs to use this text in the graph embedding methods; and by [11], who incorporated domain-specific named entity recognition for chemistry. More recently, [12] introduced significant improvements by combining more robust text extractors with embeddings derived from Large Language Models (LLMs). Despite these advances, gaps remain in fully leveraging the capabilities of LLMs, particularly for the direct extraction of biochemical properties from the full text of scientific articles.

We propose an approach based on LLMs for the automatic extraction of biochemical properties from chemistry scientific articles, exploring different evaluation strategies. We investigate the performance of General Models with Prompt-Engineering in a zero-shot scenario, using both proprietary and open-source models. Next, we adopt a few-shot strategy, relying exclusively on open-source models, to assess whether the inclusion of representative examples in the prompts enhances the extraction of the properties. We also conduct fine-tuning of an open-source model with 32 billion parameters, aiming to evaluate whether task-specific adaptation can achieve performance competitive with proprietary models with trillions of parameters. Finally, we propose an ensemble strategy of proprietary models. These strategies are applied in both single-task scenarios, where each property is extracted independently, and multi-task scenarios, where all properties are inferred within a single prompt. In short, our contributions are:

- An analysis of LLM performance on the task of biochemical information retrieval from scientific texts;
- A comparison between single-task and multi-task strategies for LLMs applied to property extraction;
- Achievement of the best performance in the Second International Biochemical Knowledge Extraction Challenge (BiKE) through the ensemble strategy.

2. Related Work

Several recent studies have explored graph-based representations using the NatUKE benchmark [5]. Initially, three graph embedding methods were investigated. One of them was DeepWalk [6], which learns latent representations of vertices in a graph by treating random walks as sentences and vertices as vocabulary. Moreover, the latent dimensions can be used to measure similarity between network nodes, enabling generalization for retrieving information from research papers in specific fields such as the NatUKE benchmark. Another method was Node2Vec [7], a variant of DeepWalk that introduces the control parameters p and q to generate structurally biased random walks. Node2Vec's flexibility proved helpful because when considering the compound name, tightly knit concepts may offer a rich exploration through a BFS approach, and different species may have similar biological activity, which can be best captured by a depth-first search (DFS) analysis. Thus, parameters p and q help provide such flexibility and enrich concept representation. Lastly, Metapath2Vec [8] was employed to learn latent representations in heterogeneous graphs using meta-path-guided random walks. This strategy allows the preservation of both structural and semantic relationships across multiple types of nodes, as found in the NatUKE knowledge graph, making it well-suited for extracting chemical properties from scientific texts. By modeling articles, clusters, and properties as distinct node types, Metapath2Vec can represent their interrelations in a shared vector space, enabling richer and more context-aware analysis of chemical information embedded in scientific texts.

do Carmo et al. [5] used the Embedding Propagation in Heterogeneous Networks (EPHEN) method in the NatUKE Benchmark. Textual content from the papers was extracted from PDFs using the PyMuPDF library and represented through BERT embeddings, which were assigned as features to the article nodes. To enable link prediction, the embeddings of property nodes were generated via the EPHEN regularization method, which propagates information across neighboring nodes, preserving the information from article nodes and generating semantic representations for those without features. Links between articles and properties were then predicted based on embedding similarity. EPHEN outperformed baseline methods such as Metapath2Vec, DeepWalk, and Node2Vec [5].

Zope, Mishra, and Tiwari [9] addressed the increasing volume and heterogeneity of biomedical data by utilizing Breadth-First Search with Word2Vec to generate node embeddings. BFS captures the structural and semantic context of biological entities by combining the benefits of semantic embedding and graph traversal techniques. BFS explores all nodes at the same level from the starting point and then proceeds to the nodes of the next level, generating a sequence of nodes. The authors evaluated the method using the NatUKE benchmark, comparing it against EPHEN. The BFS-driven Knowledge Graph Embedding approach demonstrated superior performance for the "*Compound Name*" and "*Species*" properties, while for "*Bioactivity*", "*Collection Site*", and "*Isolation Type*", EPHEN is preferable.

Fröhlich, Gwozdz, and Jooß [10] were the first to leverage LLM technology. Their contribution relied on using the ChatGPT-3.5 model in the texts extracted from PDFs via OCR to help generate better knowledge graphs as inputs to the original benchmark extraction techniques. While their results show improvement across DeepWalk, node2vec, and EPHEN, they fail to improve Metapath2Vec. Their evaluation rise is highest on EPHEN, where a +8.91% growth was observed. The authors highlight tasks such as text extraction and cleaning, prompt engineering, API calls,

and post-processing. In the results, they highlighted improvements in bioactivity, site, and isolation type properties, but the same or worse results in name and special properties.

Schmidt-Dichte and Mócsy [11] proposed an extension to the NatUKE benchmark by integrating named entity recognition (NER) into the extraction pipeline using biomedical-oriented scispaCy models. Their approach aimed to improve the precision of identifying key biochemical properties, such as bioactivity, species, and isolation type, by leveraging NER for more semantically meaningful sentence segmentation. While their method demonstrated performance gains, especially in properties like bioactivity and collection site when used with the EPHEN embedding model, challenges persisted in accurately extracting compound names and species.

do Carmo et al. [12] proposed an enhanced version of the EPHEN method. Initially, the authors focused on the text extraction process from PDF files, since a higher-quality extraction minimizes the loss of semantic information. Subsequently, they investigated variations in the initial embeddings used as input for the regularization process. In this context, they explored two alternative text extractors, Nougat and Grobid, and two different LLMs for generating embeddings. Following an empirical evaluation to identify the most effective combination, the use of Nougat/Grobid for text extraction combined with BERT-based embeddings obtained the best results, outperforming the original version of EPHEN [12].

Given the SoTA results of LLMs in information retrieval tasks, we highlight the gap in the use of LLMs in biochemical property extraction. Even though LLMs were used by [9], they were used as pre-processing and not to extract properties directly. Our approach is different since we are not using LLM technology as a means of generating better input for a knowledge graph or to extract text from PDFs. On the other hand, we evaluate the capability of more recent SoTA LLMs in performing property extraction directly from scientific text papers. In this sense, we present our LLM strategies for extracting biochemical properties in the next section.

3. LLMs for Biochemical Knowledge Extraction

The challenges and applications of LLMs have become one of the most active research topics, extending beyond the field of computer science. Besides its broad applicability in chatbots, code generation, and document summarization, specific applications in computational biology have been exhaustively explored, including protein embeddings, genomic analysis, and information retrieval [13, 14]. Considering that scientific knowledge is predominantly recorded in books and scientific journals, often in the form of PDFs, there is a need for retrieving text and other semantic information (e.g., mathematical expressions and diagrams) [15]. Existing Optical Character Recognition (OCR) engines usually detect individual characters and words. Simpler methods detect the texts but often shuffle sentences due to file format (e.g., double columns). Moreover, both approaches typically lose the text formation.

Bearing in mind that data quality influences information retrieval, besides the complexity and challenge demonstrated by the poor performance of SoTA methods in extracting features such as "*compound name*", "*species*", and "*location*" [9]; makes it clear that each detail matters. Thus, based on guidelines for data extraction from scientific data [16] and considering the specificity of scientific documents, we adopted the Neural Optical Understanding for Academic Documents (Nougat), proposed by Blecher and colleagues [15], for processing the scientific papers into a

markup language, bridging the gap between human-readable documents and machine-readable text. With the scientific documents adequately converted to text, three basic strategies were considered: zero-shot learning, few-shot learning, and fine-tuning. The first two strategies were tested using General Models with Prompt Engineering, ranging from small models with a few billion parameters (4 b) to large-scale models utilizing Mixture-of-Experts architectures. Both zero-shot and few-shot learning were conducted in accordance with the best practices of prompt engineering. The latter approach (fine-tuning) aimed to incorporate domain knowledge to optimize model performance. Each of the three strategies is detailed below.

3.1. General Models with Prompt-Engineering

Our first strategy involved leveraging several SoTA General LLMs in a zero-shot scenario. Exploring concept extraction with no prior knowledge, e.g., a training set, is usually a challenging task, especially in complex domains such as biochemistry. However, LLMs have demonstrated excellent performance in zero-shot scenarios [17], which can be useful for context awareness in academic papers. The ability to input information within the prompt is one of the advantages of LLMs as good zero-shot extractors. We take advantage of this by specifying the description and expected format of each concept, eventually through examples.

In some cases, such as for more complex properties such as "*compound names*", we decided to be even more exhaustive, by defining the entire universe of possible names to be extracted. On the one hand, it is important to note that this strategy may impair the model's ability to abstract novel concepts that are not part of previously known aspects. Conversely, we found it helpful in ensuring the identification of concepts and conformity of the outputs. Output conformity is particularly important because the *hits@k* evaluation metric proposed in the challenge considers string matches as hits, and the biochemical properties evaluated could be presented in diverging manners, yielding negative results even if semantically congruent. Despite their zero-shot capabilities, LLM performance in such contexts is heavily dependent on model size [17], with smaller models often being ineffective for larger tasks. Task decomposition could yield an opportunity for leveraging smaller models, such as querying each property independently. Considering the time constraints of the BiKE challenge, we opted for a single-prompt, multi-task approach. As a result, this work focuses on larger models.

By virtue of evaluating more than one LLM, this work explores the possibility of combining their outputs through an ensemble strategy [18]. We leverage the flexibility allowed by the nature of the BiKE benchmark task since the *hits@k* evaluation metric is $k-1$-permissive. This characteristic enables the integration of outputs from multiple models, enhancing prediction robustness without incurring additional training or fine-tuning costs. The core idea is to combine outputs from models with complementary performance. The ensemble increases the likelihood of correctly retrieving the target properties. The selection of models to be combined is informed by their individual performance in prior experiments. Therefore, we prioritized the fusion of outputs from the two most promising models, resulting in an efficient and competitive ensemble. This approach was adopted as one of the main strategies in our work, enabling us to exploit synergies between models and improve the accuracy of biochemical property extraction [18].

3.2. Few-Shot Learning

We propose a few-shot learning strategy based on SOTA LLMs for extracting biochemical properties from scientific texts [19]. Our approach relies on providing a minimal but informative set of training examples designed to contextualize the extraction task within the LLM‘s input prompt. Given the model‘s limited context windows, we aim to maximize the input informativeness by including representative samples of input-output pairs, where each input is a scientific text and the output consists of the expected biochemical properties to be extracted [19].

To guide the language model in extracting biochemical properties, we designed a task-specific prompt that simulates the role of a domain expert. The prompt instructs the model to identify key attributes from a given scientific text. For each property, we explicitly define the expected format and constraints, including closed answer sets for categorical fields and taxonomic or chemical naming conventions for more open fields. In the few-shot setting, the base prompt is extended by appending curated input-output examples extracted from the training data. These examples follow a structured format where each sample consists of a short paper excerpt and the corresponding ground-truth annotations for all properties. To increase robustness and discourage spurious generations, each output field is modeled as an array containing multiple elements, with at least one required to match the true label. This structure encourages the LLM to produce responses that cover plausible alternatives while still including the expected answer.

3.3. Fine-tuning

We also propose a fine-tuning strategy designed to extract biochemical properties from scientific texts. Fine-tuning a large language model enables the adaptation of its behavior by incorporating domain-specific knowledge and optimizing performance for specialized tasks - the model is updated using a domain-relevant dataset. By fine-tuning a pre-trained LLM with representative examples of biochemical properties, we not only enhance the model's knowledge with domain-specific information but also customize its response style. This adaptation leads to improved accuracy, relevance, and overall effectiveness in retrieving meaningful information [20].

To fine-tune the LLM, we adopted a single-task strategy rather than a multi-task approach. In this strategy, the LLM is trained to perform only one specific task, i.e., to extract a single type of information. By focusing the model on a single task, its performance on that task can be significantly improved [21]. Therefore, we trained five separate models, each dedicated to extracting a distinct biochemical property. The input to each model is the article's text, and the output is a list of items related to the property. The training process for each model is guided by a system prompt specifically tailored to bias the model toward the corresponding property.

We propose a prompt in which for each property, we replace the special token **$Property$** by an explanation of the property (each item in the bullet). We use the following prompt: *"You are a scientist trained in chemistry. You must extract information from scientific papers, identifying relevant properties associated with each natural product discussed in the academic publication. For each paper, you have to analyze the content (text) to identify the* ***$Property$****.*

- *Compound name. It can be more than one compound name."*
- *Isolation Type, i.e., Collection type of the species. Options of collection types: ['Plant Isolated', ..., 'Semisynthesis Product']."*

- *Collection Site, i.e., the place of the collection. Options of places: ['Sao Carlos/SP', ..., 'Pocos De Caldas/MG']."*
- *Collection Specie, i.e., Species from which natural products were extracted. Provide the scientific name, binomial form. The family name can be provided. For example, Tithonia diversifolia, Styrax camporum (Styracaceae), or Colletotrichum gloeosporioides (Phyllachoraceae)."*
- *Biological Activity. It can be more than one biological activity. Options of biological activities: ['Anesthetic', ...,'Inhibition of Cathepsin V']."*

We explore a Quantized Low-Rank Adapter (QLoRA) training strategy to fine-tune our LLM [22]. QLoRA combines LoRA with 4-bit quantization, enabling the fine-tuning of very LLMs using limited computational resources. This allows us to leverage a larger model without the need for high-end hardware. The QLoRA approach keeps the base model frozen in 4-bit precision (quantized) and, instead of updating all parameters, inserts small trainable neural layers which are optimized independently during training [22].

4. Experimental Evaluation

This section presents the benchmark used and experimental settings. Our research goal is to demonstrate that LLM approaches can outperform graph approaches and compare different LLM approaches, as zero-shot, few-shot, and fine-tuning. Our source codes are public available[1]

4.1. Benchmark NatUKE

The NatUKE Benchmark focuses on knowledge extraction about natural products from academic literature. It leverages a dataset derived from over two thousand annotated instances of natural product properties and evaluates four unsupervised graph embedding techniques: DeepWalk, Node2Vec, Metapath2Vec, and EPHEN. Additionally, this benchmark provides the dataset to support exploration of alternative approaches beyond graph-based embeddings [5].

The dataset, built from hundreds of peer-reviewed scientific articles, comprises a corpus of over 2,000 manually annotated entries, each reviewed by domain experts in chemistry. These annotations focus on five specific properties from the NuBBEDB database, which are essential for both training and prediction tasks: Compound Name, Bioactivity, Species from which the natural products were extracted, Collection Site of these species, and Type of Isolation [5].

After the curation process, we obtained 143 articles containing valid entries for 448 compound names, 33 bioactivities, 115 species, 51 collection sites, and five isolation types. The goal is to predict these five properties for each article. In this context, the input is the full-text PDF of the article, and the output consists of the five corresponding properties.

4.2. Experimental Settings

For our zero-shot approach, we extensively tested candidate models. In total, 13 models were initially evaluated considering four main aspects: output consistency, hit@k performance, hallucination, and prompt pricing. The exhaustive list comprises: 'deepseek-r1-distill-llama-70b',

[1]https://github.com/boasgsv/labike/

'gemini-2.0-flash-001', 'gemini-2.0-flash-lite-001', 'gemini-2.5-flash-preview:thinking', 'gemini-2.5-pro-preview-03-25', 'gemma-3-27b-it', 'gpt-4.1', 'gpt-4o-mini', 'llama-3.1-nemotron-ultra-253b-v1', 'mistral-small-24b-instruct-2501', 'phi-4-reasoning-plus', 'qwen3-235b-a22b', 'qwen3-32b'. After a careful evaluation, one of the most prevalent issues identified was inconsistent output regarding its structure. As a result, parsing their raw outputs into the evaluation set would be challenging because each model and document led to a different structure. Moreover, open-source models tested also seemed to hallucinate more when dealing with complex concepts such as "*compound names*". After a careful evaluation of Hits@k and pricing, we decided to focus our attention on evaluating the two closed-source models tested: 'gemini-2.5-pro-preview-03-25' and 'gpt-4.1'. All chosen models were evaluated using their own API default configurations, as iterating over possible customizations was expensive.

For the few-shot experiments, we employed the Qwen3 language model family, exploring variants with 4, 8, 14, and 32 billion parameters. Given the extensive number of evaluations and the context length limitations of these models, we adopted a one-shot learning setup by including only a single annotated example in each prompt. Among the tested configurations, the 32B model consistently achieved the best performance and was thus selected for final evaluation.

We selected the Qwen 2.5 model [23], with 32 billion parameters, for fine-tuning. To prevent memory overflow, we limited the input text to the first 3,000 words of each article. We use the *unsloth* library to fine-tune our model[2]. The model was fine-tuned with the following hyperparameters: max_seq_length = 16384, a learning rate of 5×10^{-4}, 25 steps (similar to epochs, but a different parameter in the *unsloth* library), weight_decay = 0.01, warmup_steps = 5, and temperature = 0.00001. Due to time constraints, we performed training using only the first stage of the benchmark splits. This simplification was necessary due to the structure of the BIKE challenge, which comprises five prediction tasks, ten training folds, and four evaluation stages, resulting in a total of 200 models if all combinations were trained. To reduce this complexity, we limited our experiments to 50 models (1 stage × 10 folds × 5 tasks).

5. Results and Discussion

Table 1 presents our results. We compare our methods with the four baselines of NatUKE, Deepwalk, Node2vec, Metapath2vec, and EPHEN. Our results show GPT4.1, Gemini 2.5, and Qwen 2.5 with zero-shot strategies, Qwen 3 with a few-shot strategy, Fine-tuned Qwen 2.5, and the ensemble of GPT4.1 and Gemini 2.5. The best results are in bold.

Our zero-shot models outperformed baseline methods across all properties. The improvements in "*compound names*", "*specie*", and "*isolation type*" were particularly relevant, considering how the previous benchmark struggled to achieve values higher than 0.25 for most of these properties. While EPHEN had already improved upon other graph-based methods in "*specie*" and "*isolation type*" (raising the bar to 0.36 and 0.75 maximum results for each property, respectively), our zero-shot LLM solution raises these benchmarks to 0.97 on both properties (ensemble model).

When evaluating these pre-trained models comparatively, we found that Gemini-2.5 greatly outperformed GPT-4.1 in "*compound names*" extraction, with a +0.24 difference. We did not observe relevant disparities between these paid models in other properties. **Gemini 2.5** achieved

[2]https://github.com/unslothai/unsloth

Table 1
Results table for extracting: compound name (C), bioactivity (B), specie (S), collection site (L), and isolation type (T). Performance metric with the `hits@k` and k is respectively: 50, 5, 50, 20, and 1.

Property	DeepWalk				Node2Vec				Metapath2Vec			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
C	0.08	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.10	0.08	0.09	0.20
B	0.41	0.12	0.10	0.07	0.41	0.07	0.03	0.03	0.27	0.17	0.13	0.12
S	0.37	0.24	0.27	0.25	0.36	0.22	0.25	0.24	0.40	0.41	0.42	0.44
L	0.56	0.41	0.38	0.29	0.57	0.36	0.28	0.23	0.40	0.42	0.42	0.40
T	0.25	0.14	0.14	0.09	0.10	0.07	0.05	0.01	0.28	0.22	0.19	0.19

Property	EPHEN				GPT 4.1 (Zero-shot)				Gemini 2.5 (Zero-shot)			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
C	0.09	0.02	0.03	0.04	0.65	0.65	0.62	0.62	0.89	0.89	0.88	0.88
B	0.55	0.57	0.60	0.64	0.77	0.78	0.80	0.81	0.77	0.78	0.80	0.81
S	0.36	0.24	0.29	0.30	0.93	0.92	0.92	0.90	0.96	0.96	0.96	0.95
L	0.53	0.52	0.55	0.55	0.71	0.72	0.72	0.73	0.71	0.71	0.73	0.71
T	0.71	0.66	0.75	0.75	0.96	0.96	0.96	0.97	0.95	0.95	0.95	0.95

Property	Qwen (Zero-shot)				Qwen (Few-shot)				Qwen (Fine-tuned)			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
C	0.09	0.09	0.08	0.10	0.13	0.15	0.14	0.17	0.27	0.28	0.27	0.25
B	0.69	0.71	0.73	0.74	0.37	0.37	0.33	0.41	0.74	0.75	0.77	0.78
S	0.46	0.47	0.49	0.46	0.50	0.49	0.48	0.51	0.45	0.45	0.45	0.42
L	0.63	0.64	0.66	0.65	0.50	0.54	0.55	0.58	0.67	0.66	0.69	0.68
T	0.95	0.95	0.96	0.97	0.77	0.80	0.79	0.79	0.94	0.95	0.95	0.95

Property	Ensemble			
	1st	2nd	3rd	4th
C	**0.92**	**0.92**	**0.90**	**0.90**
B	**0.81**	**0.82**	**0.83**	**0.83**
S	**0.97**	**0.97**	**0.97**	**0.95**
L	**0.76**	**0.76**	**0.77**	**0.77**
T	**0.96**	**0.96**	**0.96**	**0.97**

top results for **compound name** (0.89) and **species** (0.96), while **GPT-4.1** leads slightly on **bioactivity** (0.81) and **isolation type** (0.97). These results highlight the capacity of proprietary LLMs to generalize across complex biochemical properties without fine-tuning.

While the zero-shot Qwen-2.5 32b achieved lower marks when compared to proprietary models across the board, it also improved upon original benchmarks throughout all properties (with the exception of "*compound names*" where no significant increase was observed). These results show that there is room for open-source LLM applications in the realm of biochemical information extraction, especially when considering lower budget environments.

Our ensemble model based on GPT-4.1 and Gemini-2.5 proved to be an efficient strategy for raising the bar even higher, as it overcame all individual scores across all properties and splits. As previously discussed, this happened because the *hits@k* evaluation metric is $k - 1$ permissive of incorrect matches. Since the response sets for individual models were sufficiently

small, their union yields a set that covers more possible values within the k limit.

The Qwen few-shot results demonstrate competitive performance when compared to baseline methods, particularly in the extraction of most biochemical properties. Notable improvements were observed for properties such as compound name, collection species, collection site, and isolation type. However, performance on the bioactivity property remained inconsistent, with the model exhibiting high variability in its responses. In the best-performing configuration, the few-shot Qwen model achieved scores of **0.17** for compound name, **0.41** for bioactivity, **0.51** for species, **0.58** for collection site, and **0.80** for isolation type.

The fine-tuned model outperformed the few-shot strategy on 80% of the properties, demonstrating that supervised adaptation to the task can lead to significant performance gains. Compared to open-source models used in zero-shot mode, the fine-tuned model achieved superior results on 60% of the properties, consolidating its advantage even over models with similar scalability. Although proprietary models still hold the overall lead in terms of *hits@k*, the fine-tuned model proved competitive in two properties, indicating that with targeted training, open models can reach performance levels close to those of the most advanced solutions.

Despite these promising results, a clear performance gap remains when comparing the few-shot approach to the pretrained and zero-shot baselines, particularly those based on proprietary and more powerful language models. As shown in Table 1, this disparity suggests that expanding the few-shot approach to include a larger number of examples could help bridge this gap, enabling the model to better generalize and identify properties more accurately. However, reaching competitive results with state-of-the-art zero-shot models may require a substantial increase in both the number of few-shot samples and computational resources, due to the extensive pretraining, broader context windows, and larger parameter counts of those proprietary models.

6. Conclusions and Future Work

Identifying relevant biochemical properties in compounds derived from natural products remains a challenging task. Manually annotating is labor-intensive and presents high costs. On the other hand, current automation strategies to extract biochemical properties require high-quality data. The NaTUKE is a well-known benchmark that commonly explores knowledge graph methods and presents satisfactory performance in identifying some properties but exhibits poor performance on others. Additionally, SoTA methods are sensitive to data quality, and, to the best of our knowledge, there is no work exploring LLMs for this particular task.

Aiming to fill these three research gaps, regarding i) performance on critical aspects, ii) the requirement of volume of high-quality data, and iii) the negligence of LLMs, we present in this paper the efforts to improve biochemical information extraction from scientific documents, considering the NatUK Benchmark, by employing LLMs. We investigate three strategies to know: i) General Models with Prompt-Engineering in a zero-shot scenario, ii) few-shot relying on open-source models; and iii) fine-tuning of open-source models for task-specific adaptation.

The exhaustive experiments demonstrated that the ensemble model of proprietary zero-shot models achieved the best results, while the proprietary models alone achieved the second and third best results. Regarding the open-source models, some were able to be competitive with the proprietary ones but without outperforming them. On the other hand, the open-source

models outperformed the knowledge graph embedding models.

We also faced some limitations regarding the experiments conducted. First, uploading proprietary papers on proprietary LLMs might infringe on legal rights, thereby limiting the practical application of our proposed solution. Secondly, another limitation is related to the informative search employed in our experiments, as we incorporated background knowledge by listing the possible targets. Finally, proprietary solutions are also costly, considering the number of tokens in scientific papers. Moreover, we limited our experiments to the single-prompt approach to avoid the additional costs of multiple prompts per paper.

Based on the limitations, we foresee many future research paths. First, we emphasize that strategies relying on open-source models must be adapted to provide practical and legally compliant solutions. Non-informative search strategies should also be employed, aiming to avoid incorporating background knowledge or providing more general aspects. Using the divide-and-conquer approach also seems interesting and offers several research opportunities. This could be achieved by using Question-and-Answer techniques to extract one aspect from each prompt or by confronting LLMs to verify the accuracy of responses, acting as a module for identifying errors and inconsistencies. Alternatively, research papers could be scrutinized by section instead of performing a single search across the entire document. Finally, hybrid approaches, which combine graph-based solutions with LLMs, also appear promising.

Acknowledgements

This work was partially supported by CAPES (88887.671481/2022-00), CNPq (316507/2023-7, DT - 303031/2023-9, POSDOC - 101057/2024-5), FAPESP (2023/02680-0, 2022/09091-8, 2023/10100-4, 2019/07665-4, and 2013/07375-0), and Google Latin America PhD Fellowship.

Declaration on Generative AI

The authors utilized generative AI to correct grammar and enhance scientific writing. Generative AI was used to refine paragraphs written by the authors, not to create paragraphs.

References

[1] L. Zhang, J. Song, L. Kong, T. Yuan, W. Li, W. Zhang, B. Hou, Y. Lu, G. Du, The strategies and techniques of drug discovery from natural products, Pharmacology & Therapeutics 216 (2020).

[2] J. B. Calixto, The role of natural products in modern drug discovery., Anais da Academia Brasileira de Ciências 91 (2019) e20190105.

[3] C. Réda, E. Kaufmann, A. Delahaye-Duriez, Machine learning applications in drug development, Computational and structural biotechnology journal 18 (2020) 241–252.

[4] C. C. Aggarwal, Machine Learning for Text, 1st ed., Springer, 2018.

[5] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. S. e Silva, A. Pilon, Natuke: A benchmark for natural product knowledge extraction from academic literature, in: International Conference on Semantic Computing (ICSC), IEEE, 2023, pp. 199–203.

[6] B. Perozzi, R. Al-Rfou, S. Skiena, Deepwalk: Online learning of social representations, in: Proceedings of the 20th ACM SIGKDD, ACM, USA, 2014, pp. 701–710.
[7] A. Grover, J. Leskovec, node2vec: Scalable feature learning for networks, in: International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 855–864.
[8] Y. Dong, N. V. Chawla, A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, in: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, Canada, 2017, pp. 135–144.
[9] B. Zope, S. Mishra, S. Tiwari, Enhancing biochemical extraction with bfs-driven knowledge graph embedding approach., in: TEXT2KG/BiKE@ ESWC, CEUR-WS, 2023, pp. 235–243.
[10] P. Fröhlich, J. Gwozdz, M. Jooß, Leveraging chatgpt api for enhanced data preprocessing in natuke., in: TEXT2KG/BiKE@ ESWC, CEUR-WS, Greece, 2023, pp. 244–255.
[11] S. Schmidt-Dichte, I. J. Mócsy, Improving natural product automatic extraction with named entity recognition., in: TEXT2KG/BiKE@ ESWC, CEUR-WS, Greece, 2023, pp. 226–234.
[12] P. Viviurka do Carmo, M. P. Silva Gôlo, J. Gwozdz, E. Marx, R. Marcondes Marcacini, Improving natural product knowledge extraction from academic literature with enhanced pdf text extraction and large language models, in: Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, 2025, pp. 980–987.
[13] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, R. McHardy, Challenges and applications of large language models, arXiv preprint arXiv:2307.10169 (2023).
[14] C. Zhai, Large language models and future of information retrieval: opportunities and challenges, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 481–490.
[15] L. Blecher, G. Cucurull, T. Scialom, R. Stojnic, Nougat: Neural optical understanding for academic documents, arXiv preprint arXiv:2308.13418 (2023).
[16] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, M. V. Gil, S. Miret, C. T. Koch, J. A. Márquez, K. M. Jablonka, From text to insight: large language models for materials science data extraction, arXiv preprint arXiv:2407.16867 (2024).
[17] Z. Abbasiantaeb, Y. Yuan, E. Kanoulas, M. Aliannejadi, Let the llms talk: Simulating human-to-human conversational qa via zero-shot llm-to-llm interactions, in: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 8–17.
[18] Z. Chen, J. Li, P. Chen, Z. Li, K. Sun, Y. Luo, Q. Mao, D. Yang, H. Sun, P. S. Yu, Harnessing multiple large language models: A survey on llm ensemble, arXiv (2025).
[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.
[20] L. Wang, S. Chen, L. Jiang, S. Pan, R. Cai, S. Yang, F. Yang, Parameter-efficient fine-tuning in large language models: a survey of methodologies, Artificial Intelligence Review (2025).
[21] M. Gozzi, F. Di Maio, Comparative analysis of prompt strategies for large language models: Single-task vs. multitask prompts, Electronics 13 (2024) 4712.
[22] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, Advances in neural information processing systems 36 (2023).
[23] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).