Multimodal GraphRAG for the Semantic Querying of Scientific Publications

PhD Thesis Proposal

Database Team – LIP6 – Sorbonne University, Paris

Abstract

This PhD project aims to design a multimodal GraphRAG framework for the semantic querying of scientific literature. The goal is to unify the analysis of content (text, figures, tables) and contextual metadata (citations, benchmarks) to model the complex relationships between information fragments. The methodology is built upon three pillars: the adaptation of multimodal encoders, data alignment within a shared latent space, and the construction of a heterogeneous document graph augmented by external resources. This work aims to produce more accurate and explainable information retrieval tools, thereby promoting scientific reproducibility.

1 Context

Modern scientific publications are inherently multimodal: they combine text, figures, and tables [11, 15], while being embedded in a dense network of bibliographic references. To query these archives, Retrieval-Augmented Generation (RAG) systems, based on Large Language Models (LLMs), are progressively replacing traditional search engines due to their ability to represent the semantic content of articles and infer implicit relationships. However, these approaches are currently reaching their limits, particularly when faced with queries requiring a global and relational understanding of the field.

This limitation becomes particularly visible when moving beyond the document itself (often in PDF format): scientific work is now part of an open empirical ecosystem, linked to internal artifacts (tables, figures, definitions, proofs, bibliography) and external ones (GitHub repositories, cited articles, reference datasets, benchmark platforms). While the textual body provides a narrative overview, these complementary resources are crucial for exhaustive interpretation and the reproducibility of results [11]. In this perspective, GraphRAG (*Graph Retrieval-Augmented Generation*), based on knowledge graphs or documentary structure graphs, models the explicit relationships between entities and concepts, providing LLMs with a richer, more structured, and traceable context [10].

However, current GraphRAG architectures remain largely limited to textual analysis. This thesis proposes to move beyond these fragmented approaches by developing a unified representation framework to analyze and query scientific work holistically, exploiting both internal and external modalities of publications. Drawing on LIP6's expertise in graph modeling, the challenge is to effectively integrate visual and tabular *embeddings* into the core of GraphRAG systems to improve complex question answering and multimodal scientific information retrieval [11, 15].

1.1 State of the Art

Currently, the automated analysis of scientific literature relies on several families of approaches:

- The majority of systems still convert documents into plain text, relying on powerful foundation models (e.g., Nougat [3]). The emergence of Vision-Language Models (VLMs) specialized for science and visual reverse-engineering approaches (e.g., DePlot [16], ChartQA [17]) now allows for the extraction of relevant information from complex figures (diagrams, flowcharts). Nevertheless, this information often remains siloed and disconnected from the overall textual and tabular context of the document [14, 7, 19].

- Scientific archives such as arXiv and OpenAlex [18] are widely exploited for analyzing scientific dynamics [8]. While these archives allow for tracing the evolution of themes on a macroscopic scale [21], they generally treat publications as textual or semantic "black boxes." They accurately model citation topology (who cites whom and when) but ignore the exact content that actually justifies these paradigm shifts.
- An increasing number of conferences require authors to supplement their work with references to GitHub repositories containing the software resources and data necessary for experimental reproducibility. Platforms like Kaggle or *PapersWithCode* (closed in 2025) aimed to centralize model evaluation by linking publications, implementations (source code), and empirical performance on reference datasets (*benchmarks*). These repositories, however, remain under-exploited due to a lack of suitable analysis tools.

Each of these approaches has enabled substantial progress, but their juxtaposition highlights the need for a unifying framework capable of linking these heterogeneous information sources. The recent rise of RAG architectures, popularized by advanced research assistants based on restrictive source-grounding (such as NotebookLM), has proven the effectiveness of these systems in limiting LLM hallucinations on restricted personal corpora [12]. However, these classical approaches historically rely on vector search of isolated textual segments (*chunks*) [9]. While they excel at intra-document factual extraction, they fail to model the macroscopic topology of scientific knowledge and to perform multi-hop reasoning at the scale of an entire archive.

To overcome this conceptual hurdle, GraphRAG has recently emerged and demonstrated its superiority [6] by replacing flat vector databases with knowledge graphs (where nodes represent concepts, authors, or entities, and edges represent their semantic relationships). However, these graphs are currently constructed and queried almost exclusively through text. They ignore the rich and structured information vectors coming from other modalities, which nonetheless constitute the core of scientific argumentation. The primary bottleneck today is no longer the simple ability to extract a specific modality (table, figure, image), but rather the absence of a system capable of harmonizing the *embeddings* of all these modalities within a shared semantic space to feed these graph-based reasoning architectures [11].

Multiple research avenues must be explored to achieve these objectives:

1. Build a reliable document extraction pipeline: article–figure–table segmentation, metadata normalization (captions, sections, cross-references), and the creation of homogeneous documentary units for indexing.
2. Adapt specialized encoders per modality: compare variants of text, table, and vision encoders, with dedicated evaluation protocols (representation quality, robustness to OCR noise, sensitivity to the scientific domain).
3. Learn a shared latent space: test contrastive strategies, projections guided by document structure, and cross-attention mechanisms to preserve both semantic similarity and the specificity of each modality.
4. Design a multimodal document graph: define a typology of nodes and edges (citations, figure/section inclusion, methodological dependencies, code–data links), then study the impact of different weighting schemes on context retrieval.
5. Deploy and evaluate a multimodal GraphRAG: compare the approach to textual and hybrid baselines on questions requiring explicit justification, multi-hop reasoning, and fine traceability of sources.

2 Problem Statement and Scientific Objectives

The central problem of this thesis is formulated as follows: *How can representations from heterogeneous models (text, tables, figures) be semantically aligned and integrated within a unified space to multiply the reasoning capabilities of LLMs via a multimodal GraphRAG architecture?*

To concretely illustrate this need for integration, the target architecture should enable LLMs to answer highly contextual scientific queries, such as:

- "*What are the exact hyperparameter values (Table 2) responsible for the performance drop illustrated in the curve of Figure 4 and discussed in Section 3?*"
- "*Does the attention mechanism formally described in paragraph 4.1 faithfully correspond to the logical blocks in the visual architecture diagram of Figure 1?*"
- "*How do the ablation results of the proposed method (Table 1) translate visually in the scatter plots, and how does this contradict the state-of-the-art hypotheses mentioned in the introduction?*"
- "*How were the experimental limitations identified in Figure 3 of this article overcome in the architectures proposed by the articles that cited it the following year?*"
- "*How do the performances documented in the experimental tables show the progressive paradigm shift from CNNs to Transformers between 2019 and 2023 in this sub-discipline?*"

To achieve this level of reasoning, the work will be structured around three objectives:

1. Identify and adapt the best existing encoding models for complex structures (tables, figures). This will involve validating their ability to preserve topological and visual information on reference tasks to ensure the quality of the input *embeddings* for the system [22, 5].
2. Make inter-modal alignment a primary object of study, strongly coupled with downstream modeling. This objective aims to design, compare, and evaluate different fusion paradigms to identify the optimal architecture based on the topological requirements of the target graphs (e.g., spatial relations for a layout graph vs. causal relations for an experimental graph) [15].
3. Model the documentary graph integrating these modalities as enriched nodes, and deploy the GraphRAG architecture by exploiting heterogeneous graph mechanisms for inter-modal contextualization. The impact of the overall system will be measured on complex scientific understanding use cases using reference datasets [10, 14].

3 Suitability of the Subject with the Supervisory Team's Expertise

The completion of this thesis will rely on a scientific environment aligned with the targeted technological bottlenecks. It will be supervised by Bernd Amann, and co-supervised by Camélia Constantin and Rafael Angarita, within the Database team of the LIP6 laboratory.

The subject aligns directly with the competencies of the LIP6 Database team:

- Citation graph analysis and scientific dynamics [13, 20, 21];
- Representation of structured data (notably tabular) [4, 5];
- Large-scale knowledge graph engineering and semantic enrichment [1, 2].

4 Expected Results and Impact

The expected results of this thesis lie in the creation of an original methodological framework for the semantic alignment of heterogeneous data, capable of transforming static scientific archives into dynamic and interconnected knowledge graphs. Specifically, this work aims to produce a robust extraction pipeline and an innovative GraphRAG architecture enabling LLMs to perform complex "multi-hop" reasoning with increased accuracy and a significant reduction in hallucinations. Beyond technical performance, the outcomes include the development of exploration tools capable of explicitly correlating textual claims with visual (figures) or structured (tables) evidence, thereby providing researchers with fine-grained source traceability, improved answer explainability, and concrete support for scientific reproducibility.

5 Position Context

Supervision Arrangements: The doctoral student will be supervised by two members of the LIP6 Database team. Weekly working meetings will ensure scientific leadership and methodological orientation.

Monitoring of Doctoral Training: The training path will be constructed with the doctoral school, in conjunction with the project's needs: advanced scientific courses, transversal training (ethics, open science), and the development of scientific communication skills.

Research Progress Monitoring: Progress will be structured around semi-annual milestones: literature review, experimental validation, publications, and system integration. Results will be discussed during team-wide progress reviews and consolidated in intermediate reports to ensure scientific coherence, methodological quality, and progression toward the defense.

6 Profile and Skills Sought

The position is intended for a candidate motivated by interdisciplinary research at the interface of natural language processing, computer vision, and knowledge graphs.

- Scientific skills: solid foundations in machine learning, data representation, and experimental evaluation; an interest in multimodal approaches and LLMs is expected.
- Technical skills: good command of Python; experience with deep learning libraries, document processing, and graph databases is appreciated.
- Research methodology: ability to design rigorous experimental protocols, critically analyze results, and document work with a focus on reproducibility.
- Transversal skills: autonomy, initiative, ability to work collaboratively, and scientific communication in French and English (writing, presentations, international exchanges).

7 International Outlook

The project is part of an international dynamic at the intersection of scientific literature mining, multimodal models, and knowledge graphs. The scientific hurdles addressed (inter-modal alignment, traceable reasoning, GraphRAG system evaluation) concern a broad international research community. The targeted results are intended for dissemination in leading international conferences and journals in machine learning, modeling, and data querying.

Academic collaborations will be sought with teams working on scientific document analysis, graph querying, and LLM evaluation in scientific contexts. These collaborations may take the form of occasional co-supervision, research internships, co-publications, and the sharing of datasets or testbeds. The project will also encourage interactions with open science infrastructures and initiatives (bibliometric databases, code, and data repository platforms) to strengthen resource interoperability, artifact dissemination, and the international impact of the contributions.

8 Material Conditions and Specific Security Requirements

Material Conditions: The project will rely on GPU-equipped workstations and the LIP6 Convergence cluster. The software environment will be based on open-source tools for document processing, machine learning, and graph management, with systematic version control of code and experimental configurations.

Specific Security Requirements: The work will comply with institutional rules regarding information system security and data protection: access control to servers and repositories, regular backups, encryption of sensitive data, and traceability of experimental manipulations. Particular attention will be paid to respecting dataset licenses, artifact dissemination constraints (models, code, corpora), and compliance with the ethics and open science principles applicable to the project.

9 Provisional Timeline (3 years)

- **Year 1: Model Selection and Exploration of Fusion Methods**
  - Literature review on multimodal alignment and GraphRAG.

  - Evaluation, benchmarking, and familiarization with SOTA encoders for tables and figures.
  - Formal definition of the alignment problem and characterization of target graph topologies.
- **Year 2: Comparative Evaluation and Unified Semantic Space**
  - Implementation of several multimodal fusion strategies (e.g., contrastive approaches, projection networks, attention mechanisms).
  - Cross-benchmarking of these alignment methods based on different types of graphs to be constructed (knowledge graphs, structure graphs).
  - Validation on inter-modal information retrieval tasks.
- **Year 3: Heterogeneous GraphRAG and Global Validation**
  - Construction of the multimodal documentary graph and implementation of the GraphRAG pipeline.
  - Global evaluation of the ecosystem on benchmarks (e.g., *Multimodal ArXiv* https://mm-arxiv.github.io/).
  - Manuscript writing and defense.

References

[1] Yuhe Bai, Camélia Constantin, and Hubert Naacke. Leiden-fusion partitioning method for effective distributed training of graph embeddings. In *Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD*, volume 14947, pages 366–382, 2024.

[2] Yuhe Bai, Modou Gueye, and Hubert Naacke. Selective multi-hop type-aware enhancement for context-limited knowledge graph entity typing. In *IEEE International Conference on Big Data*, pages 1–10, 2025.

[3] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023.

[4] Allaa Boutaleb, Alaa Almutawa, Bernd Amann, Rafael Angarita, and Hubert Naacke. Hearts: Hypergraph-based related table search. In *ELLIS workshop on Representation Learning and Generative Models for Structured Data (RLGMSD)*, page 3, Amsterdam, NL, 2025. Poster.

[5] Allaa Boutaleb, Bernd Amann, Hubert Naacke, and Rafael Angarita. Something's Fishy In The Data Lake: A Critical Re-evaluation of Table Union Search Benchmarks. In *4th Table Representation Learning Workshop @ ACL 2025*, Vienna, Austria, 2025.

[6] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

[7] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. WikiMixQA: A multimodal benchmark for question answering over tables and charts. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24941–24958, Vienna, Austria, July 2025. Association for Computational Linguistics.

[8] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. *Science*, 359(6379):eaao0185, 2018.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[10] Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. RAG-Anything: All-in-One RAG Framework. *arXiv preprint arXiv:2510.12323*, 2025.

[11] Feng Jiang, Kuang Wang, and Haizhou Li. Bridging research and readers: A multi-modal automated academic papers interpretation system. *arXiv preprint arXiv:2401.09150*, 2024.

[12] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.

[13] Ke Li, Hubert Naacke, and Bernd Amann. An analytic graph data model and query language for exploring the evolution of science. *Big Data Research*, 26:100247, 2021.

[14] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[15] Zirui Li, Siwei Wu, Xingyu Wang, Yi Zhou, Yizhi Li, and Chenghua Lin. Docmmir: A framework for document multi-modal information retrieval. *arXiv preprint 2505.19312*, 2025.

[16] Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399, 2023.

[17] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[18] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022.

[19] Ruyi Qi, Zhou Liu, and Wentao Zhang. DataCross: A Unified Benchmark and Agent Framework for Cross-Modal Heterogeneous Data Analysis. https://arxiv.org/abs/2601.21403v1, January 2026.

[20] Hamed Rahimi, Hubert Naacke, Camélia Constantin, and Bernd Amann. ANTM: an aligned neural topic model for exploring evolving topics. In *Journées Bases de Données Avancées (BDA)*, page 11, Montpellier, France, 2023.

[21] Hamed Rahimi, Hubert Naacke, Camélia Constantin, and Bernd Amann. ATEM: A topic evolution model for the detection of emerging topics in scientific archives. In *12th International Conference on Complex Networks and their Applications*, page 10, Menton, France, 2023.

[22] Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. Representations for question answering from documents with tables and text. In *16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021.