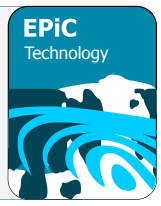
EPiC Series in Technology

Volume 3, 2026, Pages 50–56

Proceedings of AI for Accelerated Research Symposium




# Agentic Knowledge Graph Traversal in Protein-Protein Relation Grounding

Gabriel K. Reder[1], Carl Collins[2], Larisa Soldatova[2], and Ross D. King[1,3] *

[1] University of Cambridge, Cambridge CB3 0AS, United Kingdom.
[2] Goldsmiths, University of London, London SE14 6AD, United Kingdom.
[3] Chalmers University of Technology, 417 56 Göteborg, Sweden.
gr513@cam.ac.uk

**Abstract**

Automated semantic knowledge extraction from scientific literature promises to open vast quantities of scientific knowledge to formal analysis and computationally-driven discovery. In this work we investigate the promise of Large Language Model (LLM) agents in extracting structured knowledge from biomedical texts, specifically for grounding protein-protein interaction (PPI) relations to terms in the PSI-MI ontology of molecular interactions. While LLMs excel at summarization, they struggle to interface with structured knowledge representations. We equipped agents with various knowledge graph interaction strategies and measured their PPI grounding performance. Our central finding is that PageRank-guided traversal, a method rooted in graph topology, consistently outperforms embedding-based approaches such as retrieval augmented generation (RAG) and top-down traversal strategies including breadth-first search (BFS), depth-first search (DFS), and local greedy search in extracting knowledge previously missed by human curators. Our initial results indicate that the structure of a well-curated knowledge base is itself a powerful source of information, an underutilized principle in current agentic knowledge base interaction methods.

## 1 Introduction

We are at a key moment in scientific history. Advances in AI may allow the transformation of scientific papers from ambiguous natural language to semantically clear open-access formal knowledge. Such translation would revolutionize scientific discovery: papers would become more comprehensible for both humans and AIs with hypotheses, methods, results, and conclusions expressed explicitly and unambiguously, enabling direct comparison, systematic analysis, and reproducibility. Building on this foundation, AI reasoning can then transform and accelerate scientific discovery, yielding verifiably correct and hallucination-free conclusions [21].

Here we study this possibility through a well-defined challenge in biological text mining: the extraction of protein-protein interactions (PPIs). Current automated systems are proficient at

*This work was supported by grants from the UK Engineering and Physical Sciences Research Council (EPSRC) [EP/X032418/1, EP/X033740/1]

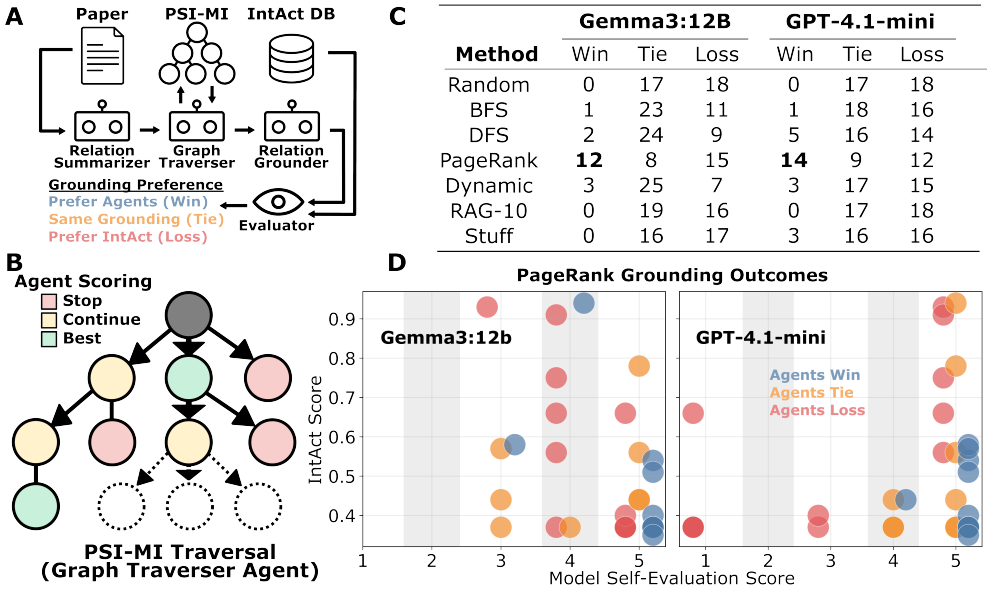
identifying protein entities in text, however grounding interactions between protein entities to a standardized vocabulary like the Proteomics Standards Initiative-Molecular Interaction (PSI-MI) ontology remains unreliable [15, 11]. Large language models (LLMs) excel at summarization and unstructured text processing, yet grounding inherently involves structured knowledge, a difficult interface for LLMs. Recent approaches, namely retrieval augmented generation (RAG), have been dominated by inference-time embedded content matching [13]. Yet knowledge bases encode valuable information in their structure which can be lost in such approaches. Scientific ontologies are not flat lists of terms; they are meticulously curated structures with expert-designed hierarchy and connectivity [5, 11].

This work investigates the role of knowledge structure and its relation to content in the context of PPI relation grounding. We developed a three-stage LLM agent flow for grounding PPIs from full-text articles to the PSI-MI ontology, equipping general-use LLMs with various knowledge graph interaction strategies and testing their performance on a small set of biomedical papers. Our results indicate that the topology-centric PageRank method outperformed embedding-based and other inference-time search methods such as breadth-first, depth-first, and greedy search in grounding relations. We also find that LLM inference-time term self-evaluation scores correlate well with grounding quality.

## 2 Related Work

The PSI-MI ontology standardizes molecular interaction data using a hierarchically structured vocabulary [16]. The IntAct database, which uses expert curators to map literature evidence to PSI-MI terms, serves as a gold-standard data source for training and evaluation [3]. Earlier PPI extraction work using traditional natural language processing techniques [18, 4] has lately been superseded by the use of transformer-based deep models, such as BioBERT [8] to directly predict relation triplets from an input text [26]. Recent studies have shown that general-purpose LLMs and task-specific prompting and tool calling can achieve competitive performance against fine-tuned BERT models without additional training [15]. We build on this observation here through the investigation of inference-time search methods for LLM knowledge base interaction.

LLM agent systems assign specific subtasks to individual LLM instances working together to achieve a larger goal [22]. Here, "agent" refers to an entity capable of perceiving its environment and dynamically making decisions based on runtime sensing and input [23] - in this case an LLM instance equipped with instructions, code-calling capabilities, and a goal. Recent agent systems have tended to rely on similarity-based embedding methods, such as RAG, during dynamic decision making. In our work, we use graph search as a form of agent "reasoning" during grounding as has been suggested for agent knowledge graph tasks [10, 12]. Proposed approaches have included using the LLM for graph traversal algorithm proposal [9], follow-up question generation [24], and action plan self-reflection [25]. Our work here compares several agent knowledge graph traversal strategies: a greedy local search (dynamic), classical static search (BFS/DFS), and a structure-based search (PageRank) which queues nodes based on connectivity. PageRank, the algorithm behind Google search [1], has been applied in biology to characterize members of core protein interaction networks and identify universal concepts within biomedical ontologies [6, 2, 17].




| Method | Gemma3:12B Win | Tie | Loss | GPT-4.1-mini Win | Tie | Loss |
|---|---|---|---|---|---|---|
| Random | 0 | 17 | 18 | 0 | 17 | 18 |
| BFS | 1 | 23 | 11 | 1 | 18 | 16 |
| DFS | 2 | 24 | 9 | 5 | 16 | 14 |
| PageRank | **12** | 8 | 15 | **14** | 9 | 12 |
| Dynamic | 3 | 25 | 7 | 3 | 17 | 15 |
| RAG-10 | 0 | 19 | 16 | 0 | 17 | 18 |
| Stuff | 0 | 16 | 17 | 3 | 16 | 16 |

Figure 1: (A) The LLM agent flow for relation grounding between two protein entities from a paper. (B) The Graph Traverser agent is tested with various strategies for searching the PSI-MI ontology graph for the best-fit relation term. PSI-MI terms are scored at inference-time. (C) Evaluation of normalized relation labels from a test set of 10 Europe PMC papers with associated IntAct entries. Wins occur when the evaluator models (O3 and Gemini-2.5-pro) preferred the agents' label to the IntAct label. Ties occur when the agents' label is the same as the IntAct label. (D) IntAct human label confidence scores vs agent self-evaluation scores for the PageRank agent runs.

# 3 Methods

The LLM agent flow (Figure 1A), consists of three sequential tasks. First "Relation Summarizer" (RS) takes the paper text and a protein entity pair as input, producing a plain-text summary of the pairwise protein interaction together with evidence quotes from the paper via a document chunk refinement chain. "Graph Traverser" (GT) uses this free-text relation summary to score terms in the PSI-MI ontology term graph, evaluating term definitions according to its equipped node traversal strategy (Figure 1B). "Relation Grounder" (RG) then picks the best term from only those scored with the max score seen during traversal. LangChain [19] and DSPy [7] were used to implement the agent flow.

Five graph traversal strategies were tested. GT evaluates PSI-MI nodes one by one, scoring them from 1 to 5 based on the free-text interaction summary. The BFS, DFS, and dynamic searches begin at the PSI-MI root term (MI:0190 'interaction type'). PageRank orders the search queue by descending node connectivity [1]. The random strategy searches the graph in a random order. The dynamic strategy is a greedy local search - the agent adds to the search queue any children of the current node that exceed a continuation evaluation score threshold (3/5). GT stops its traversal when any of the following conditions is met: (1) all graph nodes have been evaluated, (2a) the agent has not found a better-scoring candidate node in the last 10 node evaluations (BFS, DFS, and random), or (2b) the agent has no nodes left in the node queue

(dynamic). BFS, DFS, and random were given a 5-node lookahead after hitting the stopping criterion. We also tested a RAG strategy in which GT scores only the top 10 terms retrieved based on embedded definitions. The 'snowflake-arctic-embed2' and 'text-embedding-3-small' embedding models were used for Gemma3 and GPT-4.1-mini runs respectively. A "stuff" strategy was also tested in which RG is presented with every PSI-MI term and definition as grounding options. We note that this "stuff" strategy corresponds to an LLM-only baseline as no graph traversal is performed in this case.

We conducted tests on a sample set of 10 randomly selected Europe PMC (EPMC) open-access papers with linked human-curated IntAct PPI annotations. We tested with OpenAI GPT-4.1-mini and Google Gemma3:12B models as we sought to evaluate graph-based traversal strategies in models without additional pre-configured "thinking". OpenAI O3 and Google Gemini-2.5-pro were used as models for the evaluation. The evaluator models were presented with the entire paper text, agent-proposed PSI-MI term, and the IntAct PSI-MI term for a given protein pair and prompted to choose which term it preferred. Results (Figure 1C and D) were compiled for the cases in which the reasoning models agreed with each other. Comparisons were made for the most granular IntAct label for a given protein entity pair. Cases in which the reasoning models preferred the agent-produced term were considered an "agents win", cases where the agent and IntAct terms match were considered an "agents tie", and cases in which the reasoning models preferred the IntAct term were considered an "agents loss". The PMCIDs, agent codebase, intermediate outputs, and results are available in the Zenodo repository (https://doi.org/10.5281/zenodo.17235315).

# 4 Results and Discussion

Figure 1C shows performance across traversal methods for our test set. Most notably, PageRank consistently outperformed all other traversal methods in producing novel groundings (wins) over IntAct during evaluation and was the only strategy to significantly improve win rate over the 'stuff' baseline ($p=0.002$ for GPT-4.1-mini, $p=0.0001$ for Gemma3:12B by a one-sided Fisher's exact test). On manual inspection of the results, we found that PageRank was less sensitive to spurious term suggestion than other methods, notably BFS, DFS, and dynamic traversal which mimic the humanlike process of browsing an ontology graph from the top-down. PageRank had the lowest number of unique suggested terms across all groundings. In other words, PageRank constantly suggests the same highly-connected nodes because they appear earlier in its search queue. These terms are highly-connected because they capture the most central and important concepts in the ontology. None of the other traversal methods capture this aspect of knowledge graph topology, and given these initial results it appears this may be missed during human expert grounding as well. The failure of RAG-10 to produce any wins for either LLM model is curious but makes sense when considering the brittleness of vector retrieval as opposed to graph traversal. As shown previously, vector-based retrieval can only identify good candidate ontology terms when the inference-time query closely mimics the term definition [14], yet RT's free-text summaries provide no guarantees or bias toward such alignment. Indeed, it makes sense that the RAG-10 results so closely mirror the Random traversal strategy - the top 10 terms retrieved via RAG appear to be effectively random.

Human and agent confidence scores provide additional relevant grounding information. The agents self-score their groundings and the IntAct database contains human confidence scores for each annotation - the MI score [20]. Figure 1D shows the relationship between the agent self-evaluation score and the IntAct MI score in the context of agent wins, losses, and ties for the PageRank traversal runs. We note that for both LLM models, the lower right quadrant of

the resulting space is where the agent wins are concentrated. In other words, these metrics can serve as an a priori filter for potentially novel knowledge extraction generated from LLM agents. We also note that the choice of traversal (reasoning) strategy seemed to impact performance more than choice of LLM model. We deliberately chose to test modular task-specific reasoning strategies in this work and were pleased to see results translate neatly across LLMs. We hope this serves to emphasize the need for targeted and methodical reasoning strategies for agentic systems to reach their full potential. Most importantly we note that the size of our pilot study dataset of 10 papers is small, constrained by available GPU time and provider compute credits. We look forward to future work able to study such methods on a larger scale.

# 5 Conclusion

This work presents a pilot study of generative agent strategies for interacting with structured knowledge bases, focusing on the automated extraction of protein-protein interaction (PPI) networks from unstructured literature. Our findings suggest that a topology-driven approach, using PageRank for knowledge graph traversal, outperforms top-down search and embedding-based retrieval methods. Our key insight is that knowledge bases efficiently encode expert consensus not just in their terms, but in their very structure. High-centrality concepts correspond to the foundational ideas that domain experts use as anchors for annotation. By leveraging this inherent structure, we may be able to influence the behavior of LLM interaction towards domain-specific expertise and novel scientifically-grounded outputs at inference-time. Importantly, LLM systems may extract knowledge missed by human curators, accelerating the process of discovery. We see such informed knowledge extraction as crucial in the vision of translating scientific literature into explicit semantically clear knowledge.

# References

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[2] Anila Sahar Butt, Armin Haller, and Lexing Xie. Dwrank: Learning concept ranking for ontology search. *Semantic Web*, 7(4):447–461, 2016.

[3] Noemi Del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, Livia Perfetto, Karyn How, Prashansa Ratan, Gautam Shirodkar, et al. The intact database: efficient access to fine-grained molecular interaction data. *Nucleic acids research*, 50(D1):D648–D653, 2022.

[4] Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relex—relation extraction using dependency parse trees. *Bioinformatics*, 23(3):365–371, 2007.

[5] Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, et al. Intact: an open source molecular interaction database. *Nucleic acids research*, 32(suppl_1):D452–D455, 2004.

[6] Gábor Iván and Vince Grolmusz. When the web meets the cell: using personalized pagerank for analyzing protein interaction networks. *Bioinformatics*, 27(3):405–407, 2011.

[7] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

[8] J. Lee et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[9] Renjie Liu, Haitian Jiang, Xiao Yan, Bo Tang, and Jinyang Li. Polyg: Effective and efficient graphrag with adaptive graph traversal. *arXiv preprint arXiv:2504.02112*, 2025.

[10] Chuangtao Ma, Yongrui Chen, Tianxing Wu, Arijit Khan, and Haofen Wang. Large language models meet knowledge graphs for question answering: Synthesis and opportunities. *arXiv preprint arXiv:2505.20099*, 2025.

[11] Sandra Orchard, Mais Ammari, Bruno Aranda, Lionel Breuza, Leonardo Briganti, Fiona Broackes-Carter, Nancy H Campbell, Gayatri Chavali, Carol Chen, Noemi Del-Toro, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363, 2014.

[12] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

[13] Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. *arXiv preprint arXiv:2408.08921*, 2024.

[14] Gabriel K Reder, Carl Collins, Abbi Abdel Rehim, Larisa Soldatova, and Ross D King. Llm-retrieval based scientific knowledge grounding. In *Joint Proceedings of the ESWC 2025 Workshops and Tutorials co-located with 22nd Extended Semantic Web Conference (ESWC 2025)*, volume 3977 of *CEUR Workshop Proceedings*, 2025.

[15] Hasin Rehana, Nur Bengisu Çam, Mert Basmaci, Jie Zheng, Christianah Jemiyo, Yongqun He, Arzucan Özgür, and Junguk Hur. Evaluating gpt and bert models for protein–protein interaction identification in biomedical text. *Bioinformatics Advances*, 4(1):vbae133, 2024.

[16] M Sivade, Diego Alonso-López, Mais Ammari, Glyn Bradley, Nancy H Campbell, Arnaud Ceol, Gianni Cesareni, Colin Combe, Javier De Las Rivas, Noemi Del-Toro, et al. Encompassing new use cases-level 3.0 of the hupo-psi format for molecular interactions. *BMC bioinformatics*, 19(1):134, 2018.

[17] Luis E Solano, Nicholas M D'Sa, and Nikolas Nikolaidis. Prrgo: a tool for visualizing and mapping globally expressed genes in public gene expression omnibus rna-sequencing studies to pagerank-scored gene ontology terms. *bioRxiv*, 2024.

[18] Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLoS computational biology*, 6(7):e1000837, 2010.

[19] Oguzhan Topsakal and Tahir Cetin Akinci. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences*, volume 1, pages 1050–1056, 2023.

[20] Jose M Villaveces, Rafael C Jimenez, Pablo Porras, Noemi del Toro, Margaret Duesbury, Marine Dumousseau, Sandra Orchard, H Choi, Peipei Ping, NC Zong, et al. Merging and scoring molecular interactions utilising existing community standards: tools, use-cases and a case study. *Database*, 2015:bau131, 2015.

[21] H. Wang et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[22] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

[23] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 2025.

[24] Zukang Yang, Zixuan Zhu, and Xuan Zhu. Curiousllm: Elevating multi-document question an-

swering with llm-enhanced knowledge graph reasoning. *arXiv preprint arXiv:2404.09077*, 2024.

[25] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[26] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, 56(11):1–39, 2024.