PRISM: Problem Discovery via Structural Motifs in Knowledge Graphs

Anonymous Authors[1]

Abstract

While recent advancements in artificial intelligence have automated various scientific workflows, existing discovery systems primarily explore variations of user-specified inputs rather than autonomously identifying critical gaps in the broader literature. To overcome this reliance on predefined prompts, we propose a framework that formulates research problem discovery as the systematic detection of structural gaps within scientific knowledge graphs. Our approach decomposes scientific papers into structured components to build a heterogeneous graph, enabling structural-level reasoning across methods, evaluated tasks, and failure conditions. The system queries this graph for recurring inconsistencies and underexplored patterns, formulating these discoveries into structured problem objects. Through an evolutionary refinement process and localized subgraph reasoning, our framework generates novel, feasible research problems that demonstrate high traceability and robust grounding in existing scientific evidence.

1. Introduction and Motivation

Automated scientific discovery with artificial intelligence has evolved from tools that assist isolated research tasks to systems capable of supporting hypothesis generation, literature-grounded writing, and even end-to-end research workflows (Lu et al., 2024; Zhang et al., 2024b; Yamada et al., 2025; Gottweis et al., 2025; Baek et al., 2025). Recent efforts such as The AI Scientist (Lu et al., 2024) and AI Scientist-v2 (Yamada et al., 2025) aim to automate the full research pipeline, while systems like AI Co-Scientist (Gottweis et al., 2025) and ResearchAgent (Baek et al., 2025) explore hypothesis generation and literature-grounded ideation using multi-agent reasoning and structured representations.

Despite this progress, most existing systems are conditioned on a specific prompt, paper, or predefined objective. As a result, they primarily explore variations of the given input rather than identifying research opportunities that arise from gaps or inconsistencies in the broader literature. For example, AI Scientist-v2 begins from broad topical prompts and relies on human selection among generated outputs, while AI Co-Scientist depends on user-specified goals and operates under constraints such as limited access to data and incomplete integration of structured knowledge sources. These design choices limit the ability of current systems to discover problems that are not already specified or implied by the input.

Classical literature-based discovery (LBD) (Borrego et al., 2025) and knowledge graph (Xiong et al., 2024) approaches provide useful components but do not fully address this limitation. LBD focuses on uncovering connections across disparate literatures, often formulated as link prediction or graph completion, while knowledge graphs represent entities and their relationships in structured form and support reasoning over scientific artifacts. However, neither paradigm explicitly represents research problems themselves. In practice, a research problem involves identifying a gap, specifying a target, and considering a plausible approach. Without modeling these elements, existing systems remain limited in their ability to move from organizing knowledge to formulating well-defined research questions. This motivates systems that construct structured representations of scientific fields and use them to identify and formulate research problems systematically.

2. System Overview and Method

We formulate research problem discovery as the identification of structured gaps in scientific knowledge, rather than as direct idea generation. The system constructs a structured representation of a research field and identifies missing or inconsistent patterns that indicate underexplored research directions.

At a high level, the pipeline consists of four stages: (i) structured extraction from literature, (ii) scientific knowledge graph construction, (iii) gap detection via pattern queries, and (iv) problem formulation and refinement.

Given a target domain, we first retrieve a corpus of

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Submitted to the AI for Science workshop (ICML 2026)

papers from top-tier venues. Each paper is then decomposed into structured components, including methods, tasks, assumptions, results, and limitations (Zhang et al., 2024a; Li et al., 2025), using an LLM-based extraction module constrained by a predefined schema. This produces tuples such as $(\textit{method}, \textit{evaluated-on}, \textit{task})$ and $(\textit{method}, \textit{fails-under}, \textit{condition})$.

These tuples are aggregated into a heterogeneous knowledge graph, where nodes represent entities such as methods, datasets, and claims, and edges encode relations such as *evaluated-on*, *fails-under*, and *contradicts*. The graph enables reasoning across papers at a structural level.

Research problems are then identified by executing structured queries over this graph to detect recurring patterns. In particular, the system searches for motifs such as: (i) multiple methods failing under a shared condition, (ii) limitations repeatedly reported but not addressed, and (iii) conflicting claims under comparable settings (Ghafarollahi & Buehler, 2025; Borrego et al., 2025; Ofer et al., 2026). Each pattern is scored based on frequency, diversity of supporting papers, and consistency of evidence.

To enable focused reasoning, localized subgraphs are dynamically extracted while preserving links to the global graph. The system also explicitly models failure signals, treating limitations and negative results as explicit entities, which provide strong indicators of actionable gaps.

Each detected pattern is further converted into a structured problem object of the form $(\textit{gap}, \textit{target}, \textit{scope}, \textit{mechanism})$, grounded in supporting evidence. Candidate problems are then refined through an evolutionary refinement process, which modifies scope, combines compatible patterns, and filters candidates using a critic model based on novelty and feasibility.

Example. In one completed run using Qwen3.5-9B, our system retrieved 25 papers from ICML, ICLR, NeurIPS, ACL, and EMNLP on in-context learning and prompt robustness and constructed a knowledge graph with 905 nodes and 956 edges. The extracted tuples consistently showed that ICL methods are sensitive to demonstration ordering (Xiang et al., 2024), vulnerable to semantically equivalent permutation attacks—with attacks sometimes dropping LLaMA-3 performance by nearly 20 points (Chen et al., 2025)—and fail under input–label or task-distribution shifts (Kossen et al., 2023; Wang et al., 2024; Goddard et al., 2025). Although methods such as PEARL (Chen et al., 2025) and prompt embedding clustering (Pham et al., 2025) improve permutation robustness or demonstration selection, they do not fully resolve the need for stable reasoning across prompt permutations and shifted task distributions. This exposes a clear gap: existing ICL techniques improve average performance but lack guarantees of invariant reasoning and robustness across settings. Thus, the target problem is to design ICL methods whose inference remains invariant across semantically equivalent prompts and robust to distributional shifts, perhaps by extending or combining techniques like PEARL's robust optimization and clustering-based selection.

3. Evaluation and Feasibility

We propose evaluating the system along three axes: novelty, feasibility, and potential impact of the discovered research problems (Baek et al., 2025; Qiu et al., 2025). For a given domain (e.g., bandits, LLM reasoning, multimodal learning), the system generates structured problem objects that are assessed by domain experts through blinded review. Each problem is evaluated based on (i) novelty relative to existing literature, (ii) feasibility of investigation, and (iii) potential to advance the field. Reviewer agreement and ranking consistency are used to assess evaluation reliability.

To contextualize performance, we compare against two classes of baselines. The first comprises LLM-based idea generation systems, which generate proposals conditioned on prompts without explicit structural grounding (Baek et al., 2025). The second comprises literature-based discovery and knowledge graph methods, which focus on link prediction or cross-domain connections but do not explicitly formulate research problems (Xiong et al., 2024; Oarga et al., 2026). The evaluation focuses on whether the proposed system produces problems that are more grounded, non-redundant, and actionable under comparable generation budgets.

In addition to expert evaluation, we measure properties that reflect grounding in the literature, including evidence density, defined as the number and diversity of supporting elements in the graph, and traceability, defined as the ability to reconstruct the reasoning path from the graph to the final problem statement (Oarga et al., 2026). These metrics provide a quantitative view of how well the system connects structured knowledge to problem formulation.

A successful system should consistently identify problems that are both novel and well-supported by evidence, while avoiding redundancy across generated outputs. In particular, we aim to assess whether structural gap detection leads to more coherent and actionable problem formulations compared to prompt-based generation.

The system builds on existing capabilities in LLM-based information extraction and retrieval, with scalable retrieval enabling incremental graph construction (Zhang et al., 2024a). Although the global graph may grow with the literature, reasoning is restricted to bounded subgraphs, ensuring computational tractability. The pipeline is modular and parallelizable, allowing practical deployment on standard GPU resources.

References

Baek, J., Jauhar, S. K., Cucerzan, S., and Hwang, S. J. Researchagent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, 2025.

Borrego, A., Dessì, D., Ayala, D., Hernández, I., Osborne, F., Recupero, D. R., Buscaldi, D., Ruiz, D., and Motta, E. Research hypothesis generation over scientific knowledge graphs. *Knowledge-Based Systems*, 315:113280, 2025.

Chen, L., Shen, L., Deng, Y., Zhao, X., Liang, B., and Wong, K.-F. Pearl: Towards permutation-resilient llms. *arXiv preprint arXiv:2502.14628*, 2025.

Ghafarollahi, A. and Buehler, M. J. Sciagents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22): 2413523, 2025.

Goddard, C., Smith, L. M., Ngampruetikorn, V., and Schwab, D. J. When can in-context learning generalize out of task distribution? *arXiv preprint arXiv:2506.05574*, 2025.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Kossen, J., Gal, Y., and Rainforth, T. In-context learning learns label relationships but is not conventional learning. *arXiv preprint arXiv:2307.12375*, 2023.

Li, S., Sadekar, A., Self, N., Su, Y., Andersland, L., Chaplin, M., Zhang, A., Yang, H., Henderson, J. B., Wigginton, K., et al. Exploring llms for scientific information extraction using the sciex framework. *arXiv preprint arXiv:2512.10004*, 2025.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Oarga, A., Hart, M., Bran, A. M., Lederbauer, M., and Schwaller, P. Scientific knowledge graph and ontology generation using open large language models. *Digital Discovery*, 5(3):1269–1279, 2026.

Ofer, D., Linial, M., and Shahaf, D. Interfeat: a pipeline for finding interesting scientific features. *Scientific Reports*, 2026.

Pham, K., Le, H., Ngo, M., and Tran, T. Rapid selection and ordering of in-context demonstrations via prompt embedding clustering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Qiu, Y., Zhang, H., Xu, Z., Li, M., Song, D., Wang, Z., and Zhang, K. Ai idea bench 2025: Ai research idea generation benchmark. *arXiv preprint arXiv:2504.14191*, 2025.

Wang, Q., Wang, Y., Wang, Y., and Ying, X. Can in-context learning really generalize to out-of-distribution tasks? *arXiv preprint arXiv:2410.09695*, 2024.

Xiang, Y., Yan, H., Gui, L., and He, Y. Addressing order sensitivity of in-context demonstration examples in causal language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6467–6481, 2024.

Xiong, G., Xie, E., Shariatmadari, A. H., Guo, S., Bekiranov, S., and Zhang, A. Improving scientific hypothesis generation with knowledge grounded large language models. *arXiv preprint arXiv:2411.02382*, 2024.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.

Zhang, Q., Chen, Z., Pan, H., Caragea, C., Latecki, L. J., and Dragut, E. Scier: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13083–13100, 2024a.

Zhang, Y., Chen, X., Jin, B., Wang, S., Ji, S., Wang, W., and Han, J. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8783–8817, 2024b.