Not All Context Helps: Isolating Graph Expansion and Community Summaries in Scientific RAG

Filip Nový and Tariq Yousef
Department of Mathematics and Computer Science
University of Southern Denmark
finov24@student.sdu.dk yousef@sdu.dk

Abstract

Retrieval-augmented generation (RAG) systems for scientific literature face a fundamental tension: richer context does not always produce better answers. We present AlgaeBot, a GraphRAG system evaluated on a corpus of 879 peer-reviewed articles from the journal *Algae*.[1] AlgaeBot combines dense vector retrieval with two graph-based augmentation mechanisms: entity-level graph expansion and community-level thematic summaries, where community summaries are generated through a hierarchical summarization pipeline. In a controlled 2×2 ablation across 203 evaluation questions using RAGAS, we find that graph expansion alone is actively harmful to faithfulness ($\Delta = -0.021$) and context precision ($\Delta = -0.084$), while community summaries alone yield small but consistent improvements in faithfulness, answer relevancy, and context recall. The combined system achieves the highest faithfulness ($\Delta = +0.039$), exceeding the additive effect of the two mechanisms in isolation, though at a substantial cost to context precision. We argue that this trade-off reflects retrieval efficiency rather than answer quality. Results suggest that community summaries provide thematic grounding that allows entity-level graph expansion to contribute positively rather than introduce noisy relational context.

1 Introduction

Scientific literature corpora present a distinctive retrieval challenge: knowledge is relational by nature, and meaningful answers often require connecting concepts scattered across many sources. Standard retrieval-augmented generation (RAG) (Lewis et al., 2021), which retrieves the most semantically similar text chunks for a given query, handles factual lookups well but struggles with this relational and thematic structure.

This work was conducted as part of the AlgaeProBANOS project[2](Kusnick and Yousef, 2026). One outcome of the project is AlgaeBot, a scientific question-answering system designed to help researchers and practitioners explore the rapidly growing body of phycological literature. The present paper focuses on one core component of AlgaeBot, namely the retrieval architecture, and investigates how graph expansion and community summaries individually and jointly affect the performance of scientific GraphRAG systems.

GraphRAG (Edge et al., 2025) addresses this by augmenting retrieval with a knowledge graph index (**?**), offering two complementary mechanisms: local graph expansion, which injects entity-level relational triplets into the generation context, and community summaries, which provide pre-generated thematic descriptions of entity clusters detected via community detection (Traag et al., 2019). While prior work has observed performance benefits from combining these mechanisms (Lin et al., 2026), they have not been isolated as independent factors in a controlled evaluation. It therefore remains unclear whether their co-occurrence produces effects beyond additive combination, and which mechanism drives observed improvements. Moreover, existing approaches synthesize community summaries directly from raw source text, incurring substantial indexing costs that limit scalability.

To address these questions, we evaluate AlgaeBot on a corpus of 879 peer-reviewed articles from the journal *Algae*. AlgaeBot introduces a lightweight two-stage summarization pipeline in which individual documents are first compressed using a small local language model before community-level synthesis. This substantially reduces indexing costs and context-window requirements compared with direct full-text summarization. We evaluate four combinations of graph expansion and community

[1] https://www.e-algae.org/

[2] https://algaeprobanos.eu/

summaries in a controlled 2×2 ablation across 203 evaluation questions using RAGAS (Es et al., 2025). Our analysis reveals that graph expansion alone can reduce retrieval quality, whereas its combination with community summaries produces the strongest overall faithfulness.

2 Related Work

RAG (Lewis et al., 2021) has become the standard approach for grounding LLM outputs in external knowledge, retrieving relevant document chunks via dense vector similarity before generation. While effective for single-hop factual queries, standard RAG struggles with scientific corpora where knowledge is relational and distributed across documents, motivating approaches that precompute multi-level summaries for richer context (Sarthi et al., 2024; Edge et al., 2025).

Edge et al. (2025) proposed GraphRAG, which builds a knowledge graph index via LLM-based entity extraction, applying Leiden community detection (Traag et al., 2019) to generate hierarchical community summaries for global sensemaking queries. Their system treats local entity search and global community search as separate retrieval modes rather than evaluating their interaction. KG-augmented RAG approaches more broadly have shown that injecting entity-level graph triplets can introduce irrelevant relational context, particularly in dense domain-specific corpora (Lin et al., 2026).

Lin et al. (2026) propose FusionGraphRAG for geriatric medical QA, combining fine-grained entity anchoring with Leiden community summaries through a dual-granularity agent, demonstrating improvements over naive RAG and graph-only baselines. However, both mechanisms are integrated by architectural design, leaving their independent contributions unquantified. To our knowledge, no prior work has isolated the individual and combined effects of graph expansion and community summaries as independent retrieval augmentation mechanisms through a controlled ablation with statistical validation.

We evaluate using RAGAS, a reference-free framework providing metrics for faithfulness, answer relevancy, context precision, and context recall, enabling systematic comparison across retrieval conditions without requiring human-annotated ground truth.

3 System and Experimental Design

3.1 Corpus and Knowledge Graph

Beyond its availability within the AlgaeProBANOS project, this corpus is a deliberate choice of testbed: phycological literature is relationally dense, with questions naturally spanning typed entities distributed across many papers, which is precisely the retrieval regime in which graph-based augmentation is expected to help. Restricting the corpus to a single peer-reviewed journal keeps it topically coherent while retaining this relational structure, providing a controlled setting for isolating the effects of our two augmentation mechanisms.

Our corpus consists of 879 peer-reviewed articles from the journal *Algae*, covering phycological research published between 1986 and 2024. Documents were chunked using a recursive strategy, which achieved the highest HOPE score in preliminary experiments (Brådland et al., 2025). [3] This produced approximately 32k chunks, which were embedded with `bge-base-en-v1.5` (Xiao et al., 2024) and stored in ChromaDB [4] for retrieval.

A knowledge graph was constructed in Neo4j[5] from per-chunk extractions produced by DeepSeek-V3 with schema enforcement via the Instructor library: each chunk was processed independently against a Pydantic schema of six domain entity types (AlgalTaxon, Environment, Method, ChemicalCompound, GeneticMarker, and Application) and seven relation types, with each relation carrying a confidence score and relations below 0.5 discarded at ingestion. The full extraction consumed roughly 67M tokens (approximately $8) with only 3 of 31,738 chunks failing. Entity resolution merged abbreviated taxon names and annotated taxa against the GBIF backbone, consolidating the graph to approximately 60k unique entities and over 100k edges.

Community detection was performed using the Leiden algorithm (Traag et al., 2019) at resolution 8.0, yielding 1,130 non-singleton communities. Summaries were generated in two stages: each document was first summarized independently using Gemma-3-1B, after which DeepSeek-V3 synthe-

[3] We used zero overlap, as Bennani and Moslonka (2026) demonstrated it reduces indexing costs without degrading retrieval. The chunk size was set to 1000 characters to reserve context window capacity for the injection of graph triplets and community summaries, ensuring the fully assembled prompt remains safely below the context degradation cliff.

[4] `https://www.trychroma.com/`

[5] `https://neo4j.com/`

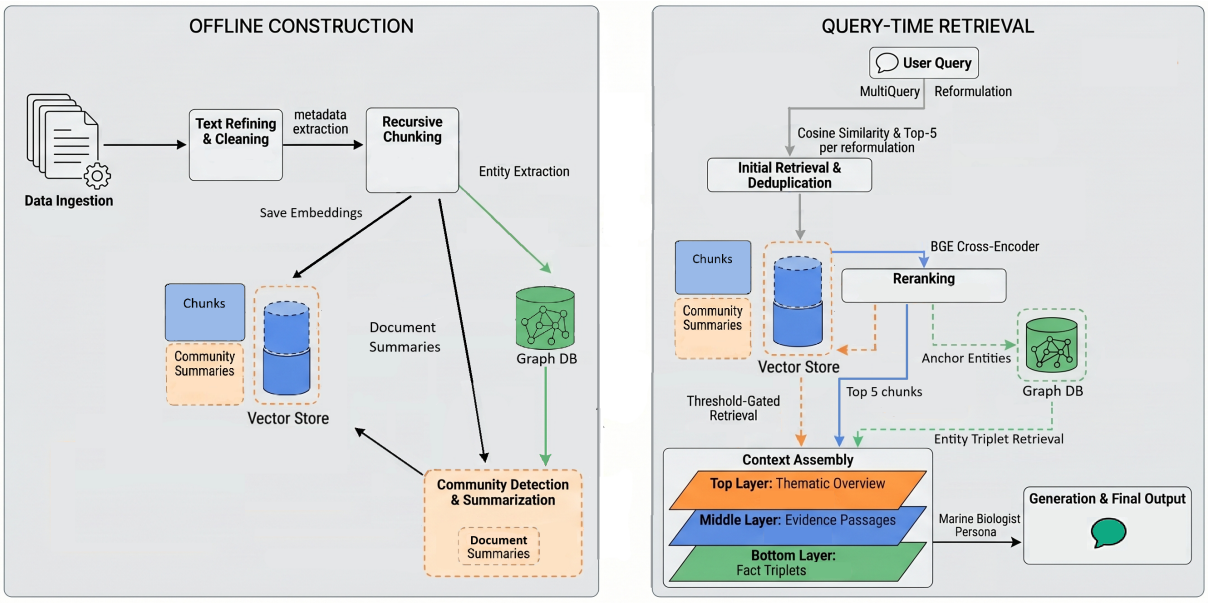

Figure 1: AlgaeBot system architecture showing offline construction (left) and query-time retrieval (right). Dashed boxes indicate optional components corresponding to the four evaluated retrieval conditions.

sized the document summaries belonging to each community into a single natural-language summary describing shared themes, entities, and relationships.

3.2 Retrieval Conditions

All conditions share a common retrieval pipeline: queries are reformulated into five query variants using DeepSeek-V3, each retrieving the top five chunks from the vector store, followed by cross-encoder reranking (`bge-reranker-base`) retaining the top-5 chunks for generation. We evaluate four conditions in a 2×2 design crossing the presence or absence of graph expansion and community summaries as independent augmentation mechanisms:

- **Vector-only**: standard dense retrieval using the pipeline described above, with no graph augmentation.
- **Graph-expanded**: vector retrieval augmented with one-hop graph expansion from anchor chunks, injecting neighboring entity triplets into the generation context. Hub nodes exceeding degree 100 are filtered to suppress noise.
- **Community-augmented**: vector retrieval augmented with up to 3 community summaries retrieved by vector similarity (max distance 0.3) for the entities present in the top-k chunks.
- **Full GraphRAG**: combines all three mechanisms: vector retrieval, graph expansion, and community summaries.

All conditions use GPT-5-nano as the generator (temperature 0.3) with a consistent system prompt. To avoid self-evaluation bias, generation and evaluation used different model families.

3.3 Evaluation

We evaluate the four conditions on a set of 203 questions generated with the RAGAS testset synthesizer (Es et al., 2025) using GPT-4o-mini. RAGAS defines four query synthesizer categories; single-hop abstract queries were not generated in our test set, leaving three types, which we refer to as *simple* (single-hop specific), *relational* (multi-hop specific), and *abstract* (multi-hop abstract). We report four metrics (Es et al., 2025): *faithfulness*, the proportion of answer claims supported by the retrieved context; *answer relevancy*, how directly the answer addresses the question; and *context precision* and *context recall*, the relevance ranking and coverage of the retrieved chunks, respectively. DeepSeek-V3 serves as the judge model for all metrics.

Statistical significance is assessed via two-way MANOVA with retrieval condition and query type as factors, followed by Tukey HSD post-hoc comparisons on faithfulness.

4 Results

Condition	Faith-fulness	Answer Relevancy	Context Precision	Context Recall
Vector	0.809	0.581	0.422	0.568
Graph-exp.	0.788	0.584	0.338	0.536
Community	0.817	0.590	0.403	0.578
Combined	0.848	0.569	0.238	0.549

Table 1: RAGAS evaluation results across four retrieval conditions ($n = 203$).

Graph expansion alone reduced faithfulness by -0.021 and substantially degraded context precision by -0.084, indicating that raw entity-level triplets actively dilute the generation context when injected without additional thematic grounding. Community summaries alone improved faithfulness consistently by $+0.008$ and produced the best answer relevancy of any condition. The full GraphRAG system achieved the highest faithfulness of all conditions, $+0.039$ over baseline, substantially exceeding the additive effect expected from the two components individually (-0.013), which indicates a positive interaction: triplets that are counterproductive in isolation contribute usefully when community context is present to anchor them.

The remaining metrics show distinct patterns. Answer relevancy is largely insensitive to retrieval condition (0.569–0.590), as all conditions share the same questions and generator. Context recall varies little, peaking with community summaries (0.578). Context precision drops sharply whenever graph triplets are injected (0.338 and 0.238, vs. 0.422 and 0.403 without).

4.1 Statistical Analysis

A two-way MANOVA (retrieval condition $\times$ query type) confirmed that both factors reached significance with $n = 203$. Retrieval condition was significant (Pillai's trace $= 0.065$, $F = 4.43$, $p < 0.001$), as was query type (Pillai's trace $= 0.216$, $F = 24.14$, $p < 0.001$), with no interaction effect ($p = 0.716$).

Factor	Pillai's trace	F	p
Condition	0.065	4.43	<0.001
Query type	0.216	24.14	<0.001
Interaction	0.024	0.819	0.716

Table 2: Two-way MANOVA results ($n = 203$). Both factors are significant; no interaction effect.

Tukey HSD post-hoc analysis on faithfulness identified one significant pairwise difference: the graph-expanded condition performed significantly worse than the full GraphRAG system ($\Delta = -0.060$, 95% CI $[-0.1125, -0.0070]$, $p = 0.019$). No other pair reached significance.

Univariate follow-ups show that the multivariate condition effect is driven primarily by context precision ($p < 0.001$). Faithfulness reaches significance at the 0.05 level ($p = 0.033$), while neither answer relevancy nor context recall shows a significant condition effect.

Metric	Condition effect p
Faithfulness	0.033
Answer relevancy	0.934
Context precision	< 0.001
Context recall	0.715

Table 3: Univariate condition effects per metric. Query type is significant across all four metrics.

Table 4 stratifies faithfulness by query type. The interaction pattern is most pronounced for abstract queries, where both mechanisms individually degrade faithfulness yet the full system produces a positive effect, suggesting that neither component is independently sufficient but their combination provides complementary context.

Query Type	Δ Graph-exp.	Δ Community	Δ Combined
Simple ($n = 69$)	$+0.015$	$+0.043$	$+0.054$
Abstract ($n = 67$)	-0.022	-0.039	$+0.013$
Relational ($n = 67$)	-0.057	$+0.019$	$+0.049$

Table 4: Faithfulness change (Δ) relative to vector-only baseline, stratified by query type.

5 Discussion

The synergistic pattern between graph expansion and community summaries suggests a semantic grounding mechanism: entity-level triplets provide structured relational facts, but without broader thematic context, LLMs may struggle to distinguish relevant relations from noise. We attribute this pattern to a *format mismatch*: community summaries are paragraph-length prose that the generator can incorporate seamlessly, whereas raw triplets are symbolic tuples that must be translated into natural-language reasoning on the fly, introducing friction that manifests as diluted faithfulness and degraded precision. Community summaries appear to resolve this friction by establishing a thematic frame

that lets the model judge which triplets are relevant evidence and which are peripheral.

A case study on Korean algae research in 2008 illustrates this effect. Community summaries surfaced thematically related studies on thermal effluent impacts and taxonomic diversity that entity-level traversal missed, while graph expansion in the full system retrieved one study through a specific entity path but displaced the other, consistent with the observed precision penalty.

The distance-threshold approach to community retrieval ($d \leq 0.3$) achieves selective augmentation without the complexity of agentic routing (Lin et al., 2026): community summaries are only injected when a sufficiently similar community exists, providing a lightweight quality gate. The dominance of query type in the MANOVA further supports this retrieval strategy. If retrieval benefit varies more by query type than by architecture, a similarity-based filter becomes a reasonable proxy for routing. This finding also suggests that reporting only aggregate metrics across heterogeneous query types may obscure differential effects, and we therefore recommend stratifying evaluation by query type as standard practice.

The context precision penalty warrants careful interpretation. Injecting additional graph context inherently lowers the precision score regardless of answer correctness. Context precision should therefore be read as retrieval efficiency rather than generation quality, with faithfulness weighted more heavily when evaluating graph-augmented systems.

6 Conclusion

We presented a controlled 2×2 ablation isolating the effects of graph expansion and community summaries in a scientific-domain GraphRAG system. Our results suggest that community-level thematic representations stabilize retrieval quality in GraphRAG systems and allow entity-level relational context to contribute positively rather than introduce noisy relational context. The per-query-type analysis reveals that this interaction is most pronounced for abstract and relational queries, where neither mechanism is sufficient on its own. Future work should investigate whether serializing graph triplets into natural prose before injection can reduce the dependency on community context, and whether these patterns generalize beyond single-domain corpora.

Limitations

This study evaluates a single, domain-specific corpus (phycological literature), and generalization of the findings to other domains or languages has yet to be tested. The evaluation set of 203 questions was generated automatically by GPT-4o-mini rather than validated by domain experts, which may introduce systematic biases in question distribution. Community detection resolution (8.0) and the distance threshold for community retrieval (0.3) were selected empirically and not ablated, leaving their optimal values unknown. While automated question generation and evaluation may not capture the full nuance of phycological expert inquiry, our primary contribution is the relative comparison across a controlled 2×2 ablation. Because the question set and judge model are held constant across all four conditions, systematic biases affect all conditions equally, preserving the validity of our comparative findings.

The central comparison of interest, full GraphRAG versus baseline on faithfulness, was directionally positive but did not reach pairwise significance ($p = 0.234$); only the graph-expanded versus full contrast was significant ($p = 0.019$). Furthermore, DeepSeek-V3 served as both knowledge graph extractor and RAGAS judge, a residual confound that three distinct model families would eliminate.

Acknowledgments

This work was supported by the AlgaeProBANOS project "Accelerating algae product development in the Baltic and North Sea", funded by the European Union's Horizon Europe research and innovation programme under grant agreement No. 101112943.

References

Sofia Bennani and Charles Moslonka. 2026. A systematic analysis of chunking strategies for reliable question answering. arXiv preprint arXiv:2601.14123.

Henrik Brådland, Morten Goodwin, Per-Arne Andersen, Alexander S. Nossum, and Aditya Gupta. 2025. A new HOPE: Domain-agnostic automatic evaluation of text chunking. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 10 pages, New York, NY, USA. ACM. Pre-print available at https://arxiv.org/abs/2505.02171.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt,

Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint*. ArXiv:2404.16130 [cs].

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2025. Ragas: Automated Evaluation of Retrieval Augmented Generation. *arXiv preprint*.

Jakob Kusnick and Tariq Yousef. 2026. Connecting the algae value chain visually: Integrating farming, processing, and product views. In *EnvirVis 2026-Workshop on Visualisation in Environmental Sciences*. The Eurographics Association.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*.

Shaofu Lin, Shengze Shao, Xiliang Liu, and Haoru Su. 2026. FusionGraphRAG: An Adaptive Retrieval-Augmented Generation Framework for Complex Disease Management in the Elderly. *Information*, 17(2):138.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive Abstractive Processing for Tree-Organized Retrieval. *arXiv preprint*.

Vincent Traag, Ludo Waltman, and Nees Jan van Eck. 2019. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources For General Chinese Embeddings. *arXiv preprint*.

A Model and Pipeline Settings

Table 5 lists the models used at each pipeline stage, and Table 6 the retrieval hyperparameters shared across all four conditions. All settings were held constant across conditions; the only difference between conditions is whether graph triplets and community summaries are appended to the generation context.

Component	Model
Embedding	`bge-base-en-v1.5`
Reranking	`bge-reranker-base`
Query rewriting	DeepSeek-V3 (`deepseek-chat`)
KG extraction	DeepSeek-V3 + Instructor schema
Document summaries	Gemma-3-1B
Community summaries	DeepSeek-V3
Testset synthesis	GPT-4o-mini
Generator	GPT-5-nano (temperature 0.3)
RAGAS judge	DeepSeek-V3

Table 5: Models used at each pipeline stage. Generation and evaluation deliberately use different model families to avoid self-evaluation bias.

Parameter	Value
Chunk size / overlap	1000 chars / 0
Query reformulations	5
Chunks per reformulation	5
Chunks after reranking	5
Community summaries injected	up to 3
Community distance threshold	0.3
Hub-node degree filter	100
Leiden resolution	8.0
Relation confidence cutoff	0.5

Table 6: Retrieval and indexing hyperparameters, identical across conditions.

B Generation Prompt Template

Placeholders are filled as follows: `{context}` assembles the retrieved material in a fixed order: community summaries first (in the community-augmented and full GraphRAG conditions), followed by the five reranked chunks, each prefixed with a numbered metadata header `[i] Authors (Year). Title`, and finally entity triplets (in the graph-expanded and full GraphRAG conditions). In the vector-only condition, `{context}` contains only the five chunks. `{history}` is empty for evaluation (single-turn setting).

```
Role: You are an experienced marine biologist and algae cultivation specialist with deep expertise in seagrass ecology, microalgae biotechnology, and industrial algae applications. You prioritize scientific precision, cite specific data when available, and clearly distinguish between established findings and uncertainties.

Task: Given retrieved passages from scientific literature on algae and marine biology, answer the user's question following these steps:

1. First, internally assess which of the provided passages are relevant to the question.

2. Synthesize information from ALL relevant passages into a coherent answer. You must draw from multiple sources where possible — do not rely on a single passage.

3. If passages contain conflicting information, acknowledge the conflict and explain both positions.

4. If the provided context is insufficient to fully answer the
```

question, state what is missing.

Domain constraint: Focus on the user's specific industry context.

Output format: Brief paragraph. You MUST cite every passage you use with its bracketed number [1], [2], etc. from the context headers. Use multiple citations to support your answer. Then, directly before listing your sources, output a Markdown horizontal rule (—) on its own line. Finally, below the line, list each source you cited on its own line in this exact format: [number] Authors (Year). Title.

The following passages are numbered [1], [2], etc. When citing, use ONLY these numbers. Do not invent or extract citations from within the passage text.

Keep your answer grounded strictly in the provided context. Do not introduce external knowledge beyond what is given. Match your answer's depth to the question's complexity.

If the provided context contains NO relevant information to answer the question, simply state that the question is outside the scope of the algae research database. Do not cite sources that weren't used.

LANGUAGE CONSTRAINT: You MUST respond exclusively in English.

Context: {context}

Question: {query}

Answer: