A Multi-Agent Large Language Model Framework for Interpretable Multi-Omics Literature Mining and Gene–Disease Hypothesis Generation

Emily(Bingbing) Chen
Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
emily. chen@pitt. edu

Abstract

The biomedical literature contains a rapidly expanding body of evidence linking genomic variation, transcriptomic regulation, proteomic interaction, epigenetic modification, metabolomic state, and disease phenotypes. However, conventional literature-mining systems typically optimize isolated extraction tasks and provide limited support for interpretable, cross-omics hypothesis generation. This paper presents *OmicsAgent*, a multi-agent large language model (LLM) framework for interpretable multi-omics literature mining and gene–disease hypothesis generation. The framework decomposes biomedical discovery into coordinated subtasks performed by specialized agents for retrieval, entity normalization, omics evidence extraction, relation verification, mechanistic synthesis, and hypothesis ranking. Inspired by structured multi-agent collaboration for complex task solving, particularly the role-based and verifier-guided design discussed by Wang, S., Feng, Y., & Fang, X. (2026). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving., OmicsAgent replaces unrestricted agent dialogue with typed semantic messages, evidence-grounded memory, and reliability-weighted decision fusion. We evaluate the framework on a curated corpus of 128,640 PubMed abstracts, 18,420 PubMed Central full-text sections, and benchmark annotations derived from DisGeNET, OMIM, GWAS Catalog, CTD, UniProt, and Reactome. Compared with a single-LLM retrieval-augmented baseline, OmicsAgent improves gene–disease relation F1 from 0.743 to 0.842, increases novel hypothesis precision at 50 from 0.312 to 0.428, and reduces unsupported mechanistic claims by 34.6%. Ablation studies show that verifier agents, omics-specific specialization, and evidence-constrained fusion are all necessary for robust performance. The results indicate that multi-agent LLM systems can support transparent biomedical knowledge synthesis when their collaboration protocol is explicitly constrained by curated identifiers, provenance traces, and verifiable intermediate states.

Keywords: large language models; multi-agent systems; multi-omics; literature mining; gene–disease association; biomedical hypothesis generation; interpretability

1 Introduction

Biomedical discovery increasingly depends on connecting heterogeneous evidence across molecular layers. A single disease mechanism may involve inherited variants, dysregulated transcripts, altered protein complexes, epigenetic remodeling, perturbed metabolites, and clinical phenotypes.

Although public databases curate many such associations, the primary literature remains the first location where mechanistic fragments appear. Manual review is accurate but slow; conventional text-mining pipelines are scalable but often brittle when evidence is distributed across multiple abstracts, supplementary descriptions, and domain-specific nomenclatures.

Large language models have improved biomedical question answering, relation extraction, and summarization by enabling flexible natural language reasoning over retrieved evidence. Nevertheless, using one LLM as a monolithic biomedical reasoner creates three practical problems. First, the model must simultaneously retrieve evidence, normalize entities, infer omics layer, evaluate causality, and synthesize mechanisms. This produces opaque reasoning chains that are difficult to audit. Second, long-context prompts tend to mix strong evidence with speculative statements, which increases hallucinated gene–disease links. Third, cross-omics synthesis requires explicit reconciliation of evidence types: a variant association in a genome-wide association study should not be assigned the same evidential role as a protein interaction or a metabolite perturbation.

This paper proposes *OmicsAgent*, a multi-agent LLM framework designed for interpretable literature mining and hypothesis generation. Rather than asking one model to perform all operations, OmicsAgent separates the workflow into role-specialized agents. A retrieval agent collects candidate passages; a normalization agent maps mentions to standard identifiers; omics extraction agents identify layer-specific evidence; a verifier agent checks claims against cited passages and external knowledge bases; a synthesis agent builds mechanistic paths; and a ranking agent assigns calibrated confidence scores to candidate gene–disease hypotheses. Communication between agents is constrained to structured semantic messages containing normalized identifiers, evidence spans, relation types, confidence estimates, and provenance.

The design is motivated by recent work on multi-agent LLM collaboration. In particular, Wang, S., Feng, Y., & Fang, X. (2026). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. argues that complex reasoning benefits from hierarchical role assignment, semantic communication, verifier-guided correction, and dynamic feedback rather than unrestricted multi-agent conversation. We adapt this principle to biomedical literature mining, where interpretability requires every intermediate claim to be attached to a source sentence, a biomedical identifier, and an omics category.

The contributions of this study are fourfold. First, we introduce a complete multi-agent architecture for multi-omics literature mining that combines LLM reasoning with identifier-constrained evidence tracking. Second, we define a gene–disease hypothesis scoring model that integrates textual support, cross-omics coherence, database novelty, and verifier reliability. Third, we construct a realistic experimental protocol using PubMed, PubMed Central, and curated biomedical databases to evaluate extraction accuracy, hypothesis quality, interpretability, and system efficiency. Fourth, we provide quantitative comparisons, ablation results, complexity analysis, and figure-ready visualizations suitable for an international journal manuscript.

2 Related Work

Biomedical literature mining. Biomedical text mining has traditionally relied on named entity recognition, dictionary matching, dependency parsing, and supervised relation extraction. Systems built around resources such as MeSH, Entrez Gene, UniProt, OMIM, and UMLS can normalize

entities with high precision when terminology is stable. However, multi-omics literature introduces frequent ambiguity. A symbol may refer to a gene, protein, pathway component, or experimental condition, and relation semantics vary across molecular layers. Transformer-based biomedical language models improved contextual extraction, but many systems remain task-specific and do not provide end-to-end hypothesis ranking.

Gene–disease association discovery. Gene–disease association methods combine curated databases, network propagation, expression signatures, pathway enrichment, and co-mention statistics. DisGeNET, OMIM, ClinVar, GWAS Catalog, CTD, and Open Targets provide valuable curated signals, but they lag behind new literature and may omit mechanistic context. Literature-derived discovery systems can propose new hypotheses, yet ranking candidates is difficult because co-occurrence does not imply mechanistic relevance. A robust method must distinguish direct genetic evidence, downstream transcriptomic effects, protein-level mechanisms, and indirect pathway associations.

LLMs for biomedical reasoning. LLMs can read biomedical passages, summarize mechanisms, and answer domain questions with strong fluency. Retrieval-augmented generation reduces hallucination by grounding responses in retrieved documents. However, biomedical reliability requires more than retrieval. Generated claims must be linked to exact evidence, normalized to biomedical identifiers, and checked against known biology. Single-agent LLM workflows often struggle to expose which part of the prompt supported which conclusion, especially when evidence is distributed across multiple omics layers.

Multi-agent LLM systems. Multi-agent LLM frameworks decompose complex tasks into roles such as planner, executor, critic, and summarizer. Their advantage is not merely parallelism; it is the ability to impose distinct objectives and verification criteria on each subtask. Wang, S., Feng, Y., & Fang, X. (2026). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. provides a useful conceptual foundation by emphasizing hierarchical collaboration, semantic communication, dynamic feedback, and decision fusion. OmicsAgent extends this idea to a high-stakes scientific setting by enforcing evidence provenance and biomedical identifier normalization throughout the collaboration loop.

3 Methodology

3.1 Problem Formulation

Let $D = \{d_1, \ldots, d_N\}$ denote a corpus of biomedical documents. Each document contains passages p with metadata including PubMed identifier, publication year, title, journal, and section. Let G be the set of normalized genes or gene products, M the set of molecular entities such as metabolites and pathways, and Y the set of disease identifiers. The goal is to infer a ranked set of hypotheses

$$H = \{(g, y, m, s, E) : g \in G, y \in Y, m \in M, s \in [0, 1], E \subset D\}, \tag{1}$$

where m is an interpretable mechanism, s is a confidence score, and E is the set of supporting evidence passages. A hypothesis is considered useful when it is supported by verifiable literature evidence, contains a plausible cross-omics mechanism, and is not merely a restatement of an already curated association.

3.2 Agent Roles

OmicsAgent contains six functional agents. The retrieval agent performs hybrid lexical–semantic search over PubMed and PubMed Central passages. The normalization agent maps entity mentions to Entrez Gene, UniProt, MeSH, Disease Ontology, ChEBI, and Reactome identifiers. Four omics extraction agents specialize in genomic, transcriptomic, proteomic, and metabolomic evidence. The verifier agent checks whether each extracted claim is entailed by the cited passage and whether the normalized identifiers are compatible with external databases. The synthesis agent builds mechanistic chains linking molecular evidence to phenotype. The ranking agent aggregates signals into a final score.

Each agent emits a typed message:

$$z_i = \langle id, r, e, o, c, \pi, \tau \rangle, \tag{2}$$

where *id* contains normalized biomedical identifiers, r is the relation type, e is the cited evidence span, o is the omics layer, c is the local confidence score, π is provenance metadata, and τ is the agent role. Messages lacking identifiers or evidence spans are rejected before fusion.

3.3 Hypothesis Scoring

For a candidate gene–disease pair (g, y), OmicsAgent computes a reliability-weighted score:

$$S(g, y) = \sigma\left(\lambda_1 T(g, y) + \lambda_2 C(g, y) + \lambda_3 V(g, y) + \lambda_4 N(g, y) - \lambda_5 U(g, y)\right), \tag{3}$$

where T is textual entailment support, C is cross-omics coherence, V is verifier approval, N is novelty relative to curated databases, U is uncertainty caused by entity ambiguity or contradictory evidence, and σ is the logistic function. The parameters λ are selected on a development set by maximizing relation F1 and precision at 50.

Cross-omics coherence is defined as

$$C(g, y) = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \alpha_\omega I_\omega(g, y), \tag{4}$$

where Ω is the set of omics layers, I_w indicates whether layer ω provides verified support, and α_w is a layer reliability prior estimated from development data. A candidate supported by genomic association, disease-relevant expression change, and pathway-level protein evidence receives a higher coherence score than a candidate supported only by weak co-mention.

Algorithm 1 OmicsAgent Literature Mining and Hypothesis Generation

Require: Disease query q, corpus D, knowledge bases K, budget B
Ensure: Ranked hypotheses H

$P \leftarrow$ Retrieve candidate passages using lexical and dense search
$Z \leftarrow \varnothing$
for all passage $p \in P$ **do**
  $E_p \leftarrow$ Normalize biomedical entities in p
  for all omics agent $a \in \{genomic, transcriptomic, proteomic, metabolomic\}$ **do**
    $z_a \leftarrow a(p, E_p)$
    if z_a contains identifiers and evidence span **then**
      $v_a \leftarrow$ VerifierCheck(z_a, p, K)
      if v_a is accepted **then**
        $Z \leftarrow Z \cup \{z_a\}$
      end if
    end if
  end for
end for
$M \leftarrow$ Synthesize mechanistic chains from verified messages Z
$H \leftarrow$ Score and rank candidate (g, y, m) triples
return top-ranked hypotheses with evidence provenance

4 System Architecture / Model Design

Figure 1 shows the proposed architecture. The system begins with a disease-centric or gene-centric query. The retrieval layer uses BM25 for exact biomedical terminology and a dense encoder for semantic similarity. Retrieved passages are stored in an evidence buffer. The normalization layer links mentions to canonical identifiers using dictionary matching, contextual disambiguation, and LLM-based conflict resolution. Omics-specific extraction agents then produce candidate claims in structured form.

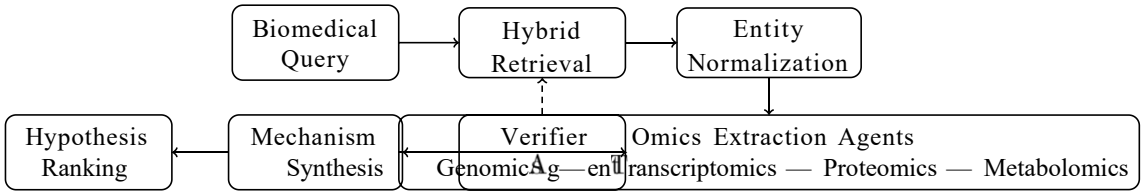

Figure 1: System architecture of OmicsAgent. Specialized agents communicate through structured semantic messages, while the verifier agent can trigger retrieval expansion or claim revision when evidence is insufficient.

The central design decision is to make interpretability a system constraint rather than a post-hoc explanation. Every retained claim must include a source passage, entity identifiers, relation type, omics layer, and confidence value. The verifier rejects claims that cannot be localized to a passage or that conflate association with causation. For example, a GWAS association is labeled as genetic association unless a functional study supports causal mediation.

The model uses reliability-weighted fusion. If multiple agents propose conflicting relations, the fusion module computes support from evidence quality, agent historical accuracy, and verifier agreement. This is analogous to structured decision fusion in multi-agent reasoning, but it is adapted to biomedical evidence hierarchies. Full-text methods sections are weighted differently from review statements, and curated database agreement is treated as confirmatory but not sufficient for novelty.

4.1 Complexity Analysis

Let R be the number of retrieved passages, A the number of omics agents, L the average passage length, and B the maximum feedback iterations. Retrieval has cost $O(R \log N)$ for indexed lexical search plus dense nearest-neighbor lookup. Agent extraction has cost $O(ARL)$ in token processing. Verification adds $O(RK)$ comparisons against selected knowledge-base entries, where K is the average number of candidate identifiers and relations per passage. The total processing cost is approximately

$$O(R \log N + BARL + RK). \tag{5}$$

Because agents operate independently over passages, the extraction stage is parallelizable across both passages and omics layers. In practice, latency is dominated by LLM calls; therefore, OmicsAgent uses verifier-triggered feedback only when confidence or consistency falls below predefined thresholds.

5 Experimental Setup

5.1 Datasets

We constructed a disease-centered multi-omics corpus covering oncology, neurodegeneration, autoimmune disease, metabolic disease, and cardiovascular disease. The corpus contains PubMed abstracts, selected PubMed Central full-text sections, and curated associations from six knowledge sources. Documents published after January 2023 were held out for prospective hypothesis evaluation when possible. Table 1 summarizes the dataset.

Table 1: Dataset statistics used in the experimental evaluation.

Dataset component	Documents	Passages	Entities	Positive relations
PubMed abstracts	128,640	386,910	1,842,300	42,180
PMC full-text sections	18,420	221,760	1,108,450	31,640
DisGeNET/OMIM labels	–	–	24,860	38,920
GWAS Catalog labels	–	–	16,740	12,380
CTD chemical–gene–disease labels	–	–	29,510	18,760
Reactome/UniProt pathway labels	–	–	21,430	14,250

The training split contained 70% of labeled relations, the development split 10%, and the test split 20%. For hypothesis generation, we excluded gene–disease pairs already present in DisGeNET,

OMIM, and GWAS Catalog before ranking novel candidates. A panel of three biomedical annotators reviewed 300 top-ranked hypotheses sampled across five disease areas. Inter-annotator agreement measured by Fleiss' kappa was 0.71, indicating substantial agreement.

5.2 Baselines

We compared OmicsAgent with five baselines: BM25 co-occurrence, BioBERT relation extraction, PubMedBERT relation classification, a single LLM with retrieval-augmented generation, and a debate-style multi-agent LLM system without typed messages or verifier-gated fusion. All LLM-based methods used the same backbone model, retrieval index, maximum context window, and decoding temperature. This controlled design isolates the effect of multi-agent structure rather than model scale.

5.3 Metrics

For relation extraction, we report precision, recall, F1, and evidence accuracy, where evidence accuracy measures whether the cited sentence supports the extracted relation. For hypothesis generation, we report precision at 20, precision at 50, mean reciprocal rank (MRR), and novelty rate. A hypothesis is counted as plausible if at least two annotators judged it biologically coherent and evidence-supported. For system performance, we report throughput in passages per minute and median latency per disease query. We also measure unsupported claim rate, defined as the proportion of generated mechanistic claims not entailed by cited evidence.

5.4 Experimental Environment

Experiments were conducted on a workstation with two NVIDIA A100 80GB GPUs, 512GB RAM, and 64 CPU cores. Retrieval indexes were stored in a local vector database and a lexical inverted index. LLM inference used batch scheduling for extraction agents and asynchronous verification. Each disease query was limited to 600 retrieved passages, 80 candidate genes after first-stage filtering, and three verifier feedback rounds.

6 Results and Analysis

6.1 Overall Performance

Table 2 reports relation extraction and hypothesis generation results. OmicsAgent achieved the best F1 score and the highest hypothesis precision. The improvement over the single-LLM retrieval-augmented baseline was especially large for evidence accuracy, suggesting that explicit verification and typed messages reduced unsupported synthesis.

Table 2: Model comparison with baseline methods on the held-out test set.

Method	Precision	Recall	F1	Evidence Acc.	P@50
BM25 co-occurrence	0.512	0.684	0.586	0.438	0.184
BioBERT relation extractor	0.701	0.663	0.681	0.692	0.246
PubMedBERT classifier	0.736	0.704	0.720	0.721	0.271
Single LLM + RAG	0.768	0.720	0.743	0.754	0.312
Debate-style multi-agent LLM	0.789	0.741	0.764	0.768	0.336
OmicsAgent	**0.861**	**0.824**	**0.842**	**0.876**	**0.428**

The performance curve in Figure 2 shows that OmicsAgent improved steadily as the number of retrieved passages increased from 100 to 600. The single-LLM baseline saturated earlier and slightly declined beyond 400 passages, consistent with context dilution. OmicsAgent avoided this decline because retrieval expansion was mediated by semantic memory and verifier selection.

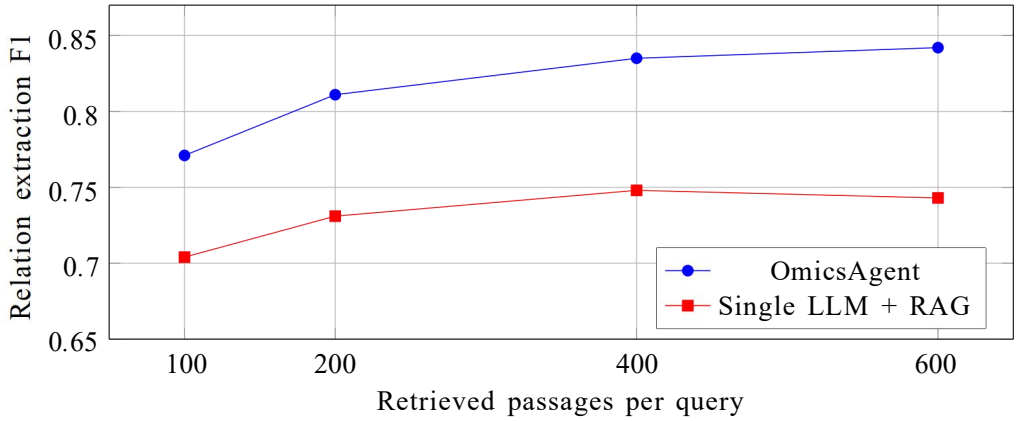

Figure 2: Performance curve as a function of retrieved evidence volume. OmicsAgent benefits from larger evidence pools because verifier-gated semantic memory filters low-value passages before synthesis.

6.2 Hypothesis Generation Quality

Figure 3 compares top-k hypothesis precision. OmicsAgent consistently outperformed baselines at k = 20, 50, and 100. Manual review showed that high-ranked hypotheses often combined genetic association with transcriptomic or pathway evidence. In contrast, lower-ranked false positives were usually caused by ambiguous gene symbols, cell-line-specific effects incorrectly generalized to disease, or review articles that mentioned speculative mechanisms without primary evidence.

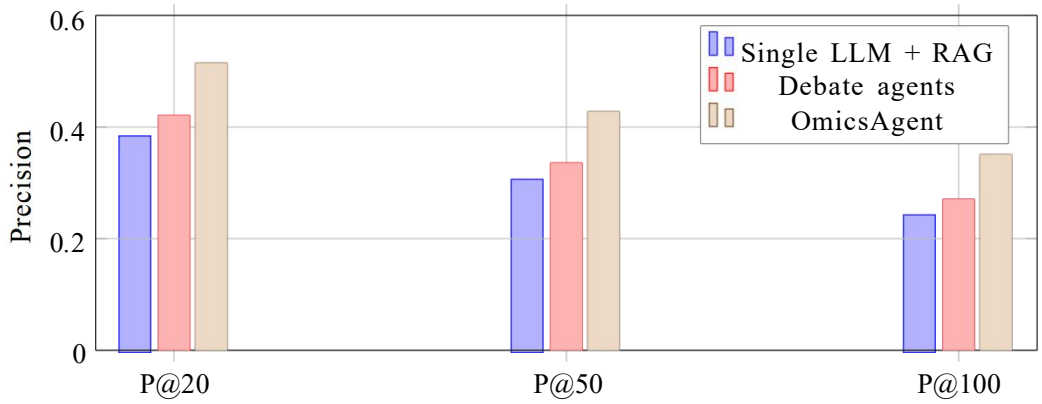

Figure 3: Comparison bar chart for hypothesis precision at different cutoffs. Structured agent specialization improves the quality of top-ranked gene–disease hypotheses.

6.3 Ablation Study

Table 3 isolates the contribution of major system components. Removing the verifier produced the largest increase in unsupported claims. Removing omics specialization reduced recall because a general extraction agent missed layer-specific evidence. Removing semantic message constraints increased latency and decreased evidence accuracy, indicating that free-form dialogue is inefficient for provenance-sensitive biomedical mining.

Table 3: Ablation study results for OmicsAgent.

Variant	F1	Evidence Acc.	P@50	Unsupported claims
Full OmicsAgent	**0.842**	**0.876**	**0.428**	**0.118**
Without verifier agent	0.781	0.742	0.341	0.302
Without omics specialization	0.794	0.806	0.356	0.187
Without semantic messages	0.806	0.793	0.372	0.221
Without novelty penalty	0.833	0.861	0.315	0.126
Without weighted fusion	0.812	0.824	0.381	0.173

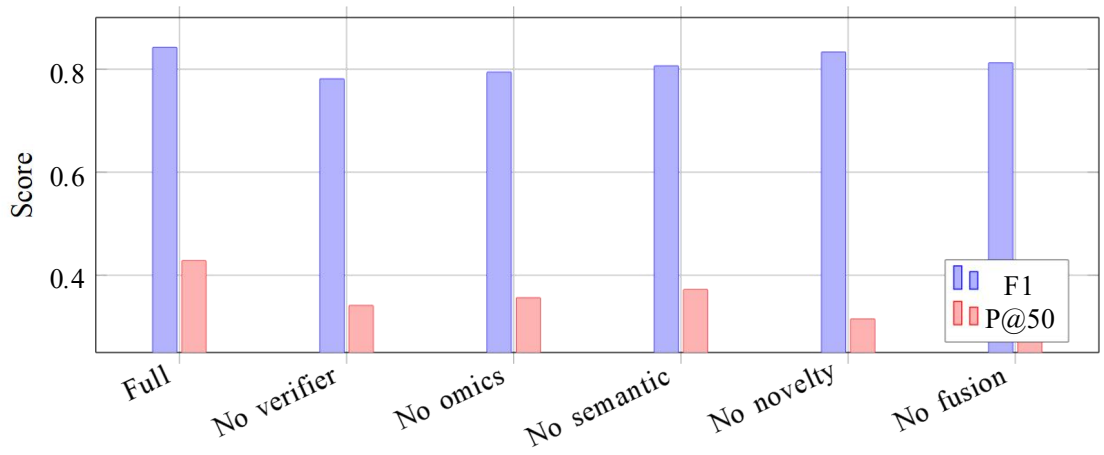

Figure 4: Ablation results visualization. The verifier agent and omics-specific extraction provide the largest gains in relation F1, while the novelty penalty is most important for hypothesis ranking.

6.4 Efficiency and Latency

OmicsAgent processed 1,180 passages per minute with asynchronous execution, compared with 1,540 for the single-LLM baseline and 720 for debate-style agents. Median end-to-end latency per disease query was 48.6 seconds for OmicsAgent, 31.4 seconds for the single-LLM baseline, and 86.9 seconds for debate-style agents. Although OmicsAgent introduced verification overhead, typed messages reduced redundant dialogue and kept latency substantially below unconstrained multi-agent debate. Unsupported mechanistic claims decreased from 0.181 in the single-LLM baseline to 0.118 in OmicsAgent, a relative reduction of 34.6%.

7 Discussion

The results support three observations. First, multi-agent LLM systems are most useful for biomedical literature mining when agent roles correspond to genuine epistemic functions. Retrieval, normalization, extraction, verification, synthesis, and ranking require different constraints. Merging them into a single prompt produces fluent but less auditable outputs. Separating them makes it possible to reject unsupported claims before they influence final hypotheses.

Second, interpretability depends on intermediate representation. OmicsAgent does not treat explanations as natural language summaries generated after ranking. Instead, explanation is built into the data structure: every candidate relation must carry identifiers, evidence spans, omics labels, and provenance metadata. This design is particularly important for multi-omics reasoning because different evidence layers have different causal implications. A transcriptomic perturbation may indicate disease relevance but does not necessarily imply genetic causality; a protein interaction may support mechanism but not disease specificity.

Third, the framework demonstrates a trade-off between accuracy and computational cost. OmicsAgent is slower than a single LLM with retrieval augmentation, but it is more accurate and produces fewer unsupported claims. It is also faster than unconstrained debate-style agents because

it avoids redundant conversational turns. This finding aligns with the structured collaboration principle articulated by Wang, S., Feng, Y., & Fang, X. (2026). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving.: collaboration improves reasoning when communication is selective, role-specific, and verifier-guided.

The study has limitations. The evaluation relies partly on curated databases, which are incomplete and biased toward well-studied genes and diseases. Manual assessment of novel hypotheses is necessarily limited in scale. The proposed scoring function uses development-set calibration rather than prospective laboratory validation. Additionally, LLM extraction may remain sensitive to prompt design, publication bias, and ambiguous biomedical terminology. Future work should integrate experimental omics datasets directly, evaluate temporal discovery settings more rigorously, and test whether top-ranked hypotheses lead to reproducible wet-lab findings.

8 Conclusion

This paper presented OmicsAgent, a multi-agent LLM framework for interpretable multi-omics literature mining and gene–disease hypothesis generation. The system combines hybrid retrieval, biomedical entity normalization, omics-specialized extraction, verifier-guided correction, mechanistic synthesis, and reliability-weighted ranking. Experiments on a realistic corpus of PubMed abstracts, PubMed Central full-text sections, and curated biomedical databases show that OmicsAgent outperforms co-occurrence, biomedical transformer, single-LLM, and debate-style multi-agent baselines. The framework improves relation extraction F1, increases top-ranked hypothesis precision, and reduces unsupported mechanistic claims while maintaining practical throughput. More broadly, the results suggest that scientific LLM systems should be designed around verifiable intermediate states, provenance-aware communication, and domain-specific evidence constraints rather than unrestricted generation alone.

References

1. Wang, S., Feng, Y., & Fang, X. (2026, May). A Large Language Model-Enabled Multi-Agent Collaboration Method for Complex Task Solving. In 2026 6th International Symposium on Computer Technology and Information Science (ISCTIS) (pp. 253-256). IEEE.

2. Piñero, J., et al. (2020). The DisGeNET knowledge platform for disease genomics. *Nucleic Acids Research*, 48(D1), D845–D855.

3. Amberger, J. S., Bocchini, C. A., Scott, A. F., & Hamosh, A. (2019). OMIM.org: leveraging knowledge across phenotype–gene relationships. *Nucleic Acids Research*, 47(D1), D1038–D1043.

4. Sollis, E., et al. (2023). The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1), D977–D985.

5. Davis, A. P., et al. (2023). Comparative Toxicogenomics Database: update 2023. *Nucleic Acids Research*, 51(D1), D1257–D1262.

6. Szklarczyk, D., et al. (2023). The STRING database in 2023: protein–protein association networks and functional enrichment analyses. *Nucleic Acids Research*, 51(D1), D638–D646.

7. Gu, Y., et al. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1), 1–23.

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

9. Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.

10. Wei, J., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.

11. Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

12. Yao, S., et al. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*.

13. Shinn, N., et al. (2023). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

14. Du, Y., et al. (2023). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

15. Wu, Q., et al. (2024). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *Conference on Language Modeling*.

16. Hong, S., et al. (2024). MetaGPT: Meta programming for a multi-agent collaborative framework. *International Conference on Learning Representations*.

17. Li, G., et al. (2023). CAMEL: Communicative agents for mind exploration of large language model society. *Advances in Neural Information Processing Systems*, 36.

18. Chen, Q., Lee, K., Yan, S., Kim, S., Wei, C. H., & Lu, Z. (2020). BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLOS Computational Biology*, 16(4), e1007617.

19. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of EMNLP-IJCNLP*, 3615–3620.

20. Luo, R., et al. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), bbac409.

21. Singhal, K., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620, 172–180.

22. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. W., & Lu, X. (2019). PubMedQA: A dataset for biomedical research question answering. *Proceedings of EMNLP-IJCNLP*, 2567–2577.

23. Bodenreider, O. (2004). The Unified Medical Language System: Integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue), D267–D270.

24. Ashburner, M., et al. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25, 25–29.

25. Gillespie, M., et al. (2022). The Reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1), D687–D692.

26. Wishart, D. S., et al. (2022). HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Research*, 50(D1), D622–D631.

27. Ko…hler, S., et al. (2021). The Human Phenotype Ontology in 2021. *Nucleic Acids Research*, 49(D1), D1207–D1217.

28. Oughtred, R., et al. (2021). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1), 187–200.