Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

BioDisco: Multi-Agent Hypothesis Generation with Dual-Mode Evidence, Iterative Feedback and Temporal Evaluation

Yujing Ke[1,2], **Kevin George**[2], **Kathan Pandya**[2,3], **Gerrit Großmann**[2], **David B. Blumenthal**[1]
Maximilian Sprang[4], **David A. Selby**[2] and **Sebastian Vollmer**[2,5]
[1]Friedrich-Alexander-Universität Erlangen-Nürnberg
[2]German Research Centre for Artificial Intelligence (DFKI)
[3]University of Saarland,
[4]University Medical Center Mainz
[5]University of Kaiserslautern–Landau (RPTU)
yujingk@chalmers.se, {first.last}@dfki.de, maspring@uni-mainz.de, david.b.blumenthal@fau.de

Abstract

Identifying novel hypotheses is essential to scientific research, yet this process risks being overwhelmed by the sheer volume and complexity of available information. Existing automated methods often struggle to generate novel and evidence-grounded hypotheses, lack robust iterative refinement and rarely undergo rigorous temporal evaluation for future discovery potential. To address this, we propose BIODISCO, a multi-agent framework that draws upon language model-based reasoning and a dual-mode evidence system (biomedical knowledge graphs and automated literature retrieval) for grounded novelty, integrates an internal scoring and feedback loop for iterative refinement, and validates performance through pioneering temporal and human evaluations and a Bradley–Terry paired comparison model for statistical assessment. Evaluations suggest improved novelty and significance relative to ablated configurations and a generalist biomedical agent. Designed for flexibility and modularity, BIODISCO allows seamless integration of custom language models or knowledge graphs, and can be run with just a few lines of code.

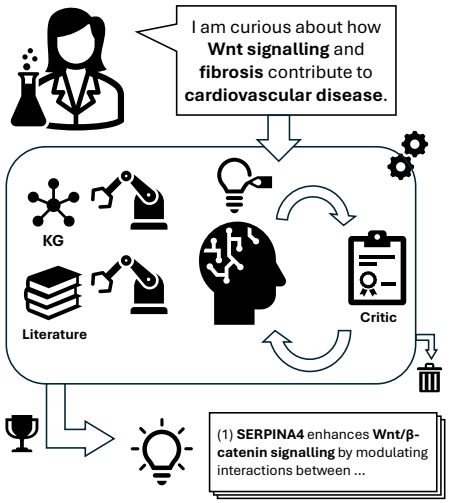

Figure 1: A high level overview of our automated framework for hypothesis generation. Overseen by a planner, agents search academic literature and query a knowledge graph to obtain articles and subgraphs relevant to a user-specified research topic. A scientist agent integrates these sources to derive initial hypotheses, which are rated by a critic, then refined with additional background, discarded or presented to the user with supporting evidence

1 Introduction

Advancing biomedical research depends on a supply of novel, testable hypotheses. However, generating such hypotheses traditionally relies on human intuition, domain expertise and manual literature reviews, all of which are increasingly strained by an exponential growth in scientific data and publications. This challenge motivates robustly evaluated, automated approaches to scientific discovery that can systematically uncover and integrate complex biological relationships.

1.1 Background

Knowledge graphs (KGs) provide a structured foundation for storing and analyzing scientific knowledge, with biomedical KGs capturing complex relationships among genes, proteins, diseases, drugs and biological pathways: e.g. HetioNet [Himmelstein *et al.*, 2017] and PharmKG [Zheng *et al.*, 2020]. While a variety of tools have been developed to mine such graphs for novel insights [Perdomo-Quinteiro and Belmonte-Hernández, 2024], construction and maintenance of KGs are time-consuming and labour-intensive tasks, leaving a significant proportion of biomedical knowledge untapped in unstructured text. Early discovery systems combined ontological resources, graph databases and literature mining for biomedical relationship discovery [Hristovski *et al.*, 2015; Jimeno-Yepes *et al.*, 2009]. Large language models (LLMs) have since emerged as a promising solution, capable of extracting relevant information and identifying latent patterns from vast corpora [Brown *et al.*, 2020], with greater capacity for contextual reasoning than earlier methods of hypoth-

esis generation [Spangler *et al.*, 2014]. However, LLMs risk returning plausible but groundless responses [Bélisle-Pipon, 2024]. To mitigate this, models can be imbued with external literature interfaces via retrieval-augmented generation [Lewis *et al.*, 2020; Singhal *et al.*, 2023].

Evaluating the output of these generative systems—particularly for novel hypothesis generation, in absence of 'ground truth'—is its own challenge. General evaluation strategies include expert-based assessments, automated semantic similarity scoring and validation against curated benchmarks [Liu *et al.*, 2025], but recent approaches rely heavily on model-based evaluation, i.e. using other LLMs as judges [Li *et al.*, 2024]. This is more accessible than consulting human domain experts, but sensitive to the relative strengths of the generator and the judge [Dorner *et al.*, 2024] and model-based evaluation alone may not fully capture a system's capacity for genuine scientific discovery.

1.2 Related Work

Recent multi-agent systems leverage LLMs with specialized roles to automate scientific discovery [Ren *et al.*, 2025]. Examples include the AI Scientist [Lu *et al.*, 2024], ResearchAgent [Baek *et al.*, 2025], and Google's Co-scientist [Gottweis *et al.*, 2025]. Narrower biomedical frameworks include SciAgents [Ghafarollahi and Buehler, 2024] and IntelliScope [Aamer *et al.*, 2025]. Biomni [Huang *et al.*, 2025b] is an generalist agent with access to various bioinformatics tools and here serves as our state-of-the-art biomedical baseline.

Evaluating generative systems at scale is challenging: human expert assessment is ideal, but difficult to perform reliably at scale [Alkan *et al.*, 2025]. Temporal evaluation tests the rediscovery of known findings using held-out historical data [Sybrandt *et al.*, 2018; Sybrandt *et al.*, 2020], while recent methods rely on LLMs as judges to score hypotheses [Qi *et al.*, 2024]. To mitigate biases inherent in direct LLM grading [Liusie *et al.*, 2023] and the lack of uncertainty quantification or model diagnostics in the Elo ratings [Elo, 1978; Gottweis *et al.*, 2025], paired comparisons models, such as the Bradley–Terry model [1952], provide more robust statistical assessment.

Finally, standardized benchmarks are available—including question-answering datasets like PubMedQA [Jin *et al.*, 2019], GPQA [Rein *et al.*, 2024] and CARDBiomedBench [Bianchi *et al.*, 2025], relational classification tasks such as TruthHypo [Xiong *et al.*, 2025] and quantitative benchmarks, e.g. HypoBench [Liu *et al.*, 2025], DiscoveryBench [Majumder *et al.*, 2024] and LAB-Bench [Laurent *et al.*, 2024]—but primarily test knowledge retrieval or classification rather than generating *de novo* multi-entity hypotheses.

1.3 Contributions

We propose BIODISCO, a multi-agent framework that uniquely integrates LLM reasoning with access to biomedical KGs and real-time scientific literature, featuring an internal multi-dimensional scoring mechanism for iterative hypothesis refinement. Our system orchestrates a pipeline of collaborative 'agents' that construct, evaluate and refine biomedical hypotheses, generating evidence-based insights and inferring verifiable relationships beyond the explicit knowledge base.

Unlike systems that treat hypothesis generation as a 'one-shot' task or do not integrate domain knowledge, BIODISCO employs a sophisticated self-critique loop grounded in dual-mode evidence. This paper presents the following:

1. BIODISCO, a modular automated hypothesis generation framework, based on LLM agents augmented with interfaces to biomedical KGs and literature search;
2. a rigorous evaluation of the system's ability to predict unpublished scientific relationships, including:
   (a) a temporal evaluation on two held-out datasets;
   (b) an ablation study, leveraging a Bradley–Terry model accounting for ties and order effects;
   (c) a human evaluation by experts in two biomedical domains, modelled using item-response theory;
3. an open-source Python package of the BIODISCO framework, available from the Python Package Index via `pip`[1], with source code and scripts for reproducing the experiments available on GitHub[2].

A detailed case study is presented in the supplement[3].

2 Multi-Agent Hypothesis Generator

The BIODISCO framework, illustrated in Figure 1, operates through a 'multi-agent' architecture. Here we use *agent* to mean a role-specific LLM-based module. Each agent has a distinct role to facilitate generation, evaluation, refinement and final selection of hypotheses. Unlike conventional retrieval-augmented generation systems, BIODISCO employs a dynamic tool selection policy, in which agents decide when and how to query KGs or scientific literature. BIODISCO is, to our knowledge, the first system to integrate KG querying, literature search, 'REVIEWER' agents and explicit numerical scoring within a unified feedback loop, wherein each hypothesis is internally evaluated and refined on multiple metrics.

The core of BIODISCO's functionality lies in the network of specialized agents, illustrated in Figure 2. The agents differ in their system prompts, workflow state, planning responsibility, permitted tool use, access to external environments and position within the feedback loop. Intermediate summaries, retrieved subgraphs, scores and feedback are routed between agents as shared working context.

The BACKGROUND agent searches an academic corpus to gather scientific evidence, while the EXPLORER queries a biomedical KG to retrieve relevant subgraphs. The SCIENTIST then formulates initial hypotheses by integrating information from the summarized literature and subgraph data, proposing novel associations. These candidates are evaluated by a CRITIC, who scores each hypothesis for novelty, verifiability, relevance and significance [following Qi *et al.*, 2024] and provides structured feedback. The REVIEWER agent uses this feedback to formulate refinement strategies, such as suggesting new entities or expanded literature review, and invokes dedicated interfaces to conduct more specialized KG queries or fine-grained literature searches. The REFINER

[1]https://pypi.org/project/biodisco/
[2]https://github.com/yujingke/BioDisco
[3]https://github.com/yujingke/BioDisco_Supplement

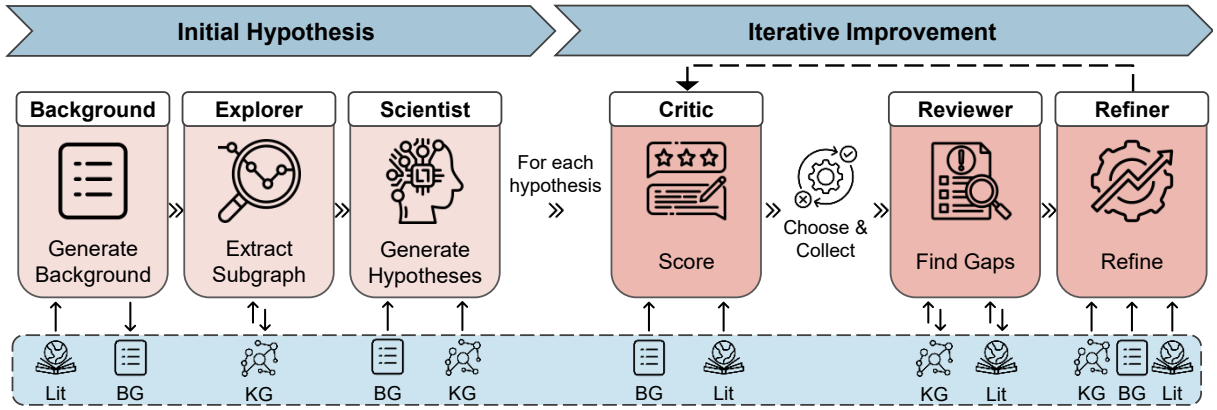

Figure 2: BIODISCO architecture. Each agent has a distinct role, some augmented with external tools. Agents interact sequentially: the user's input is first processed by the BACKGROUND agent to generate a topical summary, guiding KG extraction by the EXPLORER and initial hypothesis generation by the SCIENTIST. Each hypothesis undergoes an iterative cycle of evidence retrieval and refinement. Finally, evaluations from the CRITIC agent are used to identify the most promising hypotheses. Here, Lit refers to the literature interface, KG to the KG interface, and BG to the generated background.

amends the hypotheses based on this guidance and additional information. The decision module monitors CRITIC scores to select and collect hypotheses throughout the feedback loop. Formal agent definitions, including inputs, outputs, tool permissions, routing logic and prompt constraints, are provided in the supplement.

2.1 Dual-Mode Evidence Grounding

To avoid hallucinations, BIODISCO exploits two complementary evidence sources: a biomedical KG and academic literature retrieved via the PubMed API. These sources are dynamically queried throughout the hypothesis generation process.

The **KG interface** enables retrieval of structured biomedical knowledge from an external graph. We adopt PrimeKG [Chandak *et al.*, 2023], hosted in a Neo4j instance: queries are formulated using Cypher for efficient subgraph retrieval. For each query, agents supply a background summary and a set of keywords. These keywords are first mapped to canonical KG nodes via pretrained biomedical embeddings [`BioSimCSEBioLinkBERT-BASE`; Kanakarajan *et al.*, 2022], Candidate nodes are then selected for contextual relevance by an LLM using the supplied background summary, to reduce ambiguity and stay on topic. The resulting set of nodes defines a custom query, which is executed to retrieve evidence in the form of both direct and multi-hop relations. Query parameters, including relation types and traversal depth, are either drawn from domain defaults or overridden by the upstream agent according to the current refinement target. Returned subgraphs are summarized and converted into standardized plaintext representations for downstream use.

The **literature interface** enables agents to access biomedical literature programmatically via a context-aware query and retrieval workflow. At each invocation, agents supply a set of keywords. For background research, only keywords are provided, whereas for evaluation, the current hypothesis is included, and for revision, both the hypothesis and low-score feedback are appended. An LLM-based planning module analyzes these inputs to generate structured PubMed query strategies, grouping terms by semantic similarity and determining optimal Boolean logic for combining concepts. These strategies are converted into PubMed-API-compatible queries, including support for MeSH and TIAB fields as well as date-range constraints. If the initial strategy yields insufficient results, the interface automatically relaxes query constraints to increase recall.

2.2 Iterative Refinement & Multi-Agent Collaboration

The BIODISCO framework employs a dynamic process of cooperative evaluation and improvement. This iterative self-critique loop, unique in its integration of an internal critic and (external) dual-mode evidence, is designed to enhance hypothesis quality and credibility of initially generated hypotheses. The preliminary generation stage is driven by user input and involves three specialized LLM agents. We demonstrate this with an illustrative example.

1. The user provides a research topic or set of keywords.

   "Role of GPR153 in vascular injury and disease"

2. The BACKGROUND agent searches biomedical literature to synthesize a textual summary of the research area.

   "GPR153 plays a crucial role in vascular injury responses by modulating critical signaling pathways such as cAMP, YAP/TAZ ..."

3. The EXPLORER queries the KG to obtain context, treating user-provided terms as seed entities and retrieving a local subgraph describing their connections.

   "Nodes: GPR153 (gene/protein), CEBPB ... Direct Edges: CEBPB → protein_protein → GPR153... MultiHop Paths: PHYHIP → ... TTR → molfunc_protein → PP2CA → ..."

4. Roleplaying a biomedical researcher, the SCIENTIST integrates the plaintext summary and subgraph to formulate initial hypotheses as novel associations between two

or more entities, in natural language. Three such hypotheses are generated in parallel.

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression via the YAP/TAZ pathway ..."

Each initial hypothesis is iteratively refined to improve its quality and credibility. The feedback loop, described in steps 5–7, cycles through evaluation, information retrieval and targeted revision. This involves three additional agents:

5. The CRITIC evaluates each hypothesis alongside the retrieved background and literature summary, providing structured feedback in the form of numerical scores and comments detailing strengths and weaknesses.

> "Novelty: 4; Relevance: 5; Significance: 4; Verifiability: 4. The hypothesis can be tested using genetic manipulation ... however, the complexity of the regulatory networks may introduce challenges in isolating ... Overall Score: 17/20"

6. The REVIEWER identifies specific deficiencies (e.g. low novelty or weak evidence) and formulates targeted refinement strategies, such as suggesting new entities or an expanded literature search, invoking the EXPLORER or BACKGROUND agents, respectively.

> "Actions: Background, KG. Suggestions: Novelty - The hypothesis lacks exploration of the specific biological mechanisms ... Verifiability - The hypothesis lacks robust experimental support ..."

7. The REFINER modifies the candidate hypotheses based on the feedback and new evidence, adjusting phrasing, structure or content as needed before returning the refined versions to the CRITIC.

> "GPR153 activation in vascular smooth muscle cells enhances pro-inflammatory gene expression by promoting CEBPB-mediated YAP1 signaling, thereby potentially integrating with EGR1 and GSK3B pathways to exacerbate neointima formation following vascular injury ..."

This feedback loop repeats for up to three cycles. After each round, the CRITIC assigns four scores for novelty, relevance, significance and verifiability, and the total score is computed out of 20. If the score reaches the acceptance threshold of 19/20, the hypothesis is accepted early; otherwise, it is passed to the REVIEWER and REFINER for refinement until the iteration limit is reached. Finally, BIODISCO outputs the highest-scoring version of each hypothesis, with its broken-down scores and supporting evidence.

3 Evaluation of LLM-Generated Hypotheses

We designed a three-part empirical evaluation to assess the quality and novelty of hypotheses generated by BIODISCO. First, a temporal evaluation tests the system's ability to recover scientific relationships discovered after a fixed time cutoff. Second, an ablation study quantifies the contributions of the agentic architecture, tool use and refinement protocol. Finally, a human evaluation assesses hypothesis quality based on expert judgments from researchers in cardiovascular disease and immunology.

Except where otherwise specified, all agents were based on GPT-4.1, with knowledge cutoff at 1st June, 2024.

3.1 Can BIODISCO Predict Future Hypotheses?

First, we test if the system can generate hypotheses not present in its training corpora or knowledge interfaces.

Datasets

We used two publicly available datasets. (1) An unseen dataset from Qi *et al.* [2024], containing background–hypothesis pairs curated from biomedical literature, published in August 2023. The background text was used as the input to BIODISCO, and the paired hypothesis served as the gold-standard target for semantic comparison. We filtered this dataset to focus on core biomedical findings, removing entries with psychology or epidemiology backgrounds, resulting in 134 samples. (2) TruthHypo [Xiong *et al.*, 2025] provides keyword pairs across three categories: chemical–gene, gene–gene, and gene–disease. Each entity pair is labelled with a relationship type: positive, negative, or no relation and was extracted from literature published after 2024. Each pair was converted into a prompt asking BIODISCO to generate a hypothesis about the relationship between the two entities. We focused exclusively on positive and negative relations, excluding the 'no relation' category, as BIODISCO's primary focus is generation of plausible relationships between entities rather than confirming their absence. From this dataset, we created two subsets. We first selected 30 samples, balanced across categories, to calibrate a prompt-based classifier agent for mapping generated free-text hypotheses to positive or negative labels. From the remainder, we sampled a test set of 300 instances (limited for computational feasibility), stratified across all categories and the two classes, for evaluation.

Setup

To prevent data leakage, we enforced strict temporal cutoffs across all components of BIODISCO. Access to the PubMed interface was programmatically constrained to exclude articles published after the respective cutoff dates (Dec 2022 for Qi et al; Dec 2024 for TruthHypo) and the KG uses a fixed snapshot of PrimeKG published in April 2022. To further mitigate leakage from the underlying language model, we used GPT-3.5 Turbo (knowledge cutoff: September 1st, 2021) for the Qi *et al.* experiment and GPT-4.1 (knowledge cutoff: 1st June, 2024) for the TruthHypo experiment, ensuring that the evaluated hypotheses were not present in the LLM's pretraining data. Finally, to bridge BIODISCO's free-text hypothesis generation with the structured labels required by the TruthHypo task, we developed a classifier agent to predict the relationship class from the textual hypotheses.

Results

We evaluated the semantic alignment between the generated and 'gold-standard' hypotheses from Qi *et al.* [2024] by computing their cosine similarity using a pretrained biomedical sentence encoder, `BioBERT-mnli-snli-scinli-scitail-mednli-stsb` [Deka *et al.*, 2022]. As a baseline, we calculated pairwise similarities between unrelated 'gold' hypotheses (i.e. those

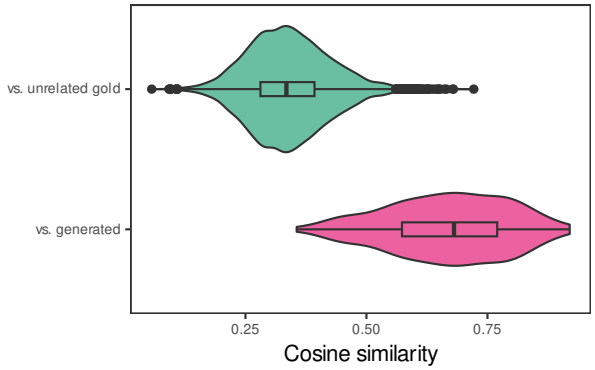

Figure 3: Violin plot, showing that hypotheses generated by BIODISCO are semantically closer to 'gold' hypotheses than gold hypotheses are to other hypotheses. Top shows pairwise similarity of unrelated gold hypotheses; bottom shows similarity of BIODISCO-generated hypotheses to gold standard for the same topics

Category	Precision	Recall	F_1	Accuracy
Chemical–Gene	0.90	0.90	0.90	0.90
Disease–Gene	0.82	0.82	0.82	0.82
Gene–Gene	0.83	0.83	0.83	0.83

Table 1: Performance of temporally-restricted BIODISCO on the TruthHypo benchmark for three categories of task

associated with different background contexts), serving as a negative control for expected similarity in semantically unrelated pairs. Figure 3 shows that generated hypotheses were more semantically similar to the gold standard (median similarity: 0.68) than unrelated gold hypotheses were to each other (median: 0.34), demonstrating that BIODISCO reliably produces hypotheses closer to human-curated ground truth than expected by chance. Cosine similarity only measures semantic relatedness, but our manual comparisons (see supplement) suggest it is appropriate in this setting.

In the second experiment, we assessed whether BIODISCO could generate a hypothesis that correctly implies the relationship for a given entity pair. We used our classifier agent to perform binary classification (positive, negative) on the generated text. Results are given in Table 1, showing the classifier achieved high precision and recall, indicating that the hypotheses generated by BioDisco contain accurate and discernible relational signals.

The TruthHypo evaluation presented in Table 1 excludes the 'no relation' class, as the primary goal of this experiment was to assess BIODISCO's generative capacity—specifically, its ability to recover valid future scientific relationships—rather than its ability to abstain. However, in practical scientific discovery settings, abstention is important for preventing hallucinations and controlling false positives. Accordingly, we reran the TruthHypo experiments with the inclusion of the 'no relation' class; results are reported in the Appendix.

These results indicate that BIODISCO, operating under strict temporal constraints, can successfully infer novel, semantically relevant and factually verifiable hypotheses.

3.2 Ablation of Knowledge Access & Refinement

To evaluate the contribution of each component to the framework, we compared BIODISCO against a series of ablated configurations. These included:

1. GPT-4.1 (baseline): a standalone single LLM, prompted without agentic structure or external tools.
2. Multi-agent system: a system mirroring our architecture, but without iterative refinement or knowledge interfaces.
3. Multi-agent + tools: a system with access to the literature and KG interfaces but without iterative refinement.
4. Multi-agent + refine: a system with an iterative refinement loop, but without access to knowledge interfaces.
5. BIODISCO: the full system described in § 2.

For a direct comparison with other biomedical agentic systems, we also generated a series of hypotheses with Biomni [Huang *et al.*, 2025b]:

6. Biomni: a general-purpose biomedical AI agent, based on GPT-4.1, with access to various bioinformatics tools but no predefined workflow or iteration loop.

LLM-based evaluation is only a proxy for expert judgement, but prior work has shown agreement between model-based evaluators and human expert judgements [Lu *et al.*, 2024]. We employed a dedicated pairwise LLM evaluator combined with a Bradley–Terry model, rather than directly scoring individual hypotheses. This statistical approach provides more stable and human-aligned assessments of hypothesis quality across multiple dimensions [Liusie *et al.*, 2024], yielding point estimates and uncertainty quantification via 95% comparison intervals, rather than direct scores only.

Each of the systems was given 100 inputs spanning a diversity of topics, including oncology, ageing and cellular senescence, generating one hypothesis per input. For every pair of generated hypotheses from different configurations, an LLM judge selected its preferred candidate (or declared a tie) for each of four metrics: *novelty*, *relevance*, *verifiability* and *significance* [from Qi *et al.*, 2024].

We fitted a Bradley–Terry [1952] paired comparison model to the results, given by

$$\log \text{odds}(i \text{ beats } j) = \beta_i - \beta_j + \alpha z_{ij}, \qquad (1)$$

where i and j denote 'players' or LLM configurations, β_i is the continuous latent *ability score* of the i^{th} LLM system and α is a 'home advantage' parameter, estimating the possible order effect bias, with $z_{ij} = 1$ if the hypothesis of i is presented to the evaluator first in the pair, and $z_{ij} = -1$ otherwise. Ties are treated as a half-win for each player. The ability scores for each configuration and each metric provide a univariate ranking that captures relative performance.

Fitted ability scores and their 95% comparison intervals (calculated using a quasi-variance approximation; Firth, 2004) are presented in Figure 4. The results show a clear, progressive improvement in LLM-judged hypothesis novelty and significance as components are added to the system. The single LLM baseline was least preferred. A multi-agent structure provided significant improvement, suggesting a benefit from role-based reasoning alone. Performance further increased

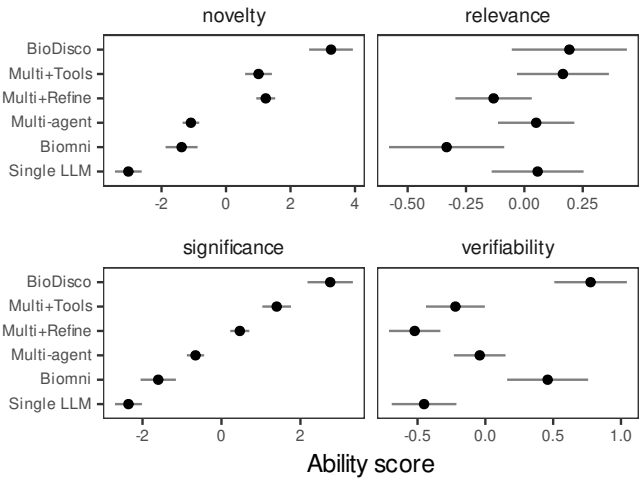

Figure 4: Centipede plot of ability scores for BIODISCO, four ablation configurations and Biomni, with 95% comparison intervals. A multi-agent system clearly outperforms a single LLM (GPT-4.1) generating novel, significant hypotheses; tool use (i.e. KG and literature search) and iterative refinement each yield further improvements. Biomni, a competing system, is mostly better than a single LLM and produces verifiable hypotheses, but less novel and significant than the full BIODISCO framework

Metric	Estimate	Std. Error	Statistic	p-value
Novelty	-0.10	0.21	-0.51	0.61
Relevance	0.26	0.11	2.38	0.02
Significance	0.09	0.17	0.52	0.60
Verifiability	1.11	0.13	8.68	0.00

Table 2: Order effect parameter estimates, $\hat{\alpha}$, from Bradley–Terry models fitted to each of the four evaluation metrics

after adding external tool use and iterative refinement. Between the two, tool use provided a greater benefit, emphasizing the value of grounding the generation process in external knowledge sources. Finally, the full BIODISCO system was most preferred for these metrics under the Bradley–Terry paired-comparison model, suggesting a complementary benefit of dual-mode evidence and iterative refinement.We compared BIODISCO against Biomni [Huang *et al.*, 2025b], a state-of-the-art generalist biomedical agent, and found that BIODISCO consistently generates hypotheses that are more novel and more significant than those from Biomni.

Interestingly, patterns for relevance and verifiability were less clearly defined: pairs of configurations have overlapping comparison intervals, with no clear winner. We find that both BIODISCO and Biomni produce more verifiable hypotheses than the other variants; wide confidence intervals suggest a statistically insignificant difference between the two.

Estimates for the order effect parameter, $\hat{\alpha}$, are given in Table 2 and indicate a statistically significant position bias for 'verifiability' evaluations (in favour of the item that appears first) but not enough evidence to indicate a statistically significant order effect exists for novelty, relevance or significance.

A direct scoring evaluation is provided in the supplement.

3.3 Human Expert Evaluation

To ascertain the practical utility of BIODISCO's outputs, our final evaluation involved two groups of human experts.

Methodology

We recruited nine experts from two biomedical fields—four in immunology and five in cardiovascular medicine—to complete an online survey about pairs of hypotheses generated by BIODISCO. Hypothesis pairs were generated as follows.

For immunology, BIODISCO was provided with the prompt: "T Cell Exhaustion Mechanisms and Therapeutic Targets in NSCLC". The system was run five times to generate five distinct hypotheses, to explore its capacity to explore a single topic in depth. For cardiovascular disease (CVD), BIODISCO was provided five unique prompts related to CVD, to assess performance across different topics. One initial and one refined hypothesis was generated for each input.

For every sub-topic, experts were presented with both the initial version (the first complete hypothesis generated) and the final version (after iterative refinement). Experts independently rated each hypothesis on a 1–5 scale according to the four metrics: novelty, relevance, significance and verifiability. They also rated their confidence in their assessments and provided qualitative feedback.

Analysis

Survey responses were modelled using a single polytomous Rasch model, implemented as a Bayesian cumulative probit mixed effects model [Bürkner, 2021] for all four metrics—one for each group of experts:

$$P(Y_{ijm} \leq k) = \Phi(\tau_k - \eta_{ijm}), \tag{2}$$

where Y_{ijm} denotes the ordinal rating ($k = 1, \ldots, 5$) given by rater i to hypothesis j on metric m, Φ denotes the standard normal distribution function and τ_k are the latent thresholds that partition the ordinal rating scale. The linear predictor η_{ijm} contains both fixed and random effects:

$$\eta_{ijm} = \underbrace{\beta_m}_{\text{metric effect}} + \underbrace{u_i}_{\text{rater effect}} + \underbrace{v_{jm}}_{\text{hypothesis-on-metric effect}}, \tag{3}$$

where β_m is a fixed effect representing the average quality or 'easiness' for metric m, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ is a rater-specific random effect capturing overall leniency or severity of rater i, and $v_{jm} \sim \mathcal{N}(0, \sigma_v^2)$ is a hypothesis- and metric-specific random effect reflecting the relative quality of hypothesis j on metric m. By disentangling rater- and metric-specific biases and accounting for the ordinal nature of responses, the model provides a more robust and unbiased estimate of the underlying quality of each hypothesis, while pooling information across the different questions.

Results

Figure 5 presents the distribution of expert ratings. In immunology, domain experts consistently provided high scores for BIODISCO-generated hypotheses. In contrast, experts in CVD had more varied opinions on hypothesis novelty.

Posterior estimates of hypothesis-on-metric effects are shown in Figure 6. A clear improvement in rated novelty is apparent following refinement. Interestingly, however, refined hypotheses are not always judged to be more

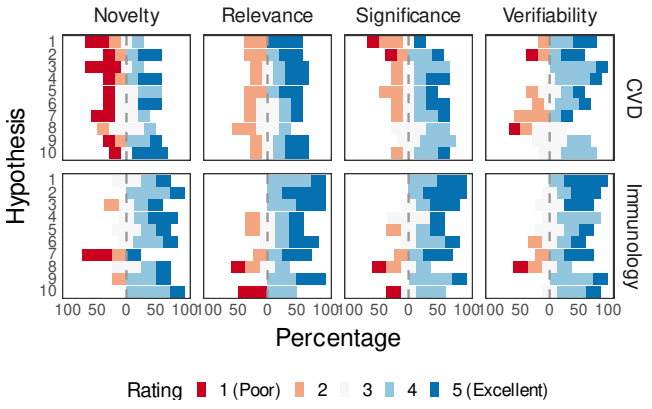

Figure 5: Ratings given by two independent groups of human experts to 10 hypotheses generated for respective topics of cardiovascular disease (CVD) and immunology

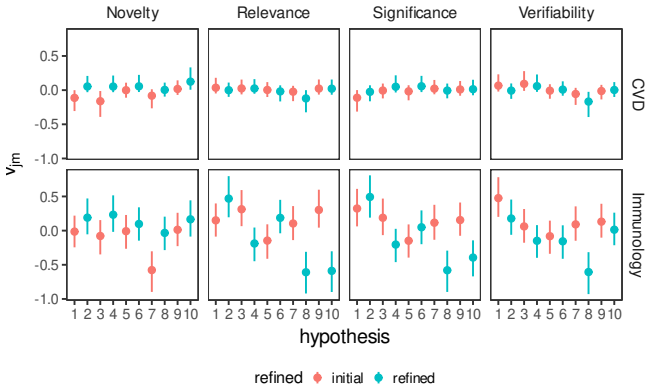

Figure 6: Posterior distributions of hypothesis-on-metric effect, $\hat{v}_{jm}$, from human evaluations of hypotheses generated by BIODISCO for cardiovascular disease and immunology

verifiable—which aligns with our ablation study. This reveals an inverse relationship between novelty and verifiability: as hypotheses move away from established literature, proposed mechanisms become more complex. This is to be expected in *ab initio* discovery: easily verifiable hypotheses represent 'low-hanging fruit', likely already harvested by researchers.

Qualitative feedback of the system's ability to generate scientifically valuable and relevant hypotheses was positive. Some hypotheses showed less novelty or proposed unvalidated connections, but experts highlighted their plausibility and potential experimental tractability. One commented, e.g.

> "This idea could be tested with targeted experiments like activating or inhibiting GPR153 in VSMCs, running transcriptomic and ChIP-seq analyses to track CEBPB/NRF1/CD7 activity, and using in vivo vascular injury models with genetic knockouts. If these links hold up, the network could point to new therapeutic targets for preventing neointima formation."

4 Discussion & Conclusion

BIODISCO is a multi-agent framework for biomedical discovery, made openly available for research use. A temporal evaluation shows the system's capability for extrapolation. Ablation suggests that iterative refinement and external knowledge each help generate hypotheses more 'novel' and 'significant'. The system was assessed by human experts using psychometric models accounting for bias and uncertainty.

Limitations

An LLM may not be the best judge of experimental feasibility; this may be explored by dedicated hypothesis *testing* frameworks, e.g. POPPER [Huang *et al.*, 2025a]. Weaker 'relevance' to an input topic may indicate breadth of exploration and diversity of knowledge, or simply a limitation of the metrics from Qi *et al.* [2024]. We compare BIODISCO against Biomni as the current state of the art. Further direct comparisons are hindered by a lack of reproducible open-source systems and the difficulty of adapting external models to our specific experimental settings. Our metrics-based evaluation is limited to the judgement of humans and LLMs of whether hypotheses *seem* novel or significant. Self-evaluation and refinement may invoke Goodhart's Law, amplifying model artifacts and exaggerating practical relevance.

Any temporal evaluation assumes discoverability equates to predictability, but even flawed or ultimately disproven hypotheses can contribute to a field. A good system should generate novel, plausible hypotheses—not just those that align with future discoveries. Finally, our comparison is only *in silico*; a wet-lab validation is beyond the scope of this paper.

Future Work

Our evaluation does not disentangle effects of architecture from data, or the relative strength of different LLMs. Any evaluation of models trained on public data is vulnerable to data leakage [e.g. Zhou *et al.*, 2023]. Another challenge lies in (interpretable) integration of different modalities, such as clinical records, omics and image data.

A Appendix

The Python package BIODISCO is free to download from the Python Package Index at https://pypi.org/project/biodisco/ using the command

```
pip install biodisco
```

An minimal example hypothesis generation program takes the form:

```
import python
BioDisco.generate(
''Role of GPR153 in vascular injury and
disease'')
```

Further documentation and details of additional experimental results are provided in the supplementary material, available online at

https://github.com/yujingke/BioDisco_Supplement

and source code for the package is available at

https://github.com/yujingke/BioDisco

References

[Aamer *et al.*, 2025] Naafey Aamer, Muhammad Nabeel Asim, Shan Munir, and Andreas Dengel. Automating AI discovery for biomedicine through knowledge graphs and LLM agents. *bioRxiv*, 2025.

[Alkan *et al.*, 2025] Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G Iyer, et al. A survey on hypothesis generation for scientific discovery in the era of large language models, 2025. arXiv:2504.05496.

[Baek *et al.*, 2025] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 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*, pages 6709–6738, April 2025.

[Bélisle-Pipon, 2024] Jean-Christophe Bélisle-Pipon. Why we need to be careful with LLMs in medicine. *Frontiers in Medicine*, 11:1495582, 2024.

[Bianchi *et al.*, 2025] Owen Bianchi, Maya Willey, Chelsea X Alvarado, Benjamin Danek, Marzieh Khani, Nicole Kuznetsov, Anant Dadu, Syed Shah, Mathew J Koretsky, Mary B Makarious, et al. CARDBiomedBench: A benchmark for evaluating large language model performance in biomedical research: A novel question-and-answer benchmark designed to assess large language models' comprehension of biomedical research, piloted on neurodegenerative diseases, 2025. bioRxiv.

[Bradley and Terry, 1952] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 12 1952.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[Bürkner, 2021] Paul-Christian Bürkner. Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software*, 100(5):1–54, 2021.

[Chandak *et al.*, 2023] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

[Deka *et al.*, 2022] Pritam Deka, Anna Jurek-Loughrey, et al. Evidence extraction to validate medical claims in fake news detection. In *International Conference on Health Information Science*, pages 3–15. Springer, 2022.

[Dorner *et al.*, 2024] Florian E Dorner, Vivian Y Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data, 2024. arXiv:2410.13341.

[Elo, 1978] Arpad E. Elo. *The Rating of Chess Players, Past and Present*. Arco Pub., New York, 1978.

[Firth, 2004] D. Firth. Quasi-variances. *Biometrika*, 91(1):65–80, March 2004.

[Ghafarollahi and Buehler, 2024] Alireza Ghafarollahi and Markus J. Buehler. SciAgents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, page 2413523, 2024.

[Gottweis *et al.*, 2025] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist, 2025.

[Himmelstein *et al.*, 2017] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726, 2017.

[Hristovski *et al.*, 2015] Dimitar Hristovski, Andrej Kastrin, Dejan Dinevski, and Thomas C Rindflesch. Constructing a graph database for semantic literature-based discovery. In *MEDINFO 2015: eHealth-enabled Health*, pages 1094–1094. IOS Press, 2015.

[Huang *et al.*, 2025a] Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Forty-second International Conference on Machine Learning*, 2025.

[Huang *et al.*, 2025b] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N. Carter, Xin Zhou, Matthew Wheeler, Jonathan A. Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical AI agent, 2025.

[Jimeno-Yepes *et al.*, 2009] Antonio Jimeno-Yepes, Rafael Berlanga-Llavori, and Dietrich Rebholz-Schuhmann. Exploitation of ontological resources for scientific literature analysis: searching genes and related diseases. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7073–7078, 2009.

[Jin *et al.*, 2019] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019. arXiv:1909.06146.

[Kanakarajan *et al.*, 2022] Kamal raj Kanakarajan, Bhuvana Kundumani, Abhijith Abraham, and Malaikannan

Sankarasubbu. BioSimCSE: BioMedical sentence embeddings using contrastive learning. In *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI)*, pages 81–86, December 2022.

[Laurent *et al.*, 2024] Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D White, and Samuel G Rodriques. Lab-bench: Measuring capabilities of language models for biology research, 2024. arXiv:2407.10362.

[Lewis *et al.*, 2020] 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. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[Li *et al.*, 2024] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods, 2024. arXiv:2412.05579.

[Liu *et al.*, 2025] Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. HypoBench: Towards systematic and principled benchmarking for hypothesis generation, 2025. arXiv:2504.11524.

[Liusie *et al.*, 2023] Adian Liusie, Potsawee Manakul, and Mark JF Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models, 2023. arXiv:2307.07889.

[Liusie *et al.*, 2024] Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 139–151, March 2024.

[Lu *et al.*, 2024] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024. arXiv:2408.06292.

[Majumder *et al.*, 2024] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models, 2024. arXiv:2407.01725.

[Perdomo-Quinteiro and Belmonte-Hernández, 2024] Pablo Perdomo-Quinteiro and Alberto Belmonte-Hernández. Knowledge graphs for drug repurposing: a review of databases and methods. *Briefings in Bioinformatics*, 25(6):bbae461, 2024.

[Qi *et al.*, 2024] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*, 2024.

[Rein *et al.*, 2024] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level Google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024.

[Ren *et al.*, 2025] Shuo Ren, Pu Jian, Zhenjiang Ren, Chunlin Leng, Can Xie, and Jiajun Zhang. Towards scientific intelligence: A survey of LLM-based scientific agents, 2025. arXiv:2503.24047.

[Singhal *et al.*, 2023] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

[Spangler *et al.*, 2014] Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter J. Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence A. Donehower, Ying Chen, and Olivier Lichtarge. Automated hypothesis generation based on mining scientific literature. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge discovery and data mining*, 2014.

[Sybrandt *et al.*, 2018] Justin Sybrandt, M. Shtutman, and Ilya Safro. Large-scale validation of hypothesis generation systems via candidate ranking. *2018 IEEE International Conference on Big Data (Big Data)*, pages 1494–1503, 2018.

[Sybrandt *et al.*, 2020] Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. Agatha: automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2757–2764, 2020.

[Xiong *et al.*, 2025] Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models, 2025. arXiv:2505.14599.

[Zheng *et al.*, 2020] Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. PharmKG: a dedicated knowledge graph benchmark for biomedical data mining. *Briefings in Bioinformatics*, 22(4):bbaa344, 12 2020.

[Zhou *et al.*, 2023] Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your LLM an evaluation benchmark cheater, 2023. arXiv:2311.01964.