BioVerge: A Comprehensive Benchmark and Study of Self-Evaluating Agents for Biomedical Hypothesis Generation

Fuyi Yang*
University of California, Los Angeles
fyyang@ucla.edu

Chenchen Ye*
University of California, Los Angeles
ccye@cs.ucla.edu

Mingyu Derek Ma
University of California, Los Angeles
ma@cs.ucla.edu

Yijia Xiao
University of California, Los Angeles
yijia.xiao@cs.ucla.edu

Matthew Yang
University of California, Los Angeles
ymatt24@ucla.edu

Wei Wang
University of California, Los Angeles
weiwang@cs.ucla.edu

Abstract

Hypothesis generation in biomedical research has traditionally centered on uncovering hidden relationships within vast scientific literature, often using methods like Literature-Based Discovery (LBD). Despite progress, current approaches typically depend on single data types or predefined extraction patterns, which restricts the discovery of novel and complex connections. Recent advances in Large Language Model (LLM) agents show significant potential, with capabilities in information retrieval, reasoning, and generation. However, their application to biomedical hypothesis generation has been limited by the absence of standardized datasets and execution environments. To address this, we introduce BIOVERGE, a comprehensive benchmark, and BIOVERGE AGENT, an LLM-based agent framework, to create a standardized environment for exploring biomedical hypothesis generation at the frontier of existing scientific knowledge. Our dataset includes structured and textual data derived from historical biomedical hypotheses and PubMed literature, organized to support exploration by LLM agents. BIOVERGE AGENT utilizes a ReAct-based approach with distinct Generation and Evaluation modules that iteratively produce and self-assess hypothesis proposals. Through extensive experimentation, we uncover key insights: 1) different architectures of BIOVERGE AGENT influence exploration diversity and reasoning strategies; 2) structured and textual information sources each provide unique, critical contexts that enhance hypothesis generation; and 3) self-evaluation significantly improves the novelty and relevance of proposed hypotheses. Our dataset and code are accessible here: https://github.com/Fuyi-Yang/BioVerge/

1 Introduction

Traditional hypothesis generation tasks centered around finding hidden relationships among concepts within the scientific domain. A widely recognized method was **Literature Based Discovery** (LBD), which highlights that undiscovered knowledge exist within scientific literature and are obtainable through automated extraction systems Bhasuran et al. [2023]. Importantly, the ABC principle depicted in Figure 1 tackles the LBD task Swanson [1986], stating that if distinct literature connects

*Equal Contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Science.

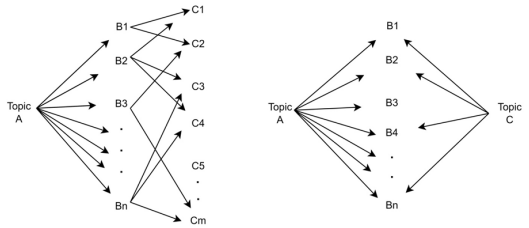

Figure 1: The ABC principle as a method to explore Literature Based Discovery (LBD).

concepts A/B and concepts B/C, then the intersections of concept B infers potential links between concepts A/C. This technique harnessed many novel relationships that were proven experimentally, including connections between magnesium deficiency and migraine, Raynaud's disease and fish oil, significantly contributing to scientific research Swanson [1988], Bhasuran et al. [2023]. Despite earlier successes, these traditional co-occurrence, semantic, and graph-based methods often relied on a single data type, limiting their ability to discover subtler connections Smalheiser et al. [2009], Fleuren and Alkema [2015], Jha et al. [2018], Wilson et al. [2018], Tshitoyan et al. [2019], Sybrandt et al. [2020]. Furthermore, with little reasoning or refinement, these approaches fail to develop intricate reasoning paths that lead to more discoveries Hristovski et al. [2006].

Recently, Large Language Models (LLM) agents have garnered significant interest from its performance on exploration, reasoning, and generation tasks Tong et al. [2023], Wang et al. [2024], Ye et al. [2024]. LLM-based agents can utilize tools to gather information, perform thoughtful reasoning, and create novel proposals Qi et al. [2023], Baek et al. [2024], Zhang et al. [2024], Kumbhar et al. [2025], Su et al. [2025]. However, despite its potential, LLM agents have limited applications in the biomedical hypothesis generation, due to a lack of standardized benchmarks and databases.

To address this gap, we introduce BIOVERGE, a benchmark that stores biomedical knowledge as structured hypotheses triplets from PubTator3 and textual PubMed literature Wei et al. [2024], PubMed [2025]. The knowledge base contains data published before Jan. 1, 2024 to prevent test set data contamination, and both datasets perform cleaning and filtering to ensure experiment and evaluation fairness. We rank candidate test hypotheses by the impact factor (IF) of their publishing journals to select a credited, reliable test set SCImago [2025].

We also propose BIOVERGE AGENT, an LLM agent framework leveraging BIOVERGE to perform biomedical hypothesis generation. We construct an API interface to support agent tool calling to extensively explore and retrieve context in natural language. By alternating between *Generation* and *Evaluation* modules, BIOVERGE AGENT can propose, self-evaluate, and iteratively refine hypothesis. We further compare Single and Double Agent architectures, which differ in memory sharing between the modules, to explore their impacts on hypothesis generation reasoning and performance.

Finally, we perform extensive experiments on our constructed test dataset. Our results reveal that Single and Double Agent clearly present different reasoning strategies and suggests a tradeoff between exploration and accuracy performance. Additionally, the access to diverse structural and textual sources, and self-evaluation of past proposals, all encourage BIOVERGE AGENT to refine its reasoning trajectory and notably improve the novelty and relevance of hypothesis proposals.

Overall, our main contributions are three-fold:

1. We introduce the BIOVERGE benchmark, which enables tool-augmented reasoning over both structured and unstructured biomedical data.
2. We demonstrate the capability of BIOVERGE AGENT to generate hypotheses that push the boundaries of existing biomedical knowledge.
3. We present a systematic study of self-evaluating agent architectures and information sources, identifying key design choices that enhance biomedical hypothesis generation.

2 BIOVERGE benchmark

We define **hypothesis** as a triplet (s, r, o), where $s, o \in \mathcal{E}$ are biomedical entities and $r \in \mathcal{R}$ is one of twelve relation types in Table 2, all defined in the PubTator3 ontology Wei et al. [2024]. We denote **hypothesis description** d as a natural language explanation for (s, r, o).

Biomedical hypothesis generation task. Given entities $s, o \in \mathcal{E}$, propose relation $r \in \mathcal{R}$ and description d such that $(s, r, o) \notin H$ (knowledge base), $(s, r, o) \in T$ (test set), d is distinct from historical articles with hypotheses of the same s, o entities and aligns with ground truth articles.

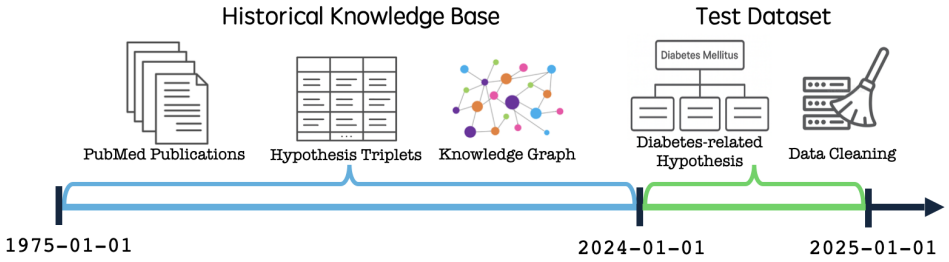

Figure 2: The cutoff date separation between the historical knowledge base and test dataset candidates for dataset construction. The knowledge base contains article corpus, knowledge graph, and triplet representations. The test dataset is cleaned to ensure all test queries are proposed after Jan. 1, 2024, related to Diabetes Mellitus, and ranked by the impact factor (IF) score of its publishing journal.

Table 1: Dataset statistics for the knowledge base and test set.

Dataset	Triplets	Publications	Unique Entities
Knowledge Base	10,566,798	8,896,514	558,006
Test Set	177	135	174

Table 2: The relations, descriptions, and valid entity pairs for each relation defined in Pubtator3.

Relation	Description	Valid Entity Pairs
Associate	Complex or unclear relationships	(Chemical, Disease), (Chemical, Gene), (Chemical, Variant), (Disease, Gene), (Disease, Variant), (Variant, Variant)
Cause	Triggering a disease by a specific agent	(Chemical, Disease), (Variant, Disease)
Compare	Comparing the effects of two chemicals or drugs	(Chemical, Chemical)
Cotreat	Simultaneous administration of multiple drugs	(Chemical, Chemical)
Drug interact	Pharmacodynamic interactions between two chemicals	(Chemical, Chemical)
Inhibit	Reduction in amount or degree of one entity by another	(Chemical, Variant), (Gene, Disease)
Interact	Physical interactions, such as protein-binding	(Chemical, Gene), (Chemical, Variant), (Gene, Gene)
Negative correlate	Increases in the amount or degree of one entity decreases the amount or degree of the other entity	(Chemical, Gene), (Chemical, Variant), (Gene, Gene)
Positive correlate	The amount or degree of two entities increase or decrease together	(Chemical, Chemical), (Chemical, Gene), (Gene, Gene)
Prevent	Prevention of a disease by a genetic variant	(Variant, Disease)
Stimulate	Increase in amount or degree of one entity by another	(Chemical, Variant), (Gene, Disease)
Treat	Treatment of a disease using a chemical or drug	(Chemical, Disease)

2.1 Dataset construction

In following sections we describe the construction and cleaning of the historical knowledge base and test set, along with evaluation metrics for experiment studies. Dataset statistics are shown in Table 1.

2.1.1 Knowledge base

BIOVERGE contains structured triplets, literature articles, and knowledge graph, shown in Figure 2. Specifically, the knowledge base contains hypotheses and articles published before Jan. 1, 2024.

Structured sources. PubTator3 provides abundant article-extracted hypothesis triplets, but limited natural language entity names. Therefore, we standardize triplet entities from the official databases: Medical Subject Headings or Mesh (chemical and disease), NCBI Gene (gene), NCBI Taxonomy (species), and Cellosaurus (cellline) National Library of Medicine [2025], National Center for Biotechnology Information [2025a,b], Bairoch [2018]. All entities are appended with entity names, except "mutation" types due to lack of interpretability. We further construct a searchable knowledge graph with entity nodes and relation edges. Each unique triplet is a transition $(s \xrightarrow{r} o)$, where $(s \xrightarrow{r} o)$ and $(o \xrightarrow{r} s)$ are distinct transitions, possibly with different relations.

Textual sources. PubMed articles identified with titles and abstracts map to all triplets in the knowledge base, providing textual historical contexts. Articles missing the publication date or its title and abstract are discarded from the knowledge base.

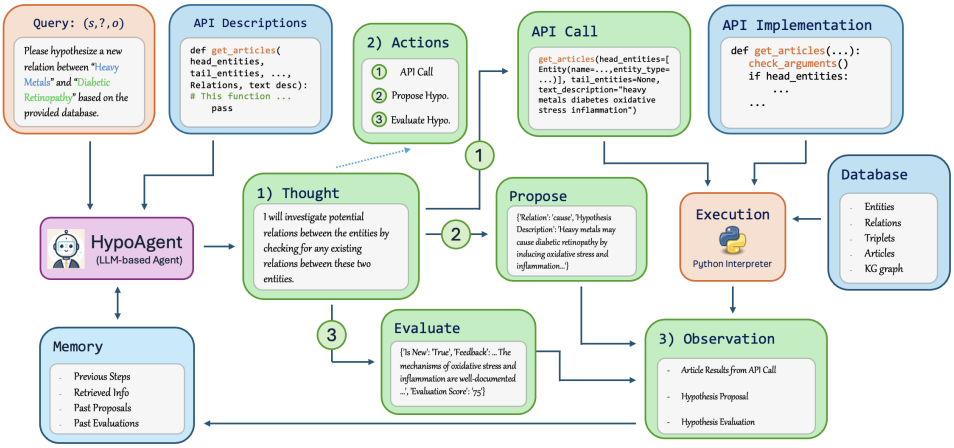

Figure 3: *Generation* and *Evaluation* modules' ReAct workflow. Given a query, the agent iteratively reflects on its current state, executes an action, and stores the observed output until termination.

Data cleaning and processing. We remove triplets that either have missing entity names or is not a valid triplet (s, r, o) combination defined in the PubTator3 report, shown in Table 2. We then merge triplets that appear in multiple articles, append all unique PMIDs, and assign the earliest publication date as the discovery date. Finally, we discard any triplets with no associated articles to ensure every triplet links to verifiable literature.

2.1.2 Test set

The test dataset contains hypotheses discovered between Jan. 1 and Dec. 31, 2024 involving "Diabetes Mellitus" (MeSH: $D003920$), thereby developing a **diabetes-related** test set. We then apply similar data cleaning as the knowledge base to filter suboptimal triplets and ensure they are absent prior to Jan. 1, 2024. Finally, we rank the remaining hypotheses with the **journal impact factor (IF)** score documented in Scimago's Scientific Journal Rank (SJR) metrics SCImago [2025]. Taking the top 50 highest ranked journals, we curate a test set of 177 hypotheses spanning diverse relation types, triplet combinations, and diabetes entities under "Diabetes Mellitus."

2.2 Answer evaluation metric

We evaluate each query's hypothesis proposal by comparing its relation r and description d against the ground truth using two metric categories:

1) **Novelty**: whether the hypothesis is absent from knowledge base. We define $\texttt{novelty}_r$ to check if $(s, r, o) \notin H$, where H is the historical knowledge base, and $\texttt{novelty}_d$ to assesses whether the generated hypothesis description d is semantically different to the historical articles that have proposed some hypothesis of the same entities s or o.

2) **Alignment**: whether the hypothesis exists in the test set. We define $\texttt{alignment}_r$ to check if $(s, r, o) \in T$, where T is the test set, and $\texttt{alignment}_d$ to evaluate the similarity of the generated hypothesis description d to the ground truth articles that support the test hypothesis.

Evaluation metrics for relations, $\texttt{novelty}_r$ and $\texttt{alignment}_r$, are computed as binary scores (1 = true, 0 = false) and expressed as percentages over the test set. Description-based metrics $\texttt{novelty}_d$ and $\texttt{alignment}_d$ are scored by an LLM evaluator out of 100 and averaged over the test set.

3 BioVerge Agent

Considering the individual benefits of co-occurrence, semantic, and graph-based methods, we introduce BioVerge Agent, an LLM agent framework that combines these approaches for hypothesis generation. We develop an API interface for agent tool calling to retrieve diverse historical information and introduce two modules, *Generation* and *Evaluation*, to iteratively generate, self-evaluate, and refine hypothesis proposals. Furthermore, we examine the effects on hypothesis generation outcomes when applying these modules in distinct Single and Double agent architectures.

3.1 API usage

We construct a comprehensive API interface for agent tool calling to access our knowledge base. The interface standardizes historical knowledge retrieval into *data classes* (Entity, Relation, PMID, Triplet, Article) and exposes a set of *functions* for the agent to call and retrieve relevant contexts.

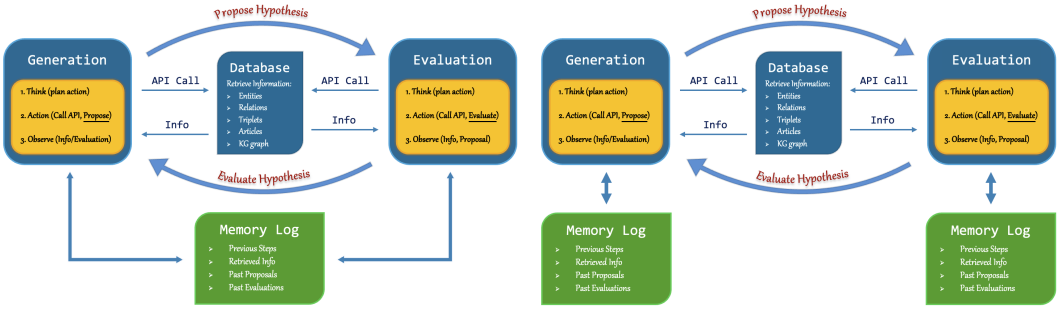

Figure 4: Single Agent (left) and Double Agent (right) architectures. Memory is shared between *Generation* and *Evaluation* modules in Single Agent and separated in Double Agent.

Table 3: The list of data classes and queryable API functions.

Dataclass / Function	**Description**
`PMID`	Represents the unique identifier of a PubMed article.
`Entity`	Represents a named biomedical entity and its type.
`Triplet`	Represents a hypothesis (subject, object, relation).
`Article`	Represents a PubMed article including PMID, title, and abstract.
`get_entities`	Retrieves a list of entities matching specified filters.
`get_relations`	Retrieves a list of relations matching specified filters.
`get_triplets`	Retrieves a list of triplets matching specified filters.
`get_articles`	Retrieves a list of PMIDs matching specified filters.
`browse_articles`	Returns metadata (title, abstract) for given PubMed IDs.
`get_shortest_entity_paths`	Finds shortest paths between two entities in the knowledge graph.
`get_mesh_parents`	Returns parent entities in MeSH for a disease or chemical.
`get_mesh_children`	Returns child entities in MeSH for a disease or chemical.
`get_mesh_siblings`	Returns sibling entities in MeSH for a disease or chemical.

API functions. The API interface covers all data types, providing both structural and textual historical contexts. Core APIs such as `get_entities`, `get_relations`, and `get_articles` return the most relevant entities, relations, and articles for a given query, while auxiliary functions like `get_relation_description` and `get_entity_description` provide additional semantic detail. Together, these APIs enable the agent to flexibly combine structured triplets with article context. The complete API list is in Table 3.

Input parameters. Each API accepts optional parameters that guide query results. Common parameters include `head_entities`, `tail_entities`, `relations`, `PMIDs`, and `text_description`. Varying these parameters enable BIOVERGE AGENT to specialize the query criteria, constrain the search space, and return results that best match the provided inputs.

3.2 Generation and evaluation agent architectures

Proposing a new relation between two entities is inherently difficult because no prior literature or triplets exist to document this connection. Even when past hypotheses mention the same entities, they rarely provide direct support for new proposals. Thus, hypothesis generation requires extensive exploration and reasoning over historical data, with no guarantee of correctness.

Modules. To address these challenges, BIOVERGE AGENT introduces *Generation* and *Evaluation* modules (system prompts in Appendix section C). We also propose two agent architectures, shown in Figure 4, and explain a practical execution example in Figure 5.

The ***Generation*** module proposes new hypotheses, including the relation and hypothesis description, and refines previously proposed hypotheses. The ***Evaluation*** module acts as a critic that accesses the most recently proposed hypothesis. The evaluation specifically addresses 3 metrics: 1) **Is New**, a boolean value stating the novelty of the hypothesis from the historical dataset; 2) **Feedback**, a natural language explanation that justifies the reasonability of the proposed hypothesis and offers suggestions for improvement; and 3) **Evaluation Score**, a numeric score from 0 to 100 assessing the reasonableness of the proposed hypothesis based on **Feedback**.

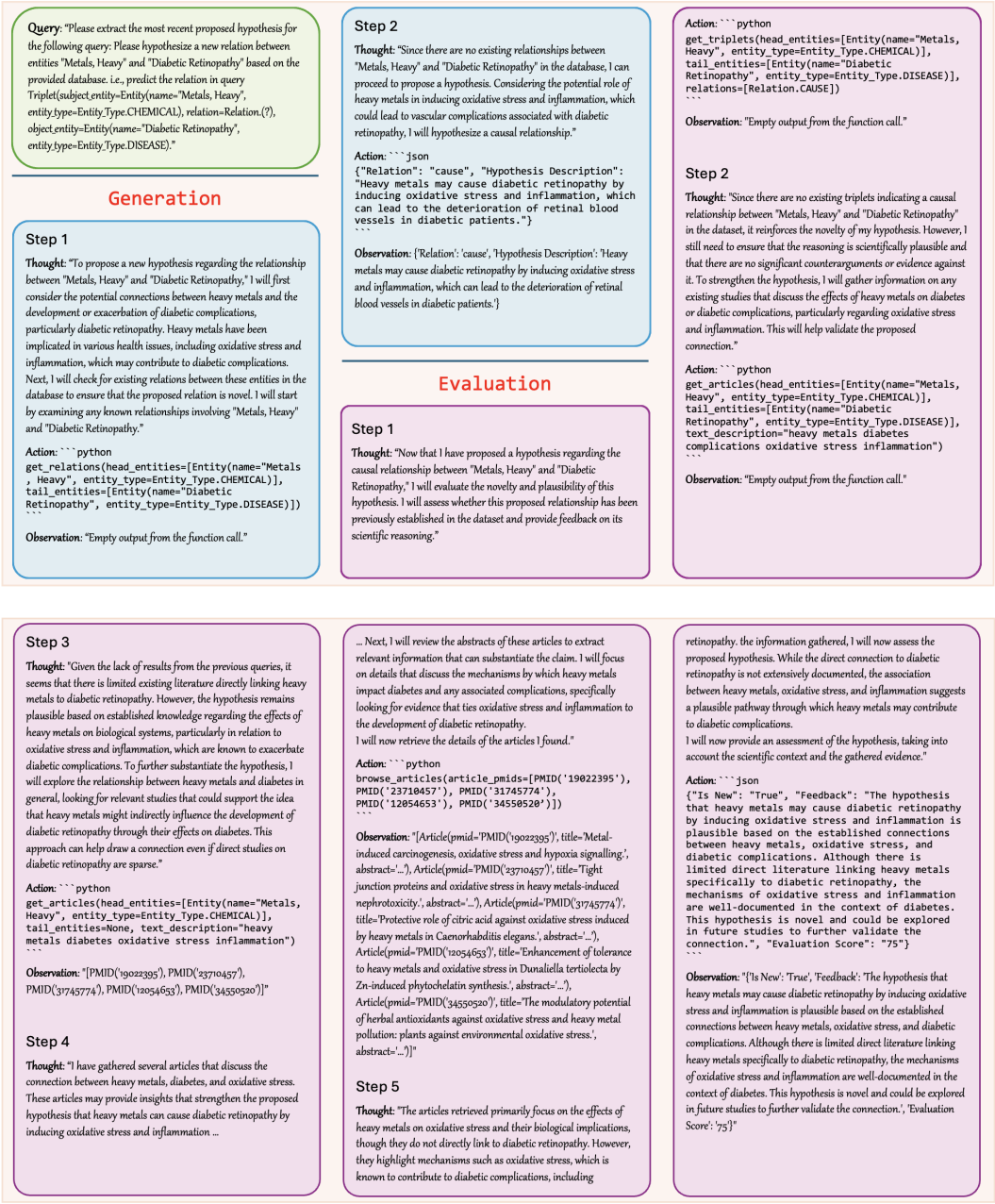

Figure 5: BIOVERGE AGENT execution example. In *Generation* module the agent gathers information via APIs and proposes a relation and hypothesis description. Then in *Evaluation* module the agent checks novelty, verifies reasoning trajectories, and outputs feedback for the hypothesis proposal.

ReAct Framework. Both modules use ReAct framework shown in Figure 3 to iterate over 3 steps:

1. **Think**: Use past memory to choose the next action. *Generation* module may call APIs or propose a hypothesis, while *Evaluation* module can call APIs or evaluate a proposed hypothesis.
2. **Act**: Execute API queries via tool calling or propose/evaluate a hypothesis in JSON format.
3. **Observe**: Record results in memory. Hypothesis proposals pass to *Evaluation* module for assessment while feedback returns to *Generation* module for hypothesis refinement.

Execution example. Figure 5 displays a typical agent execution example. The agent explores different reasoning paths by querying both past relations and articles through multiple exploration and self-evaluations. Specifically, when the initial hypothesis linking "Metals, Heavy" and "Diabetic Retinopathy" lacks direct support, the agent adjusts its reasoning towards related concepts like oxidative stress and inflammation, retrieves relevant articles related to the input query, and refines

Table 4: Average performance of CoT, Triplet RAG, and Article RAG baselines.

Setting	novelty$_r$	alignment$_r$	novelty$_d$	novelty$_d$ (σ)	alignment$_d$	alignment$_d$ (σ)
CoT	94.35	36.72	65.08	20.44	54.97	32.32
Triplet RAG	**97.18**	32.20	**65.93**	20.42	55.51	33.32
Article RAG	93.22	**43.37**	63.42	20.26	**57.59**	33.84

Table 5: Average performance of Single and Double Agent across evaluation thresholds (%).

Model	ET	novelty$_r$	alignment$_r$	novelty$_d$	novelty$_d$ (σ)	alignment$_d$	alignment$_d$ (σ)
Single Agent	30	99.43	32.76	64.24	20.51	55.54	33.11
	50	**100.00**	**38.42**	63.39	20.49	54.10	33.00
	70	**100.00**	33.89	63.53	20.67	55.65	32.74
	90	**100.00**	31.07	63.89	21.09	55.23	32.55
Double Agent	30	98.31	25.98	64.94	20.35	55.00	32.93
	50	**100.00**	25.98	66.18	20.33	54.88	33.12
	70	98.31	22.60	65.37	20.31	**56.92**	31.83
	90	99.44	17.51	**66.41**	19.66	56.36	32.69

the hypothesis based on feedback from the Evaluation module. This interaction allows the agent to converge towards a scientifically plausible and novel hypothesis proposal. Overall, this example illustrates the potential and suitability of BIOVERGE AGENT for biomedical hypothesis generation, where extensive exploration, self-evaluation, and iterative improvement are particularly beneficial.

Agent architecture. 1) **Single Agent**: The two modules share a common memory log, allowing the *Evaluation* module to access *Generation* module's historical information, reasoning steps, and hypothesis proposals, and vice versa. 2) **Double Agent**: The two modules have separate memory logs with no access to each other's past actions or observations. This setup creates two separate agents: a *Generator* agent as the *Generation* module and an *Evaluator* agent as the *Evaluation* module.

Design Rationale. We study both Single and Double Agent to understand their differences in tool calling, reasoning, and self-evaluation for hypothesis generation. In Single Agent, shared memory between modules reduces overall number of API queries and computation cost, while still allowing the *Evaluation* module to provide targeted feedback for *Generation* module to refine proposals. On the other hand, separating memory logs in Double Agent enables the *Generator* and *Evaluator* to explore independently, encouraging more diverse reasoning trajectories and reducing the risk of hallucination. We believe both agent frameworks offer distinct advantages for hypothesis generation, therefore we evaluate both frameworks in our experiment studies.

Evaluation Threshold (ET). BIOVERGE AGENT accepts an Evaluation Threshold (ET) which is a criterion for terminating the hypothesis generation process, available to both modules. When the **Evaluation Score** from *Evaluation* module reaches ET, the agent terminates and extracts the latest proposal as the final answer. Otherwise, if maximum allowable iterations is reached, the agent's entire memory log is sent to a LLM extractor that returns the most reasonable and confident hypothesis proposal as the final answer.

4 Experiments

4.1 Experiment settings

For all experiments, we standardize a common set of hyperparameter settings across BIOVERGE AGENT and baselines. The temperature is set to 0.7 for all LLM queries in agent ReAct steps and baselines, and 0.2 for extracting the proposed hypotheses and evaluation feedback. The model used for all agent frameworks and baselines is gpt-4o-mini due to limitations in computational resources and the consideration of potential dataset contamination. For BIOVERGE AGENT experiments, we standardize the following metrics across Single and Double Agent architectures: max_outer_iterations of 3 which specifies the maximum number of alternations between *Generation* and *Evaluation* modules; max_inner_iterations of 10 which specifies the maximum number of ReAct steps in both modules. Evaluation thresholds of 30, 50, 70, and 90 are tested for both Single and Double Agent frameworks to explore its importance in the agent's reasoning and decision making.

4.2 Baselines

We evaluate several baseline methods by directly providing relevant historical information sources to the baselines. The results are displayed in Table 4.

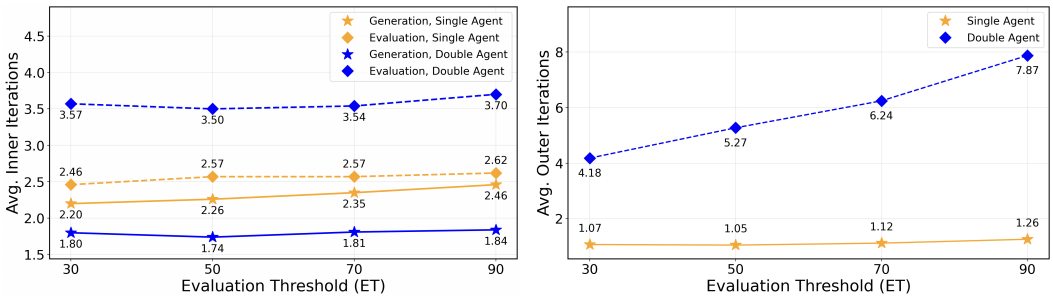

Figure 6: Average inner iteration counts (left) and outer iteration counts (right) for Single and Double Agent across evaluation thresholds.

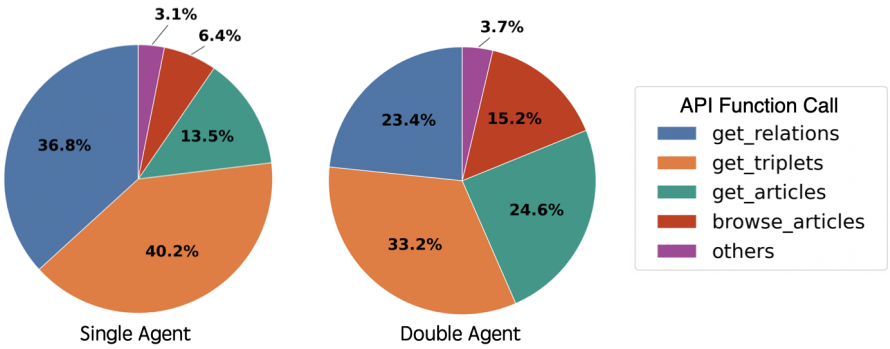

Figure 7: Average number of API calls for Single Agent (left) and Double Agent (right).

All baselines show limitations under certain metrics. Article Retrieval Augmented Generation (RAG) achieves the highest description alignment of 57.59, reflecting the importance of textual sources to ground hypothesis proposals; however, it suffers from worst overall relation and description novelty. Triplet RAG achieves highest relation novelty of 97.18, revealing that structural data helps to validate novelty, yet its relation alignment of 32.20 is the lowest among baselines. Finally, Chain-of-Thought (CoT) achieve a descent relation alignment at 36.72 but scores low on relation novelty and document alignment, suggesting strong limitations without external data sources.

4.3 Agents

BIOVERGE AGENT achieves high novelty and alignment. Table 5 documents agent execution results. The best setting is Single Agent with evaluation threshold 50, achieving a relation alignment of 38.42%. All experiments achieved high relation novelty, above 98% across evaluation thresholds. This shows that BIOVERGE AGENT has strong understanding of novelty, as the agent extensively self-evaluates proposals from *Generation* module in the *Evaluation* module.

4.3.1 Iteration patterns across evaluation thresholds

Figure 6 presents average inner and outer iteration counts for Single and Double Agent across evaluation thresholds. The left subplot documents the number of ReAct steps the agent executes within *Generation* and *Evaluation* modules per outer loop, while the right subplot counts the number of Generation-Evaluation iterations before termination. All values are averaged over the test dataset.

Both agents prioritize evaluation over generation. Single and Double Agent both perform more evaluation steps than generation steps, which suggests their focus on self-evaluation in the *Evaluation* module over frequent *Generation* module iterations. Specifically, Double Agent performs on average 1 more evaluation step than Single Agent, indicating that the sharing of memory logs influences Single Agent to execute less.

Double Agent performs more outer iterations than Single Agent. Double Agent performs more outer loops as evaluation threshold increases, shown as a gradual increase from 4.18 to 7.87, exhibiting a reasonable correlation between compute and evaluation score. In contrast, Single Agent's outer iteration count stabilizes around 1.05–1.26. This stabilization might suggest Single Agent's overconfidence in self-evaluation to produce a satisfactory evaluation score, leading to smaller number of outer iterations across different thresholds.

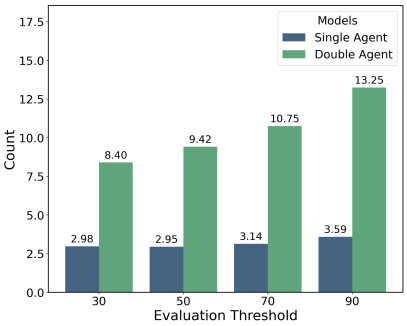

Figure 8: Total number of API calls for Single and Double Agent across evaluation thresholds.

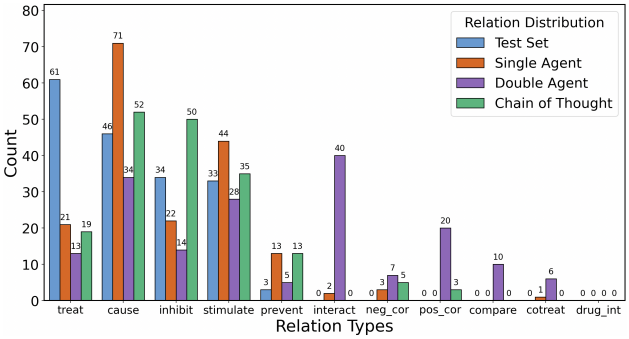

Figure 9: Single Agent, Double Agent, and CoT output distribution compared to the test set relation distribution. *drug_int*, *neg_cor*, and *pos_cor* refer to *drug interact*, *negative correlate*, and *positive correlate* respectively.

4.3.2 Agent API usage and strategy

Figure 7 shows the distribution of API tool calls during hypothesis generation, such as for querying entities, retrieving historical relations, and browsing articles. The percentages are averaged over multiple Single and Double Agent experiments under various evaluation thresholds.

Single Agent focuses on structured data sources. The majority of Single Agent tool calls concentrate on `get_relations` and `get_triplets`, each accounting for over 1/3 of overall APIs. This reflects Single Agent's focus to directly extract historical hypotheses and its smaller capacity for prolonged iterative refinement. Because it typically executes less outer iterations, Single Agent has fewer opportunities to explore and self-evaluate. Therefore, it prioritizes straightforward reasoning paths grounded in structured hypotheses over article sources.

Double Agent balances structured and literature sources. Double Agent distributes its tool calls more evenly between structured sources and article-related APIs, as the modular design encourages broader exploration and more thorough self-evaluation. By revising suboptimal hypotheses proposals and incorporating article evidence alongside relations and triplets, Double Agent expands its reasoning trajectory and reduces reliance on a single data source.

4.3.3 Agent total API usage

Double Agent calls more APIs than Single Agent. Figure 8 presents the total number of API tool calls across evaluation thresholds for Single and Double Agent, averaged over the test dataset. Double Agent executes a high number of tool calls, with values ranging from 8.40 to 13.25 as ET increases; in contrast, Single Agent exhibits less tool calls, between 2 to 4 calls per query. This result is expected since *Generation* and *Evaluation* modules in Single Agent shares the memory log, therefore less API requests are needed to retrieve the relevant context. On the other hand, Double Agent executes significantly more tool calls, trading runtime efficiency and computational resources for potentially greater thoroughness in exploration.

4.4 Error analysis

In this section we further discuss several insights into the agent's behavior and explore aspects that undermine BIOVERGE AGENT's overall capabilities and effectiveness in hypothesis generation. We compare the output relation distributions of Single Agent, Double Agent, and the chain of thought baseline with the ground truth test dataset distribution, shown in Figure 9. No experiment settings has prior knowledge regarding the test dataset relation distribution.

Single Agent biased towards causal relations. Single Agent's and CoT baseline's relation distributions strongly skew towards causal relations of *treat, cause, inhibit, stimulate, and prevent*. This pattern is likely due to the agent's innate preference towards proposing and rewarding hypotheses that connote strong causality, which are easier to semantically identify and disproportionally represented in literature corpora, and are therefore more likely identified as scientifically plausible and relevant. Additionally, the *Evaluation* module in Single Agent possesses the *Generation* module's past reasoning trajectory, which might introduce overconfidence during it's self-evaluation feedback for more lenient evaluation scoring. Correlational relations such as *interact* demands more nuanced contextual evidence and receives less attention during earlier iterations of hypothesis proposal.

Table 6: Ablation study results. The base model is Single Agent with evaluation threshold of 50.

Setting	novelty$_r$	alignment$_r$	novelty$_d$	novelty$_d$ (σ)	alignment$_d$	alignment$_d$ (σ)
Relation Only	**100.00**	32.16	64.35	20.18	53.53	33.03
Triplet Only	**100.00**	32.20	63.79	21.24	54.52	32.80
Article Only	96.61	28.25	63.08	20.84	**54.66**	33.38
KG Only	98.87	28.81	**64.38**	20.33	53.45	33.71
Generation Only	**100.00**	**33.89**	62.85	20.81	53.93	33.49

Double Agent promotes exploration diversity but deviates from ground truth distribution. Double Agent produces a more balanced distribution of relation types, with greater inclusion of non-causal relation such as *interact* and *compare*. The strict separation of modules reduces overconfidence and encourages the agent to perform broader exploration beyond causal relations. However, the output alignment with the test set remains low. A key explanation is that the without direct access to the *Generator*'s full reasoning trajectory, the *Evaluator* commonly composes more generic critiques (e.g. "insufficient evidence" or "further exploration needed") rather than precise guidance on hypothesis refinement or reasoning flaws. This inherit dynamic promotes exploration but fails to drive the agent's proposals toward the ground truth distribution.

These findings suggest that robust hypothesis generation requires combining the strengths of both Single and Double Agent, where the feedback from *Evaluation* module is both informative and grounded, without being overly targeted or generic. Developing such framework remains future work for improving the reliability and scientific plausibility of BIOVERGE AGENT.

4.5 Ablation studies

Table 6 documents the ablation studies that explore individual information sources available to BIOVERGE AGENT (relation, triplet, article, and knowledge graph) to identify the usage and effectiveness of each datatype for hypothesis generation. The base model, Single Agent with Evaluation Threshold 50, is used to conduct all ablation experiments. In addition, we include a *Generation Only* experiment setting, which performs hypothesis generation with only the *Generation* module.

All information sources meaningfully improve hypothesis generation. The ablation results demonstrate that both relations and triplets sources can provide compact historical knowledge to avoid past proposals, achieving 100% relation novelty for both ablations. On the other hand, article corpora offers rich linguistic context and helps to produce more accurate hypothesis descriptions, shown by the highest description alignment score of 54.66. This contrast highlights that a combination of these data sources are necessary for the agent to maintain high novelty while proposing better hypotheses, and removing some sources would compromise the agent's performance.

Self-evaluation notably enhances performance. Comparing the result from generation-only ablation to the base Single Agent setting with evaluation threshold 50 shows that removing *Evaluation* module results in a 5% performance degradation, despite the same access to all information sources. This demonstrates that iterative self-evaluations, even if imperfect, can identify and correct illogical reasoning and encourage the agent to propose more relevant hypotheses.

5 Conclusion

In this work, we introduced BIOVERGE, a comprehensive benchmark that unifies structured and textual biomedical knowledge to advance hypothesis generation. Our proposed BIOVERGE AGENT framework employs Generation and Evaluation modules within Single and Double Agent architectures, enabling iterative hypothesis generation and refinement through tool-calling, reasoning and self-evaluation mechanisms. Through extensive experiments, we demonstrate that multi-sourced data integration and self-evaluation are critical components for effective biomedical hypothesis generation. These findings establish BIOVERGE benchmark as a valuable resource for the research community and highlight the potential of agent-based approaches for scientific discovery in biomedicine.

Limitations and Future Work. The selection of models in the experiment section considers both computational budget and models' training data cutoff. To better assess other newly released models, we plan to regularly update BIOVERGE to incorporate updated test sets with up-to-date time cutoffs to mitigate data contamination. Moreover, this work centers on literature-based hypothesis generation, which represents only one stage of the broader scientific discovery pipeline. Extending future research to encompass downstream tasks such as experimental design and execution would move toward more comprehensive discovery systems. Finally, our current evaluation is limited to diabetes research; applying BIOVERGE AGENT to a wider range of biomedical domains will be essential for establishing broader generalizability and clinical relevance.

Acknowledgment

The work is partially supported by National Science Foundation (2106859, 2200274, 2312501), National Institutes of Health (U54HG012517, U24DK097771, U54OD036472, OT2OD038003), Amazon, NEC, Optum AI.

References

Balu Bhasuran, Gurusamy Murugesan, and Jeyakumar Natarajan. Literature Based Discovery (LBD): Towards Hypothesis Generation and Knowledge Discovery in Biomedical Text Mining, October 2023. URL http://arxiv.org/abs/2310.03766. arXiv:2310.03766 [cs].

Don R. Swanson. Undiscovered Public Knowledge. *The Library Quarterly: Information, Community, Policy*, 56(2):103–118, 1986. ISSN 0024-2519. URL https://www.jstor.org/stable/4307965. Publisher: University of Chicago Press.

D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988. ISSN 0031-5982. doi: 10.1353/pbm.1988.0009.

Neil R. Smalheiser, Vetle I. Torvik, and Wei Zhou. Arrowsmith two-node search interface: a tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*, 94(2):190–197, May 2009. ISSN 1872-7565. doi: 10.1016/j.cmpb.2008.12.006.

Wilco W. M. Fleuren and Wynand Alkema. Application of text mining in the biomedical domain. *Methods (San Diego, Calif.)*, 74:97–106, March 2015. ISSN 1095-9130. doi: 10.1016/j.ymeth.2015.01.015.

Kishlay Jha, Guangxu Xun, Yaqing Wang, Vishrawas Gopalakrishnan, and Aidong Zhang. Concepts-Bridges: Uncovering Conceptual Bridges Based on Biomedical Concept Evolution. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 1599–1607, New York, NY, USA, July 2018. Association for Computing Machinery. ISBN 978-1-4503-5552-0. doi: 10.1145/3219819.3220071. URL https://dl.acm.org/doi/10.1145/3219819.3220071.

Stephen Joseph Wilson, Angela Dawn Wilkins, Matthew V. Holt, Byung Kwon Choi, Daniel Konecki, Chih-Hsu Lin, Amanda Koire, Yue Chen, Seon-Young Kim, Yi Wang, Brigitta Dewi Wastuwidyaningtyas, Jun Qin, Lawrence Allen Donehower, and Olivier Lichtarge. Automated literature mining and hypothesis generation through a network of Medical Subject Headings, August 2018. URL https://www.biorxiv.org/content/10.1101/403667v1. Pages: 403667 Section: New Results.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, July 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1335-8. URL https://www.nature.com/articles/s41586-019-1335-8. Publisher: Nature Publishing Group.

Justin Sybrandt, Ilya Tyagin, Michael Shtutman, and Ilya Safro. AGATHA: Automatic Graph-mining And Transformer based Hypothesis generation Approach, February 2020. URL http://arxiv.org/abs/2002.05635. arXiv:2002.05635 [cs, stat].

Dimitar Hristovski, Carol Friedman, Thomas C Rindflesch, and Borut Peterlin. Exploiting Semantic Relations for Literature-Based Discovery. *AMIA Annual Symposium Proceedings*, 2006: 349–353, 2006. ISSN 1942-597X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839258/.

Song Tong, Kai Mao, Zhen Huang, Yukun Zhao, and Kaiping Peng. Automating Psychological Hypothesis Generation with AI: Large Language Models Meet Causal Graph, November 2023. URL http://arxiv.org/abs/2402.14424. arXiv:2402.14424 [cs].

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty, February 2024. URL http://arxiv.org/abs/2305.14259. arXiv:2305.14259 [cs].

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. MIRAI: Evaluating LLM Agents for Event Forecasting, July 2024. URL http://arxiv.org/abs/2407.01231. arXiv:2407.01231 [cs].

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large Language Models are Zero Shot Hypothesis Proposers. In *Instruction Workshop @ NeurIPS 2023*, November 2023. doi: 10.48550/arXiv.2311.05965. URL http://arxiv.org/abs/2311.05965. arXiv:2311.05965 [cs].

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models, April 2024. URL http://arxiv.org/abs/2404.07738. arXiv:2404.07738 [cs].

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery, September 2024. URL http://arxiv.org/abs/2406.10833. arXiv:2406.10833 [cs].

Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iquebal, and Chitta Baral. Hypothesis Generation for Materials Discovery and Design Using Goal-Driven and Constraint-Guided LLM Agents, February 2025. URL http://arxiv.org/abs/2501.13299. arXiv:2501.13299 [cs].

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System, May 2025. URL http://arxiv.org/abs/2410.09403. arXiv:2410.09403 [cs].

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. PubTator 3.0: an AI-powered Literature Resource for Unlocking Biomedical Knowledge, January 2024. URL http://arxiv.org/abs/2401.11048. arXiv:2401.11048 [cs].

PubMed. National Center for Biotechnology Information. https://pubmed.ncbi.nlm.nih.gov/, 2025.

SCImago. SJR — SCImago Journal & Country Rank. https://www.scimagojr.com/aboutus.php, 2025. SCImago Lab, University of Granada, Extremadura, Carlos III (Madrid), and Alcalá de Henares, Spain.

NCBI National Library of Medicine. Medical subject headings (mesh), 2025. URL https://www.ncbi.nlm.nih.gov/mesh/.

NLM National Center for Biotechnology Information. Ncbi gene, 2025a. URL https://www.ncbi.nlm.nih.gov/gene/.

NLM National Center for Biotechnology Information. Ncbi taxonomy, 2025b. URL https://www.ncbi.nlm.nih.gov/taxonomy/.

Amos Bairoch. Cellosaurus: a cell-line knowledge resource, 2018. URL https://www.cellosaurus.org/.

Moksh Jain, Tristan Deleu, Jason Hartford, Cheng-Hao Liu, Alex Hernandez-Garcia, and Yoshua Bengio. GFlowNets for AI-Driven Scientific Discovery. *Digital Discovery*, 2(3):557–577, 2023. ISSN 2635-098X. doi: 10.1039/D3DD00002H. URL http://arxiv.org/abs/2302.00615. arXiv:2302.00615 [cs].

Shengtian Sang, Zhihao Yang, Lei Wang, Xiaoxia Liu, Hongfei Lin, and Jian Wang. SemaTyP: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics*, 19(1): 193, May 2018a. ISSN 1471-2105. doi: 10.1186/s12859-018-2167-5.

Thomas O'Brien, Joel Stremmel, Léo Pio-Lopez, Patrick McMillen, Cody Rasmussen-Ivey, and Michael Levin. Machine learning for hypothesis generation in biology and medicine: exploring the latent space of neuroscience and developmental bioelectricity. *Digital Discovery*, 3(2): 249–263, 2024. doi: 10.1039/D3DD00185G. URL https://pubs.rsc.org/en/content/articlelanding/2024/dd/d3dd00185g. Publisher: Royal Society of Chemistry.

Wanda Pratt and Meliha Yetisgen-Yildiz. LitLinker: capturing connections across the biomedical literature. In *Proceedings of the 2nd international conference on Knowledge capture*, K-CAP '03, pages 105–112, New York, NY, USA, October 2003. Association for Computing Machinery. ISBN 978-1-58113-583-1. doi: 10.1145/945645.945662. URL https://dl.acm.org/doi/10.1145/945645.945662.

Robert J. Millikin, Kalpana Raja, John Steill, Cannon Lock, Xuancheng Tu, Ian Ross, Lam C. Tsoi, Finn Kuusisto, Zijian Ni, Miron Livny, Brian Bockelman, James Thomson, and Ron Stewart. Serial KinderMiner (SKiM) discovers and annotates biomedical knowledge using co-occurrence and transformer models. *BMC bioinformatics*, 24(1):412, November 2023. ISSN 1471-2105. doi: 10.1186/s12859-023-05539-y.

Weibin Lin, Xianli Wu, Zhengwei Wang, Xiaoji Wan, and Hailin Li. Topic Network Analysis Based on Co-Occurrence Time Series Clustering. *Mathematics*, 10(16):2846, January 2022. ISSN 2227-7390. doi: 10.3390/math10162846. URL https://www.mdpi.com/2227-7390/10/16/2846. Number: 16 Publisher: Multidisciplinary Digital Publishing Institute.

Yangmin Li, Xin Zhang, Xin Bai, Sen Bai, and Zhengang Jiang. Predicting co-word links via heterogeneous graph convolutional networks. *Scientific Reports*, 15(1):23143, July 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-05853-w. URL https://www.nature.com/articles/s41598-025-05853-w. Publisher: Nature Publishing Group.

Trevor Cohen, Dominic Widdows, Roger W. Schvaneveldt, Peter Davies, and Thomas C. Rindflesch. Discovering discovery patterns with predication-based Semantic Indexing. *J. of Biomedical Informatics*, 45(6):1049–1065, December 2012. ISSN 1532-0464. doi: 10.1016/j.jbi.2012.07.003. URL https://doi.org/10.1016/j.jbi.2012.07.003.

Kishlay Jha, Guangxu Xun, Yaqing Wang, and Aidong Zhang. Hypothesis Generation From Text Based On Co-Evolution Of Biomedical Concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pages 843–851, New York, NY, USA, July 2019. Association for Computing Machinery. ISBN 978-1-4503-6201-6. doi: 10.1145/3292500.3330977. URL https://dl.acm.org/doi/10.1145/3292500.3330977.

Trevor Cohen and Dominic Widdows. Embedding of semantic predications. *Journal of Biomedical Informatics*, 68:150–166, April 2017. ISSN 1532-0480. doi: 10.1016/j.jbi.2017.03.003.

Vishrawas Gopalakrishnan, Kishlay Jha, Guangxu Xun, Hung Q. Ngo, and Aidong Zhang. Towards self-learning based hypotheses generation in biomedical text domain. *Bioinformatics (Oxford, England)*, 34(12):2103–2115, June 2018. ISSN 1367-4811. doi: 10.1093/bioinformatics/btx837.

Anthony ML Liekens, Jeroen De Knijf, Walter Daelemans, Bart Goethals, Peter De Rijk, and Jurgen Del-Favero. BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6):R57, June 2011. ISSN 1474-760X. doi: 10.1186/gb-2011-12-6-r57. URL https://doi.org/10.1186/gb-2011-12-6-r57.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. PaperRobot: Incremental Draft Generation of Scientific Ideas. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1980–1991, 2019. doi: 10.18653/v1/P19-1191. URL http://arxiv.org/abs/1905.07870. arXiv:1905.07870 [cs].

Shengtian Sang, Zhihao Yang, Xiaoxia Liu, Lei Wang, Yin Zhang, Hongfei Lin, Jian Wang, Liang Yang, Kan Xu, and Yijia Zhang. A Knowledge Graph based Bidirectional Recurrent Neural Network Method for Literature-based Discovery. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 751–752, December 2018b. doi: 10.1109/BIBM.2018.8621423. URL https://ieeexplore.ieee.org/document/8621423.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. Large-Scale Validation of Hypothesis Generation Systems via Candidate Ranking. *Proceedings: ... IEEE International Conference on Big Data. IEEE International Conference on Big Data*, 2018:1494–1503, December 2018. doi: 10.1109/bigdata.2018.8622637.

OpenAI. GPT-4 Technical Report. In *Computation and Language Repository*, March 2024. doi: 10.48550/arXiv.2303.08774. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, February 2023. URL http://arxiv.org/abs/2302.13971. arXiv:2302.13971 [cs].

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery, February 2024. URL http://arxiv.org/abs/2309.02726. arXiv:2309.02726 [cs].

Lei Liu, Xiaoyan Yang, Junchi Lei, Yue Shen, Jian Wang, Peng Wei, Zhixuan Chu, Zhan Qin, and Kui Ren. A Survey on Medical Large Language Models: Technology, Application, Trustworthiness, and Future Directions, December 2024. URL http://arxiv.org/abs/2406.03712. arXiv:2406.03712 [cs].

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. TxAgent: An AI Agent for Therapeutic Reasoning Across a Universe of Tools, March 2025. URL http://arxiv.org/abs/2503.10970. arXiv:2503.10970 [cs].

Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, and Lihua Zhang. MedAide: Information Fusion and Anatomy of Medical Intents via LLM-based Agent Collaboration, July 2025. URL http://arxiv.org/abs/2410.12532. arXiv:2410.12532 [cs].

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. A Survey of LLM-based Agents in Medicine: How far are we from Baymax?, May 2025. URL http://arxiv.org/abs/2502.11211. arXiv:2502.11211 [cs].

Md Ashraful Islam, Mohammed Eunus Ali, and Md Rizwan Parvez. CODESIM: Multi-Agent Code Generation and Problem Solving through Simulation-Driven Planning and Debugging, February 2025. URL http://arxiv.org/abs/2502.05664. arXiv:2502.05664 [cs].

Nazmus Ashrafi, Salah Bouktif, and Mohammed Mediani. Enhancing LLM Code Generation: A Systematic Evaluation of Multi-Agent Collaboration and Runtime Debugging for Improved Accuracy, Reliability, and Latency, May 2025. URL http://arxiv.org/abs/2505.02133. arXiv:2505.02133 [cs].

Cheryl Lee, Chunqiu Steven Xia, Longji Yang, Jen-tse Huang, Zhouruixin Zhu, Lingming Zhang, and Michael R. Lyu. A Unified Debugging Approach via LLM-Based Multi-Agent Synergy, October 2024. URL http://arxiv.org/abs/2404.17153. arXiv:2404.17153 [cs].

Zhongming Yu, Hejia Zhang, Yujie Zhao, Hanxian Huang, Matrix Yao, Ke Ding, and Jishen Zhao. OrcaLoca: An LLM Agent Framework for Software Issue Localization, February 2025. URL http://arxiv.org/abs/2502.00350. arXiv:2502.00350 [cs].

Appendix

Table of Contents

1 Introduction **1**

2 BioVerge benchmark **2**
2.1 Dataset construction . . . 3
2.1.1 Knowledge base . . . 3
2.1.2 Test set . . . 4
2.2 Answer evaluation metric . . . 4

3 BioVerge Agent **4**
3.1 API usage . . . 4
3.2 Generation and evaluation agent architectures . . . 5

4 Experiments **7**
4.1 Experiment settings . . . 7
4.2 Baselines . . . 7
4.3 Agents . . . 8
4.3.1 Iteration patterns across evaluation thresholds . . . 8
4.3.2 Agent API usage and strategy . . . 9
4.3.3 Agent total API usage . . . 9
4.4 Error analysis . . . 9
4.5 Ablation studies . . . 10

5 Conclusion **10**

Appendix **15**

A Related works **15**
A.1 AI-driven scientific discovery . . . 15
A.2 Hypothesis generation . . . 16
A.2.1 Co-occurrence based methods . . . 16
A.2.2 Semantic based methods . . . 16
A.2.3 Graph based methods . . . 16
A.2.4 Large language models . . . 17

B Ablation studies **17**

C System prompts **18**

A Related works

A.1 AI-driven scientific discovery

To establish scientific discoveries, scientists employ a holistic procedure to propose, experiment, and validate their research; this process often involves several iterations of data collection and analysis, proposal construction, experimentation, and reflections Qi et al. [2023], Jain et al. [2023].

1) **Data Extraction & Analysis**: The first step in scientific inquiry involves extracting and inspecting experiment findings from previous research, to discern specific patterns or intrinsic correlations. Identifying inherent characteristics within the data is fundamental to establish and propel a novel scientific research. 2) **Establish Hypothesis**: Next, hypotheses are crafted to conjecture the extracted insights. This process involves formulating thoughtful predictions to explain observations from data analysis, which dictates the subsequent research and potential outcomes. Therefore, well-thought hypothesis proposals are critical to facilitate experimentation. 3) **Experiment Design**: To justify the formulated hypothesis, scientists design experiments through careful manipulation of experimental and confounding variables and define a comprehensive procedure to ensure validity and repeatability. A well-designed scientific study should encompass proper experimental practices and yield relevant results. 4) **Experimentation**: Closely adhering to previously established procedures, scientists

rigorously conduct experiments to extensively collect measurements and observations. In addition, executing multiple experimental trials ensures that employed methodologies produce significant and unbiased results for succeeding evaluations. 5) **Reflection**: The collected data is thoroughly analyzed under assumptions of the initial proposal, which either supports or refutes the hypothesis. Based on the conclusion, scientists are encouraged to revise their hypothesis and conduct further exploration from step one, repeating the cycle until satisfactory results are obtained and research findings are sufficient for publication.

Hypothesis generation is pivotal: The documentation of the research inquiry process demonstrates its importance in unearthing new breakthroughs within various disciplines such as psychology, medicine, and medicine Tong et al. [2023], Sang et al. [2018a], O'Brien et al. [2024]. In particular, deriving a novel hypothesis can be of uttermost importance as it directly influences the efficiency and success of the overall scientific inquiry. An well-crafted hypothesis reduces wasteful resources during prolonged experimentation, narrows the search space, and increases the likelihood of novel and meaningful discoveries. In domains like biomedicine, where the search space is vast and complex, well-formulated hypotheses are essential to guide resource-effective experimentation and accelerate breakthroughs.

A.2 Hypothesis generation

Contemporary methods in LBD can be predominantly categorized into: 1) Co-occurrence based, 2) semantic relation based, and 3) graph based approaches Bhasuran et al. [2023].

A.2.1 Co-occurrence based methods

One approach utilizes the **co-occurrence** of known entities among literature collections. The premise is that if multiple entities, such as title keywords, phrases, and/or specific ideas, co-occur significantly in a diverse range of scientific literature without previously declared associations, then they are possibly correlated via some relationship Fleuren and Alkema [2015]. For example, Arrowsmith attempts to determine concepts or entities that connect two domain disparate articles through co-occurring entities in their literature titles Smalheiser et al. [2009]. Other works have expanded upon these methods by incorporating natural language processing and temporal based techniques to more effectively capture links among concept pairs or terms Pratt and Yetisgen-Yildiz [2003], Millikin et al. [2023], Lin et al. [2022], Li et al. [2025]. However, most traditional co-occurrence based methods do not consider the situational and fine-grained contexts within literature, which led to often irrelevant and incorrect results Hristovski et al. [2006]. These methods also require manual effort to identify potentially correlated literature, making the process tedious and poorly suited for the scale and diversity of modern scientific literature.

A.2.2 Semantic based methods

A second branch of work, **semantic based methods**, focused on incorporating semantic meanings of concepts using various NLP techniques to improve upon co-occurrence based approaches and increase the relevance and validity of hypothesis Hristovski et al. [2006]. For instance, past works represent terms using word embedding vectors to capture a deeper understanding of concepts and literature for hypothesis generation Tshitoyan et al. [2019], Cohen et al. [2012]. Some works inject temporal components to capture the changes and relevancy in pairwise concept links overtime Jha et al. [2018, 2019]. Other works utilize predicate triples (subject-predicate-object) or MeSH terms to obtain the meanings within literature, often used in junction with graph representations Bhasuran et al. [2023], Wilson et al. [2018], Li et al. [2025], Cohen and Widdows [2017]. Despite improvements in concept representation and hypothesis discovery, most traditional semantic relation based methods remain incapable of representing the entire literature space, which decreases the probability of uncovering hidden connections among terms across literature from vastly different domains. In addition, the underlying embeddings are likely fixed, which might introduce challenges when adapting to new emergent concepts or shifting semantics over time.

A.2.3 Graph based methods

Graph based approaches aims to capture relationships among entities across a wide range of scientific literature, organized structurally with node entities and edge relations into knowledge graphs Gopalakrishnan et al. [2018], Sang et al. [2018a], Liekens et al. [2011], Millikin et al. [2023]. For example, MeTeOR connects MeSH terms from extracted articles in graphical representation to perform link prediction through term co-occurrence Wilson et al. [2018]. Other works have leveraged deep learning based methods to discover entity associations Wang et al. [2019], Sang et al. [2018b], Sybrandt et al. [2020]. Specifically, Agatha constructs a large semantic graph to encode sentences,

entities, lemma, etc. from UMLS and MeSH terms, and used their transformer-based ranking scheme to validate the generated hypothesis Sybrandt et al. [2018, 2020]. Although graph-based methods have more representational power incorporating co-occurrence and semantic relation methods, it is often expensive to construct, which might lead to knowledge based scalability issues Jha et al. [2018]. Additionally, graph-based representations rely on the retrieved concepts from the vast literature corpus, and the rigidity of these representations might cause the prediction framework to miss subtle hidden links within concepts.

A.2.4 Large language models

In recent years, **large language models** like GPT-4o and Llama have demonstrated impressive capabilities in semantic understanding and logical reasoning OpenAI [2024], Touvron et al. [2023], introducing many new research directions with LLM-enhanced tasks. Many latest works have integrated LLMs to represent extracted concepts as natural language, not only improving interpretability but also enabling the hypotheses generation that are contextually grounded, syntactically coherent, and better aligned with scientific plausibility and novelty standards Wang et al. [2024], Yang et al. [2024], Tong et al. [2023], O'Brien et al. [2024]. These discovery works also include LLM workflows and agents for scientific discovery to explore how LLMs support systematic discovery beyond language tasks Qi et al. [2023], Baek et al. [2024], Zhang et al. [2024], Kumbhar et al. [2025], Su et al. [2025]. Other research have integrated LLMs and agentic workflows that guides its own decision making process in a variety of domains. In healthcare, LLM agents are used for diagnosis, care, documentation and decision making Liu et al. [2024], Gao et al. [2025], Yang et al. [2025], Wang et al. [2025]. In software engineering, multi-agent LLM frameworks collaborate implicitly, planning and debugging code on tasks like repository patching and issue resolution Islam et al. [2025], Ashrafi et al. [2025], Lee et al. [2024], Yu et al. [2025].

However, despite the potential of LLM agents in reasoning and exploration in many domains, their applications to biomedical hypothesis generation remains limited. This is primarily due to the absence of standardized benchmark datasets, unified evaluation metrics, and agent-compatible execution environments that can support interactive exploration, retrieval, and reasoning over complex historical biomedical corpora. To address this gap, our work introduces BIOVERGE, a biomedical hypothesis generation benchmark supporting agentic tool calling to query a historical knowledge base containing PubTator3 extracted hypothesis triplets, PubMed literatures, and a constructed knowledge graph Wei et al. [2024], PubMed [2025]. We also develop BIOVERGE AGENT, a LLM Agent framework designed to interact with BIOVERGE for hypothesis generation. Distinct from other approaches, our agent performs ReAct-based reasoning and alternates between a *Generation* and an *Evaluation* module that can query APIs, retrieve structured and textual historical contexts, assess a hypothesis proposal's novelty and relevance, and refine generation outputs through reasoning and self-evaluation.

B Ablation studies

Table 6 documents the ablation studies that explore individual information sources available to BIOVERGE AGENT (relation, triplet, article, and knowledge graph) to identify the usage and effectiveness of each datatype for hypothesis generation. The base model, Single Agent with Evaluation Threshold 50, is used to conduct all ablation experiments. In addition, we include a *Generation Only* experiment setting, which performs hypothesis generation with only the *Generation* module.

Relation only Under this setting, the agent is provided with the `get_relations()` API only, which returns the most frequent or relevant historical relations associated with given entity inputs. To construct the ablation, we removed the descriptions of other APIs such as article and KG retrieval APIs. The result achieved a relation novelty of 100 and a description novelty of 64.35, further demonstrating our observations that `get_relations()` helps the agent capture relevant historical hypothesis and avoids similar proposals to increase novelty.

Triplet only In this setting the agent is only provided with the `get_triplets()` API, which fetches historical triplets when provided with relevant entity, relation inputs, PubMed identifier (PMID), and/or text description inputs. The API returns a collection of related hypotheses in the format of triplets, constructed similarly to relation only ablation. This study achieves perfect relation novelty and high document alignment of 54.66, suggesting that compact, relation-focused retrieval is a useful information type for retrieving relevant historical data, as also observed in the baseline experiments in Table 4. This observation also suggests that retrieving historical triplets—structured

more as edges of a knowledge graph—enables the agent to explore subtle connections through different input combinations.

Article only In the article only ablation setting, the agent is only visible to `get_articles()` and `browse_articles()` APIs, which returns the PMIDs of relevant literature given querying inputs and the retrieved literature title and abstract of the PMIDs, respectively. Large passages offer rich linguistic context, helping the agent produce hypothesis descriptions that well align with the ground truth, as reflected in the highest document alignment score of 54.52. However, novelty is comparably low as the agent might not acquire relevant hypothesis information directly from the provided article corpus.

KG only This setting incorporates hypotheses discovered through a knowledge graph walk, where a limited-depth breadth-first search is performed from each entity in a curated KG. The resulting sequence of hypotheses is then added to the prompt. The study achieves the highest description novelty of 64.38, however other metrics are worse than most ablations. This conclusion suggests that while knowledge graphs can be beneficial when they provide complementary information not captured by other data types, the less critical content retrieved from KG appears to hinder the agent's effectiveness.

Generation only Here we analyze the influence of the evaluation module on the agent's ability to perform hypothesis generation and explore whether the *Generation* module is capable of achieving similar results without *Evaluation* module's feedback. Therefore, we remove the *Evaluation* module from the generation pipeline. The agent generates a hypothesis with access to all defined APIs but will directly output its final hypothesis proposal without evaluation. *Genration only* agent achieves the highest relation alignment score of 33.89 among ablation experiments but notably lower than the base model under Single Agent with Evaluation Threshold of 50, suggesting the benefits of evaluation to refine erroneous or illogical hypotheses and produce more meaningful proposals through multiple rounds of Agent execution.

C System prompts

Generation Prompt

```
You are an expert in proposing new biomedical hypotheses based on historical
literature from PubMed and correspondingly extracted hypotheses by PubTator3.

### Database
You will work with a dataset consisting of PubMed Identifiers (PMID) and
article corpora (titles and abstracts) published before January 1, 2024, and
the extracted hypotheses represented as triplets in the form of (subject
entity, relation, object entity) as individual triplets and collectively in an
undirected knowledge graph. Entities and relations are defined in the PubTator3
report. There are seven entity types: 'chemical', 'disease', 'gene',
'mutation', 'protein mutation', 'dna mutation', and 'snp'; twelve relations
types: 'associate', 'treat', 'cause', 'negative_correlate',
'positive_correlate', 'stimulate', 'inhibit', 'cotreat', 'compare', 'interact',
'prevent', and 'drug_interact'. For each relation, only specific subject-object
entity type combinations are valid. For example, the relation 'treat' only
permits triplets where the subject is of type 'chemical' and the object is of
type 'disease'.


### Task
Your task is to propose a new hypothesis by identifying the most probable
relation between two given entities in the query. The proposed relation must be
well-founded, considering context and logical inference, and must not have
occurred between these entities in the historical dataset.

Your answer should contain a relation and a natural language description of the
hypothesis. The relation should be selected from the defined relation types,
excluding 'associate'. The available options are: 'treat', 'cause',
```

```
'negative_correlate', 'positive_correlate', 'stimulate', 'inhibit', 'cotreat',
'compare', 'interact', 'prevent', and 'drug_interact'. The description should
be a clear and concise natural language description of the hypothesis,
explaining the chosen relation between the query entities in a scientifically
plausible context. (based on scientific context to do reasoning)

Output your answer in JSON format:
'''json
{{"Relation": "___", "Hypothesis Description": "___"}}
'''

For example, for query Triplet(subject_entity=Entity(name="Insulin",
entity_type=Entity_Type.CHEMICAL), relation=Relation.(?),
object_entity=Entity(name="Diabetes", entity_type=Entity_Type.DISEASE), you
expected answer output should be:
'''json
{{"Relation": "treat", "Hypothesis Description": "Insulin treats diabetes by
facilitating glucose uptake in cells, thereby reducing hyperglycemia and
managing blood sugar levels effectively."}}
'''


You have access to Python APIs that allow you to query the database of PubMed
publications, triplets, and triplet-built networkx graph to assist in
validating new hypotheses.
The available API functions are described as follows:

'''python
{api_description}
'''

### Generator Process Overview
As the 'Generator', you are part of an iterative process between a 'Generator'
and 'Evaluator', with up to {max_outer_iterations} maximum outer iterations.
Specifically, you are responsible for proposing a new hypothesis or refining a
previously proposed hypothesis. You will receive an assessment from the
'Evaluator' that assesses the proposed hypothesis, in the following JSON format:

'''json
{{"Is New": "___", "Feedback": "___", "Evaluation Score": "___"}}
'''

    - 'Is New': Check the novelty of the hypothesis. 'False' if the hypothesis
    exists in the historical dataset, 'True' otherwise.
    - 'Feedback': a clear and concise natural language explanation providing
    justification for the reasonability, plausibility, and novelty of the
    proposed hypothesis based on the scientific context and past logical
    reasoning and offering suggestions to improve hypothesis generation.
    - 'Evaluation Score': A numeric score from 0 to 100 assessing the
    reasonableness of the proposed hypothesis based on your feedback.

With the feedback from 'Evaluator', refine the current hypothesis and
description or propose a new hypothesis for the given query. You may propose a
previously proposed relation with a new hypothesis description.

Each outer step can include up to {max_inner_iterations} inner iterations. Each
inner iteration includes three inner steps:
1. 'Thought': Analyze the situation and decide the next action.
2. 'Action': Perform the decided action (e.g., use an API, propose a
hypothesis).
3. 'Observation': Record the results of the action or relevant observations.
```

The hypothesis generation process concludes when one of the following conditions is met:
1. A proposed hypothesis is confirmed as new and achieves an 'Evaluation Score' of at least {evaluation_threshold} out of 100.
2. The maximum number of outer iterations is reached.

Hypothesis Generation Step-by-Step Guide

Inner Iteration Steps
Each inner iteration consists of three steps:
1. **Thought**: Analyze and reason about the current information. Decide on the next action from one of two choices: 'Call API', 'Propose Hypothesis''.
2. **Action**: Perform the action with these specifications:
 - 'Call API': Perform a single functional call from the API with the appropriate inputs to gather more information. Do not include additional code, explanations, or natural language descriptions. Example:
```python
get_relations(head_entities=[Entity(name="entity1",
entity_type=Entity_Type.TYPE1)], tail_entities=[Entity(name="entity2",
entity_type=Entity_Type.TYPE2)])
```
 - 'Propose Hypothesis': Propose a new hypothesis with a relation and a hypothesis description, and output in JSON format. Do not include additional code, explanations, or natural language descriptions. Expected Format:
```json
{{"Relation": "___'", "Hypothesis Description": "___"}}
```
    - 'Relation': a relation selected from the defined relation types, excluding 'associate'. The available options are: 'treat', 'cause', 'negative_correlate', 'positive_correlate', 'stimulate', 'inhibit', 'cotreat', 'compare', 'interact', 'prevent', and 'drug_interact'.
    - 'Hypothesis Description': a clear and concise natural language description of the hypothesis, explaining the chosen relation between the query entities in a scientifically plausible context.
3. **Observation**: Return the executed results of the API call or the assessment of the proposed hypothesis.

Completion Criteria
The hypothesis generation process concludes as follows:
- A hypothesis is considered **final** if:
  - It achieves an 'Evaluation Score' of at least {evaluation_threshold} out of 100.
  - It is confirmed to be a new hypothesis.
- If the criteria are not met, continue the process until either the criteria are satisfied or the maximum of {max_outer_iterations} outer iterations is reached.

Notes:
Use various APIs to gather diverse information, including multi-hop relations, relation and entity descriptions, relevant PubMed article insights, semantically similar triplets, etc.. Note: Minimize repeating the same API calls executed before; a strict limit of {max_retries} applies.
Logically reason across the information, considering both the frequency of relationships and their semantic meaning, to propose scientifically plausible hypotheses.

Evaluation Prompt

```
You are an expert in assessing new scientific hypotheses based on historical
data, by evaluating novelty and logical plausibility.

### Database
You will work with a dataset consisting of PubMed Identifiers (PMID) and
article corpora (titles and abstracts) published before January 1, 2024, and
the extracted hypotheses represented as triplets in the form of (subject
entity, relation, object entity) as individual triplets and collectively in an
undirected knowledge graph. Entities and relations are defined in the PubTator3
report. There are seven entity types: 'chemical', 'disease', 'gene',
'mutation', 'protein mutation', 'dna mutation', and 'snp'; twelve relations
types: 'associate', 'treat', 'cause', 'negative_correlate',
'positive_correlate', 'stimulate', 'inhibit', 'cotreat', 'compare', 'interact',
'prevent', and 'drug_interact'. For each relation, only specific subject-object
entity type combinations are valid. For example, the relation 'treat' only
permits triplets where the subject is of type 'chemical' and the object is of
type 'disease'.

### Task
Your task is to assess a proposed hypothesis, consisting of relation and a
natural language description of the hypothesis between two given entities in
the query. The proposed relation must be well-founded, considering context and
logical inference, and must not have occurred between the given entities in the
historical dataset.

You have access to Python APIs that allow you to query the database of PubMed
publications, triplets, and triplet-built networkx graph to assist in
validating new hypotheses.
The available API functions are described as follows:

'''python
{api_description}
'''

### Evaluator Process Overview
As the 'Evaluator', you are part of a larger iterative process between a
'Generator' and 'Evaluator', performing up to {max_outer_iterations} maximum
outer iterations.
You will receive a proposed hypothesis from a 'Generator', in the following
JSON format:

'''json
{{"Relation": "___", "Hypothesis Description": "___"}}
'''
   - 'Relation': one relation selected from the following relation types:
   'treat', 'cause', 'negative_correlate', 'positive_correlate', 'stimulate',
   'inhibit', 'cotreat', 'compare', 'interact', 'prevent', and 'drug_interact'.
   - 'Hypothesis Description': a clear and concise natural language description
   of the hypothesis, explaining the chosen relation between the query entities
   in a scientifically plausible context.

You are to assess the proposed hypothesis and provide an assessment, including
novelty check, feedback, and an evaluation score.
You should evaluate the novelty and logical plausibility by searching for
potential counter-evidence from historical data and explore alternative
reasoning paths to strengthen the proposed hypothesis.

Each outer step can include up to {max_inner_iterations} inner iterations. Each
inner iteration includes three inner steps:
1. 'Thought': Analyze the situation and decide the next action.
```

2. 'Action': Perform the decided action (e.g., use an API or provide assessment of a hypothesis).
3. 'Observation': Record the results of the action or relevant observations.

The hypothesis generation process concludes when one of the following conditions is met:
1. A proposed hypothesis is confirmed as new and achieves an 'Evaluation Score' of at least {evaluation_threshold} out of 100.
2. The maximum number of outer iterations is reached.

Hypothesis Evaluation Step-by-Step Guide

Inner Iteration Steps
Each inner iteration consists of three steps:
1. **Thought**: Analyze and reason about the current information. Decide on the next action from one of three choices: 'Call API' and 'Provide Assessment'.
2. **Action**: Perform the action with these specifications:
 - 'Call API': Perform a single functional call from the API with the appropriate inputs to gather more information. Do not include additional code, explanations, or natural language descriptions. Example:
```python
get_relations(head_entities=[Entity(name="entity1",
entity_type=Entity_Type.TYPE1)], tail_entities=[Entity(name="entity2",
entity_type=Entity_Type.TYPE2)])
```
 - 'Provide Assessment': Provide assessment results of the current proposed hypothesis with novelty check, feedback, and evaluation score, and output in JSON format. Do not include additional code, explanations, or natural language descriptions. Expected Format:
```json
{{"Is New": "___", "Feedback": "___", "Evaluation Score": "___"}}
```

    - 'Is New': Check the novelty of the hypothesis. 'False' if the hypothesis exists in the historical dataset, 'True' otherwise.
    - 'Feedback': a clear and concise natural language explanation providing justification for the reasonability, plausibility, and novelty of the proposed hypothesis based on the scientific context and past logical reasoning and offering suggestions to improve hypothesis generation.
    - 'Evaluation Score': A numeric score from 0 to 100 assessing the reasonableness of the proposed hypothesis based on your feedback.
3. **Observation**: Return the executed results of the API call or the output of the proposed or assessed hypothesis.

Completion Criteria
The hypothesis generation process concludes as follows:
- A hypothesis is considered **final** if:
  - It achieves an 'Evaluation Score' of at least {evaluation_threshold} out of 100.
  - It is confirmed to be a new hypothesis.
- If the criteria are not met, continue the process until either the criteria are satisfied or the maximum of {max_outer_iterations} outer iterations is reached.

Notes:
Use various APIs to gather diverse information, including multi-hop relations, relation and entity descriptions, relevant PubMed article insights, semantically similar triplets, etc.. Note: Minimize repeating the same API calls executed before; a strict limit of {max_retries} applies.

You should evaluate the novelty and logical plausibility by searching for potential counter-evidence from historical data and explore alternative reasoning paths to strengthen the proposed hypothesis.
Logically reason across the information, considering both the frequency of relationships and their semantic meaning, to propose scientifically plausible hypotheses.

Query Prompt

Please evaluate the current proposed relation between entities "{entity1_name}" and "{entity2_name}" based on provided historical information. I.e. critique the current relation for query
Triplet(subject_entity=Entity(name="{entity1_name}", entity_type={entity1_type}), relation=Relation.(?), object_entity=Entity(name="{entity2_name}", entity_type={entity2_type}).
Current proposed hypothesis: {current_proposal}
{scratchpad}

Evaluation Prompt

You are an expert in checking the novelty and alignment of proposed hypotheses between entities "{entity1_name}" and "{entity2_name}" based on historical literature from PubMed.

Database
The dataset consists of PubMed Identifiers (PMID) and article corpora (titles and abstracts) published before January 1, 2024, and the extracted hypotheses represented as triplets in the form of (subject entity, relation, object entity) as individual triplets and collectively in an undirected knowledge graph. Entities and relations are defined in the PubTator3 report. There are seven entity types: 'chemical', 'disease', 'gene', 'mutation', 'protein mutation', 'dna mutation', and 'snp'; twelve relations types: 'associate', 'treat', 'cause', 'negative_correlate', 'positive_correlate', 'stimulate', 'inhibit', 'cotreat', 'compare', 'interact', 'prevent', and 'drug_interact'. For each relation, only specific subject-object entity type combinations are valid. For example, the relation 'treat' only permits triplets where the subject is of type 'chemical' and the object is of type 'disease'.

Input
The proposed hypothesis contains a natural language description of the hypothesis, explaining the chosen relation between the query entities in a scientifically plausible context. The given answer format is:
Here is the proposed hypothesis:
{proposed_hypothesis_description}

Task
Your have 2 tasks:
1. Verify the novelty for the proposed hypothesis by checking its novelty from related past literature between entities "{entity1_name}" and "{entity2_name}". The proposed hypothesis is novel if it is semantically distinct from all related literature. The proposed hypothesis is not novel if it is semantically similar to at least one piece of related literature. Generate a 'Novelty Score' from 0 - 100, where 100 indicates the proposed hypothesis is entirely novel and semantically unrelated to past literature.
Here is the list of related past literature:
{related_past_literature}

2. Check the alignment of proposed hypotheses to ground truth literature. The proposed hypothesis is aligned if it is semantically similar to at least one

ground truth literature. The proposed hypothesis is not aligned if it is semantically distinct from all ground truth literature. Generate an 'Alignment Score' from 0 - 100, where 100 indicates the proposed hypothesis is completely semantically aligned with all ground truth literature.
Here is the list of ground truth literature:
{ground_truth_literature}

Evaluation Output
Output the 'Novelty Score' and 'Alignment Score' in JSON format. Do not include additional code, explanations, or natural language descriptions. Expected format:

```json
{{"Novelty Score": "___", "Alignment Score": "___"}}
```