Biomedical Hypothesis Explainability with Graph-Based Context Retrieval

Ilya Tyagin TYAGIN@UDEL.EDU
Center for Bioinformatics and Computational Biology
University of Delaware
Newark, DE, USA

Saeideh Valipour SVALIPOU@UDEL.EDU
Computer and Information Sciences
University of Delaware
Newark, DE, USA

Aliaksandra Sikirzhytskaya SIKIRZHA@MAILBOX.SC.EDU
Drug Discovery and Biomedical Sciences (DDBS) College of Pharmacy
University of South Carolina
Columbia, South Carolina, USA

Michael Shtutman SHTUTMANM@COP.SC.EDU
Drug Discovery and Biomedical Sciences (DDBS) College of Pharmacy
University of South Carolina
Columbia, South Carolina, USA

Ilya Safro ISAFRO@UDEL.EDU
Computer and Information Sciences
University of Delaware
Newark, DE, USA

Abstract

We introduce an explainability method for biomedical hypothesis generation systems, built on top of the novel Hypothesis Generation Context Retriever framework. Our approach combines semantic graph-based retrieval and relevant data-restrictive training to simulate real-world discovery constraints. Integrated with large language models (LLMs) via retrieval-augmented generation, the system explains hypotheses with contextual evidence using published scientific literature. We also propose a novel feedback loop approach, which iteratively identifies and corrects flawed parts of LLM-generated explanations, refining both the evidence paths and supporting context. We demonstrate the performance of our method with multiple large language models and evaluate the explanation and context retrieval quality through both expert-curated assessment and large-scale automated analysis. Our code is available at: `https://github.com/IlyaTyagin/HGCR`.

1. Introduction

Automated biomedical hypothesis generation (HG, also known as Literature-Based Discovery), aims to uncover implicit biomedical connections from large-scale literature corpora. Originating from Swanson's idea of "undiscovered public knowledge" (Swanson, 1986), HG identifies non-trivial links between biomedical concepts to support testable scientific insights (Popper, 1959). Over time, HG systems have evolved from simple keyword overlap methods to those leveraging knowledge graphs and deep learning (Henry and McInnes, 2017; Cesario et al., 2024).

However, most HG systems often lack interpretability, offering unexplainable numeric predictions which limits their practical utility and weakens user trust. Integrating HG with predictive modeling can improve biomedical discoveries (e.g., identifying gene-disease links (Sybrandt et al., 2018; Aksenova et al., 2020; Cummings et al., 2022)), but explainability remains a critical bottleneck.

Recent work explores Large Language Models (LLMs) for HG explainability. However, LLMs, while performing well for language generation, often hallucinate facts and lack domain-specific reasoning, leading to scientifically implausible hypotheses. They are also unaware of temporal constraints and often fail to align with structured biomedical knowledge.

To address this, we propose the Hypothesis Generation Context Retrieval (HGCR), a retrieval-augmented framework for hypothesis generation with explainability. HGCR constructs a dynamic biomedical co-occurrence graph from MedLine abstracts, where nodes represent biomedical concepts using the Universal Medical Language System (UMLS CUIs, (Bodenreider, 2004b)) and edges represent co-occurrences of these concepts in abstracts. The system then identifies and ranks semantic paths between concept pairs in this graph based on their plausibility. Retrieved path is associated with the specific Medline abstracts/literature where the concepts co-occur, creating path-literature pairs and used as context for LLM-generated explanations.

To ensure scientific soundness of generated hypothesis explanations, we introduce a feedback loop architecture that uses AGATHA (Sybrandt et al., 2020) and SemRep (Rindflesch and Fiszman, 2003) to extract and validate semantic predicates (i.e., subject–verb–object triples) representing biomedical relationships between UMLS terms from the LLM output. Unsupported claims trigger context update and iterative explanation refinement.

Our contribution:

(1) We propose HGCR, a graph-based biomedical context retrieval system that constructs semantic paths from the whole MEDLINE dataset of biomedical abstracts and citations to explain potential relationships between concepts of interest.

(2) We introduce a feedback mechanism that validates LLM-generated explanations using biomedical hypothesis generation system (in our experiments we use AGATHA) by iteratively refining context to minimize the number of LLM-generated semantic predicates identified by HG system as potentially wrong.

(3) We perform a retrospective temporal evaluation of both retrieval and explanation quality, benchmarking HGCR and the feedback loop pipeline against existing retrieval-augmented generation (RAG) systems and demonstrate the performance of the proposed method by human expert evaluation.

Generalizable Insights about Machine Learning in the Context of Healthcare

Our study presents key insights for developing more effective ML systems in healthcare workflows. First, automatic explainability is crucial especially in domains like biomedicine, where interpretable reasoning behind predictions is as important as their accuracy and clinical adoption. Second, the combination of knowledge graphs into retrieval-augmented generation frameworks shows promise, as integrating LLMs with structured knowledge and external validation tools (like AGATHA) results in more scientifically sound hypotheses than using LLMs alone. Finally, iterative refinement based on feedback and validated evidence enhances the plausibility and utility of ML-generated hypotheses, highlighting socio-technical benefits relevant for deployment in real-world biomedical settings, usefulness of generated hypotheses and their explanations.

2. Related Work

Hypothesis Generation. Scientific HG was pioneered by Swanson (Swanson, 1986) who proposed to discover unknown connections by reasoning over semantically disconnected literature. Subsequent systems evolved to graph-based and topic modeling methods (e.g., MOLIERE (Sybrandt et al., 2017) or BioLDA (Wang et al., 2011), which applied LDA topic modeling and used multi-modal semantic graphs to extract plausible biomedical associations. Many frequency-based approaches such as co-occurrence frequency of term (Srinivasan, 2004) other frequencies such as relative frequency (Lindsay and Gordon, 1999), tf-idf (Srinivasan, 2004) were also investigated to complement Swanson work. The high co-occurrence frequencies do not necessarily guarantee meaningful relationships. More recently, a series of text mining technologies, such as random walking (Shi et al., 2015), Latent Semantic Indexing (LSI) (Gordon and Dumais, 1998), ranking (Rastegar-Mojarad et al., 2015), association rules (Yetisgen-Yildiz and Pratt, 2006) have been investigated to learn more implicit semantics of terms, inferring complex semantic associations. Despite the improvements in modeling semantic associations, graph theoretic machine learning approaches have also made swift inroads into HG task for complex association generation, such as constructed a subgraph to provide for deep understanding of the associations (Cameron et al., 2015), extracting the graph pattern features from graphs to infer treatment and causative relations (Bakal et al., 2018), applying the direct edge searches and meta-paths in knowledge graphs (KG) for hypothesis generation (Taneja et al., 2023). While biomedical literature is growing exponentially with new knowledge being discovered every day, The temporal dynamics of scientific term relations has been studied by several recent works (Jha et al., 2019; Xun et al., 2017; Akujuobi et al., 2020; Zhou et al., 2022).

AGATHA. AGATHA (Sybrandt et al., 2020; Tyagin et al., 2022) introduced a transformer-based hypothesis generation pipeline over a structured semantic graph derived from MEDLINE, producing link plausibility scores. Later works explored generative approaches, such as BioGPT (Luo et al., 2022) and CBAG (Sybrandt and Safro, 2021), as well as recent attempts to directly use LLMs for hypothesis formulation (Qi et al., 2024; Iser, 2024). While promising, LLM-only methods often hallucinate connections lacking mechanistic or temporal validity, that is, they do not correspond to known biological or causal processes. This motivates retrieval-augmented or hybrid approaches, such as RUGGED (Pelletier et al., 2024) and the proposed HGCR framework, which use graph-structured paths to improve hypotheses based on plausible biomedical context.

LLMs and Transformers in Biomedical Research. Biomedical-specific LLMs such as BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2021), and SciFive (Phan et al., 2021) enable named entity recognition, question answering, and classification on MEDLINE scale corpora. However, generative models like PaLM (Chowdhery et al., 2022) or GPT-4, while capable of synthesizing plausible hypotheses, frequently lack interpretability and scientific validity (Elbadawi et al., 2024; Park et al., 2024). HGCR addresses this limitation by treating hypothesis explanation as a path ranking problem, enriched with literature-derived context and embeddings. Furthermore, it introduces a feedback loop for explanation refinement, correcting unsupported claims through external validation with hypothesis generation systems like AGATHA.

Biomedical Question Answering (QA) and RAG. Medical QA systems have progressed from BERT-based models (Devlin et al., 2019; Jin et al., 2022) to LLMs fine-tuned on biomedical datasets (Singhal et al., 2023; Wu et al., 2023). RAG architectures (Zuheros et al., 2021) mitigate hallucination by providing relevant documents to generate LLM responses. In the biomedical domain, systems like Clinfo.ai (Lozano et al., 2023), BiomedRAG (Li et al., 2024), and MedRAG (Xiong et al., 2024a) retrieve MEDLINE abstracts to support QA or relation extraction. iMedRAG (Xiong et al., 2024b) adds iterative query refinement for multi-step reasoning.

The proposed HGCR framework complements these approaches by ranking semantically coherent graph paths rather than flat documents, enabling structured retrieval of mechanistic explanations. Unlike existing RAG systems, it incorporates a contrastive training objective to distinguish meaningful reasoning chains from corrupted or unrelated paths, and includes a refinement loop to iteratively align generated explanations with contextual evidence.

3. Methods

The schema of the proposed hypothesis self-adjusting algorithm (called *feedback loop*) is shown in Figure 1. It integrates the biomedical knowledge network, context retriever, LLMs and automatic evaluator to generate and refine indirect explanations between biomedical concept pairs. It bridges the gap between machine-generated predictions and human interpretability by converting ranked entity relationships into natural language descriptions supported by literature evidence.

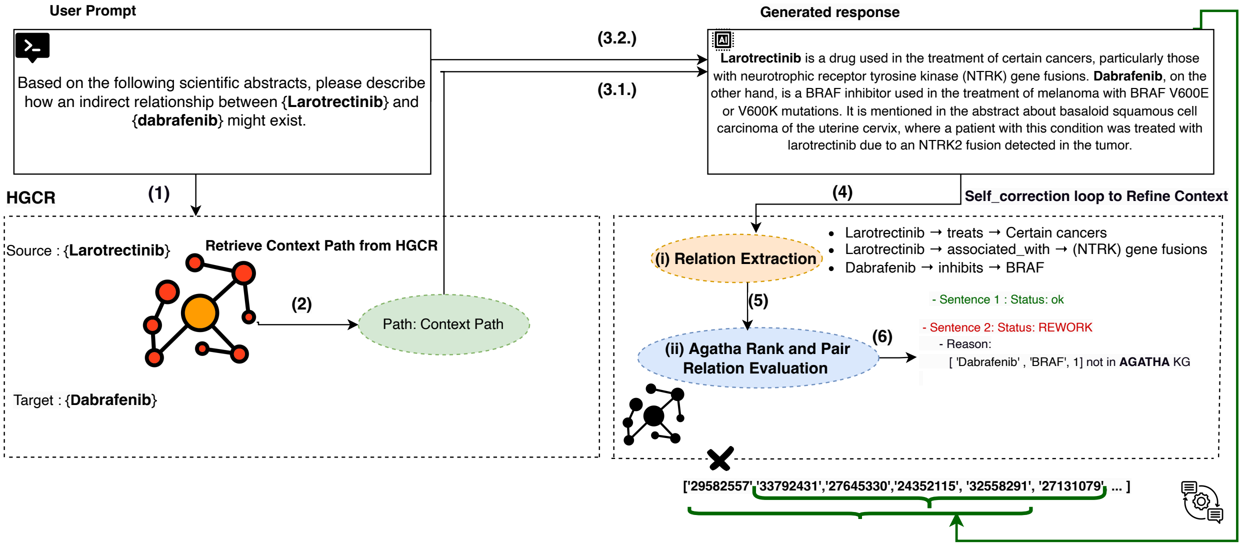
3.1. Context Path Retrieval with HGCR

Graph structure and temporal context. The system operates over a dynamic biomedical semantic network $G = \{G_t\}_{t=1}^{T}$, where each snapshot represents the state of the network at time t:

$$G_t = (N_t, E_t). \tag{1}$$

Nodes N_t correspond to unique UMLS Concept Unique Identifiers (CUIs), and edges E_t represent validated biomedical associations between them (Tyagin and Safro, 2023). An association is defined as the co-occurrence of two UMLS concepts within the same MEDLINE abstract, without requiring them to appear within the same sentence. Each edge $(u, v) \in E_t$ is supplied with a set of PubMed identifiers (PMIDs) referencing biomedical literature that supports the association and a corresponding publication year t.

Temporal evolution is defined by cumulative literature support: for each year t, the edge set E_t includes all associations observed in the literature up to that year with PMIDs that correspond to the paper abstract:



Based on the following scientific abstracts, please describe how an indirect relationship between {**Larotrectinib**} and {**dabrafenib**} might exist.

Larotrectinib is a drug used in the treatment of certain cancers, particularly those with neurotrophic receptor tyrosine kinase (NTRK) gene fusions. **Dabrafenib**, on the other hand, is a BRAF inhibitor used in the treatment of melanoma with BRAF V600E or V600K mutations. It is mentioned in the abstract about basaloid squamous cell carcinoma of the uterine cervix, where a patient with this condition was treated with larotrectinib due to an NTRK2 fusion detected in the tumor.

- Larotrectinib → treats → Certain cancers
- Larotrectinib → associated_with → (NTRK) gene fusions
- Dabrafenib → inhibits → BRAF

- Sentence 1 : Status: ok

- Sentence 2: Status: REWORK
  - Reason:
    ['Dabrafenib' , 'BRAF', 1] not in **AGATHA** KG

['29582557','33792431','27645330','24352115', '32558291', '27131079', ...]

Figure 1: Overview diagram of the proposed pipeline based on Hypothesis Generation Context Retriever (HGCR) and self-correction loop for context refinement.

$$E_t = \{(u, v) \mid \exists \, \text{PMID associated with } (u, v) \text{ and year} \leq t\} \tag{2}$$

The graph dynamically grows as new associations are discovered and published, with subsequent snapshots G_t expanding upon earlier ones. This structure allows for modeling temporal biomedical knowledge with fine-grained control over historical context.

Path-finding and dataset generation. A path p_k is defined within a specific snapshot $G_t = (N_t, E_t)$ of the dynamic graph $G = \{G_t\}_{t=1}^{T}$. It consists of an ordered sequence of nodes connecting a source node $m_i \in N_t$ to a target node $m_j \in N_t$, traversing intermediate nodes $m_{k_1}, m_{k_2}, \ldots, m_{k_n} \in N_t$, such that each consecutive pair of nodes in the sequence is connected by an edge in E_t. Formally, a path is defined as:

$$p_k = (m_i, m_{k_1}, m_{k_2}, \ldots, m_{k_n}, m_j), \quad \text{with } (m_{k_\ell}, m_{k_{\ell+1}}) \in E_t \text{ for all } \ell \tag{3}$$

Additionally, each edge (m_{k_i}, m_{k_j}) within the path is associated with one or more PMIDs, indicating the source of its literature context, which is an integral part of the overall path representation.

Positive and negative samples. For each source-target pair (m_i, m_j), where a validated association appears at timestamp t, we construct positive training samples by identifying sets of paths by finding shortest paths from the network G at timestamp $t - 1$.

To generate positive samples, we extract a set of future reference terms $\mathcal{F}^t_{(m_i, m_j)}$ from PubMed abstracts published at timestamp t that describe the association (m_i, m_j). Specifically, we collect all terms m_q mentioned in the abstracts associated with the PMIDs that refer to the discovery (m_i, m_j) at timestamp t:

$$\mathcal{F}^t_{(m_i,m_j)} = \{m_q \mid \text{PMID}(m_i, m_j) \text{ at } t \text{ mentions } m_q\} \tag{4}$$

We define a set of positive paths $\mathcal{P}^+$. A path $p_k = (m_i, m_{k_1}, m_{k_2}, \ldots, m_{k_n}, m_j) \in \mathcal{P}^+$ is labeled as positive if its set of intermediate nodes is contained in the set of future reference terms $\mathcal{F}^t_{(m_i,m_j)}$. Formally:

$$p_k \in \mathcal{P}^+ \implies (m_{k_1}, m_{k_2}, \ldots, m_{k_n}) \subseteq \mathcal{F}^t_{(m_i,m_j)} \tag{5}$$

The positive label indicates that the sequence of nodes in the path (and the associated context) includes the future reference terms, suggesting that this path is associated with the discovery. The main idea is that if intermediate concepts in a path appear in future scientific literature discussing the target association, then the path likely reflects a valid reasoning trajectory.

We generate three types of negative samples to enhance the model's discriminative ability.

Hard Negative Samples. Hard negative samples are constructed by sampling paths from the network snapshot G_{t-1}, which represents the state of the biomedical co-occurrence graph prior to the discovery at timestamp t; they are intended to represent trajectories that are not related to the future discovery (m_i, m_j).

For a given source-target pair (m_i, m_j), a path p_k is labeled as negative if it includes intermediate node(s) that are not part of the future reference set $\mathcal{F}^t_{(m_i,m_j)}$:

$$p_k \in P^-_{\text{hard}} \quad \implies \quad \exists\, m_{k_r} \in \{m_{k_1}, m_{k_2}, \ldots, m_{k_n}\} : m_{k_r} \notin \mathcal{F}^t_{(m_i,m_j)}$$

Hard negatives P^-_{hard} are sampled from a pool of paths between m_i and m_j that are structurally valid (i.e., shortest paths in the co-occurrence graph) but were not part of future-relevant literature.

Corrupted Paths. Corrupted paths P^-_{corr} are generated by *introducing noise* into positive paths through *node replacement*:

$$P^-_{\text{corr}} = (m_i, m_{k_1}, \ldots, m_{k_{r-1}}, m'_{k_r}, m_{k_{r+1}}, \ldots, m_j)$$

where m'_{k_r} is a randomly sampled term from a valid set of terms N_{t-1} such that $m'_{k_r} \neq m_{k_r}$. This corruption simulates *near-miss scenarios*, where only a single term is altered.

Corrupted Contexts. Negative path samples with corrupted contexts $P^-_{\text{corr\_ctx}}$ are constructed by pairing a structurally valid path $p_k \in P^+$ with deliberately simulated corrupted contextual information. Specifically, we retrieve PMIDs related to individual nodes of a positive path P^+, but *without maintaining their co-occurrence relationships*:

$$C^-_{\text{corr}} = \{c_i \mid i \in P^+,\, c_i \in \mathcal{C}_{\text{node}}\}$$

where $\mathcal{C}_{\text{node}}$ refers to context abstracts retrieved by querying each node independently, breaking their relational structure.

We define the corrupted context path sample as:

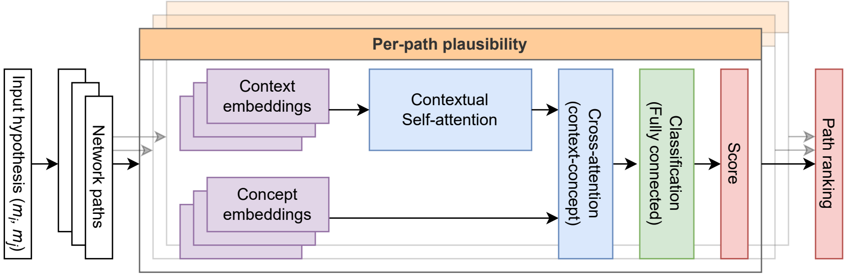

Input hypothesis (m_i, m_j)
Network paths
Per-path plausibility
Context embeddings
Contextual Self-attention
Concept embeddings
Cross-attention (context-concept)
Classification (Fully connected)
Score
Path ranking

Figure 2: Overview of the proposed HGCR framework. Input is a hypothesis: a pair of biomedical concepts (m_i, m_j). The system samples paths from the network G and ranks them based on their predicted alignment with the future reference context.

$$P^-_{\text{corr\_ctx}} = \left\{(p_k, C^-_{\text{corr}}) \,\middle|\, p_k \in P^+,\, C^-_{\text{corr}} \not\subseteq C^+\right\}$$

This type of negative sampling forces the model to distinguish between truly meaningful sequences of evidence and those composed of isolated, unrelated abstracts, despite otherwise valid path structure.

As a result, for every positive sample $p_k \in P^+$ we have the following set of negative samples:

$$P^- = P^-_{\text{hard}} \cup P^-_{\text{corr}} \cup P^-_{\text{corr\_ctx}}$$

3.2. Ranking Hypothesis Context

The proposed framework is designed to evaluate network paths between a given source term m_i and a target term m_j. The goal is to score these paths based on their relevance to a future discovery, inferred from the past data. The model is trained using a contrastive learning approach, optimizing the ability to distinguish positive paths from various types of negative paths discussed earlier.

3.2.1. Objective Function

The model is trained using a margin ranking loss to encourage the model to assign higher scores to plausible (future-validating) context-path pairs than to negative (implausible or corrupted) ones. Given a score for a positive sample $\hat{S}^+$ and a set of scores for negative samples $\{\hat{S}_i^-\}_{i=1}^N$, the loss is defined as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N} \max\left(0, \delta - \left(\hat{S}^+ - \hat{S}_i^-\right)\right),$$

where δ is a predefined margin, which we set to 0.3 for the best balance between classes separation and performance.

3.2.2. Model Architecture

The HGCR model computes a plausibility score for a candidate network path connecting two biomedical terms (m_i, m_j). It leverages two sources of information: the structural information encoded in concept embeddings $P = \{\mathbf{e}_1, \ldots, \mathbf{e}_k\}$, where each $\mathbf{e}_i \in \mathbb{R}^{d_n}$ represents a sequence of k nodes in the path; and the contextual information encoded in abstract embeddings $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$, where each $\mathbf{c}_j \in \mathbb{R}^{d_p}$ is derived from a corresponding MEDLINE abstract. The model architecture is shown in Figure 2.

The model first applies multi-head self-attention to the context embeddings to capture the interactions among different elements of the abstract context. The resulting context representations are then used as queries in a cross-attention block, while the concept embeddings are used as keys and values, such that each context position can attend to all path nodes. This produces a sequence-level path representation that integrates graph-based structural information into the contextual evidence. The resulting sequence is mean-pooled and passed through a fully connected classification layer with sigmoid activation to produce a final plausibility score $\hat{S} \in [0, 1]$:

$$\begin{aligned} C' &= \mathrm{SelfAttn}(C) \\ A &= \mathrm{CrossAttn}(Q = C',\ K = P,\ V = P) \\ \hat{S} &= \sigma(W \cdot \mathrm{MeanPool}(A) + b) \end{aligned}$$

This architecture is conceptually related to cross-modal models such as LXMERT and VilBERT (Tan and Bansal, 2019; Lu et al., 2019), which use self-attention followed by cross-attention to align different input modalities. Similarly, HGCR aligns structured (node-based) and unstructured (text-based) information via stacked attention mechanism, and produces path-level predictions based on their joint representation.

At inference time, paths are ranked based on their scores and are then passed to a downstream pipeline along with their context to serve as a basis for hypothesis explainability framework.

3.3. Explainability Framework

In this phase of the system, the goal is to transform latent, graph-based biomedical connections into interpretable explanations. To achieve this, we introduce a structured explainability pipeline that consists of four main components: (1) prompt construction with appended retrieved context from HGCR, (2) explanation generation via LLM, (3) validation with relation extraction and HG system ranking, and (4) iterative feedback loop for context refinement. The following subsections detail each component of this explainability pipeline, starting with how prompts are constructed and submitted to LLM for generating explanations.

Prompt Construction and LLM Generation. Step 1 in Figure 1 illustrates the prompt provided to LLM, which integrates scientific abstracts retrieved by HGCR as contextual input. The prompt is framed as follows: “Based on the following scientific abstracts, please describe how an indirect relationship between {SOURCE} and {TARGET} might exist.” In Step 2 we append context retrieved from HGCR to the prompt that is finally executed by LLM.

Structured Relation Extraction. Once LLM generates an explanation from the provided contextualized prompt (Figure 1, Steps 3.1 and 3.2), the output is passed to a relation extraction module (Step 4), which parses the explanation into structured biomedical predicates using SemRep. Each predicate is extracted in the form of *(subject, verb, object)*. All terms are normalized to UMLS Concept Unique Identifiers (CUIs) to ensure consistency and compatibility with downstream modules.

The extracted predicates serve as the basis for validating the factual accuracy of the generated explanation. At this point, the system transitions from generation to evaluation and refinement (i.e., the feedback loop).

Relation Evaluation via AGATHA Evaluator. Each extracted predicate is evaluated using the AGATHA Evaluator (Figure 1, Step 5), which performs both direct validation and plausibility ranking.

In the direct validation phase, the extracted predicate is matched against a precomputed set of biomedical relationships in the AGATHA semantic network. If no exact match is found, the system treats it as a hypothesis and computes its plausibility score using AGATHA predictor. This score is then compared against the scores of semantically related, but randomly generated biomedical concept pairs. The predicate is considered valid if its score ranks within the top 10% among this set. Predicates failing both checks are flagged as scientifically implausible. When this occurs, the corresponding sentence in the explanation is labeled as a *rework sentence*. These sentences identify problematic parts in the explanation that lack scientific support and trigger the context refinement process described next.

Feedback Loop and Context Refinement. The feedback loop mechanism (Step 6) allows the system to iteratively refine the prompt context in response to rework sentences. For each flagged sentence, the system identifies which PMID in the current context most likely contributed to the unsupported relationship.

The system then uses MedCPT embeddings to embed all candidate abstracts (PMIDs) associated with the same edge as the flagged sentence. Cosine similarity between the rework sentence and each candidate abstract is computed, and the top-ranked abstract that has not been already used is selected as a replacement. This updated "rework-aware context" is then used in the prompt, and LLM is queried again with the new context (returning to Step 3.1).

This feedback loop continues iteratively, with the explanation being regenerated, reevaluated, and further refined as needed. The loop terminates either when all sentences pass validation or when a maximum number of iterations is reached (in our experiments $n = 5$). Each source-target path is refined independently, ensuring that explanations are tailored to their specific biomedical context.

3.4. Baseline Approach

The baseline experiment evaluates how well LLMs generate coherent, scientifically plausible explanations for indirect biomedical relationships using a fixed context without feedback loop. For each path p_k, retrieved by the HGCR system, we collect abstracts for each edge and rank them by semantic similarity to the node pair of the corresponding edge using MedCPT embeddings. Then an explanation is generated by prompting an LLM as follows:

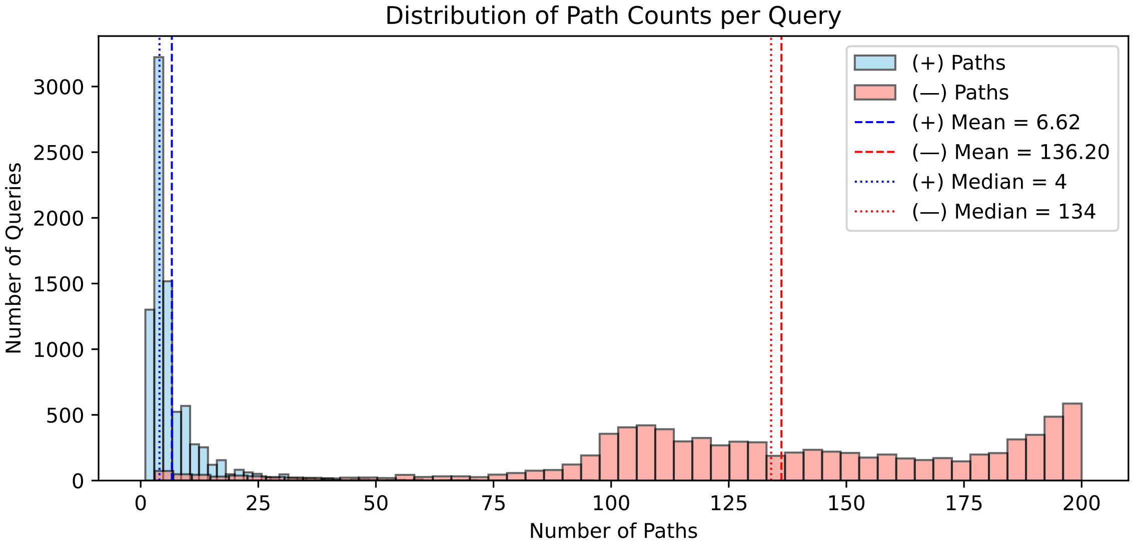


Figure 3: Distribution of positive and negative path counts per query in the test set. Queries q were extracted from the timestamp t corresponding to 2022 (and onwards) and the shortest paths between them were sampled from G_{t-1}, representing the graph state of 2021.

"Based on the following scientific abstracts, please describe how an indirect relationship between {SOURCE} and {TARGET} might exist. Consider the key findings, underlying mechanisms, and any intermediate entities or processes mentioned in the abstracts. Your explanation should connect these elements to form a coherent narrative that illustrates the possible indirect linkage between {source} and {target}."

4. Experiments

Experimental setup. We describe the experimental setup for two key components: graph-based context retrieval (HGCR) and explainability framework. The explainability framework is then run in two different settings: baseline and feedback loop. The baseline experiment evaluates the performance of the entire system in generating hypotheses from fixed context retrieved by the HGCR system, while the feedback loop experiment iteratively refines the hypotheses by leveraging generated responses and incorporating feedback to improve the quality of the generated explanations.

Test dataset. The evaluation of biomedical hypothesis generation and discovery-driven pipelines is challenging (Sybrandt et al., 2018), particularly in terms of assessing the quality of their explainability. Our evaluation methodology is an attempt to reflect on how scientific discoveries are made and validated over time.

Our test samples are extracted from the hypothesis generation benchmarking framework Dyport (Tyagin and Safro, 2023). It provides a dynamic biomedical knowledge graph G_t

(defined in Equation 1), where nodes $m_k \in N_t$ are UMLS concepts and edges $(m_i, m_j) \in E_t$ represent validated biomedical discoveries at timestamp t, annotated with contextual metadata. We use a temporal split where data prior to 2022 defines the subgraph G_{t-1}, used for HGCR training, and edges introduced in 2022 and later define our test set at timestamp t.

Each test instance is a query $q = (m_i, m_j) \in Q$, where (m_i, m_j) is an experimentally verified connection that exists in G_t (but not in G_{t-1}), which is also identified as a co-occurrence in the MEDLINE corpus with first appearance of this term pair in the literature at time t. The future reference set $FR(q)$ for a query is the collection of scientific abstracts from timestamp t that mention both m_i and m_j (abstract-level co-occurrence).

For each query q, we sample candidate network paths $p \in P(q)$ from G_{t-1}, where each path $p_k = (m_i, \ldots, m_j)$ is a shortest path between m_i and m_j that we limit to length 4 due to computational constraints. Note that multiple shortest paths in G are possible. Paths are labeled as positive or negative based on criteria described in the Methodology section. This results in a strongly imbalanced dataset which histogram is shown on Figure 3. We limit number of negative paths per query to 200 (100 paths of length 3 and 100 paths of length 4) to keep the task computationally feasible and yet this corresponds to a mean negative-to-positive ratio of approximately 20:1, though this varies across queries due to the underlying graph structure.

HGCR is evaluated as an information retrieval system over labeled paths $(p, l(p))$, while the explainability framework is assessed based on its ability to generate explanations that align with their future references $FR(q)$. For explainability evaluation, we use a subset of queries $Q_{\text{expl}} \subset Q$, selected such that each $q \in Q_{\text{expl}}$ has strong semantic connection to its future reference set, i.e., we keep only queries where both m_i and m_j appear together in the title of at least one future reference abstract $FR(q)$, or where a semantic predicate of the form (m_i, verb, m_j) is identified via the SemRep relation extraction system (Rindflesch and Fiszman, 2003). This ensures that $FR(q)$ provides a meaningful reference for evaluating the generated explanations.

4.1. Evaluation Metrics

4.1.1. Information Retrieval metrics

The metrics we use to report the HGCR performance are commonly used in the information retrieval field and describe the ranking performance: the area under receiver-operating characteristic curve (AUC ROC) and average precision (AP).

AUC ROC measures the trade-off between true positive rate and false positive rate across all thresholds, capturing the model's ability to distinguish between positive and negative samples independently of their absolute scores. AP summarizes the precision–recall curve, and is sensitive to the ranking of positive instances making it more appropriate in highly imbalanced settings, where correctly identifying rare positives is critical.

We report both micro-averaged and macro-averaged scores. In our context, where the number of paths per query is highly variable and class imbalance is significant, macro-averaging better reflects consistency across queries, while micro-averaging is dominated by high-volume queries and highlights global performance.

4.1.2. Explainability Evaluation Metrics

To evaluate the quality and factual alignment of generated explanations with novel scientific knowledge, we employ three complementary metrics. First, *lexical similarity* is measured using the Jaccard Index over UMLS terms extracted from the generated explanation and the reference abstract. Second, *semantic similarity* is assessed via the dot product between latent embeddings of the explanation and reference abstract, using two domain-specific models: MedCPT and SciNCL. They were selected for their strong performance in biomedical text representation learning (Jin et al., 2023b; Ostendorff et al., 2022a).

Finally, we report the error rate, defined as the proportion of low-ranked statements among all extracted statements in the generated explanation. This number reflects how frequently the system produces claims that are not supported by the hypothesis predictor. To compute the ranking criteria, we apply the same strategy outlined in Step 5 of the feedback loop (Section 3.3) that identifies wrong statements in the generated explanations.

5. Results

In this section, we present the results of our experiments and report evaluation metrics for the methods described earlier. We evaluate the retriever and the explainability framework separately, each within its respective task.

5.1. HGCR Context Ranking Evaluation

			ROC AUC		AP	
			Micro	Macro	Micro	Macro
Model	Text Encoder	Terms Encoder				
HGCR	MedCPT	AGATHA	**0.895**	**0.932**	**0.36**	0.536
		MedCPT	0.893	0.927	0.318	0.519
		SapBERT	0.894	0.927	0.329	0.516
	PubMedNCL	AGATHA	0.891	0.929	0.358	0.525
		MedCPT	0.891	0.929	0.33	0.531
		SapBERT	0.893	0.929	0.352	**0.536**
MedCPT	Article Encoder	Query Encoder	0.705	0.696	0.11	0.153

Table 1: Evaluation of HGCR model variants and comparison with MedCPT IR model using micro- and macro-averaged ROC AUC and Average Precision (AP). All HGCR models use the same architecture with different combinations of term and text encoders. Best values are hignlighted in **bold**, second best are underlined.

Table 1 summarizes the performance of HGCR across multiple encoder configurations using micro- and macro-averaged AUC and average precision metrics. The best results are consistently achieved when combining MedCPT for context encoding with AGATHA for term encoding, showing the benefit of aligning fine-tuned biomedical text representations with graph-based concept embeddings.

The HGCR model using AGATHA Term Encoder outperforms SapBERT and MedCPT variants across macro metrics, suggesting slightly better generalization to rare or diverse biomedical connections. Replacing the text encoder with PubMedNCL yields similar performance, confirming the robustness of the HGCR architecture with respect to contextual input.

We also include one baseline method based on zero-shot biomedical information retrieval model MedCPT (Jin et al., 2023b). It includes article and query encoders, which are utilized to compare its performance in the ranking evaluation setting. MedCPT computes query–document similarity directly without modeling paths, and we evaluate it by pooling context from positive and negative paths (ensuring that there is no overlap) and computing per-query scores. As expected, it does not perform as good as path-based models, because it lacks the ability to model multi-hop relational network structure and is tuned to work as an information retrieval system and not a predictor.

5.2. Explainability Framework Evaluation

We evaluate our proposed explainability approach by measuring how well the generated explanations align with corresponding future reference abstracts and how factually correct they are. To this end, we use multiple metrics described in Section 4.1.2. The results are presented in Table 2.

To ensure a fair comparison across systems, we controlled the information available to each model. All queries from Q_{expl} were first submitted to the VAIV system (the only system requiring interaction with a proprietary online platform[1]). After obtaining the VAIV outputs, we filtered out any query q which explanation e_{VAIV} included at least one abstract published in 2022 or later. This resulted in a filtered query set of 106 examples, which was then used across all systems. For MedCPT-based retrieval systems, we restricted the document corpus to abstracts published prior to 2022, aligning with the constraints applied to HGCR.

To account for variability, we generate explanations three times for non-retrieval models (with non-zero temperature), HGCR-based retrievals (based on the top-3 scored paths per query), and VAIV (which produced different outputs for repeated queries). Scores are averaged across runs. In contrast, MedRAG and iMedRAG were executed once with default parameters (including the LLM model), as they both internally use deterministic retrieval and fixed near-zero LLM temperature.

For reference, we also report results for standalone LLMs without retrieval. However, this comparison is inherently unfair, as these models were trained on data containing the future reference abstracts as well as works that cite them or mention the same results (from 2022 onward), which is evident from the reported knowledge cutoff[2]. Applying the same temporal restrictions as for retrieval systems is not feasible for pre-trained LLMs.

As shown in Table 2, the proposed HGCR-based approach yields explanations that better align with future references, both lexically and in latent embedding spaces. Notably, MedRAG and iMedRAG achieve relatively high MedCPT similarity despite lower Jaccard and SciNCL scores.

[1] https://bio.vaiv.kr/

[2] https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

Table 2: Performance comparison between the proposed explainability framework and similar systems. We report average values and standard deviation across 3 runs unless specified otherwise. (·) represents dot product between the explanation embedding and reference abstract embedding in the corresponding latent space. Best values are highlighted in **bold**, second best are underlined.

			Jaccard	MedCPT	Scincl	Error Rate
LLM	Explainer	Retriever	Index	(·)	(·)	(mean ± std)
Phi-4	Prompt	N/A	0.070 ± 0.027	61.098 ± 2.884	488.002 ± 27.264	0.044 ± 0.074
	BL	HGCR	**0.085 ± 0.034**	**62.052 ± 2.991**	**497.184 ± 25.293**	0.046 ± 0.063
	FL	HGCR	0.080 ± 0.032	61.982 ± 2.976	496.042 ± 26.718	0.020 ± 0.044
Llama-3.1 8B	Prompt	N/A	0.067 ± 0.026	60.573 ± 2.747	488.665 ± 29.073	0.045 ± 0.072
	BL	HGCR	0.075 ± 0.033	61.514 ± 2.736	494.373 ± 23.093	0.045 ± 0.072
	FL	HGCR	0.073 ± 0.033	61.169 ± 2.695	491.247 ± 23.344	0.015 ± 0.043
Llama-3.3 70B	Prompt	N/A	0.071 ± 0.029	60.711 ± 2.742	490.788 ± 27.637	**0.012 ± 0.032**
	BL	HGCR	0.081 ± 0.030	61.859 ± 2.895	495.750 ± 24.507	0.045 ± 0.073
	FL	HGCR	0.078 ± 0.029	61.692 ± 2.869	494.702 ± 25.124	0.021 ± 0.064
ChatGPT-3.5	MedRag	MedCPT	0.059 ± 0.029	61.471 ± 3.537	478.173 ± 34.952	0.079 ± 0.214
	iMedRag	MedCPT	0.062 ± 0.024	61.518 ± 3.234	484.051 ± 31.241	0.060 ± 0.180
	VAIV	Custom	0.065 ± 0.022	60.904 ± 3.247	480.932 ± 30.222	0.044 ± 0.061

The main goal of the feedback loop is to reduce hallucinations and unsupported claims, which is reflected in consistently lower error rates across models when the feedback mechanism is applied (rows marked with FL in Table 2). While the Llama-3.3 70B model without retrieval performs best overall in terms of error rate, the feedback loop-enabled Llama-3.1 8B closely follows. This can be explained by the ability of larger models like Llama-3.3 70B to manage its own parametric knowledge since there is likely an overlap between these models' knowledge cutoff and our test set, as was mentioned earlier. Other models equipped with the feedback loop also achieve comparable gains, demonstrating its general effectiveness in enhancing explanation reliability.

5.3. Case study

The interpretability of LLM-generated explanations is critical for fostering trust in hypothesis generation systems, particularly among biomedical researchers and clinicians. Building on prior work that highlights the value of human feedback (Tyagin et al., 2022), we conducted an expert evaluation to assess the explanations of five pairwise concept associations proposed by domain experts in drug discovery area. After the explanations were generated, the experts reviewed them for biological plausibility and interpretability to gauge the real-world utility of our system.

For the GABRA5 ↔ carbamazepine association, the explanation effectively connects two concepts by highlighting their roles in neural excitability and psychiatric conditions; Isavuconazole ↔ sunitinib pair is supported by its shared interaction with CYP3A4, where isavuconazole's inhibition of the enzyme can raise sunitinib levels and risk of toxicity, as shown by (Hu et al., 2024); and SLC6A3 ↔ bupropion, where the drug's inhibition of the dopamine transporter explains its behavioral effects. Icariin ↔ vascular dementia is linked through icariin's neuroprotective effects and modulation of PI3K/Akt and MEK/ERK pathways,

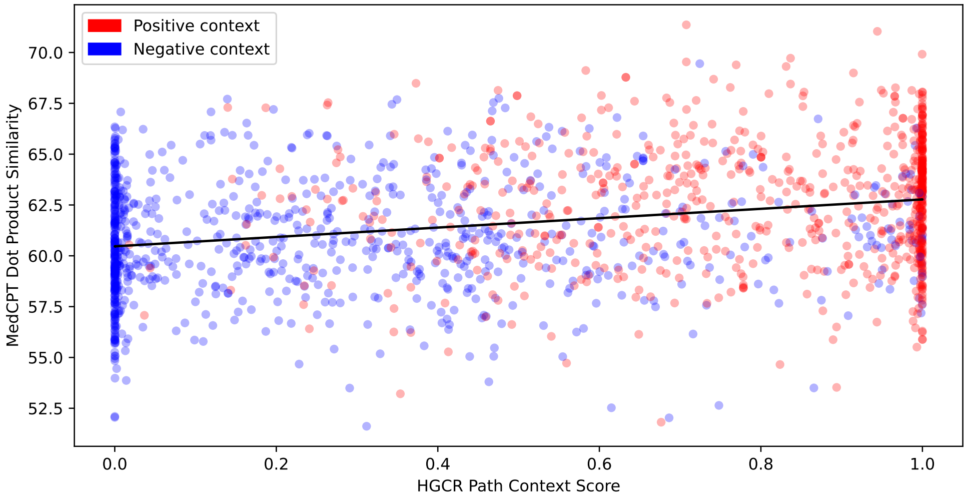


Figure 4: Relationship between HGCR Context Score (horizontal axis) and semantic similarity (MedCPT Dot Product, vertical axis) between final LLM-generated explanations (with feedback loop) and future reference scientific abstracts.

potentially mitigating apathy as a dopamine-related symptom via enhanced VTA–NAcc signaling, while SEZ6L2 ↔ calcitriol highlights overlapping pathways in neural development and glioma biology. These explanations demonstrate the effectiveness of the proposed path-based RAG framework in providing targeted biomedical context, enabling LLMs to generate novel and plausible hypotheses that are consistent with expert understanding and advance scientific discovery (See Appendix A for full explanations).

5.4. Retrieval Score and Explanation Alignment with Future Reference Context

In this section, we evaluate whether the HGCR path score reflects alignment between generated explanations and future reference abstracts. Specifically, we use a subset of queries Q_{expl} and, for each query $q \in Q_{\text{expl}}$, sample an equal number of positive and negative context paths $p_k \in P^+(q) \cup P^-_{\text{hard}}(q)$. We fix the number of sampled paths to 3 per class.

Each path p_k is assigned an HGCR context score $\hat{S}(p_k) \in [0, 1]$, computed as described in Methodology section. The explanation $e(p_k)$ corresponding to p_k is generated by iteratively applying the feedback loop until convergence. We measure the semantic alignment between $e(p_k)$ and the future reference abstract $FR(q)$ using MedCPT model, computing their similarity via dot product like it is done in the previous section.

We then analyze the relationship between $\hat{S}(p_k)$ and $sim_{\text{MedCPT}}(e(p_k), FR(q))$ across all samples, where $sim_{\text{MedCPT}}(e(p_k), FR(q))$ denotes the dot product similarity between the MedCPT embeddings of the generated explanation and the future reference abstract. Figure 4 plots this relationship for the explanations generated with Llama-3.3-70B model. Red points indicate paths sampled from positive (future-aligned) contexts, while blue points represent negative (non-aligned) contexts. A linear regression line clearly indicates a positive

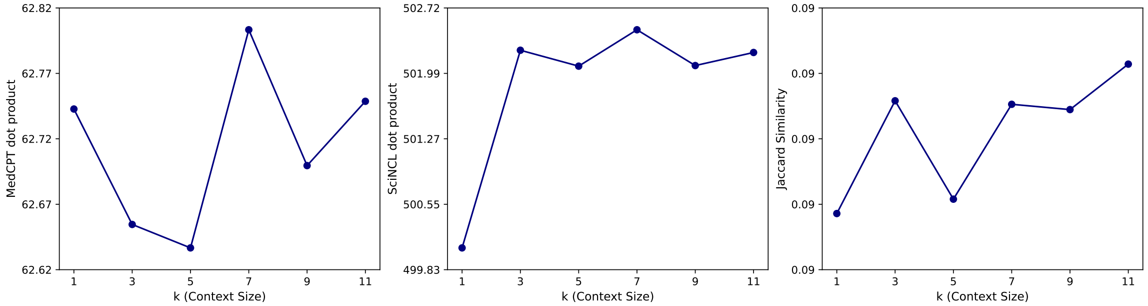


Figure 5: Different metrics across context sizes (k) in the ablation study. Larger context size tends to improve similarity metrics.

correlation, which suggests that the HGCR path ranking criteria can indicate the degree of semantic alignment between generated explanations and future scientific knowledge, as captured by reference abstracts.

5.5. Ablation Study: Context Size Influence on Explanation Quality

To assess the effect of increasing context length on LLM performance, we analyze key evaluation metrics across different context sizes $k \in \{1, 3, 5, 7, 9, 11\}$ for Phi-4 model. We keep the dataset the same as in previous experiment (Section 5.4), but we select only positively-labeled paths to simulate best-case scenario for the language model and reduce noise. The results (Figure 5) indicate that increasing k generally improves semantic similarity, particularly in the range $k = 3$ to $k = 7$. Based on these observations, the value $k = 7$ was used in our main experiment in the Results section.

Limitations. Evaluating free-text explanations is inherently difficult, especially in biomedical domains. Most prior work focuses on structured outputs (Xiong et al., 2024b,a; Soman et al., 2024), with limited attention to the quality and scientific plausibility of open-ended explanations. Moreover, identifying appropriate scientific context at scale remains challenging. This is an issue we attempt to address, but we should admit that it is still an open problem.

We also acknowledge that large language models may use their internal knowledge beyond HGCR's cutoffs, which became especially evident in case of Llama-3.3 70B model without any retrieval. Our primary retrieval-focused explanation evaluation strategy is supported by the results of ClashEval study (Wu et al., 2024) showing that LLMs tend to override their parametric knowledge with supplied context, which addresses this limitation, although not completely. Constructing the evaluation dataset with non-overlapping knowledge cutoffs between LLMs and recent scientific discoveries is not quite feasible due to the rapid models development and the nature of their training, especially using high-quality scientific data. Lastly, the feedback loop context update rule relies on local similarity-based heuristics, which may reinforce context drift or amplify biases introduced by early generation errors.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint arXiv: 2412.08905*, 2024.

Marina Aksenova, Justin Sybrandt, Biyun Cui, Vitali Sikirzhytski, Hao Ji, Diana Odhiambo, Matthew D Lucius, Jill R Turner, Eugenia Broude, Edsel Peña, Sofia Lizarraga, Jun Zhu, Ilya Safro, Michael D Wyatt, and Michael Shtutman. Inhibition of the dead box rna helicase 3 prevents hiv-1 tat and cocaine-induced neurotoxicity by targeting microglia activation. *Journal of Neuroimmune Pharmacology*, 15(2):209–223, 2020.

U Akujuobi, M Spranger, S.K. Palaniappan, and X Zhang. T-pair: Temporal node-pair embedding for automatic biomedical hypothesis generation. In *IEEE Transactions on Knowledge and Data Engineering*, pages 2988–3001. IEEE, 2020. doi: 10.1109/TKDE.2020.3017687.

Alan R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: The metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

G Bakal, P Talari, EV Kakani, and R Kavuluru. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. In *Journal of Biomedical Informatics*, pages 189–199. Elsevier, 2018. doi: 10.1016/j.jbi.2018.05.003.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. In *Proceedings of Nucleic Acids Research*, pages D267–D270. Oxford University Press, 2004a.

Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004b.

D Cameron, R Kavuluru, T.C. Rindflesch, A.P. Sheth, K Thirunarayan, and O Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. In *Journal of Biomedical Informatics*, pages 141–157. Elsevier, 2015. doi: 10.1016/j.jbi.2015.01.014.

Eugenio Cesario, Carmela Comito, and Ester Zumpano. A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. 568:127079, 2024.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben

Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Tammy H Cummings, Joseph Magagnoli, Sasha Sikirzhytskaya, Ilya Tyagin, Ilya Safro, Michael D Wyatt, Michael Shtutman, and S Scott Sutton. Exposure to angiotensin-converting enzyme inhibitors that cross the blood-brain barrier and the risk of dementia among people living with hiv. *AIDS*, pages 10–1097, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

M. Elbadawi, H. Li, A. W. Basit, and S. Gaisford. The role of artificial intelligence in generating original scientific research. *International Journal of Pharmaceutics*, 652:123741, 2024.

M.D. Gordon and S Dumais. Using latent semantic indexing for literature based discovery. In *Journal of the American Society for Information Science*, pages 674–685. Wiley, 1998. doi: 10.1002/(SICI)1097-4571(199806)49:8⟨674::AID-ASI2⟩3.0.CO;2-T.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, and many others. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3:1–23, 2021.

Sam Henry and Bridget T McInnes. Literature based discovery: models, methods, and trends. *Journal of biomedical informatics*, 74:20–32, 2017.

J Hu, H Xia, X Chen, X Xu, HL Wu, Y Shen, RA Xu, and W Wu. Effect of isavuconazole on the pharmacokinetics of sunitinib and its mechanism. *BMC Cancer*, 24(1):1131, Sep 11 2024. doi: 10.1186/s12885-024-12904-4.

Markus Iser. Automated explanation selection for scientific discovery, 2024.

K Jha, G Xun, Y Wang, and A Zhang. Hypothesis generation from text based on coevolution of biomedical concepts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 843–851. Association for Computing Machinery, 2019. doi: 10.1145/3292500.3330977.

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36, 2022.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39, 2023a. ISSN 1367-4811.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. Medcpt: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023b.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234–1240, 2020.

Mingchen Li, Halil Kilicoglu, Hua Xu, and Rui Zhang. Biomedrag: A retrieval augmented large language model for biomedicine. 2024.

R.K. Lindsay and M.D. Gordon. Literature-based discovery by lexical statistics. In *Journal of the American Society for Information Science*, pages 574–587. Wiley, 1999. doi: 10.1002/(SICI)1097-4571(1999)50:7⟨574::AID-ASI3⟩3.0.CO;2-Q.

Jerry Liu. LlamaIndex, 11 2022.

A. Lozano, S. L. Fleming, C.-C. Chiang, and N. Shah. Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature. *Patterns*, 8, 2023.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: Generative pre-trained transformer for biomedical text generation and mining. In *Proceedings of Briefings in Bioinformatics*, page bbac409. Oxford University Press, 2022.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. *arXiv preprint arXiv:2202.06671*, 2022a.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings, 2022b.

Y. J. Park, D. Kaplan, Z. Ren, C.-W. Hsu, C. Li, H. Xu, S. Li, and J. Li. Can chatgpt be used to generate scientific hypotheses? *Journal of Materiomics*, 10(3):578–584, 2024. DOI: 10.1016/j.jmat.2023.08.007.

Alexander R. Pelletier, Joseph Ramirez, Irsyad Adam, Simha Sankar, Yu Yan, Ding Wang, Dylan Steinecke, Wei Wang, and Peipei Ping. Explainable biomedical hypothesis generation via retrieval augmented generation enabled large language models, 2024. URL https://arxiv.org/abs/2407.12888.

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: A text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*, 2021.

Karl Popper. The logic of scientific discovery. In *Routledge Classics*, 1959.

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, 2024.

M Rastegar-Mojarad, RK Elayavilli, D Li, R Prasad, and H Liu. A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. In *Proceedings of the 15th IEEE Conference on Bioinformatics and Biomedicine (BIBM)*, pages 669–674. IEEE, 2015. doi: 10.1109/BIBM.2015.7359766.

Thomas C. Rindflesch and Marcelo Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. In *Proceedings of the Journal of Biomedical Informatics*, pages 462–477. Elsevier, 2003. doi: 10.1016/j.jbi.2003.11.003.

F Shi, J.G. Foster, and J.A. Evans. Weaving the fabric of science: dynamic network models of science's unfolding structure. In *Social Networks*, pages 73–85. Elsevier, 2015. doi: 10.1016/j.socnet.2015.02.006.

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.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, Sharat Israni, Charlotte A Nelson, Sui Huang, and Sergio E Baranzini. Biomedical knowledge graph-optimized prompt generation for large language models. *Bioinformatics*, 40(9):btae560, 09 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae560. URL https://doi.org/10.1093/bioinformatics/btae560.

P Srinivasan. Text mining: generating hypotheses from medline. In *Journal of the American Society for Information Science and Technology*, pages 396–413. Wiley, 2004. doi: 10.1002/asi.10389.

Don R. Swanson. Undiscovered public knowledge. *The Library Quarterly*, 56:103–118, 1986.

Justin Sybrandt and Ilya Safro. Cbag: Conditional biomedical abstract generation. *PLOS ONE*, 16:e0253905, 2021. doi: 10.1371/journal.pone.0253905.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. Moliere: Automatic biomedical hypothesis generation system. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1633–1642, 2017. doi: 10.1145/3097983.3098057.

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

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*, pages 2757–2764. Association for Computing Machinery, 2020.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

S.B. Taneja, T.J. Callahan, M.F. Paine, S.L. Kane-Gill, H. Kilicoglu, M.P. Joachimiak, and R.D. Boyce. Developing a knowledge graph for pharmacokinetic natural product-drug interactions. In *Journal of Biomedical Informatics*, page Article 104341. Elsevier, 2023. doi: 10.1016/j.jbi.2023.104341.

Ilya Tyagin and Ilya Safro. Dyport: Dynamic importance-based hypothesis generation benchmarking technique. In *Proceedings of the Conference on Artificial Intelligence (AAAI)*. AAAI Press, 2023. URL https://arxiv.org/abs/2312.03303. arXiv:2312.03303.

Ilya Tyagin, Ankit Kulshrestha, Justin Sybrandt, Krish Matta, Michael Shutman, and Ilya Safro. Accelerating covid-19 research with graph mining and transformer-based learning. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*. AAAI, 2022.

Huijun Wang, Ying Ding, Jie Tang, Xiao Dong, Bing He, Judy Qiu, and David J Wild. Finding complex biological relationships in recent pubmed articles using bio-lda. *PloS one*, 6(3):e17243, 2011.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2023.

Kevin Wu, Eric Wu, and James Y Zou. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in Neural Information Processing Systems*, 37:33402–33422, 2024.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024a.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. Improving retrieval-augmented generation in medicine with iterative follow-up questions. *arXiv preprint arXiv:2408.00727*, 2024b.

G Xun, K Jha, V Gopalakrishnan, Y Li, and A Zhang. Generating medical hypotheses based on evolutionary medical concepts. In *Proceedings of the IEEE International Conference on Data Mining (ICDM'17)*, pages 535–544. Institute of Electrical and Electronics Engineers Inc., 2017. doi: 10.1109/ICDM.2017.63.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

M Yetisgen-Yildiz and W Pratt. Using statistical and knowledge-based approaches for literature-based discovery. In *Journal of Biomedical Informatics*, pages 600–611. Elsevier, 2006. doi: 10.1016/j.jbi.2005.11.010.

H Zhou, H Jiang, W Yao, and X Du. Learning temporal difference embeddings for biomedical hypothesis generation. In *Bioinformatics*, pages 5253–5261. Oxford University Press, 2022. doi: 10.1093/bioinformatics/btac660.

Cristina Zuheros, Eugenio Martínez-Cámara, Enrique Herrera-Viedma, and Francisco Herrera. Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. case study of restaurant choice using tripadvisor reviews. *Information Fusion*, 68:22–36, 2021.

Appendix A. Case study

Here we provide the full version of five source–target concept pairs alongside their corresponding explanations, which were evaluated by a domain expert to assess biological plausibility and interpretability.

A.1. GABRA5 gene and carbamazepine

Explanation: To explore the potential indirect relationship between the GABRA5 gene and carbamazepine, we need to consider the information provided in the abstracts and connect the dots through genetic, pharmacological, and neurobiological pathways. GABRA5 Gene and Suicidal Behavior: The first abstract discusses genetic variants associated with suicidal behavior, including the GABRA5 gene. GABRA5 is part of the GABA A receptor complex, which is involved in inhibitory neurotransmission in the brain. Variants in this gene have been linked to psychiatric conditions and behaviors, including those related to suicide risk. Carbamazepine and Seizure Control: The third abstract mentions the use of carbamazepine in treating seizures in a patient with Dyke-Davidoff-Masson syndrome (DDMS). Carbamazepine is an anticonvulsant and mood-stabilizing drug that works by stabilizing hyperexcited nerve membranes, inhibiting repetitive neuronal firing, and reducing synaptic propagation of excitatory impulses. It primarily acts on sodium channels but also has effects on GABAergic neurotransmission. Indirect Link via Neurotransmission: The connection between GABRA5 and carbamazepine can be hypothesized through their roles in neurotransmission. GABRA5, as part of the GABA A receptor, influences inhibitory signaling in the brain. Carbamazepine, while primarily affecting sodium channels, may also modulate GABAergic activity indirectly. This modulation could influence the overall excitatory-inhibitory balance in the brain, which is crucial in conditions like epilepsy and potentially in mood disorders linked to suicidal behavior. Potential Intermediate Mechanisms: The interaction between GABRA5 variants and carbamazepine's effects might involve changes in GABAergic signaling pathways. If GABRA5 variants alter GABA receptor function, this could affect how carbamazepine modulates neuronal excitability and seizure control. Additionally, since GABAergic dysfunction is implicated in various psychiatric disorders, carbamazepine's influence on this system might also impact mood and behavior, indirectly linking it to the genetic risk factors for suicidal behavior. Gene-Environment Interactions: The first abstract also mentions gene-environment interactions, such as the interaction between NTRK2 genetic variation and childhood abuse. While not directly related to GABRA5 or carbamazepine, this highlights the complexity of genetic influences on behavior and how environmental factors can modulate genetic risk, potentially affecting how individuals respond to medications like carbamazepine. In summary, the indirect relationship between the GABRA5 gene and carbamazepine may exist through their shared involvement in modulating neurotransmission, particularly GABAergic signaling. Variants in GABRA5 could influence the efficacy or side effects of carbamazepine by altering the inhibitory-excitatory balance in the brain, which is crucial for both seizure control and mood regulation. Further research would be needed to elucidate these pathways and their clinical implications.

Expert assessment: Explanation correrctly identifies the role of GABRA5 in GABAergic neurotransmission and links it to carbamazepine's effects on neuronal excitability. Additionally, it considers gene environment interactions and psychiatric implications, providing a more comprehensive view of their potential connection. The GABRA5 gene encodes the $\alpha5$ subunit of the GABA A receptor, a key component of inhibitory neurotransmission in the brain. It plays a role in neurodevelopment, synaptic plasticity, and psychiatric disorders. Carbamazepine is an anticonvulsant and mood stabilizer that primarily acts by blocking voltage gated sodium channels, reducing neuronal excitability. It is commonly used to treat epilepsy, bipolar disorder, and trigeminal neuralgia. The indirect relationship between GABRA5 and carbamazepine likely involves their shared involvement in neural excitability and plasticity. GABRA5 regulates inhibitory neurotransmission, while carbamazepine modulates excitatory activity by stabilizing sodium channels. Through the balance of excitatory inhibitory signaling, alterations in GABRA5 expression or function could influence carbamazepine's effectiveness or mechanisms of action.

A.2. SLC6A3 gene and bupropion

Explanation: To explore the potential indirect relationship between the SLC6A3 gene and bupropion, we need to consider the roles and interactions of these elements based on the provided abstracts. SLC6A3 Gene: The SLC6A3 gene encodes the dopamine transporter (DAT), which is crucial for regulating dopamine levels in the brain by facilitating the reuptake of dopamine from the synaptic cleft back into presynaptic neurons. This gene is part of the dopaminergic system, which is implicated in cognitive processes and intelligence, as mentioned in the first abstract. Bupropion: Bupropion is an antidepressant and smoking cessation aid that primarily acts as a norepinephrine-dopamine reuptake inhibitor (NDRI). By inhibiting the reuptake of dopamine and norepinephrine, bupropion increases the levels of these neurotransmitters in the synaptic cleft, enhancing dopaminergic and adrenergic signaling. Indirect Relationship: Dopaminergic System: The SLC6A3 gene, through its product (DAT), plays a role in modulating dopamine levels. Bupropion, by inhibiting DAT, indirectly affects the function of the SLC6A3 gene product, leading to increased synaptic dopamine levels. Cognitive and Behavioral Effects: The dopaminergic system, influenced by SLC6A3, is involved in cognitive processes and mood regulation. Bupropion's action on dopamine reuptake can alter these processes, potentially impacting cognitive functions and mood, which are areas where SLC6A3 is also relevant. Side Effects and Psoriasis: The third abstract highlights a case where bupropion exacerbated erythrodermic psoriasis, a severe skin condition. While this is not directly related to SLC6A3, it underscores the complex interactions and side effects that can arise from altering neurotransmitter systems, which are influenced by genes like SLC6A3. In summary, the indirect relationship between the SLC6A3 gene and bupropion can be understood through their shared involvement in the dopaminergic system. Bupropion affects dopamine levels by inhibiting the dopamine transporter encoded by SLC6A3, thereby influencing cognitive and mood-related processes. This interaction highlights the broader implications of pharmacological interventions on genetic and neurobiological pathways.

Expert assessment: Explanation offers a particularly direct and relevant connection between the SLC6A3 gene and bupropion. It focuses on their shared role in the dopaminergic system, clearly explaining how the gene encodes the dopamine transporter (DAT) and how bupropion inhibits DAT, leading to increased synaptic dopamine levels. This direct mechanism is central to understanding their interaction. This connection is further supported by its logical extension to cognitive and behavioral effects, which are well-established areas of influence for both dopamine regulation and bupropion. It also touches on the potential for side effects, showing a more comprehensive view.

A.3. Isavuconazole and Sunitinib

Explanation: An indirect relationship between isavuconazole and sunitinib can be proposed based on the provided abstracts, focusing on the role of CYP3A inhibitors and their impact on the pharmacokinetics of both drugs. Here's a coherent narrative illustrating this possible indirect linkage: Isavuconazole and CYP3A Interaction: Isavuconazole is known to be a moderate inhibitor of the CYP3A enzyme. Although the abstracts do not directly discuss isavuconazole's interaction with sunitinib, understanding its role as a CYP3A inhibitor is crucial.Sunitinib Metabolism and CYP3A: Sunitinib is primarily metabolized by the CYP3A enzyme. The study on the effects of CYP3A inhibitors (ketoconazole, voriconazole, and itraconazole) on the pharmacokinetics of sunitinib shows that strong CYP3A inhibitors like ketoconazole and voriconazole can significantly increase the exposure of sunitinib by inhibiting its metabolism. This suggests that any drug that inhibits CYP3A could potentially affect sunitinib's pharmacokinetics.Indirect Relationship: Given that isavuconazole is a CYP3A inhibitor, albeit moderate, it could potentially influence the pharmacokinetics of sunitinib if co-administered. The mechanism would involve isavuconazole inhibiting the CYP3A enzyme, which in turn could reduce the metabolism of sunitinib, leading to increased exposure of sunitinib in the body. This increased exposure could enhance the efficacy of sunitinib but also potentially increase the risk of toxicity. Intermediate Entities or Processes: The key intermediate entity in this relationship is the CYP3A enzyme. The process involves the inhibition of this enzyme by isavuconazole, which then affects the metabolism of sunitinib. Another intermediate process could be the alteration in the pharmacokinetic profile of sunitinib, including its absorption, distribution, metabolism, and excretion (ADME), primarily due to the inhibition of its metabolism. Clinical Implications: Understanding this indirect relationship is crucial for clinical practice. If isavuconazole and sunitinib are to be used concomitantly, healthcare providers should be aware of the potential for increased sunitinib exposure and monitor patients closely for signs of toxicity or enhanced efficacy. This might necessitate dose adjustments of sunitinib or careful monitoring of its levels to maintain therapeutic efficacy while minimizing adverse effects. In summary, the indirect relationship between isavuconazole and sunitinib is mediated through their interaction with the CYP3A enzyme. Isavuconazole, as a CYP3A inhibitor, can potentially increase the exposure of sunitinib by reducing its metabolism, thereby indirectly affecting its pharmacokinetics and potentially its efficacy and safety profile.

Expert assessment: The link between isavuconazole and sunitinib has the potential to be true because both drugs interact with the same enzyme, CYP3A4. Isavuconazole is a moderate inhibitor of CYP3A4, which slows down the enzyme's activity. Sunitinib, on the other hand, relies almost entirely on CYP3A4 to break down and leave the body. It is known that strong CYP3A4 inhibitors like ketoconazole can double sunitinib levels in the blood by blocking its metabolism and increasing its levels, which was shown recently by (Hu et al., 2024). This could mean stronger effects from sunitinib and also a higher risk of side effects like high blood pressure, fatigue, or liver dysfunction. It is important to monitor patients closely and adjust the dose of sunitinib if the two are co-administered.

A.4. Icariin and Dementia, Vascular

Explanation: Based on the provided abstracts, an indirect relationship between icariin and vascular dementia (VaD) can be proposed through the following mechanisms and intermediate entities: Icariin's neuroprotective effects: Icariin has been shown to have neuroprotective properties, particularly in the context of Parkinson's disease (PD). It can protect dopaminergic neurons and modulate signaling pathways such as PI3K/Akt and MEK/ERK, which are involved in cell survival and neuroprotection. Dopamine pathway and apathy: The studies on MPTP-lesioned primates suggest that dopaminergic dysfunction, particularly in the ventral tegmental area (VTA)-nucleus accumbens (NAcc) pathway, contributes to the manifestation of apathetic behaviors. Apathy is a common symptom in both PD and VaD. Apathy and social withdrawal in VaD: The study on social withdrawal in patients with VaD found that apathy, impaired social judgment, and disturbance of basic daily activities are associated with social withdrawal. Apathy, in particular, was a significant predictor of social withdrawal. Link between icariin and VaD: While there is no direct evidence of icariin's effects on VaD, its neuroprotective properties and potential to modulate dopamine pathways could indirectly influence the development or progression of VaD. Icariin might help mitigate apathy and social withdrawal in VaD patients by protecting dopaminergic neurons and promoting dopamine signaling. The proposed indirect relationship between icariin and VaD can be summarized as follows: Icariin → Neuroprotection of dopaminergic neurons → Modulation of dopamine pathways (e.g., VTA-NAcc) → Reduction of apathy → Decreased social withdrawal → Potential improvement in VaD symptoms This hypothetical relationship highlights the potential for icariin to influence VaD symptoms indirectly, through its effects on dopamine pathways and apathy. However, further research is necessary to explore this relationship and determine the therapeutic potential of icariin in VaD.

Expert assessment: This proposed connection is logical as a hypothesis because it links established properties of icariin with a key symptom of vascular dementia through a shared biological pathway. It is known that icariin demonstrates neuroprotective effects, particularly on dopaminergic neurons, and can modulate critical cell survival pathways like PI3K/Akt and MEK/ERK. The apathy, which is one of the debilitating symptoms in vascular dementia, is understood to involve dysfunction in dopamine signaling, specifically pathways like the VTA-NAcc connection. Therefore, it is biologically plausible that icariin's ability to protect dopamine neurons and enhance dopamine-related signaling could indirectly improve vascular dementia symptoms by mitigating apathy and its downstream consequences, like social withdrawal.

A.5. SEZ6L2 gene and calcitriol

Explanation: To explore the potential indirect relationship between the SEZ6L2 gene and calcitriol, we need to consider the roles and interactions of these entities within the context of neural development and glioma biology, as described in the provided abstracts. Key Findings and Mechanisms: SEZ6L2 Gene in Neural Development: The SEZ6L2 gene is identified as differentially expressed during mouse neural tube development, with alternative splicing observed. This suggests its involvement in the complex regulation of neural development processes. SEZ6L2 is among the genes dysregulated in a retinoic acid-induced neural tube defects (NTDs) model, indicating its potential role in neural development and associated pathologies. Calcitriol in Glioma Biology: Calcitriol, the active form of vitamin D3, promotes differentiation of glioma stem-like cells (GSCs) and increases their susceptibility to temozolomide, a common chemotherapeutic agent for glioblastoma. The study suggests that calcitriol can reduce the stemness of GSCs, potentially impacting tumor growth and treatment resistance. Indirect Relationship: To hypothesize an indirect relationship between SEZ6L2 and calcitriol, we can consider the following narrative: Neural Development and Glioma Initiation: SEZ6L2's role in neural development suggests it may influence the cellular environment or pathways that are also relevant in glioma initiation or progression. Neural progenitors and stem cells, which are implicated in both normal neural development and glioma formation, could be a connecting point. Calcitriol's Impact on Glioma Stem Cells: Calcitriol affects glioma stem-like cells by promoting differentiation and reducing stemness, which could alter the cellular landscape in a way that impacts genes involved in neural development, such as SEZ6L2. Potential Intermediate Pathways: Both SEZ6L2 and calcitriol may interact with common signaling pathways or cellular processes, such as those involving transcription factors or signaling molecules that regulate cell differentiation and proliferation. For instance, pathways involving retinoic acid (mentioned in the context of SEZ6L2 dysregulation) and vitamin D signaling (related to calcitriol) could intersect, influencing gene expression profiles and cellular behaviors in neural and glioma contexts. Conclusion: While the abstracts do not provide direct evidence of an interaction between SEZ6L2 and calcitriol, the indirect relationship could be mediated through shared pathways in neural development and glioma biology. Further research would be needed to elucidate specific molecular mechanisms and validate this hypothesis.

Expert assessment: A notable strength of the explanation lies in a tight focus on the neural context, directly addressing the roles of the SEZ6L2 gene in neural development and calcitriol in glioma biology. By meticulously outlining the potential connections through neural progenitor cells and shared signaling pathways, it provides a logical and coherent description. The explanation doesn't just state a possibility; it researches into potential mechanisms, such as the interplay between retinoic acid and vitamin D signaling, demonstrating a deeper understanding of the subject matter. Furthermore, it strikes a balance between acknowledging the absence of direct evidence and presenting a plausible hypothesis, making it a well-reasoned and insightful response.

Appendix B. Incorporated Technologies

The proposed pipeline relies on multiple key technologies and biomedical databases to conduct experiments effectively.

B.1. UMLS (Unified Medical Language System)

UMLS (Bodenreider, 2004a), developed by the National Library of Medicine (NLM), integrates and standardizes biomedical terminologies, taxonomies, and coding systems. It provides concept unique identifiers (CUIs) that unify terms from diverse vocabularies, facilitating interoperability across biomedical resources.

B.2. MetaMap and SemRep

MetaMap (Aronson, 2001) is an NLM-developed software that identifies biomedical concepts in text using Named Entity Recognition (NER) and maps them to UMLS concepts, enabling standardized representation. SemRep (Rindflesch and Fiszman, 2003), also developed by NLM, leverages MetaMap's capabilities to extract structured semantic predicates from biomedical literature, representing relationships as subject-verb-object predicates and capturing explicit biomedical knowledge through these entities.

B.3. AGATHA

AGATHA (Sybrandt et al., 2020) (Automatic Graph-mining And Transformer-based Hypothesis generation Approach) is a versatile hypothesis generation (HG) system that integrates a multi-layered semantic graph with transformer-based deep learning techniques. It constructs a large-scale semantic network from biomedical literature, applies advanced NLP techniques, and utilizes transformer-based architecture to predict and rank plausible connections efficiently. Unlike traditional link prediction systems, AGATHA offers a comprehensive framework where predicting links is just one part of a broader pipeline, making it suitable for diverse hypothesis generation tasks across biomedical literature.

B.4. BERT-derivative Models

The pipeline employs BERT-based models such as SciNCL/PubMedNCL (Ostendorff et al., 2022b) and MedCPT (Jin et al., 2023a), which are optimized for scientific and biomedical text processing, respectively.

B.5. VLLM

VLLM (Kwon et al., 2023) is an optimized inference and serving framework for large language models (LLMs), enabling efficient model execution with reduced memory overhead. It supports fast decoding and inference for transformer-based models, making it particularly useful for large-scale biomedical text processing.

Appendix C. Technical details for Large Language Models

For the feedback loop experiment, various Large Language Models (LLMs) were utilized to generate hypotheses. To ensure fairness and consistency across experiments, we uniformly used Phi-4 (Abdin et al., 2024), Llama-3.1 8B (Grattafiori et al., 2024), and Llama-3.3 70B for all LLM-related components, including both the baseline and feedback loop experiments. For consistency and reproducibility in the feedback loop experiment, the temperature parameter was set to 1×10^{-19}, and the top-p value was set to 1×10^{-9}. These extreme values were selected to minimize randomness and ensure deterministic output across multiple iterations, helping maintain a high degree of consistency in the LLM responses. This was particularly important for iteratively refining hypotheses based on feedback.

Models Llama-3.1 8B with 8 billion parameters, Llama-3.3 70B with 70 billion parameters, and Phi-4 with 14 billion parameters were downloaded and deployed locally using VLLM (Kwon et al., 2023), enabling efficient parallel processing and inference for hypothesis generation.

In our entire experiment, the prompts were designed to use a zero-shot approach, where the LLMs generated hypotheses without any task-specific fine-tuning. Furthermore, the max_completion_tokens parameter was set to 1000 to limit the response length during inference using VLLM, and the following parameters were applied across all models: max-model-len was set to 16384, temperature was set to 1×10^{-19} and top-p was set to 1×10^{-9} to ensure coherence and reduce variability. For the prompt experiment, where no retriever was used, we set the temperature to 0.5 and top-p to 0.6.

Appendix D. Explainability Correctness Evaluation

In addition to our main results, we attempt to evaluate factual correctness using an LLM-as-a-judge approach (Chiang and Lee, 2023) implemented with LlamaIndex's `Correctness Evaluator` (Liu, 2022), which scores contextual relevance and accuracy on a 1–5 scale. This score is intended to capture both terminological overlap and deeper semantic and factual alignment. To reduce bias, we incorporate 3 strong local LLMs serving as evaluation experts: Llama-3.3 70B, Gemma 3 27B (Kamath et al., 2025) and Qwen 3 32B (Yang et al., 2025). We calculate correctness for each explanation independently and then take the average of 3 to get the final score.

The result of this experiment is shown in Table 3. Scores indicate that the best factual correctness is achieved with models that do not use any retrieval. This can be explained by the limitation we mentioned earlier: applying the appropriate temporal knowledge cutoff restrictions as we did for tested retrieval systems is not feasible for pre-trained LLMs and their exposure to our reference texts from the test set potentially gave them an unfair advantage. Also, we hypothesize that since LLMs rely on latent knowledge beyond the provided context, the observed difference in the correctness scores could be a result of the epistemic uncertainty impact.

LLM	Prompt	FL	BL
Phi-4	3.874 ± 0.403	3.808 ± 0.566	3.795 ± 0.561
Llama 3.1_8B	3.541 ± 0.563	3.291 ± 0.672	3.360 ± 0.663
Llama 3.3_70B	3.996 ± 0.413	3.666 ± 0.624	3.650 ± 0.633

Table 3: Mean ± standard deviation of correctness scores across different models and experiments.

Appendix E. Prompt Complexity and Operational cost of Pipeline

Table 4: Mean response time and token statistics (Mean ± std) across different models and experiments.

Model	Experiment	Mean Response Time (sec)	Total Input Tokens	Total Output Tokens	Iterations
Phi-4	Prompt	8.18 ± 0.87	74.17 ± 7.09	567.34 ± 57.34	1
	FL	67.89 ± 59.66	11598.92 ± 8423.15	1215.65 ± 894.46	1.93 ± 1.39
	BL	10.72 ± 2.55	6273.74 ± 1623.09	599.45 ± 76.08	1
Llama-3.1 8B	Prompt	6.47 ± 1.32	74.15 ± 7.06	672.68 ± 138.58	1
	FL	76.04 ± 74.22	11944.56 ± 8325.30	1455.44 ± 1056.21	2.03 ± 1.39
	BL	8.24 ± 2.95	6268.51 ± 1620.32	684.48 ± 136.35	1
Llama-3.3 70B	Prompt	37.02 ± 4.65	74.15 ± 7.06	687.21 ± 87.33	1
	FL	117.90 ± 91.48	11719.58 ± 8169.37	1198.43 ± 838.37	1.97 ± 1.37
	BL	35.75 ± 4.69	6268.51 ± 1620.32	593.94 ± 81.20	1

Prompt complexity and operational cost of the proposed pipeline are presented in Table 4. We report mean values for response times as well as input and output tokens per model and experiment type.

Appendix F. Convergence Statistics of the Feedback Loop Experiment

The convergence report, which quantifies how often the feedback loop stops early and how often it hits the iteration limit, is shown in Table 5. In this table, the total number of runs is 318 for every language model, which represents 106 queries that were selected for the results section (Table 2) and run with 3 different context paths.

Table 5: Iteration counts required for convergence across models. The last row indicates cases that required more than 5 iterations and did not converge.

Iterations completed and converged	Phi-4	Llama-3.1 8B	Llama-3.3 70B
1	185	171	177
2	61	60	61
3	21	34	27
4	11	14	18
5	5	13	11
>5 (did not converge, more iterations needed)	35	26	24