Building a literature knowledge base towards transparent biomedical AI

Yuanhao Huang[1] Zhaowei Han[1] Xin Luo[1] Xuteng Luo[2] Yijia Gao[2]
Meiqi Zhao[3] Feitong Tang[2] Yiqun Wang[1] Jiyu Chen[2] Chengfan Li[2]
Xinyu Lu[2] Jiahao Qiu[2] Feiyang Deng[4] Tiancheng Jiao[2] Dongxiang Xue[5]
Fan Feng[1] Thi Hong Ha Vu[1] Lingxiao Guan[2] Jean-Philippe Cartailler[6]
Michael Stitzel[7,8,9] Shuibing Chen[5,10] Marcela Brissova[11,12] Stephen Parker[1,4,13]
Jie Liu[1,2]

[1]Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, MI, USA
[2]Department of Computer Science & Engineering, University of Michigan, Ann Arbor, MI, USA
[3]School of Information, University of Michigan, Ann Arbor, MI, USA
[4]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA
[5]Department of Surgery, Weill Cornell Medical College, 1300 York Avenue, New York, NY 10065, USA
[6]Creative Data Solutions, Vanderbilt Center for Stem Cell Biology, Vanderbilt University, Nashville, TN, USA
[7]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032
[8]Department of Genetics and Genome Sciences, University of Connecticut, Farmington, CT 06032
[9]Institute of Systems Genomics, University of Connecticut, Farmington, CT 06032
[10]Center for Genomic Health, Weill Cornell Medicine, 1300 York Avenue, New York, NY 10065, USA
[11]Division of Diabetes, Endocrinology and Metabolism, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN
[12]Human Pancreas Analysis Program, Nashville, TN; Philadelphia, PA; and Gainesville, FL
[13]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA

Abstract

Knowledge graphs have recently emerged as a powerful data structure to organize biomedical knowledge with explicit representation of nodes and edges. The knowledge representation is in a machine-learning ready format and supports explainable AI models. However, PubMed, the largest and richest biomedical knowledge repository, exists as free text, limiting its utility for advanced machine learning tasks. To address the limitation, we present LiteralGraph, a computational framework that rigorously extracts biomedical terms and relationships from PubMed literature. Using this framework, we established Genomic Literature Knowledge Base (GLKB), a knowledge graph that consolidates 263,714,413 biomedical terms, 14,634,427 biomedical relationships, and 10,667,370 genomic events from 33 million PubMed abstracts and nine well-established biomedical repositories. The database is coupled with RESTful APIs and a user-friendly web interface that make it accessible to researchers for various usages, including machine learning using the semantic knowledge in PubMed articles, reducing hallucination of large language models (LLM), and helping experimental scientists explore their data using vast PubMed evidence.

1 Introduction

Knowledge graphs have recently emerged as a powerful data structure to organize biomedical knowledge explicitly, integrating disparate data resources through the representation of nodes and edges. These knowledge graphs adopt widely accepted ontologies to model functional relationships between genomic (e.g., genes, proteins, pathways, and GO terms) and biomedical terms (e.g., drugs and diseases) [1, 2, 3, 4]. For example, the Clinical Knowledge Graph (CKG) integrates proteomics datasets to enhance the clinical decision-making process [5], while Bioteque collects over 150 data sources into an expansive knowledge graph for drug prediction [6]. Another example is Genomic Knowledge Base (GenomicKB), which gathers 347 million genomic terms from over 30 high-quality genomic datasets and annotations to facilitate cross-domain query [7].

In these knowledge graphs, domain knowledge is explicitly represented in a machine-learning (ML) ready format (e.g., nodes, edges, and properties). This structured approach allows for more explainable biomedical AI models, facilitating clear interpretation and traceable decision-making processes, such as reasoning and hypothesis generation [8, 9]. By contrast, PubMed, the largest and richest biomedical knowledge repository exists as free-text [10], and therefore ML cannot be implemented on the explicit knowledge level. While PubMed has been recently utilized to train large-language models (LLMs) that support many powerful AI tools, the captured knowledge is not easily accessible by humans, and these models cannot provide knowledge provenance in practice, and sometimes suffer from hallucination [11]. Some work extracts structured knowledge from PubMed literature as an explicit knowledge graph [12], but it focuses on authors, funding, and topics of the articles, ignoring the actual knowledge between the biomedical terms, limiting their utilities in biomedical knowledge retrieval. So far, there is no explicit knowledge graph extracted from PubMed that supports biomedical ML tasks, and this gap limits the utilities of PubMed literature.

To address the limitation, we present *LiteralGraph*, a unified framework that rigorously extracts biomedical terms and relationships from PubMed literature. The framework rigorously resolves extracted knowledge using a controlled vocabulary collected from standardized OBO ontologies, and deploys a well-defined BioCypher schema. Using this framework, we established *Genomic Literature Knowledge Base* (GLKB), a knowledge graph database that consolidates 263,714,413 biomedical terms, 14,634,427 biomedical relationships, and 10,667,370 genomic events from 33 million PubMed abstracts and nine well-established biomedical repositories. The database is coupled with RESTful APIs and a user-friendly web interface (`https://glkb.dcmb.med.umich.edu/`) that makes it accessible to researchers with different backgrounds. Using semantic embeddings of biomedical terms, GLKB seamlessly connects biomedical knowledge from PubMed literature with various ML tasks such as functional predictions of biomedical terms. In addition, GLKB helps grounding LLMs with biomedical contexts from PubMed articles, and reduces LLM hallucination in retrieval augmented generation (RAG) frameworks [13]. With gene RFX6 in type 2 diabetes as an example, we demonstrate GLKB's utility for experimental scientists to systematically and efficiently explore and explain their experimental data using vast evidence from PubMed literature.

2 Results

2.1 LiteralGraph curates PubMed literature with existing biomedical repositories and ontologies

To connect PubMed with existing structured biomedical repositories, especially these well-established knowledge graphs and ontologies, we developed the LiteralGraph framework that automatically extracts semantic knowledge in accordance with existing biomedical repositories and ontologies. In total, the framework contains three modules (Fig. 1), namely an *ontology resolution module* which integrates biomedical ontologies and produces a controlled vocabulary, a *data repository resolution module* which consolidates heterogeneous biomedical data repositories and produces a curated network, and *a knowledge extraction module* which extracts structured knowledge from PubMed and produces a semantic network. The details of the three modules are described as follows.

The *ontology resolution module* establishes a controlled vocabulary to standardize biomedical terms from different ontologies, which guarantees that the yielded structured knowledge uses the unique identifiers and ontologies in existing biomedical repositories (Fig. 1A). This is to deal with a practical challenge — many biomedical terms are defined repetitively in different ontologies, and the mappings between them are sometimes missing. For example, the term "myeloma" in the vocabulary refers to 0070004 in Disease Ontology, 0001378 in Experimental Factor Ontology, and D009101 in MeSH, but manually curated mappings are only available between the first two sources. To ensure the uniqueness of the vocabulary, the module resolves redundant biomedical terms between ontologies using AML [14] and LogMap [15]. These ontology mappers consider not only the textual similarity between different terms, but also the structural similarity between them. After resolution, input ontologies are merged hierarchically using the Biolink model, making it extendable and reusable [16]. A *controlled vocabulary* is then generated using the merged ontology, and defines the scope of the knowledge graph.

The *data repository resolution module* rigorously integrates biomedical data repositories into a unified curated network in accordance with the controlled vocabulary (Fig. 1B). One of the main challenges in this step is to precisely connect heterogeneous biomedical repositories that use different data schemas and identifiers. For example, although both REACTOME and FamPlex contain relationships between proteins and protein complexes, REACTOME describes the relationships as "component", whereas FamPlex uses the name "partof". Meanwhile, REACTOME and FamPlex use two different sets of identifiers, which makes direct data integration more difficult. To efficiently align these data repositories, LiteralGraph first maps the biomedical terms to the controlled vocabulary, and then transforms the associations between biomedical terms into corresponding relationships in the Biolink model using BioCypher adapters. This *curated network* forms the backbone of the resulting knowledge graph.

The *knowledge extraction module* is the core component of the framework, which extracts structured knowledge from PubMed free-text into a semantic network (Fig. 1C). In this module, three types of semantic knowledge are extracted, namely biomedical entities, biomedical relationships, and biomedical events. Specifically, biomedical entities are first extracted using named entity recognition (NER) models, and the extracted entities are fed into relationship extraction (RE) models to identify the pairwise relationships between them. Meanwhile, relationships between multiple entities are extracted in the form of events, namely event extraction (EE). These events are organized hierarchically, and capture the complicated relationships between different events. For example, the formation of protein complexes usually involves multiple proteins, which cannot be precisely described by pairwise relationships. Finally, the extracted entities are mapped to the controlled vocabulary, and a semantic network is constructed. The *semantic network*, extracted from PubMed literature, enriches the curated network from the data repository resolution module.

2.2 GLKB consolidates diverse biomedical knowledge from biomedical repositories and PubMed articles

Using LiteralGraph framework, we connected PubMed literature and a large collection of well-established biomedical repositories. To maximize the utilities of the consolidated knowledge graph, we configured the three modules in the LiteralGraph framework as follows.

To best leverage the topics of PubMed articles, we fed the MeSH descriptor, a widely accepted article annotation in biomedical research [17], into the ontology resolution module. These descriptors are hierarchically arranged into sixteen primary categories, and the prevalence of MeSH categories inherently reflects their popularity [17]. Among the sixteen MeSH categories, five frequent categories, namely chemicals and drugs, organisms, phenomena and processes, diseases, and anatomy, appear in 28,725,371 PubMed articles, accounting for 86% of articles in PubMed. To further extend the coverage of the biomedical terms, twelve biomedical ontologies from OBO were also fed into the ontology resolution module so that they are integrated into the controlled vocabulary (Fig. 2A). The remaining eleven MeSH categories were adopted directly without further enrichment. In total, the controlled vocabulary has 205,124 biomedical terms, in-

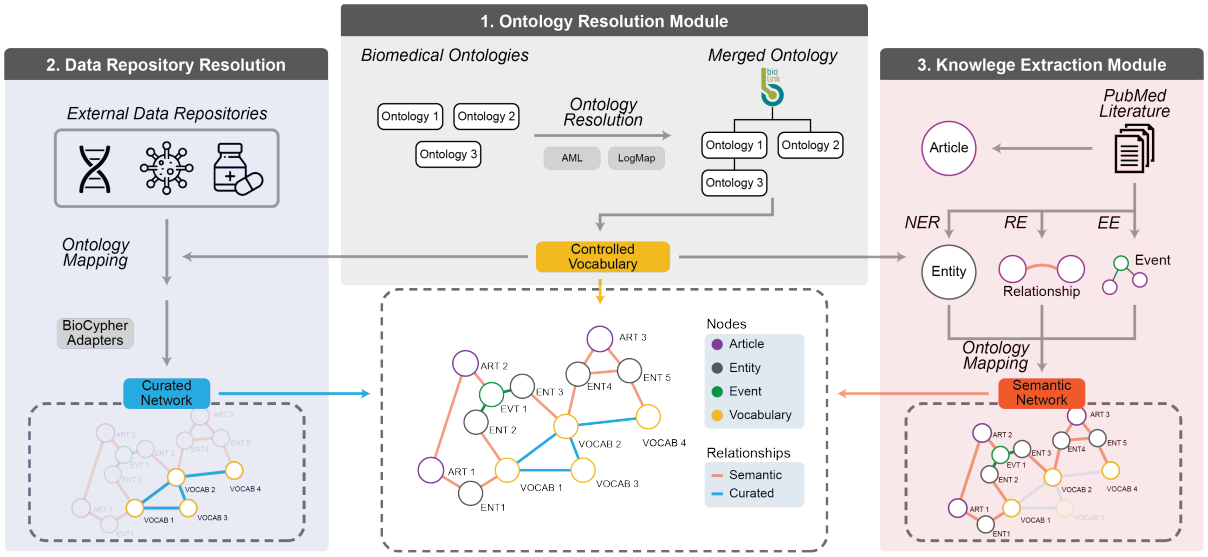


Figure 1: An overview of the LiteralGraph framework. A knowledge graph is constructed using three modules sequentially. **Step 1**. The ontology resolution module integrates biomedical ontologies and produces a controlled vocabulary. **Step 2**. The data repository resolution module consolidates heterogeneous biomedical repositories via BioCypher adaptors and produces a curated network. **Step 3**. The knowledge extraction module extracts structured knowledge from PubMed and produces a semantic network. The semantic network contains biomedical entities extracted by an NER pipeline, biomedical relationships extracted by a RE pipeline, and genomic events extracted by an EE pipeline.

cluding 108,508 from MeSH descriptors and 96,616 from OBO ontologies.

To include reliable domain knowledge, we fed nine manually curated databases into the biomedical repository resolution module (Fig. 2B). These well-established databases connect the five aforementioned categories, forming a curated network that includes both *structural* relationships and *functional* relationships. Structural relationships describe the hierarchical structure of the collected ontologies, while functional relationships include functional associations annotated by domain experts. Specifically, eight types of associations are provided by the incorporated databases, including gene-gene interactions, gene-chemical interactions, chemical-chemical interactions, gene-disease associations, gene-GO associations, gene-pathway associations, tissue-specific gene expressions, and the ontology hierarchy. With the two types of relationships, the yielded curated network provides precise and comprehensive background knowledge about the topics in PubMed. The curated network connects 56,930 biomedical terms using 3,795,248 structural and functional relationships from manually curated databases.

To configure the knowledge extraction module, we let the module extract *three levels* of semantic knowledge from PubMed, including biomedical terms, biomedical relationships, and genomic events. Firstly, biomedical terms were extracted with two NER models, namely PIQN and BERN2 [18, 19], which are proficient in identifying different types of biomedical terms (Method Section 4.3). Combining the two models, the NER pipeline achieves an impressive overall F1 score of 0.99 on the GENIA corpus test set. Next, biomedical relationships were extracted by a state-of-the-art RE model KnowPrompt [20]. The automatically extracted semantic relationships are validated on the BioRED dataset and real-world PubMed data. On the BioRED test set, the relationship extractor achieves state-of-the-art performance, with an overall F1 score of 0.755 [20, 21]. In addition, the number of PubMed articles related to the extracted relationships is significantly higher than that of random biomedical terms, suggesting that the extracted biomedical term pairs are indeed relevant (Fig. 2C). At last, genomic events were extracted with DeepEventMine [22]. These events illustrate the connections between two or more genes and gene products, such as the forming of protein complexes, and the co-regulation of multiple genes. The model performance is evaluated on the test set

of the GENIA event corpus, with an F1 score of 0.69 [23]. In total, the module extracts 99,046,429 terms, 8,527,387 relationships, and 10,523,535 events, which all together constitute the semantic networks.

Cumulatively, the resulting knowledge graph, which was named Genomic Literature Knowledge Graph (GLKB), aligns 55,074,746 nodes and 627,641,483 relationships from biomedical literature and databases, rigorously describing the structural and functional associations among 205,124 biomedical terms of seven different research domains, namely genes and gene products, chemicals and drugs, diseases, anatomy, organisms, pathways, and GO terms.

2.3 GLKB bridges PubMed literature and machine learning models via semantic embeddings

Due to its unstructured nature, semantic knowledge in biomedical literature is under-utilized in biomedical ML tasks. Because GLKB represents the semantic knowledge in PubMed as an explicit graph, the embeddings of the biomedical terms allow an ML approach to access the literature knowledge effectively. Specifically, GLKB systematically encodes three types of biological terms (i.e., genes, chemicals, and diseases) into 1,536-dimensional embeddings using the semantic network (Method Section 4.5). These semantic embeddings capture the functions of the corresponding biomedical terms (Figure 3A). For example, gene embeddings cluster according to their GO terms and KEGG pathways (Figure 3B). We further validated these semantic embeddings with two experiments. In the first experiment, we utilized the embeddings to predict different types of links in the GLKB curated network. If a classifier can be trained to predict curated network edges from semantic embeddings, then we can conclude that these embeddings from the semantic network are faithful. In the second experiment, we used the embeddings to separate real PubMed sentences from randomly selected irrelevant or randomly generated incorrect sentences. If we can build an accurate classifier with the semantic embeddings, we can also conclude that these embeddings are faithful. The details of the two experiments are as follows.

We first investigated whether the semantic embeddings capture real-world biological functions of the biomedical terms, which are represented by the relationships in the GLKB curated network (Figure 3C, Supplementary Figure **??**). Specifically, we trained multiple logistic regression models to predict six types of relationships in the curated network using semantic embeddings. These relationships cover different fields of study and are comprised of different types of biomedical terms, including gene regulatory networks, ligand-receptor pairs, hierarchical relationships between biomedical term pairs, tissue-specific gene expressions, gene-disease associations, and gene-phenotype associations. The semantic embeddings well predicted all real-world biological relationships, indicating that they are consistent with the manually curated knowledge in biomedical databases.

Second, we evaluated whether the embeddings accurately captured semantic relationships in PubMed articles through a sentence classification task, focusing on both relevance and correctness (Figure 3D). We began by sampling 1,246 sentences from PubMed abstracts and collecting the embeddings of biomedical term pairs within these sentences. To test semantic relevance, we randomly selected an additional 1,246 PubMed sentences that did not contain the chosen term pairs. We then attempted to distinguish the related sentences from the random ones using cosine similarity. For semantic correctness, we generated a fake sentence for each sampled sentence by altering the relationships and conclusions using GPT-3.5-turbo, and attempted to classify the real sentences from the fake ones, again using cosine similarity. We expected higher similarity between biomedical embeddings and true sentences compared to the unrelated or fake ones if the embeddings capture precise knowledge from PubMed articles. The embeddings demonstrated an AUROC of 0.99 in classifying real versus irrelevant PubMed sentences, and an AUROC of 0.77 in differentiating real from fake PubMed sentences, outperforming biomedical language model BlueBERT in both tasks (Figure 3E) [24]. These results indicate that the embeddings accurately and precisely capture semantic knowledge from PubMed sentences.

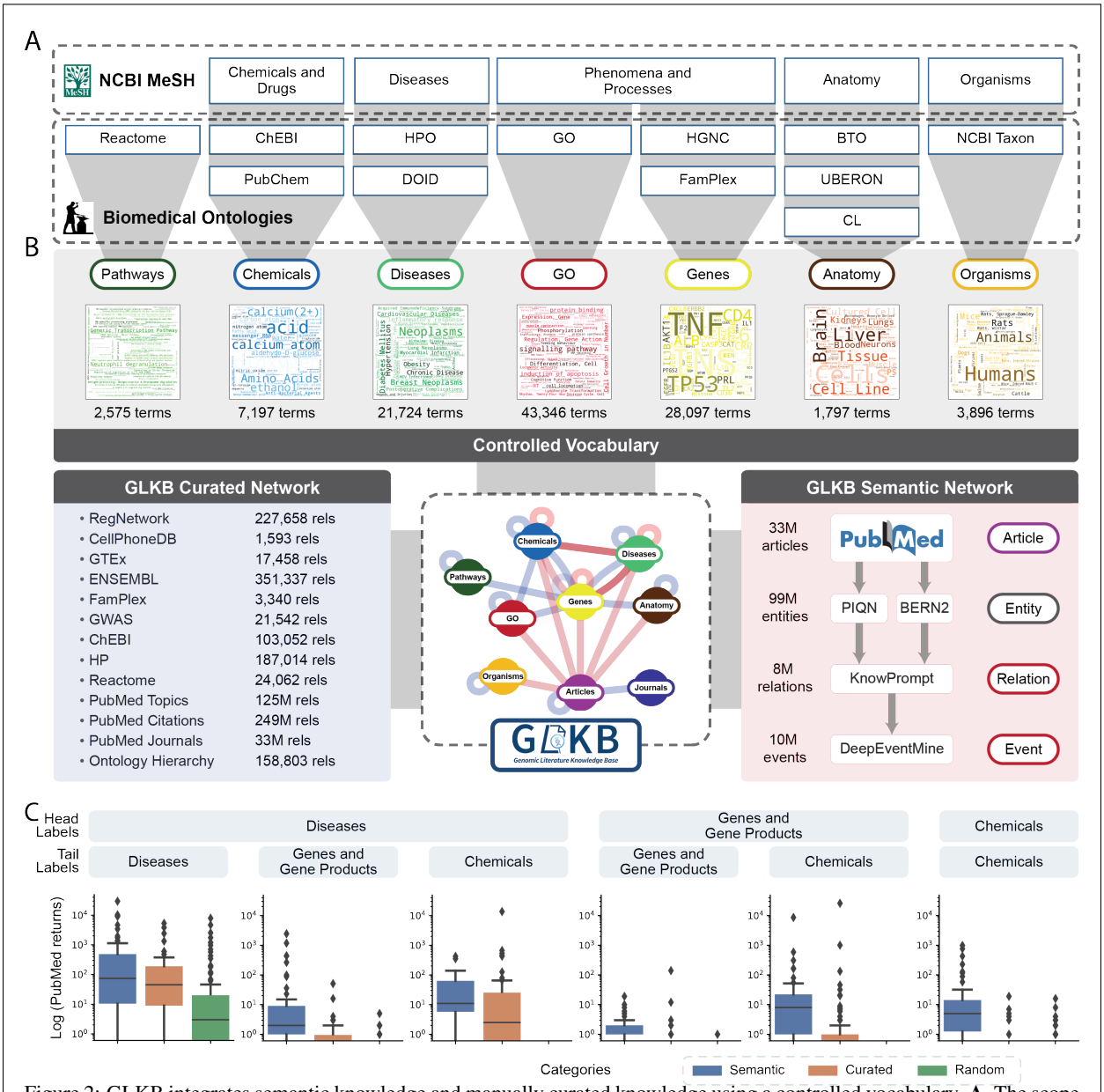


Figure 2: GLKB integrates semantic knowledge and manually curated knowledge using a controlled vocabulary. **A**. The scope of GLKB is defined based on five categories of MeSH descriptors, namely chemicals and drugs, organisms, phenomena and processes, diseases, and anatomy. Twelve ontologies are integrated to provide alternative terms for individual domains. **B**. (Top) The controlled vocabulary containing standardized genomic terms consolidates multiple biomedical ontologies. (Left) The GLKB curated network contains manually curated knowledge annotated by domain experts from biomedical repositories. (Right) Semantic knowledge from PubMed abstracts enriches the relationships between biomedical terms using NLP methods. **C**. The extracted relationships between different types of biomedical terms are validated using relevant publications in PubMed. In all relationship types, the number of PubMed articles related to the extracted relationships is significantly higher than that of randomly sampled irrelevant biomedical terms.

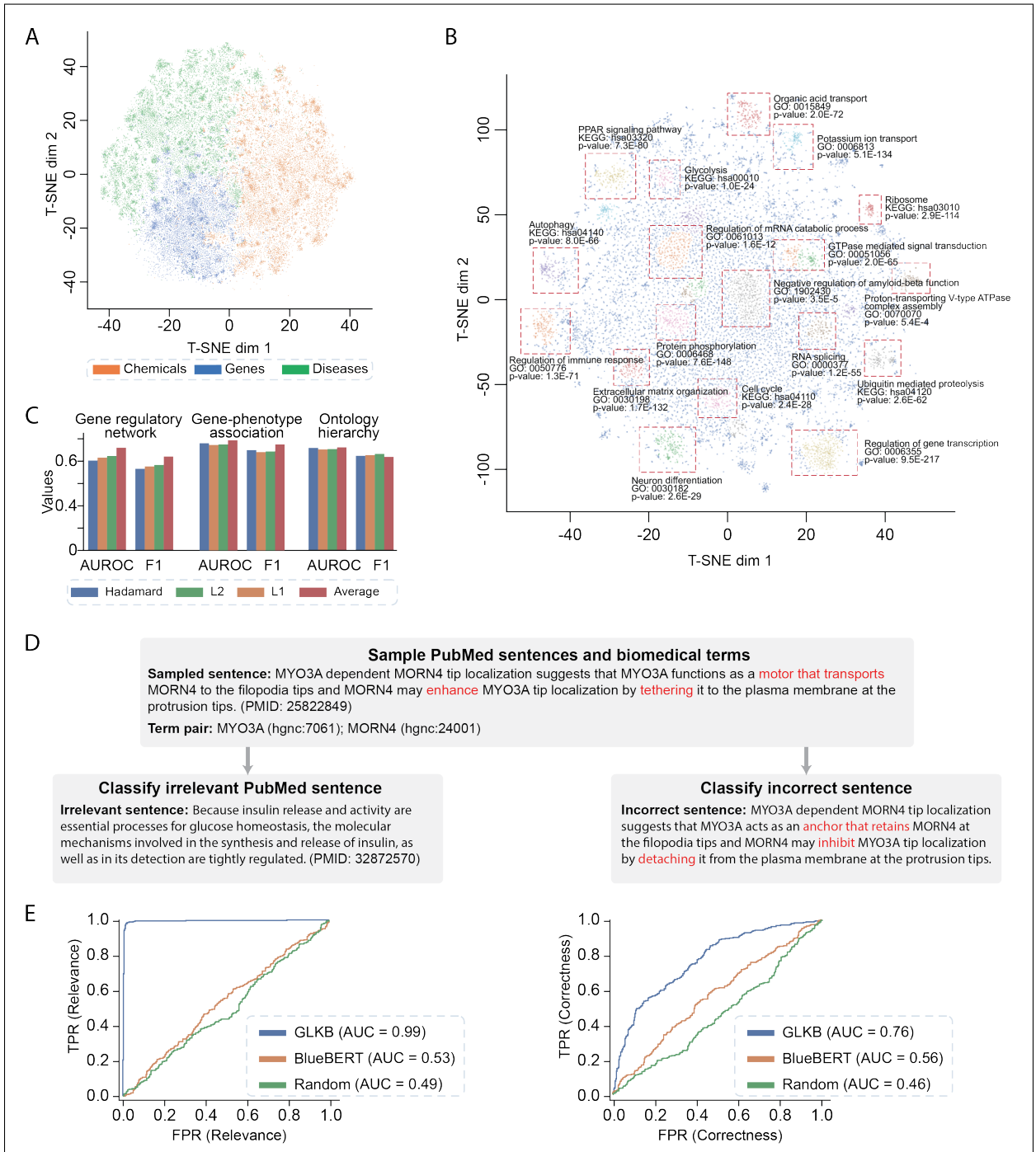


Figure 3: Semantic knowledge agrees with curated knowledge in GLKB. **A**. Semantic embeddings of genomic terms in GLKB visualized by T-SNE colored with genomic term types. **B**. Clusters in gene embeddings are highly associated with the biological functions of corresponding genes. Dot colors indicate gene clusters identified by DBSCAN. **C**. The semantic embeddings effectively predict curated biological relationships in genomic databases. Performance of predicting different types of manually curated knowledge using the semantic embeddings is evaluated by AUC ROCs and F1 scores. Four pooling methods, including hardamard product, L1 norm, L2 norm, and average pooling, are used for prediction. **D**. Examples of the sentence classification task. In this experiment, we use the cosine similarity between biomedical term embeddings and different sentences to identify real relevant PubMed sentences against randomly sampled irrelevant sentences or fake sentences with incorrect knowledge. **E**. The semantic embeddings successfully distinguish real PubMed sentences against irrelevant sentences (left) and fake sentences (right), evaluated by AUC ROCs.

2.4 GLKB reduces LLM hallucination in biomedical domains

Despite the success of LLMs in various domains, they face two significant limitations when applied to biomedical research. Firstly, LLMs sometimes hallucinate in the sense that they generate fabricated results and sources in their answers [11, 25]. Secondly, LLMs require retraining to stay updated with new knowledge [26]. GLKB addresses these limitations by providing biomedical contexts from PubMed literature via retrieval-augmented generation (RAG) frameworks [27, 28]. GLKB is indexed at both sentence and abstract levels, supporting both vector-based and full-text search-based retrieval approaches. This multi-level indexing, together with the graphical structure enables GLKB to integrate seamlessly into popular RAG methods such as GraphRAG [29] in common frameworks, enhancing LLMs' reliability and currency in biomedical research.

GLKB's utility in RAG is first tested in a biomedical question-answering task [30]. To test whether the domain knowledge from GLKB improves the answers, we used GPT-3.5-turbo to generate answers to given questions with or without the biomedical context provided by GLKB. The results were evaluated based on accuracy and F1 score in two biomedical question-answering datasets, PubMedQA and BioASQ, in a zero-shot setting [31, 32]. The two datasets contain 500 and 1,271 research questions, respectively, that can be answered with "yes", "no", or "maybe". In addition to correct answers, PubMedQA also provides the original PubMed articles from which the questions were derived, making it suitable for evaluating GLKB's retrieval performance. On the PubMedQA dataset, GLKB correctly retrieves the corresponding PubMed abstracts as the most relevant results for all 500 test set questions using cosine similarities and significantly improves GPT-3.5-turbo's performance with the retrieved context. Specifically, the accuracy increases from 0.62 to 0.76 after integrating the context provided by GLKB, and the F1 score rises from 0.44 to 0.54 when using the related abstracts and relationships retrieved from GLKB, approaching human performance levels (accuracy of 0.78 and F1 score of 0.72). Similar improvements are observed with the BioASQ dataset. Incorporating related biomedical abstracts and relationships from GLKB enhances the accuracy from 0.80 to 0.83 and the F1 score from 0.70 to 0.78 in answering biomedical questions. These experimental results indicate that GLKB effectively provides essential biomedical domain knowledge for question answering, significantly enhancing the performance of LLMs across various biomedical question-answering tasks.

We further standardized the retrieval and answering process in a question-answering agent on GLKB (GLKB QA agent) (Figure 4). The agent consists of four components, each designed to handle specific tasks. The first component, the planner, decomposes input questions into discrete steps. Each step is addressed using either the semantic knowledge from individual PubMed articles or curated biomedical repositories and ontologies within GLKB. This decomposition is critical, as not all questions can be directly answered by single articles, ensuring that each subtask contributes toward the final solution. The second component, the semantic retriever, extracts detailed semantic knowledge from PubMed articles using GLKB's multi-level indexes to solve subqueries. The third component, the graph retriever, retrieves relevant information from curated biomedical repositories and ontologies stored in GLKB's node and relationship properties. This complements the semantic retriever by enabling access to a broader range of knowledge sources. Finally, the answer generator synthesizes the retrieved knowledge from multiple subqueries to produce well-grounded responses, complete with appropriate references. Together, these components enable the GLKB QA agent to generate interpretable and reliable content for biomedical questions on PubMed articles and the knowledge they contain, effectively reducing LLM hallucinations in biomedical research.

2.5 A use case of GLKB: systematically elucidating RFX6 mechanisms in type 2 diabetes (T2D) from PubMed evidence

Experimental scientists often face the challenge of interpreting experimental data and narrowing their hypotheses for downstream experiments. This requires a scientist to review the vast PubMed literature to select target genes from their data, and this process can be both labor-intensive and incomplete. GLKB systemat-

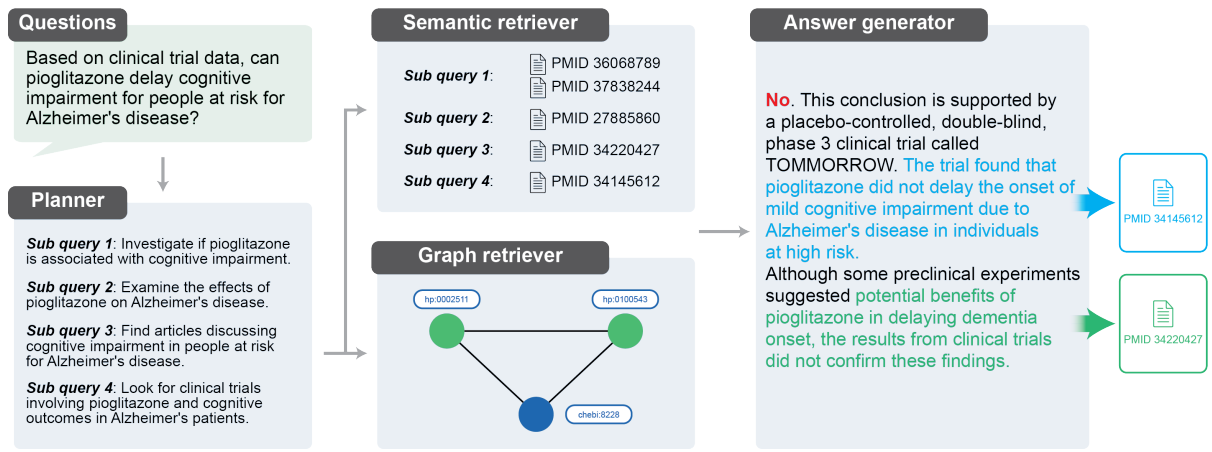


Figure 4: GLKB reduces LLM hallucination in biomedical question answering. An GLKB QA agent is implemented to enhance LLM in biomedical question answering using PubMed articles and biomedical relationships. Utilizing domain knowledge retrieved from GLKB, the LLM agent is able to interpret the result with real-world biomedical evidence and pointers to two PubMed articles.

ically retrieves associations between biomedical terms from PubMed and summarizes domain knowledge. Here, we demonstrate how an experimental scientist can use GLKB to elucidate a gene's function in human disease by leveraging extensive PubMed evidence. For example, *RFX6* has recently been identified as a hub gene for beta cell function, but its relationship with other T2D-related genes is not well understood [33]. We show that a scientist can use GLKB to conveniently identify T2D-related genes associated with *RFX6* and explore their functional groups in specific cell lines.

The whole process can be done in two steps (Figure 5A). In the first step, we screened 2,675 differentially expressed genes (DEGs) from an *RFX6*-knock-down (*RFX6*-KD) dataset [33] with the hope of identifying key genes modulated by *RFX6*. With GLKB, we identified ten significant genes based on strong literature evidence with a co-occurrence test (Method Section 4.6). Next, we explored how these *RFX6*-related genes associate with T2D. Specifically, we tested the co-occurrence of the ten *RFX6*-related genes with the 7,018 highly expressed DEGs from 20 T2D-related gene-KO cell lines using GLKB [34]. This revealed 68 T2D-related genes associated with the ten *RFX6*-related genes, with seven genes overlapping between the two steps. Combining the results, we identified a final set of 72 genes (i.e., the combined set of *RFX6*, *RFX6*-related genes, and T2D-related genes) potentially related to both *RFX6* and T2D (Supplementary File 2). We then investigated whether these identified genes were implicated in T2D phenotypes. In particular, we selected 36 putative T2D causal genes for which the "curated approach probability" is "causal", from 132 effector genes in the Type 2 Diabetes Knowledge Portal (T2DKP) [35] and assessed their overlap with GLKB-identified genes. Over half ($n = 19/36$) of T2DKP "causal" genes were included in our final set, demonstrating the effectiveness of GLKB in identifying relevant biomedical associations.

To further understand how these 72 *RFX6*-related genes function in T2D, we investigated the potential interactions among them. We included the remaining 17 *RFX6*-independent causal genes as controls and hypothesized that these genes are involved in mechanisms independent of *RFX6*. The biological functions of the 89 genes were partially reflected by their co-occurrences in PubMed literature, where genes with similar functions were frequently mentioned together. (Figure 5B). Genes with four types of functions were found neighboring *RFX6*, highlighting different T2D mechanisms. The first group includes four *RFX6*-related genes (*PAX6*, *ISL1*, *MNX1*, and *NKX6-2*) and nine T2D-related genes (*ARX*, *NKX6-1*, *NKX2-2*, *NEUROG3*, *PAX4*, *PDX1*), *HNF4A*), *HNF1A*, and *GLIS3*)), all involved in the development and differentiation of beta cells [36, 37]. The second group contains three genes, including *SLC19A22*, *SLC16A1*, and *SLC5A1*, all

involved in glucose metabolism [38, 39, 40]. The third group consists of three causal genes (*GIP*, *GIPR*, and *GLP1R*) that mediate incretin effects and stimulate insulin secretion [41]. The fourth group contains eight genes, including *ABCC8*, *KCNJ11*, *HNF1B*, *GCK*, *INS*, *WFS1*, *SLC30A8*, and *MTNR1B*, critical for beta cell function and insulin secretion [42]. Especially, mutations in *GCK*, *KCNJ11*, *WFS1*, and *INS* directly lead to maturity-onset diabetes of the young (MODY) [42, 43]. Another group that is less related to *RFX6* was also identified. This group includes six genes (*GATA6*, *CDX2*, *SOX2*, *SOX1*, *PROX1*, and *MAFA*) that regulate stem cell maintenance and early cell development [44, 45]. As expected, the majority of the remaining T2D causal genes ($n = 15/17$) were not included in the closest gene groups of *RFX6*. The observed gene-gene associations suggest that *RFX6* is more likely to influence insulin secretion and beta cell development. It is worth noting that the co-occurrence of different genes in the literature does not necessarily indicate similar biological functions but is influenced by other factors, such as their popularity in the scientific community or, at times, opposing biological functions. As a result, only the "early development" and "incretin effect" groups form statistically significant in DBSCAN clusters. Combined with human curation, we believe these associations and functional groups remain useful for scientists to refine their hypotheses for further experiments.

2.6 GLKB is conveniently accessible via web interface and APIs

GLKB offers two access options tailored to different user needs: a user-friendly web interface and RESTful command-line APIs. The web interface allows users to visualize PubMed literature within specific research domains, while the APIs enable users with coding expertise to retrieve and analyze PubMed literature at scale.

The GLKB web interface organizes knowledge from PubMed literature in research domains defined by input triplets. Each triplet includes one or two biomedical terms and an optional relationship between them. For example, if a researcher is interested in the association between the gene *HNF4A* and T2D, and also considers *TGFB1* as a potentially relevant gene, the research domain can be defined with three triplets: (1) (HNF4A)-[association]-(T2D), (2) (HNF4A)-[any relationship]-(TGFB1), and (3) (T2D)-[any relationship]-(TGFB1). The most relevant biomedical terms and PubMed articles are then displayed in a subgraph. The interface also provides detailed information on these nodes and edges, including their names, synonyms, definitions, and external identifiers, along with key PubMed articles. For relationships, the interface shows their types, sources, and associated PubMed articles. Users can export their results as spreadsheets, which include comprehensive details on the biomedical terms and their connections.

In addition to the web interface, GLKB offers a RESTful command-line API for large-scale literature retrieval and advanced graph analysis. This API allows researchers to systematically retrieve details of nodes and relationships, as well as subgraphs in GLKB, using various properties. Nodes and relationships can be queried by their names, synonyms, or identifiers from external biomedical repositories. Researchers can customize and extend their GLKB subgraphs using the APIs, which support various graph generation algorithms, including the k-shortest path algorithm [46], the hypergeometric test [47], and the Louvain community detection method [48]. These algorithms automatically identify related biomedical terms and relationships, enabling researchers to gather relevant knowledge without needing to know all the terms in a given research field.

3 Discussion

We developed a principled computational framework LiteralGraph which resolves existing ontologies, connects existing biomedical data repositories, extracts knowledge from PubMed, and produces a consolidated knowledge graph to capture the comprehensive knowledge explicitly. To our best knowledge, it is the first time that PubMed is systematically connected to existing data repositories, ontologies, and knowledge graphs at the detailed knowledge level. Because knowledge is explicitly represented in a machine-learning ready format, these nodes, edges, and properties can be directly fed into computers for advanced AI-powered

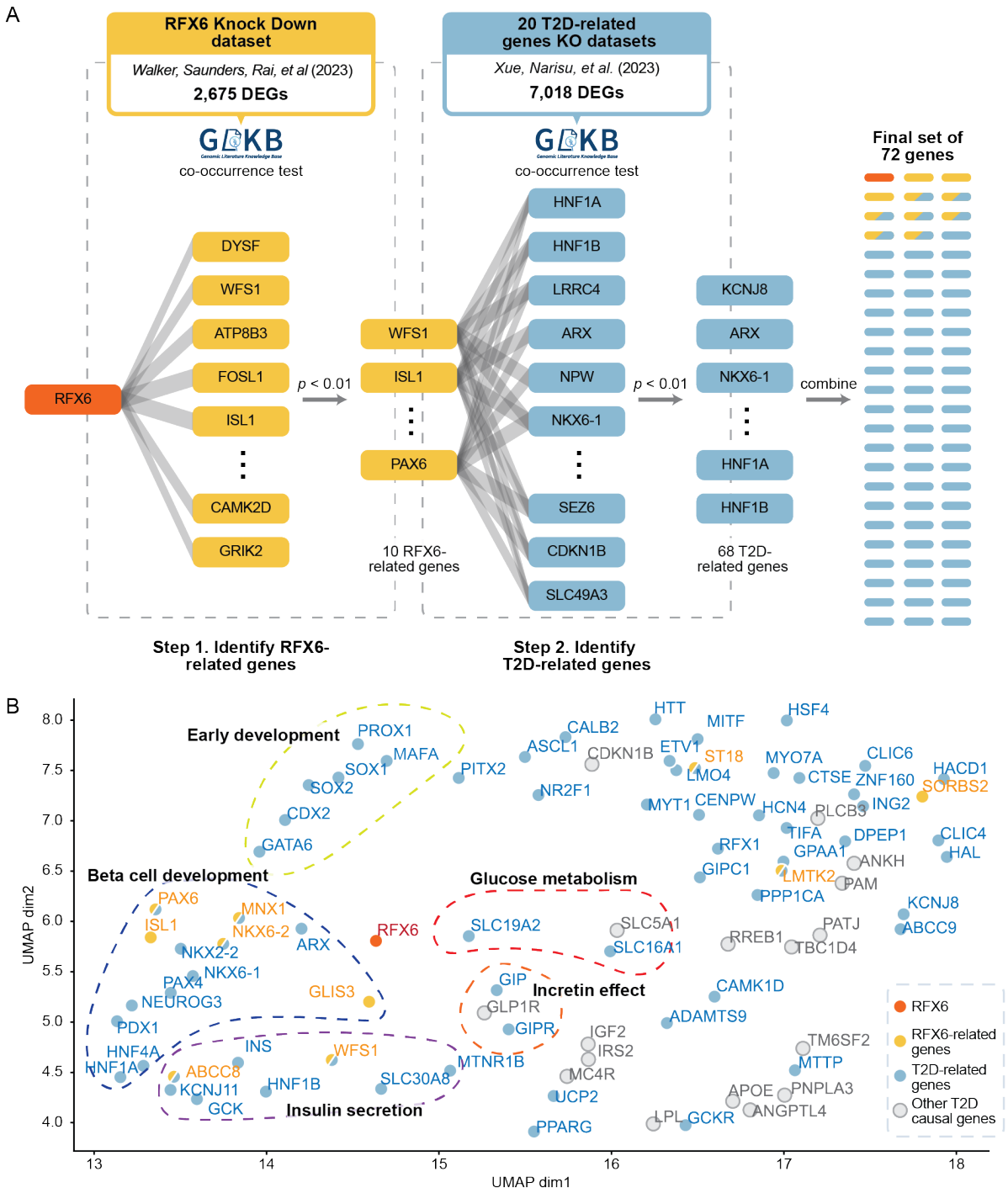


Figure 5: GLKB identified *RFX6*-related genes from PubMed articles and revealed their potential relationships to T2D. **A**. 10 *RFX6*-related genes (in yellow) were identified by GLKB from 2,675 DEGs in the *RFX6*-KD cell line, and 68 T2D-related genes were further found (in blue), including seven overlapped with the ten *RFX6*-related genes, associated with these *RFX6*-related genes from 7,018 DEGs in 20 KO cell lines where T2D-related genes were knocked out. In total, a combined final set of 72 genes related to the *RFX6*-T2D mechanisms was selected using GLKB. **B**. GLKB reflects the underlying mechanisms of the *RFX6*-related genes to T2D. The 72-gene final set is combined with the remaining 17 T2D causal genes from the Type 2 Diabetes Knowledge Portal (in grey), and visualized according to their co-occurrence in PubMed articles. Genes with different biological functions are located accordingly, and the distances between different genes indicate the strength of associations between them. According to the result, *RFX6* is associated with beta cell functions (e.g., insulin secretion and glucose metabolism) and beta cell development.

downstream applications, such as reasoning and hypothesis generation [8, 9]. Because of the interface with large-language models, GLKB allows LLMs to reduce hallucination and provides real facts, knowledge, and references in the process of generating text.

When we design and implement LiteralGraph, we aim to meet the FAIR principles, namely findable, accessible, interoperable, and reusable, benefiting the research community in two ways [49]. First, we followed a well-defined schema based on BioCypher, which allows easy adaptation to different research domains and connecting with external data repositories and knowledgebases [50]. Second, the framework utilizes a modular architecture, which supports convenient data reusing and pipeline updating. The framework is dynamic and automatic in the sense that all input data and computational models are described via JSON files and the pipelines are shared with the community. When any component needs to be updated in the future, we only have to update the JSON files and rerun the corresponding modules.

Errors and noise in the GLKB semantic network can be self-controlled and refined. For example, one effective approach to refining relationships in the semantic network is to summarize relationships from all relevant sentences. Unlike conventional methods that extract semantic relationships from a single sentence, summarizing a consensus relationship from all relevant sentences provides a more comprehensive understanding of the scientific consensus regarding a specific relationship and presumably contains less bias. GLKB is particularly valuable for such approaches because it includes multi-level embeddings of abstracts, sentences, and biomedical terms. This structure provides machine learning models with a complete context of specific biomedical domains, enabling them to generate more precise results. Utilizing the rich, multi-level context, GLKB has the potential of self-enrichment and self-correctness via advanced LLM agents [51].

In this paper, we tested fundamental use cases of GLKB, which can be further improved. In the LLM integration task, GLKB, as both a PubMed vector store and a graphical database, can seamlessly connect with LLM agents in multiple ways, such as generation-augmented retrieval (GAR) [52], hybrid retrieval [53], chain-of-thoughts reasoning [54], and tree-of-thoughts reasoning [55]. These strategies enhance LLM performance in more complex tasks such as clinical question answering, where the query typically involves clinical records and requires reasoning from multiple articles to generate the final answer. GLKB also supports more advanced hypothesis generation frameworks. For instance, instead of merely identifying significant biomedical term pairs, machine learning agents like reinforcement learning (RL) models can be employed to generate more comprehensive hypotheses based on multi-hop trajectories in GLKB [56]. By employing a structure that handles both structured and unstructured data, we believe that GLKB can become an even more valuable resource for machine learning models to utilize the knowledge contained in PubMed articles.

4 Methods

4.1 System design

Software systems: GLKB was designed as a property graph that supports heterogeneous nodes, including articles, entities, and genomic terms, as well as edges such as citation, hierarchical relations, and gene-gene regulations, each with properties. To implement the database, we chose the *Neo4j* graph management system [57], which allows us to efficiently implement the property graph model directly down to the storage level. Additionally, the system offers an efficient query language, Cypher [58], which supports constant-time traversals for both depth and breadth queries over data graphs. With these features, GLKB provides a powerful tool for querying and analyzing complex genomic relationships.

Database update: GLKB tracks the latest releases of all input data resources and updates accordingly every twelve months to coincide with the PubMed annual update (current version release 2023).

Web interface: The GLKB database is hosted on a web server at the Michigan Academic Computing Center (MACC) at the University of Michigan, providing web-based access to users. The system backend

is implemented using the Flask web framework and the socket library, with the graph database powered by Neo4j. To communicate with the Neo4j database, the server program uses the Neo4j Python driver. For interactive network visualization of the graph database, we used the *Cytoscape* framework in the frontend [59], while the user interface is built using the *React* framework. The NGINX web server acts as the reverse proxy, routing internet requests to our server, and no user cookies are collected. To support concurrent access from multiple users, the GLKB web server is designed with multi-threading and subjected to stress testing to ensure reliable service.

4.2 Data sources

Data retrieval: The articles in GLKB are obtained in XML format from the PubMed database. From each article, we extract various information such as the PubMed ID, title, abstract, keywords, authors, publication date, ISSNs of the published journal, citations, and publication types. Ontologies in GLKB are retrieved from NCBO BioPortal in CSV format, and for each genomic term, we extract its name, synonyms, URI, source ontology, MeSH tree location, cross-references, and NCBO LOOM mappings. Journals in GLKB are retrieved from the Medline database in TXT format, and we extract details such as JrID, NLM ID, title, NLM abbreviation, ISO abbreviation, online ISSNs, and printed ISSNs. Genomic databases in GLKB are obtained from GenomicKB or their respective data portals and web servers in various formats. We implemented a Python pipeline to automatically convert the raw tabular data to CSV format, which can be easily loaded into the Neo4j data loader [7].

Data preprocessing: To prepare the PubMed abstracts for entity and relation extraction, we performed two preprocessing steps: sentence splitting and coreference resolution. First, we identified sentences in the original abstracts using the NLTK sentence splitter [60]. Next, we applied the NeuralCoref 4.0 model in spaCy [61, 62] to identify and resolve coreferences in the separated sentences. Coreferences are tokens that refer to other tokens or sets of tokens (e.g., "they", "it", and "their") in the same or another sentence.

Vocabulary construction: To construct the controlled vocabulary of GLKB, biomedical ontologies are first downloaded from the OBO foundry using the Python package PyOBO 0.10.6 [63]. We resolved redundant gene, chemical, and disease terms in the vocabulary. For genes, we used the NCBI gene dataset as a reference, revealing 28,493 general gene matches, 14,406 pseudogene matches, and 767 other matches from a total of 43,794 gene entries, with only 128 entries requiring manual alignment. The alignment of the NCBI gene dataset with the MeshTerm dataset revealed 117 gene matches and 3 pseudogene matches, indicating some contamination within the MeshTerm dataset. These multi-referencing issues primarily arose from synonyms or outdated gene symbols, with mismatched terms including ontology terms which should be relocated to an ontology vocabulary. For chemicals, a direct alignment in the MeshTerm subset successfully identified 7,846 duplicated entries. For diseases, all entries were matched with the UMLS system using the provided [64], identifying 21,694 duplicated entries, which were remapped to properly represent the connected concept. This initiative ensures that gene, chemical, and disease terms in GLKB are accurately linked and updated, enhancing its utility and reliability for bioinformatics research.

4.3 Information retrieval

Named entity recognition: To extract entities from PubMed abstracts, we employed two named entity recognition (NER) models: a pretrained BERN2 model and a fine-tuned Parallel Instance Query Network (PIQN) model [19, 18]. The PIQN model, which we fine-tuned on the GENIA term annotation corpus, extracted five types of biomedical entities, including genes and gene products, diseases, chemicals and drugs, organisms, and anatomical structures [65]. To assess entity extraction confidence, another pipeline, BERN2, is also utilized to extract the same types of biomedical entities. Entities extracted by both PIQN and BERN2 are treated as highly confident entities and the others as less confident entities. Finally, we utilized the Gilda normalizer to map the extracted entities to the GLKB's controlled vocabulary [66]. To ensure correct entity mapping, we set up a constraint that the entity and the genomic term it mapped to must have

the same label. For example, an entity labeled "ER" would be mapped to the term "endoplasmic reticulum" instead of "erdosteine" if it was recognized as an anatomical structure by the entity extractor.

Relationship extraction: GLKB utilized Knowprompt with a SciBERT language model to extract semantic relationships from PubMed abstracts [20, 67]. The pretrained language model is fine-tuned on the BioRED corpus, which provides eight types of annotated relationships, including positive and negative correlations, associations, bindings, drug interactions, cotreatments, comparisons, and conversions [21].

Event extraction: We utilized a Deep Event Mine (DEM) model to extract genomic events that involve multiple entities in the text [22]. The pretrained DEM model was fine-tuned on the GENIA event annotation corpus [23]. Before fine-tuning, we selected nine genomic event labels from the corpus, including "Gene expression", "Transcription", "Protein catabolism", "Phosphorylation", "Localization", "Binding", "Regulation", "Positive regulation", and "Negative regulation". We also selected three relation labels, namely "theme of", "cause of", and "site", based on their number of training examples in the corpus.

4.4 Database integration

We selected nine biomedical databases that contained diverse biomedical knowledge, including gene-gene interactions, gene-chemical interactions, chemical-chemical interactions, gene-disease associations, and tissue-specific gene expression. These databases were integrated into GLKB as relationships between biomedical terms in the controlled vocabulary. We initially mapped the entities from the original data to GLKB's controlled vocabulary based on their external identifiers. Next, we defined a relationship schema and categorized the relationships from the biomedical databases based on the types of biomedical terms they connected to (Supplementary File 3). By doing so, we translated knowledge from various biomedical databases into well-defined relationships in GLKB, allowing for more comprehensive and accurate biomedical literature analysis.

4.5 GLKB embedding

Biomedical term embedding: We utilized the Heterogeneous Graph Transformer (HGT) to generate 1,536-dimensional embeddings of biomedical terms [68]. Due to computational constraints, we were compelled to produce embeddings only for 87,624 biomedical terms including genes, chemicals, and diseases. Firstly, we generated the input feature from the top 20 abstracts and the definition of each biomedical term using the OpenAI text-embedding-3-small model. The semantic network containing the 87,624 terms was then fed into a two-layer HGT model and trained on a link prediction task that predicts the co-occurrence of these terms. The final biomedical term embeddings are generated by the trained model.

Abstract and sentence embedding: Semantic embeddings of 33,403,054 PubMed abstracts and 108,536,203 sentences were generated using the gte-large-en-v1.5 model [69]. Due to computational constraints, a subset of sentences was selected for embedding. Specifically, we trained a spaCy sentence classification pipeline on the PubMed 200K RCT dataset, selecting only sentences labeled as "Background", "Results" or "Conclusion" for embedding [61, 70].

4.6 Pearson's chi-square test

Pearson's chi-square tests are performed to identify significant associations between biomedical terms. Specifically, for two given terms T_i, T_j, their expected co-occurrence frequency is

$$E(T_i, T_j) = \frac{N(T_i) * N(T_j)}{N(articles)},$$

where $N(T_i)$ and $N(T_j)$ are the relevant articles of T_i and T_j respectively, and $N(articles)$ is the total number of articles in GLKB. The test statistic is then calculated from the co-occurrence contingency table of the two terms. Clusters of the biomedical term co-occurrence are calculated by agglomerative clustering in python.

References

[1] A. García, A. L. Palacio, J. F. R. Román, J. C. Casamayor, and O. Pastor, "Towards the understanding of the human genome: a holistic conceptual modeling approach," *IEEE Access*, vol. 8, pp. 197111–197123, 2020.

[2] B. Walsh, S. K. Mohamed, and V. Nováček, "Biokg: A knowledge graph for relational learning on biological data," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 3173–3180, 2020.

[3] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *Elife*, vol. 6, p. e26726, 2017.

[4] P. Chandak, K. Huang, and M. Zitnik, "Building a knowledge graph to enable precision medicine," *Nature Scientific Data*, 2023.

[5] A. Santos, A. R. Colaço, A. B. Nielsen, L. Niu, M. Strauss, P. E. Geyer, F. Coscia, N. J. W. Albrechtsen, F. Mundt, L. J. Jensen, *et al.*, "A knowledge graph to interpret clinical proteomics data," *Nature biotechnology*, vol. 40, no. 5, pp. 692–702, 2022.

[6] A. Fernández-Torras, M. Duran-Frigola, M. Bertoni, M. Locatelli, and P. Aloy, "Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque," *Nature Communications*, vol. 13, no. 1, p. 5304, 2022.

[7] F. Feng, F. Tang, Y. Gao, D. Zhu, T. Li, S. Yang, Y. Yao, Y. Huang, and J. Liu, "GenomicKB: a knowledge graph for the human genome," *Nucleic Acids Research*, vol. 51, no. D1, pp. D950–D956, 2023.

[8] Y. Liu, M. Hildebrandt, M. Joblin, M. Ringsquandl, R. Raissouni, and V. Tresp, "Neural multi-hop reasoning with logical rules on biomedical knowledge graphs," in *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*, pp. 375–391, Springer, 2021.

[9] J. Sybrandt, I. Tyagin, M. Shtutman, and I. Safro, "Agatha: Automatic graph mining and transformer based hypothesis generation approach," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2757–2764, 2020.

[10] J. White, "Pubmed 2.0," *Medical Reference Services Quarterly*, vol. 39, no. 4, pp. 382–387, 2020.

[11] J. Gravel, M. D'Amours-Gravel, and E. Osmanlliu, "Learning to fake it: limited responses and fabricated references provided by chatgpt for medical questions," *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 3, pp. 226–234, 2023.

[12] J. Xu, S. Kim, M. Song, M. Jeong, D. Kim, J. Kang, J. F. Rousseau, X. Li, W. Xu, V. I. Torvik, *et al.*, "Building a pubmed knowledge graph," *Scientific data*, vol. 7, no. 1, p. 205, 2020.

[13] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[14] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, "The agreementmakerlight ontology matching system," in *On the Move to Meaningful Internet Systems: OTM 2013 Conferences: Confederated International Conferences: CoopIS, DOA-Trusted Cloud, and ODBASE 2013, Graz, Austria, September 9-13, 2013. Proceedings*, pp. 527–541, Springer, 2013.

[15] E. Jiménez-Ruiz and B. Cuenca Grau, "Logmap: Logic-based and scalable ontology matching," in *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*, pp. 273–288, Springer, 2011.

[16] D. R. Unni, S. A. Moxon, M. Bada, M. Brush, R. Bruskiewich, J. H. Caufield, P. A. Clemons, V. Dancik, M. Dumontier, K. Fecho, *et al.*, "Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science," *Clinical and translational science*, vol. 15, no. 8, pp. 1848–1855, 2022.

[17] C. E. Lipscomb, "Medical subject headings (mesh)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, p. 265, 2000.

[18] Y. Shen, X. Wang, Z. Tan, G. Xu, P. Xie, F. Huang, W. Lu, and Y. Zhuang, "Parallel instance query network for named entity recognition," *arXiv preprint arXiv:2203.10545*, 2022.

[19] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "Bern2: an advanced neural biomedical named entity recognition and normalization tool," *arXiv preprint arXiv:2201.02080*, 2022.

[20] X. Chen, N. Zhang, X. Xie, S. Deng, Y. Yao, C. Tan, F. Huang, L. Si, and H. Chen, "Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction," in *Proceedings of the ACM Web Conference 2022*, pp. 2778–2788, 2022.

[21] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, and Z. Lu, "Biored: a rich biomedical relation extraction dataset," *Briefings in Bioinformatics*, vol. 23, no. 5, p. bbac282, 2022.

[22] H.-L. Trieu, T. T. Tran, K. N. Duong, A. Nguyen, M. Miwa, and S. Ananiadou, "Deepeventmine: end-to-end neural nested event extraction from biomedical texts," *Bioinformatics*, vol. 36, no. 19, pp. 4910–4917, 2020.

[23] J.-D. Kim, N. Nguyen, Y. Wang, J. Tsujii, T. Takagi, and A. Yonezawa, "The genia event and protein coreference tasks of the bionlp shared task 2011," in *BMC bioinformatics*, vol. 13, pp. 1–12, 2012.

[24] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets," *arXiv preprint arXiv:1906.05474*, 2019.

[25] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[26] R. T. McCoy, E. Pavlick, and T. Linzen, "Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference," *arXiv preprint arXiv:1902.01007*, 2019.

[27] Z. Hu, Y. Xu, W. Yu, S. Wang, Z. Yang, C. Zhu, K.-W. Chang, and Y. Sun, "Empowering language models with knowledge graph reasoning for open-domain question answering," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9562–9581, 2022.

[28] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *arXiv preprint arXiv:2306.08302*, 2023.

[29] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, and J. Larson, "From local to global: A graph rag approach to query-focused summarization," *arXiv preprint arXiv:2404.16130*, 2024.

[30] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu, "Biomedical question answering: a survey of approaches and challenges," *ACM Computing Surveys (CSUR)*, vol. 55, no. 2, pp. 1–36, 2022.

[31] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[32] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, *et al.*, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC bioinformatics*, vol. 16, no. 1, pp. 1–28, 2015.

[33] J. T. Walker, D. C. Saunders, V. Rai, H.-H. Chen, P. Orchard, C. Dai, Y. D. Pettway, A. L. Hopkirk, C. V. Reihsmann, Y. Tao, *et al.*, "Genetic risk converges on regulatory networks mediating early type 2 diabetes.," *Nature*, 2023.

[34] D. Xue, N. Narisu, D. L. Taylor, M. Zhang, C. Grenko, H. J. Taylor, T. Yan, X. Tang, N. Sinha, J. Zhu, *et al.*, "Functional interrogation of twenty type 2 diabetes-associated genes using isogenic human embryonic stem cell-derived β-like cells," *Cell Metabolism*, vol. 35, no. 11, pp. 1897–1914, 2023.

[35] M. C. Costanzo, M. von Grotthuss, J. Massung, D. Jang, L. Caulkins, R. Koesterer, C. Gilbert, R. P. Welch, P. Kudtarkar, Q. Hoang, *et al.*, "The type 2 diabetes knowledge portal: An open access genetic resource dedicated to type 2 diabetes and related traits," *Cell metabolism*, vol. 35, no. 4, pp. 695–710, 2023.

[36] Y. Gosmain, L. S. Katz, M. H. Masson, C. Cheyssac, C. Poisson, and J. Philippe, "Pax6 is crucial for β-cell function, insulin biosynthesis, and glucose-induced insulin secretion," *Molecular endocrinology*, vol. 26, no. 4, pp. 696–709, 2012.

[37] X. Wang, M. Sterr, I. Burtscher, A. Böttcher, J. Beckenbauer, J. Siehler, T. Meitinger, H.-U. Häring, H. Staiger, F. M. Cernilogar, *et al.*, "Point mutations in the pdx1 transactivation domain impair human β-cell development and function," *Molecular metabolism*, vol. 24, pp. 80–97, 2019.

[38] L. Han, Q. Qu, D. Aydin, O. Panova, M. J. Robertson, Y. Xu, R. O. Dror, G. Skiniotis, and L. Feng, "Structure and mechanism of the sglt family of glucose transporters," *Nature*, vol. 601, no. 7892, pp. 274–279, 2022.

[39] M. Xue, M. Song, D. Yan, S. Sun, Y. Wang, T. Fu, H. Cai, H. Xu, G. Sun, K. Wang, *et al.*, "Effect of slc16a1 on hepatic glucose metabolism in newborn and post-weaned holstein bulls," *Frontiers in Genetics*, vol. 13, p. 811849, 2022.

[40] V. M. Veetil, D. Pachat, K. Nikitha, and J. M. Kutty, "Thiamine-responsive megaloblastic anaemia.," *The National medical journal of India*, vol. 36, no. 5, 2023.

[41] Y. Seino, M. Fukushima, and D. Yabe, "Gip and glp-1, the two incretin hormones: similarities and differences," *Journal of diabetes investigation*, vol. 1, no. 1-2, pp. 8–23, 2010.

[42] D. K. Demirci, F. Darendeliler, S. Poyrazoglu, A. D. K. Al, N. Gul, Y. Tutuncu, G. Gulfidan, K. Y. Arga, C. Cacina, O. Ozturk, *et al.*, "Monogenic childhood diabetes: dissecting clinical heterogeneity by next-generation sequencing in maturity-onset diabetes of the young," *OMICS: A Journal of Integrative Biology*, vol. 25, no. 7, pp. 431–449, 2021.

[43] L. K. Billings, Z. Shi, W. K. Resurreccion, C.-H. Wang, J. Wei, T. I. Pollin, M. S. Udler, and J. Xu, "Statistical evidence for high-penetrance mody-causing genes in a large population-based cohort," *Endocrinology, diabetes & metabolism*, vol. 5, no. 6, p. e372, 2022.

[44] K. M. Goudarzi, J. A. Espinoza, M. Guo, J. Bartek, M. Nistér, M. S. Lindström, and D. Hägerstrand, "Reduced expression of prox1 transitions glioblastoma cells into a mesenchymal gene expression subtype," *Cancer research*, vol. 78, no. 20, pp. 5901–5916, 2018.

[45] C. Y. Chia, P. Madrigal, S. L. Denil, I. Martinez, J. Garcia-Bernardo, R. El-Khairi, M. Chhatriwala, M. H. Shepherd, A. T. Hattersley, N. R. Dunn, *et al.*, "Gata6 cooperates with eomes/smad2/3 to deploy the gene regulatory network governing human definitive endoderm and pancreas formation," *Stem cell reports*, vol. 12, no. 1, pp. 57–70, 2019.

[46] J. Y. Yen, "An algorithm for finding shortest routes from all source nodes to a given destination in general networks," *Quarterly of applied mathematics*, vol. 27, no. 4, pp. 526–530, 1970.

[47] W. T. Adams and T. R. Skopek, "Statistical test for the comparison of samples from mutational spectra," *Journal of molecular biology*, vol. 194, no. 3, pp. 391–396, 1987.

[48] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[49] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[50] S. Lobentanzer, P. Aloy, J. Baumbach, B. Bohar, V. J. Carey, P. Charoentong, K. Danhauser, T. Doğan, J. Dreo, I. Dunham, *et al.*, "Democratizing knowledge representation with biocypher," *Nature Biotechnology*, pp. 1–4, 2023.

[51] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen, "Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph," *arXiv preprint arXiv:2402.11163*, 2024.

[52] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, "Generation-augmented retrieval for open-domain question answering," *arXiv preprint arXiv:2009.08553*, 2020.

[53] D. Sanmartin, "Kg-rag: Bridging the gap between knowledge and creativity," *arXiv preprint arXiv:2405.12035*, 2024.

[54] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions," *arXiv preprint arXiv:2212.10509*, 2022.

[55] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models, 2023," *URL https://arxiv. org/pdf/2305.10601. pdf*, 2023.

[56] W. Xiong, T. Hoang, and W. Y. Wang, "Deeppath: A reinforcement learning method for knowledge graph reasoning," *arXiv preprint arXiv:1707.06690*, 2017.

[57] J. Webber, "A programmatic introduction to neo4j," in *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pp. 217–218, 2012.

[58] N. Francis, A. Green, P. Guagliardo, L. Libkin, T. Lindaaker, V. Marsault, S. Plantikow, M. Rydberg, P. Selmer, and A. Taylor, "Cypher: An evolving query language for property graphs," in *Proceedings of the 2018 international conference on management of data*, pp. 1433–1445, 2018.

[59] M. E. Smoot, K. Ono, J. Ruscheinski, P.-L. Wang, and T. Ideker, "Cytoscape 2.8: new features for data integration and network visualization," *Bioinformatics*, vol. 27, no. 3, pp. 431–432, 2011.

[60] E. Loper and S. Bird, "Nltk: The natural language toolkit," *arXiv preprint cs/0205028*, 2002.

[61] M. Honnibal and I. Montani, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *To appear*, vol. 7, no. 1, pp. 411–420, 2017.

[62] M. Neumann, D. King, I. Beltagy, and W. Ammar, "Scispacy: Fast and robust models for biomedical natural language processing," *arXiv preprint arXiv:1902.07669*, 2019.

[63] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, *et al.*, "The obo foundry: coordinated evolution of ontologies to support biomedical data integration," *Nature biotechnology*, vol. 25, no. 11, pp. 1251–1255, 2007.

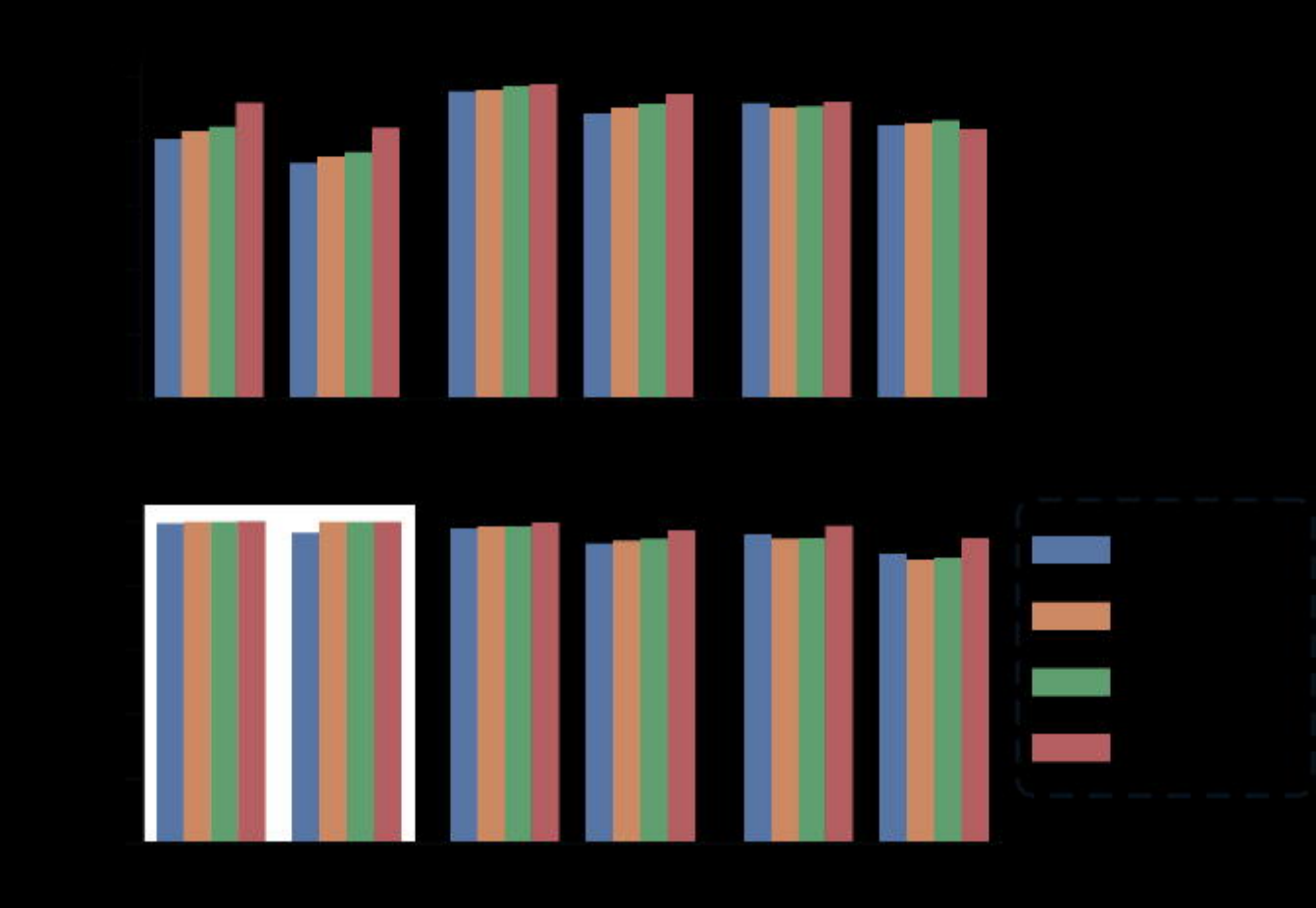
[64] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, Jan. 2004.

[65] J.-D. Kim, T. Ohta, and J. Tsujii, "Corpus annotation for mining biomedical events from literature," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–25, 2008.

[66] B. M. Gyori, C. T. Hoyt, and A. Steppi, "Gilda: biomedical entity text normalization with machine-learned disambiguation as a service," *Bioinformatics Advances*, vol. 2, no. 1, p. vbac034, 2022.

[67] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," *arXiv preprint arXiv:1903.10676*, 2019.

[68] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of the web conference 2020*, pp. 2704–2710, 2020.

[69] Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang, "Towards general text embeddings with multi-stage contrastive learning," *arXiv preprint arXiv:2308.03281*, 2023.

[70] F. Dernoncourt and J. Y. Lee, "Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts," *arXiv preprint arXiv:1710.06071*, 2017.