BMC Bioinformatics

DATABASE

Open Access

VAIV bio-discovery service using transformer model and retrieval augmented generation

Seonho Kim[1*] and Juntae Yoon[2*]

*Correspondence:
shkim.lex@gmail.com;
jtyoon@vaiv.kr

[1] Department of Computer Science, Sogang University, 35, Baekbeom-Ro, Mapo-Gu, Seoul, Korea
[2] VAIV Company Inc, 97, Dokseodang-Ro, Yongsan-Gu, Seoul, Korea

Abstract

Background: There has been a considerable advancement in AI technologies like LLM and machine learning to support biomedical knowledge discovery.

Main body: We propose a novel biomedical neural search service called 'VAIV Bio-Discovery', which supports enhanced knowledge discovery and document search on unstructured text such as PubMed. It mainly handles with information related to chemical compound/drugs, gene/proteins, diseases, and their interactions (chemical compounds/drugs-proteins/gene including drugs-targets, drug-drug, and drug-disease). To provide comprehensive knowledge, the system offers four search options: basic search, entity and interaction search, and natural language search. We employ T5slim_dec, which adapts the autoregressive generation task of the T5 (text-to-text transfer transformer) to the interaction extraction task by removing the self-attention layer in the decoder block. It also assists in interpreting research findings by summarizing the retrieved search results for a given natural language query with Retrieval Augmented Generation (RAG). The search engine is built with a hybrid method that combines neural search with the probabilistic search, BM25.

Conclusion: As a result, our system can better understand the context, semantics and relationships between terms within the document, enhancing search accuracy. This research contributes to the rapidly evolving biomedical field by introducing a new service to access and discover relevant knowledge.

Keywords: Natural language processing, Text mining, LLM, Transformer, RAG, Biomedical interaction extraction, Neural search, T5, Embedding

Background

Due to the ongoing progress and the emergence of new discoveries in biomedical areas such as genetics, pharmacology, public health and life science topics, the repository of the biomedical literature is continually expanding with a remarkable growth rate. This proliferation of biomedical data makes it challenging for researchers and professionals to keep up with the latest findings, useful knowledge and research trends. Thus, information retrieval and automatic text mining systems have become extremely important.

As one of the systems, PubMed is a widely used, well-organized public database that supports the advanced search and retrieval of biomedical and life sciences literature. It

contains more than 36 million citations and abstracts of biomedical literature from several NLM (National Library of Medicine) literature resources, including MEDLINE, life science journals, clinical studies and online books, with new content continually being added. Utilizing Medical Subject Headings (MeSH) terms correctly can enhance search accuracy on PubMed. However, identifying and applying MeSH terms effectively may be difficult for general users. In addition, users should select relevant keywords or phrases to retrieve precise search results or understand the various filters and operators to refine the search. Furthermore, it is hard to search for hidden knowledge, such as associations or interactions between entities, on PubMed.

Hence, there has been a significant demand for a more advanced search and knowledge discovery system to reduce the time to develop biomedical hypothesis or curate databases. Such a system should not only perform basic search based on queries and keywords but also provide capabilities for summarizing the search results, QA (question answering), and delivering more enriched complex information about interactions between biomedical entities from unstructured text data.

Related works

Recently, promising advancements in transformer-based LLM (large language models) [1], utilizing billions of parameters trained on extensive text corpora, such as GPT-4 [2] or T5 [3], have achieved state-of-the-art (SOTA) performance in many NLP (natural language processing) tasks. In practice, ChatGPT [2, 4] demonstrates excellent capabilities in language understanding and generation, and it continues to evolve rapidly to understand and respond to a wide range of complex query requests including image or voice.

In the biomedical field, transformer-based methodologies like BioBERT [5], PubMedBERT [6], BioLinkBERT [7], SciFive [8], and $T5_{slim\_dec}$ [9] have demonstrated the potential of LLMs in various text mining tasks. For instance, SciFive [8] and T5-MTFT [10], pretrained on biomedical texts using T5 architecture [4] have shown good performance in RE (relation extraction). Specifically, SciFive attempted a domain-specific T5 model which was pretrained on C4 [11], PubMed abstracts, and PMC full-text articles. It outperformed other encoder-only models in biomedical domain. BioLinkBERT [7] captured document links, such as hyperlinks and citation links. It was pretrained by feeding the linked documents by PubMed citation links into the same context as inputs to include knowledge that spans across multiple documents. It proposed document relation prediction, which classifies two linked segments as 'contiguous', 'random', or 'linked', as an alternative to the next sentence prediction objective in BERT [12]. This approach enables the incorporation of cross-document knowledge that is not available in single documents.

Some recent studies [13, 14], have directly applied LLMs such as ChatGPT to their problem domains without extensive finetuning via few-shot learning or simple prompting. According Chen Q. et al.'s study [13], the average performance of SOTA systems in biomedical NER (named entity recognition) was about 0.86 (F1-score), and that in RE was approximately from 0.80 to 0.82 on average. However, ChatGPT based on prompts obtained a BLURB (Biomedical Language Understanding & Reasoning Benchmark) score of 0.595, which was significantly lower than the SOTA systems. Only in the QA task, ChatGPT (0.825) showed a competitive performance compared to the systems like PubMedBERT (0.717), BioLinkBERT-Base (0.808), and BioLinkBERT-Large (0.835). That

means that the biomedical domain is a still significantly challenging and complex area to handle directly using prompts in ChatGPT.

Although ChatGPT is capable of responding to various biomedical questions, it often encounters hallucination problem, where the model generates sentences containing incorrect information. Verifying references accurately, is another important issue particularly in conducting scientific research based on precise facts. In this study, we combine neural retrieval system with advanced language generation technique for a more informed and reliable text mining.

System objectives and effects

This study aims to provide a comprehensive knowledge database and intensive search & QA system on biomedical articles. We suggest a novel biomedical search service called 'VAIV Bio-Discovery' which incorporates transformer-based large-scale biomedical text mining with neural search [15]. The foremost objective is to facilitate understanding of the complex interactions between chemical compounds/drugs and other entities from scientific literature.

Our system exhibits the following strengths: (1) it provides four types of user-friendly interfaces with query options including a basic search, entity and interaction search, and natural language query search. (2) The intended use of this database is to provide enriched complex data sources about biomedical entities, MeSH terms, interactions between entities extracted from scientific literature such as PubMed abstracts by using a transformer-based deep learning method [9]. The target entities are chemical compounds/drugs, genes/proteins, and diseases. The target interactions are drug-drug interactions (DDI), chemical compounds and protein/gene relations (CPR), and chemical compounds/drug and diseases relations (CDR). (3) The system presents meaningful insights into research trends by offering statistically ranked quantitative information on how frequently named entities and their relations are mentioned in research publications. (4) It assists in interpreting research findings by summarizing the retrieved search results into natural language using an LLM. We provide a functionality which allows users to start the natural language query related to entities of interest. To this end, we employ the Retrieval Augmented Generation (RAG) method [16] which first conduct a deep learning-based neural search to identify articles that are likely to contain answers in response to the given natural language. Then, it creates prompts using the query and extracted passages for LLM to generate a summarized answer text.

This process effectively highlights valuable information extracted from biomedical texts rather than simply listing up the retrieved documents with keywords from a query. There are several database systems publicly available, such as UniProt,[1] which provides information on protein sequences and functions; the Comparative Toxicogenomics Database (CTD) [17, 18], offering customized data related to a set of chemicals, diseases, genes, Gene Ontology terms, pathways, and references; and the Therapeutic Target Database (TTD) [19] as well as GenBank[2] and Protein Data Bank (PDB).[3]

[1] https://www.uniprot.org/

[2] https://www.ncbi.nlm.nih.gov/genbank/

[3] https://www.rcsb.org/

Table 1 Search types

Query type	Example
(1) Basic query search (operator/filter)	• Search for documents with calcium and anoxia in the title → (calcium[Title]) AND (anoxia[Title]) [PubMed]
(2)Entity or relation search	• What diseases co-occur most with calcium in the document collection? I would like to search for documents with calcium and related diseases • What diseases are associated with calcium in the papers containing the MeSH term 'child'? • What gene/proteins have the "ACTIVATOR" interaction with calcium? List them by frequency order and show the documents
(3) Natural language query search	• What methods can be used for detecting beta thalassemias? • What chemical compounds inhibit the activation of BRCA1?

Specialized curated databases in a targeted area and their integration are still necessary to ensure data currency and completeness for knowledgebase. However, there has been a considerable advancement in AI technologies like LLM and machine learning to support biomedical knowledge discovery. Our system specializes in fully automatically discovering enriched information mentioned in unstructured text data without relying on external databases or resources. In addition, it distinguishes itself from other DBs by enabling natural language queries for user friendly searches and summarizing the content of the retrieved research articles.

Consequently, it supports the development of biomedical hypotheses and the appropriate curation of databases. It can facilitate for researchers or curators/DB constructors to access and discovery more complex and enriched information from unstructured biomedical publication text in a comprehensible way and follow up the latest research findings.

The interactions on our database can support to understand unpredictable changes in pharmacological effects of drugs, mechanisms of diseases, and to develop therapeutic drugs. In addition to, they help in designing drugs that modulate gene expressions or interact with specific proteins involved in disease pathways. Since many diseases have a genetic basis, understanding the relationships between chemical compounds and specific genes or proteins can lead to the development of targeted therapies.

Construction and content

Motivation example

We first describe the intended use of this database system, providing with motivation examples. Consider Table 1: most search engines straightforwardly retrieve results for basic queries like 1) using Boolean operators and title filters. In the example, to perform the specified search on PubMed, the 'AND' operator and the 'Title' filter are necessary. However, addressing search requirements such as examples in 2) and 3) can be very challenging because we practically have no prior knowledge about which genes/proteins are most closely associated with a specific entity. Although previous biomedical text-mining approaches are helpful in finding documents including specific biomedical terms, they often fail to provide statistically ranked quantitative information or meaningful relation information between biomedical entities.

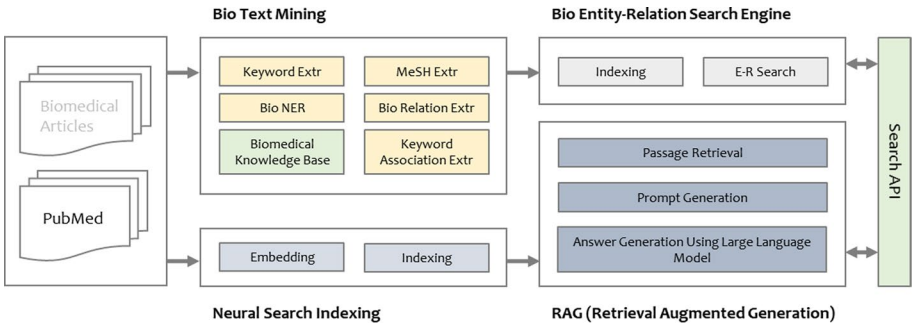


Fig. 1 System architecture

In addition, researchers seek to obtain answers through natural language queries from a large number of documents. Classical term-based search systems rank a set of documents by computing a relevance score for each document based on a given query. However, a document can be actually relevant to a query even without matching terms. Thus, we attempt neural search which uses vector embedding to represent document and query more semantically. It enables to capture context of a term in a document and semantic relations with other terms. Moreover, since the answers to a given query are often scattered across multiple documents, summarization is very helpful. For search and summary generation, we adopt the retrieval augmented generation based on the large language model.

Database content

In this section, we describe a system design including indexing and implementation along with data sources and the informatics of data generation. For convenience, we explain the details with each module. The system architecture can be depicted as shown in Fig. 1.

NE (named entity) recognition and relation extraction module

Using the specialized NLP module, biomedical named entities, general keywords, and biomedical relations between the entities are gradually extracted from the documents. We first collected 219,317 publication abstracts from 2023 PubMed[4] baseline and 6,924 from PubMed daily updates. The National Library of Medicine (NLM) offers an annual baseline set of PubMed citation records as well as daily update records, both available for free download in XML format. In the MEDLINE PubMed XML, certain mandatory elements such as the article title, abstract, author name, publication date and journal title are essential for a record to be complete. On the other hand, optional supplementary elements like author affiliation, grant support, reference, keywords, chemical lists, MeSH terms, tags, and other metadata provide additional information that enhances the record's comprehensiveness and usefulness. In this study, two supplementary elements, chemical list and MeSH term lists are considered in addition to the essential elements. Additionally, documents related to the functions of 8,499 targets from TTD

[4] https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/

[19] which describe therapeutic protein and nucleic acid targets, related diseases, pathways, and corresponding drugs were added. These documents were indexed, but interactions between entities were not extracted. The biomedical named entity tagger[5] [20] we used recognizes entities of chemicals/drugs, genetics (genes/proteins), and diseases/symptoms from the abstract texts. Its core engine for text entity recognition is based on BioBERT [4]. According to the study [20], this tagger achieved a micro-average F1-score of 0.86 for partial matches on the PGxCorpus [21].

In biomedical literature, accurate recognition and category mapping of entities are very challenging since they often exhibit inconsistencies and ambiguities in expressions due to synonyms, abbreviations, and diverse nomenclature of terms. In addition, some entities can belong to both chemical compound/drug and protein/gene categories. For example, 'interferon alfa-2b' is a form of recombinant human interferon used to treat 'hepatitis B and C infection', 'genital warts', 'hairy cell leukemia', 'follicular lymphoma', 'malignant melanoma', and 'AIDs-related Kaposi's sarcoma'. It exhibits biological characteristics as a recombinant protein and is used as a drug to mimic the action of the protein in some contexts. In scientific and medical contexts, a comprehensive understanding requires considering these diverse perspectives.

To address potential errors of the NE tagger, we also include terms from the chemical list provided by the PubMed XML as entities if they appear in the abstract. However, since PubMed's chemical list encompasses genes and proteins besides chemical compounds, we constructed an additional dictionary to differentiate them under the gene/protein category. It was compiled using the entities listed in DrugBank [22], CTD (Comparative Toxicogenomics Database) [17, 18] and UniProt [23] databases along with the entities annotated in the ChemProt [24], DDI [25] and DrugProt [26] Corpus.

After identifying entities in the text, their co-occurrences are indexed to investigate their relatedness. Similarly, general key terms are extracted and indexed after removing stop words for keyword search. The keywords in our search engine correspond to general terms, named entities, and MeSH terms. For each keyword, co-occurrences with named entities and MeSH terms are indexed and sorted by frequency. Table 2 presents terms associated with 'mcp-1', a protein that plays a crucial role in the immune response and inflammatory processes in the human body. These terms can be categorized into gene/protein, chemical compound, disease entities, MeSH terms and general terms. In our work, MeSH terms are treated as entities. Consequently, the search engine has indexed various pairs of entities, including the entity–entity, general term-entity, and entity-interactions.

Our system classifies the type of interaction into the relevant category, as presented in Table 3, if a sentence contains a pair of entities that exhibit potential interaction. For instance, in the case of DDI, when two distinct chemical compounds or drugs are mentioned within the same sentence, they are considered as candidates for interaction classification. To analyze interaction candidates, approximately 2.12 million sentences from abstracts were processed using a sentence splitter.

[5] https://github.com/librairy/bio-ner

Table 2 Associated terms for 'mcp-1'

Keywords	Freq	Category
Inflammation	127	Gene/Protein
il-6	113	Gene/Protein
il-8	63	Gene/Protein
monocyte chemoattracant protein-1	58	Gene/Protein
cytokines	54	Gene/Protein
rantes	40	Gene/Protein
atherosclerosis	40	Disease
tnf-alpha	29	Gene/Protein
infection	27	Disease
ip-10	26	Gene/Protein
il-10	26	Gene/Protein
stress	25	Disease
ccl2	25	Gene/Protein
icam-1	23	Gene/Protein
Chemokines	22	Gene/Protein
Monocyte chemotactic protein-1	20	Gene/Protein
Influenza	18	Disease
:	:	
Keywords	**Freq**	**Category**
Humans	127	MeSH
Animals	113	MeSH
Chemokine ccl2	63	MeSH
Male	58	MeSH
Mice	54	MeSH
Inflammation	40	MeSH
Cytokines	40	MeSH
Female	29	MeSH
Cultured cells	27	MeSH
Inbred c57bl mice	26	MeSH
Tumor necrosis factor-alpha	26	MeSH
Macrophages	25	MeSH
Messenger rna	25	MeSH
nf-kappa b	23	MeSH
Middle aged	22	MeSH
Signal transduction	20	MeSH
Biomarkers	18	MeSH
:	:	

To extract relations between entities, we adopted the $T5_{slim\_dec}$ model proposed in our previous study [9], which is a modified version of the original T5 [4] specifically designed for interaction generation. In the relation generation task, the transformer model generates a single interaction string such as "DDI-effect" or "AGONIST," as its output for each given sentence input. In this task, the self-attention mechanism in decoder block primarily functions as an identity function and the multi-head does not effectively capture the connections between target tokens due to the presence of only a single target token. Thus, the $T5_{slim\_dec}$ model removes the self-attention layer in the general transformer's decoder and integrates the target interaction labels directly into the vocabulary.

Table 3 Target interactions

Relation	Interaction types		Semantic meaning
Chem/Drug-Protein/gene Interaction	CPR:1	PART-OF	Part-of
	CPR:2	DIRECT-REGULATOR, INDIRECT-REGULATOR, REGULATOR	Regulator
	CPR:3	ACTIVATOR, INDIRECT-UPREGULATOR, UPREGULATOR	Upregulator or activator
	CPR:4	DOWNREGULATOR, INHIBITOR, INDIRECT-DOWNREGULATOR,	Downregulator or inhibitor
	CPR:5	AGONIST, AGONIST-ACTIVATOR, AGONIST-INHIBITOR	Agonist
	CPR:6	ANTAGONIST	Antagonist
	CPR:7	MODULATOR, MODULATOR-ACTIVATOR, MODULATOR-INHIBITOR	Modulator
	CPR:8	COFACTOR	Cofactor
	CPR:9	SUBSTRATE, SUBSTRATE_PRODUCT-OF, PRODUCT-OF	Substrate or product-of
	CPR:10	NOT	Non-interacting entities
Drug-Drug Interaction	DDI-Mechanism		a pharmacokinetic interaction mechanism
	DDI-Effect		the effect of an interaction
	DDI-Advice		a recommendation or advice regarding the concomitant use of two drugs
	DDI-Int		the sentence mentions that interaction occurs and does not provide any detailed information about the interaction
	DDI-False		Non-interacting entities
Chem/Drug-Disease Interaction	Potential		the case where certain type of interaction is expected
	NOT		Non-interacting entities

Consequently, $T5_{slim\_dec}$ constrains its outputs (target labels) to generate complete whole tokens, rather than predicting a sequence of separated tokens in an autoregressive manner. It utilizes the pretrained parameters of SciFive [8] which were further finetuned on specific training datasets, namely ChemProt [24] and DrugProt [26] for BioCreative RE tasks. The model has demonstrated improved relation classification performance compared to SOTA models in the ChemProt and DDI tasks. It achieved an F-score accuracy of 0.92 in the DDI dataset and 0.943 in the ChemProt dataset.

In the ChemProt BioCreative task [24], interactions were grouped into 10 semantically related classes, labeled from CPR:1 to CPR:10. However, only five relation types were utilized to evaluate system performance. The types of interest correspond to CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9. In contrast to ChemProt evaluation, this work considers all CPR interaction types as target interactions. From a granularity perspective, these groups pose challenges of the practical utility in biomedical applications and add complexity into the classification procedure. Moreover, the training datasets for CPR:7(modulator) and CPR:8 (cofactor) are quite limited in size. This indicates that the categories are difficult to classify accurately. Nevertheless, the

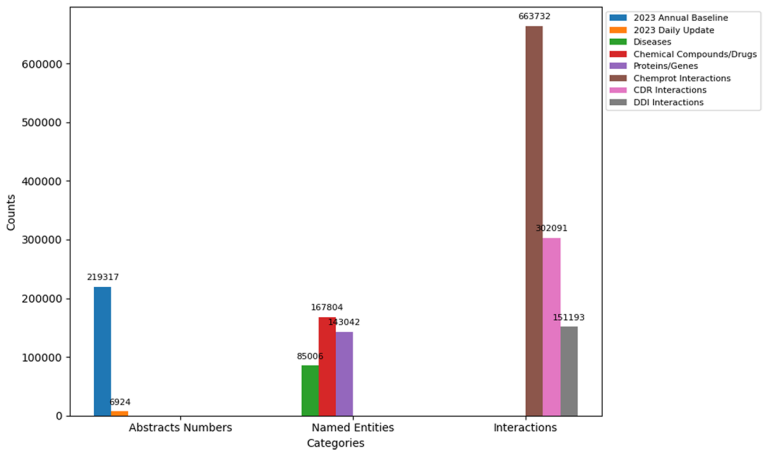


Fig. 2 Database contents

Relations	Freq
POTENTIAL	2,221
REGULATOR	1,815
DDI-effect	1,578
DDI-mechanism	779
SUBSTRATE/PRODUCT_OF	739
ACTIVATOR/UPREGULATOR	326
DOWNREGULATOR/INHIBITOR	195
ATAGONIST	62
AGONIST	38
COFACTOR	27
:	:

Entities	Freq
calmodulin	42
Parathyroid hormone	25
alkaline phosphatase	23
pth	22
insulin	22
albumin	20
protein kinase c	18
thrombin	16
endothelin-1	15
pkc	15
calcitonin	14
renin	14
amylase	13
:	:

Fig. 3 Some interactions associated with 'calcium'

ChemProt and DrugProt training datasets are widely recognized as Gold Standard datasets due to their comprehensive coverage and manually annotation by experts. For more details, please refer to the study [9]. The target interaction types considered in this work are presented in Table 3.

For the CDR task, the training dataset available for the $T5_{slim\_dec}$ transformer model is extremely limited. The datasets for the BioCreative V Chemical-Disease Relation (CDR) task [27] comprised 1,500 PubMed abstracts, which were equally divided into 500 each for training, development, and testing, and were focused on chemical-induced disease (CID) relations. At the current stage, in case of CDR, only potential interactions are recognized, due to the insufficient datasets for training more detailed and specific interaction types.

Table 4 Indexing structure for entity and relation database

Entities	Associated entites	Freq	Doc List
(1)			
Calcium	nifedipine	643	D1, D2, …
Calcium	acetylcholine	613	D1, D101, …
Calcium	pkc	158	D15, D26, …
Calcium	insulin	140	D2, D64, …

Entities	Relations	Freq
(2)		
Calcium	SUBSTRATE/PRODUCT_OF	739
Calcium	DDI-effect	1578
Calcium	DDI-mechanism	779
Calcium	REGULATOR	1815

Entities	Relations	Entities	Freq	Doc List
(3)				
calcium	SUBSTRATE	endothelin	11	D1, D2, …
calcium	SUBSTRATE	calcitonin	11	D1, D101, …
calcium	SUBSTRATE	ryr (ryanodine receptor)	11	D15, D26, …
calcium	SUBSTRATE	trpv6	9	D2, D64, …

(1) Entity–entity inverted index table	**(entity, association entity, frequency, doc list)**
(2) Entity-relation index table	(entity, relations, frequency)
(3) Entity-relation-entity inverted index table	(entity, relation, association entity, frequency, doc list)

Consequently, from 226,241 abstracts, 85,006 diseases, 167,804 chemical compounds/drugs, 143,042 proteins/genes were recognized. Additionally, 663,732 (CPR), 151,193 (DDI), and 302,091 (CDR) pairs were ultimately identified from 2.12 million sentences as exhibiting specific interactions between the entities after excluding recognized false interactions such as 'NOT' in CPR and 'DDI-false' in DDI, as shown in Fig. 2.

Figure 3 shows various interactions associated with 'calcium' including 'POTENTIAL', 'REGULATOR', 'DDI-effect', 'DDI-mechanism', 'SUBSTRATE/PRODUCT-OF' and so on. Additionally, specific entities that interact with 'calcium' are identified. It has 'REGULATOR' relationships with proteins such as 'calmodulin', 'parathyroid hormone (pth)', 'alkaline phosphatase', 'insulin', 'albumin' and others.

Indexing module

The indexing process makes it easier to access the content related to each entity. As stated earlier, the associated terms covered by our search engine include gene/protein, chemical compound, disease entities and MeSH terms. For efficient retrieval in entity and relation searches, three distinct types of index tables are employed: (1) an entity–entity inverted index table, designed to find associated terms for each entity, (2) an entity-relation index table to discover associated interaction for each entity, and (3) an entity-relation-entity inverted index table which facilitates the identification of associated terms for a given entity and interaction, as detailed in Table 4. In a given document, if interactions between the same entities occur in multiple sentences and belong to the same interaction category, they are indexed only once and counted as one.

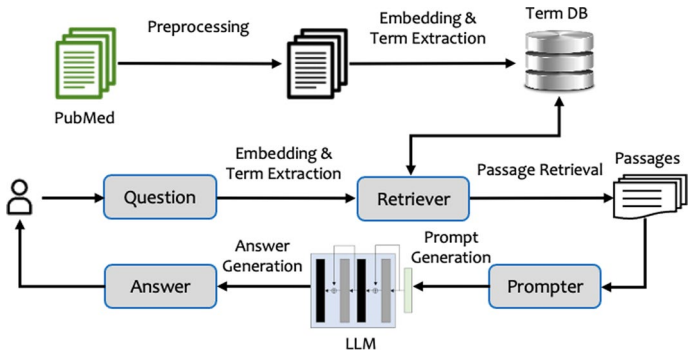


Fig. 4 System flow for Retrieval Augmented Generation

As a result, the index size is substantial as all entity pairs and entity-relation-entity triples are stored with their associated document information in the database. The size is expected to increase significantly as more documents are added. To efficiently handle this expanding volume of data, we employ Hadoop Distributed File System and a Hadoop-based NoSQL database, HBase [28].

Search and answer generation module

We provide answers derived from papers in response to natural language queries based on the Retrieval Augmented Generation (RAG) method [16]. It initially retrieves articles likely to contain relevant information, and then generates prompts based on the query and retrieved passages, rather than merely extracting documents containing query keywords. Finally, large language model generates answers by using the prompts.

To retrieve abstracts, we adopt a hybrid method combining neural search with keyword-based probabilistic retrieval model BM25 for passage retrieval, as shown in Fig. 4. This can leverage the strengths of both approaches: neural search retrieves documents containing answers to natural language queries by identifying semantically related texts through embedding vectors, while BM25, a keyword-based model, emphasizes important keywords in documents relevant to the query. In the neural search, both documents and queries are converted into vector embeddings, and answers are located based on the vector relatedness by comparing the query and the document embeddings [15]. The system indexes the text with its vector embeddings in a vector index.

To this end, we utilize RoBERTa model [29], where text is first tokenized by Byte Pair Encoding (BPE). It begins by splitting the text into individual characters and then progressively merges the most frequently occurring pairs of characters or character sequences. Consequently, it creates a vocabulary of the most common character combinations which consist of whole words and subwords. This method is particularly effective in handle rare words and out-of-vocabulary terms. The tokenized text is transformed into vectors through embedding. The RoBERTa model combines token embeddings, which capture the semantic meanings of words, with positional embeddings that highlight their order and position in the sequence. This allows the model to effectively understand the context, semantics, and structural relationships of tokens within a document, thereby enhancing the relevance and accuracy of the search results.

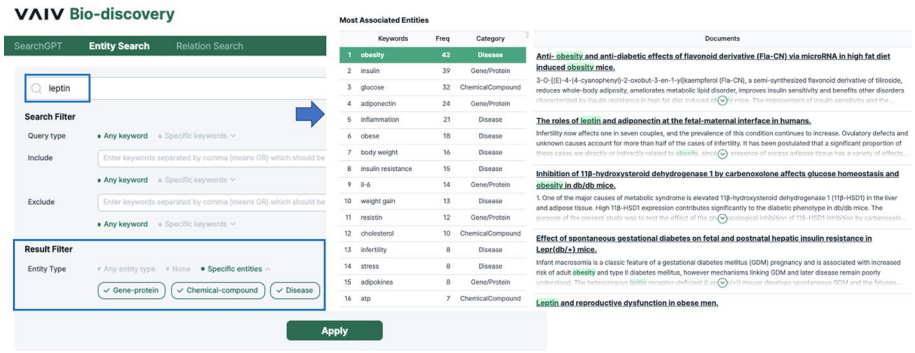


Fig. 5 Keyword and entity search

To generate answer for a given natural language query, the ChatGPT model [3] is adopted as the LLM. By combining search capabilities with the LLM, we can mitigate the hallucination problem, which is one of major issues in LLMs. By conditioning on retrieved relevant documents, the RAG architecture can generate more accurate and contextually appropriate answers, especially for questions requiring factual knowledge. Furthermore, this integration enhances a comprehensive understanding of the contents within the search results from papers rather than simply presenting them in a list.

Utility and discussion

In this section, we describe user interface and the intended uses of the database. In addition, we introduce the benefits of functionality on provisioned module and improvement of similar existing databases. A case study of the use of the database and future plan are also presented.

User interface and utility

This database supports three types of search. As shown in Fig. 5, it offers functionality for both keyword and entity searches. It can invoke a search filter function to either include or exclude specific keywords and a result filter to limit the types of associated entities for a given query. The figure displays entity search results related to 'leptin'. It identifies 'obesity' as the most closely related disease and presents links to related publications, along with their abstracts, on the right-hand side.

Figure 6 presents the results of a relation search related to 'leptin'. As shown in the figure, if the relation-centered filter option is selected, it lists associated interactions and then displays the entities related to any chosen interaction. For instance, chemical compounds like 'glucose', 'cholesterol', 'fatty acid', 'nitric oxide', 'triglyceride', and 'plasminogen' are identified as having regulatory interactions with 'leptin'. Documents related to these regulatory interactions are then collectively displayed on the right side of the interface. Conversely, when the entity-centered button is selected, interactions involving leptin are displayed with a focus on the entity.

Figure 7 shows the system returns a response to a natural language query, 'What is leptin and how is it related to glucose?'. It presents the search results similar to general search engines. The key differences are: (1) the documents, that are likely to contain answers to the question, are retrieved not just because they include keywords in query,

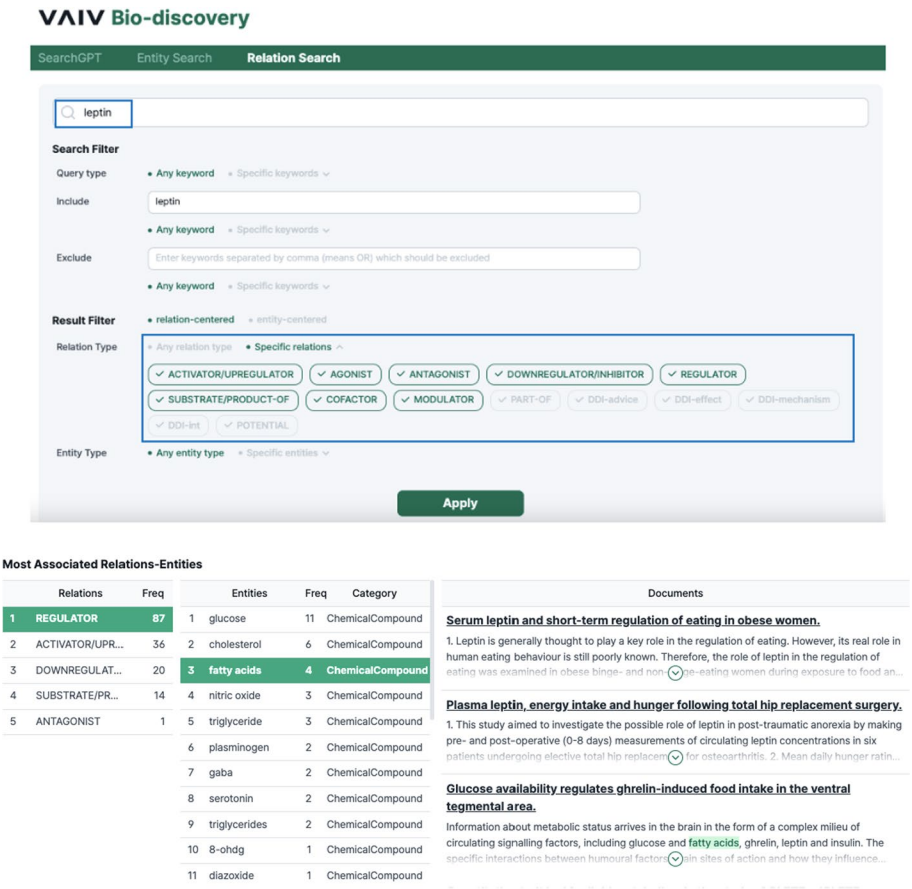


Fig. 6 Relation search

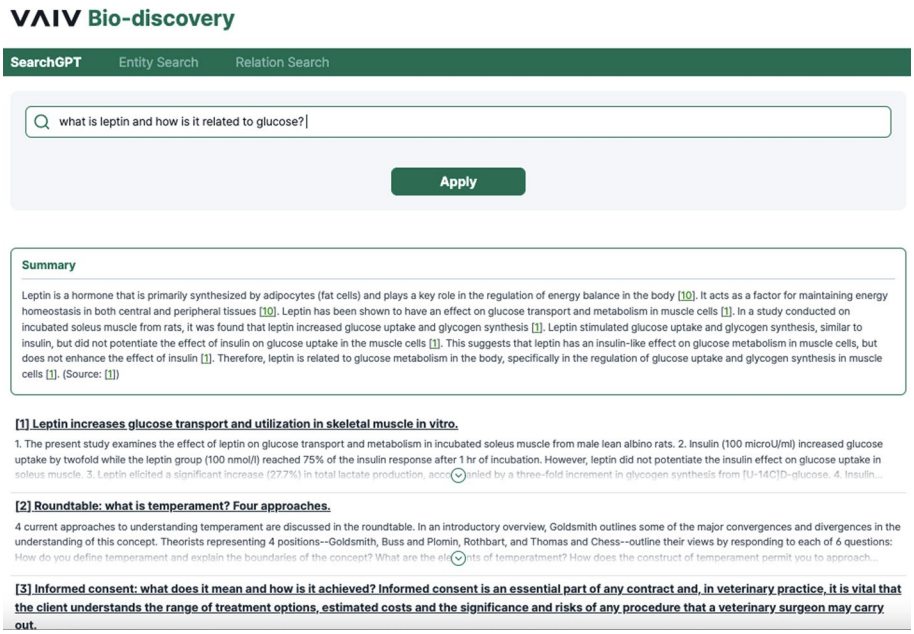


Fig. 7 Search and summary with natural language query

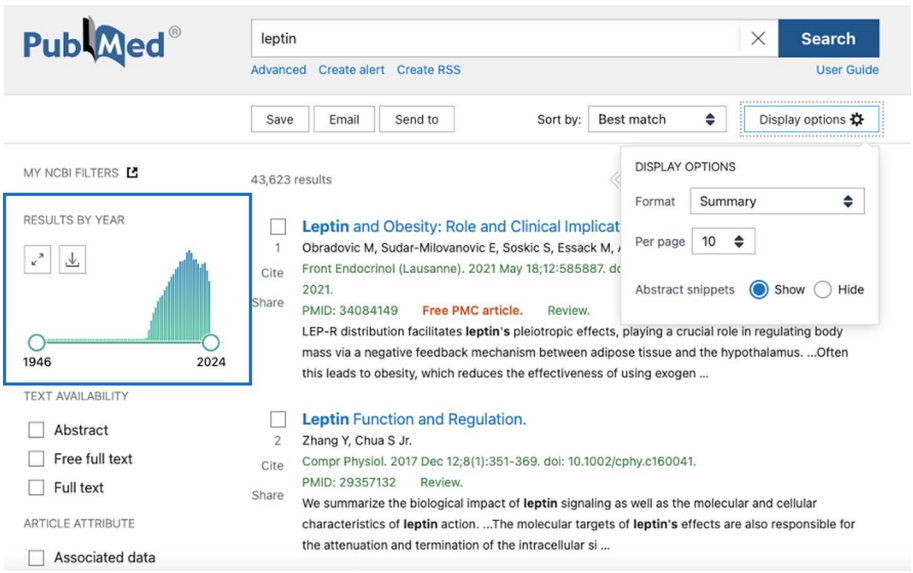


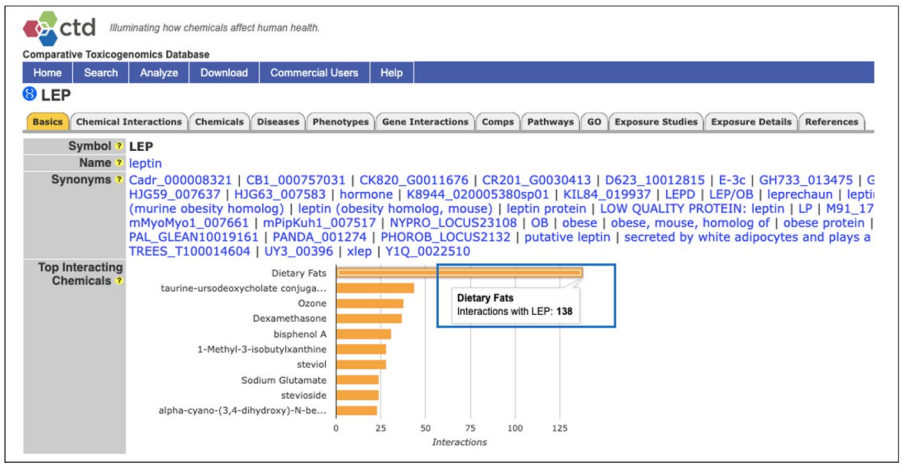
Fig. 8 PubMed search results for 'leptin'. This figure shows the search result from PubMed (https://pubmed.ncbi.nlm.nih.gov/?term=leptin). Permission to use the screenshot has been granted by the PubMed team

and (2) the LLM generates the response based on the retrieved results. It summarizes the contents of retrieved results as "leptin is a hormone secreted by fat cells that exerts significant effects on the brain, glucose metabolism, and muscle cells. It has insulin-like properties, enhancing glucose uptake and metabolism in muscle cells. Additionally, leptin increases glucose uptake and stimulates the synthesis of glycogen, akin to insulin's effects". Furthermore, references to the abstracts related to the generated summary are included.

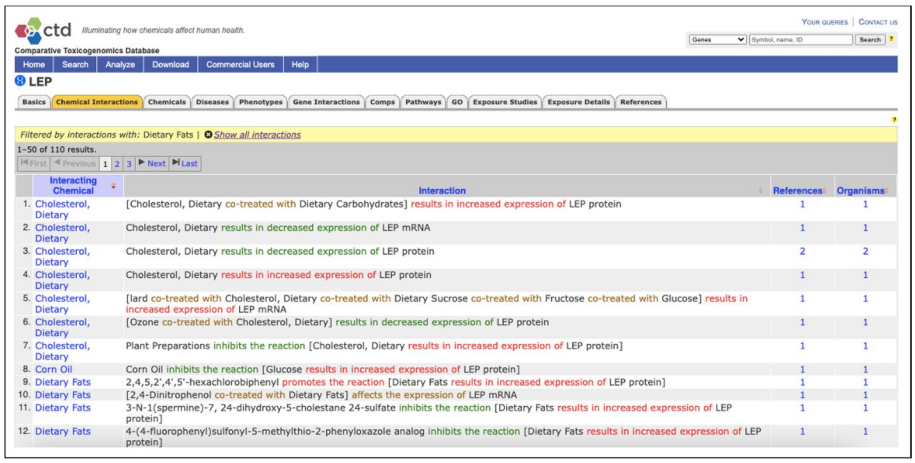
Comparison with other databases and text mining systems

In this section, we first compare our database with other database search systems such as PubMed and CTD [17, 18]. Figure 8 shows PubMed's search results for 'leptin'. To control the results, PubMed offers filters related to text availability (such as abstract, free full text, full text), article attributes, article types, and publication dates. Additionally, there is an option to display the abstracts of the retrieved documents. There is no further enriched information related to the keyword. On the other hand, CTD supports improved information about 'leptin' as shown in Fig. 9. It displays top-ranked interacting chemicals. CTD integrates data from diverse resources such as BioGRID, ChemIDplus, CL, GO, KEGG, MeSH, and PubMed, providing manually curated data relating chemical exposures with their genetic, molecular, and biological outcomes.

Some curation application tools for data entry invoke functions with automatic quality control to help annotations of CTD biocurators and interactions are translated into readable sentences. For example, structured interaction notation such as 'C1/n+act G1/p' is displayed as "bisphenol A analog results in increased activity of ESR1 protein" by conjoining terms from vocabularies, MeSH, 4 chemical qualifiers, 4 action term degrees, 55 action terms, NCBI gene symbol and gene qualifiers. Figure 9 shows interaction sentences between 'leptin' and 'dietary fat'. However, these expressions tend to be



(a) information related to 'leptin'



(b) evidence sentences of interactions between leptin and 'dietary fats'

Fig. 9 CTD search results for 'leptin'. **a** This figure shows the search results from Comparative Toxicogenomics Database (https://ctdbase.org/detail.go?type=gene&acc=3952). Permission to use the screenshot has been granted by the CTD team. **b** This figure shows the search results from Comparative Toxicogenomics Database (https://ctdbase.org/detail.go?type=gene&acc=3952&view=ixn&chemAcc=D004041). Permission to use the screenshot has been granted by the CTD team

overly rigid and lack contextual depth, resulting in evidence sentences that are so structured they limit their ability to provide unique insights.

Our system has discovered that 'glucose' interacts with 'leptin' most frequently, acting as a "regulator". Moreover, chemical compounds such as 'glucose', 'atp (adenosine triphosphate)', 'cholesterol', 'k+(potassium ion)', 'nitric oxide', and 'fatty acid' are identified as having relationships with 'leptin'. Figure 10 shows the interacting chemical compounds or drugs with 'leptin'. Their interactions are quantified based on individual sentences instead of on a document level to facilitate comparison. 'Leptin' often regulates 'k-atp channel', 'k+', 'k(atp)', and 'atp'.

Although the database has yet to accumulate a large quantity of publication documents, it is noticeable that it has discovered quite new interesting findings not present in the Comparative Toxicogenomics Database (CTD) such as the inhibition of 'leptin'

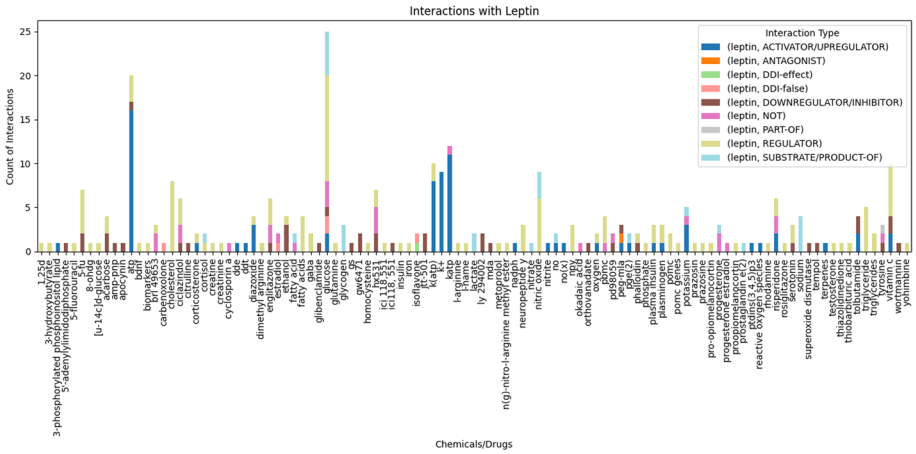


Fig. 10 Interacting chemicals with 'leptin' and their interaction types

secretion by 'vitamin C' and its impact on glucose levels, the reduction in '5-FU (fluorouracil)' cytotoxicity through leptin treatment, and the activation of 'ATP'-sensitive 'K+channels' by 'leptin'. Moreover, the interaction types are more sophisticated as shown in the figure. Figure 11 displays common interacting chemicals of both CTD and our database such as 'glucose', 'cholesterol', 'fatty acids', 'acarbose', 'diazocide', 'nitric oxide', 'tempol', and so on. CTD contains data on 474 chemical compounds interacting with 'leptin', which is a substantial volume compared to our system that found a total of 99 interacting entities. Figure 12 shows some example sentences that convey interactions between 'leptin' and chemical compounds/drugs. As seen in the figure, interaction sentences encompass a wide range of contexts, making it important to provide accompanying literature information for the specific interaction of interest.

Currently, CTD [18] consists of 17,117 chemicals, 54,355 genes, 6,187 phenotypes, 954 anatomical terms, 7,274 diseases, 202,000 exposure statements, and over 3.4 million evidence-based, manually curated interactions including chemical–gene, chemical–phenotype, chemical–disease, gene–disease, and chemical–exposure interactions. Additionally, it generates over 31 million inferred gene-disease interactions and 2.9 million statistically ranked chemical-disease predictive interactions from the internal integration of curated direct interactions. External integration with imported annotations from other databases produces an additional 13 million inferences. In CTD, if chemical A interacts with gene C, and gene C is associated with disease B, then interaction between chemical A and disease B are inferred to be related via gene C. In total, CTD includes over 50 million toxicogenomic relationships for computational analysis and hypothesis development. Our system identified 167,804 chemical compounds/drugs, 143,042 proteins/genes, and 85,006 diseases from 2.12 million sentences, which encompass a significantly broader range of entities than those covered by CTD. However, as illustrated in Fig. 13, the number of recognized interactions is significantly smaller than in CTD because CTD includes inferred interactions, and we only consider abstracts, not the full texts of publications.

In practice, even with a large collection of papers, some crucial knowledge may only be mentioned in a very few papers, making it difficult to discover. Thus, there are

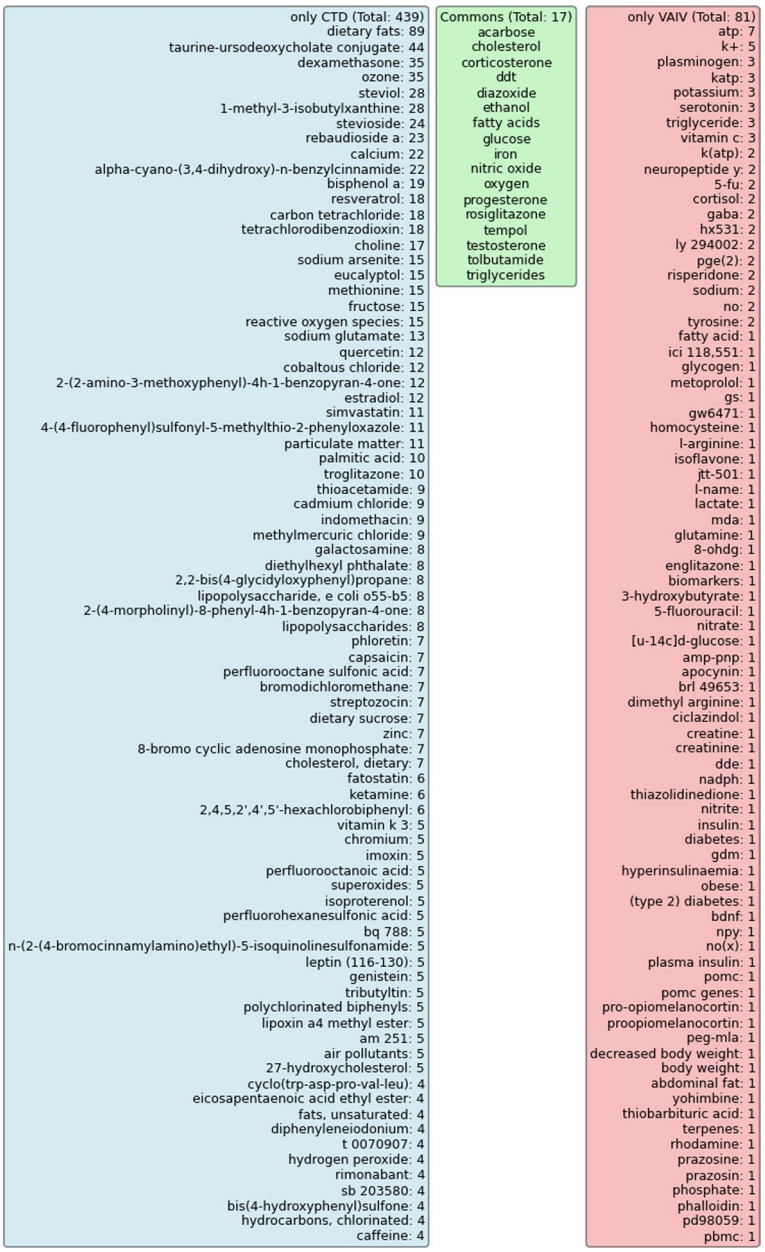


Fig. 11 Comparison of the top-50 ranked interacting chemical compounds/drugs

inherent limitations in relying solely on research papers to extract important knowledge. This highlights the importance for human-curated databases to complement the gaps in knowledge extraction from academic literature. Furthermore, to ensure accuracy, all data entry must be carefully verified.

Nevertheless, the development of such automatic knowledge construction and mining systems should proceed simultaneously with the creation of curated databases. In this context, even if our database lacks completeness, users are still capable of discovering new insights, provided that a range of relevant information is available. The most significant benefit of our system is that it can also perform QA and summarization for more

Sentence	Entity1	Entity2	Interaction
however, ## leptin ## did not potentiate the insulin effect on ** glucose ** uptake in soleus muscle.	glucose	leptin	NOT
## leptin ## per se exerts an insulin-like effect stimulating ** glucose ** uptake, glycogen synthesis, and lactate formation and also seems to potentiate the effect of insulin on glucose incorporation into glycogen in incubated soleus muscle.	glucose	leptin	SUBSTRATE
information about metabolic status arrives in the brain in the form of a complex milieu of circulating signalling factors, including ** glucose ** and fatty acids, ghrelin, ## leptin ## and insulin.	glucose	leptin	INDIRECT-REGULATOR
## leptin ## correlated modestly with serum ** creatinine ** in non-dialysis subjects.	creatinine	leptin	INDIRECT-REGULATOR
eight weeks of ** vitamin c ** supplementation restores the lost correlation between serum ## leptin ## and c-reactive protein(crp) in patients with type 2 diabetes; a randomized, double-blind, parallel-group, placebo-controlled clinical trial.	vitamin c	leptin	INDIRECT-REGULATOR
after eight weeks of follow-up, ## leptin ## level was significantly increased in the ** vitamin c ** group (md=3.48 change=24%, p-value=0.001) but did not change in the placebo group.	vitamin c	leptin	INDIRECT-UPREGULATOR
also, the correlation between serum crp and ## leptin ## became significant in the ** vitamin c ** group after eight weeks of follow-up but not in the placebo group. (rs =0.730, p<0.001 vs rs =0.286, p-value=0.266 in placebo group).	vitamin c	leptin	
the significant changes in the ## leptin ## level among the vitamin c group also remained after controlling for age, bmi, blood pressure (bp), ** triglyceride ** (tg), and cholesterol.	triglyceride	leptin	INDIRECT-REGULATOR
previously, we have demonstrated that in normal rats ## leptin ## has a time-dependent effect on renal na(+)/k(+)-atpase that drives tubular ** sodium ** reabsorption.	sodium	leptin	SUBSTRATE
in conclusion, ## leptin ## may act as a mediator linking body adiposity with changes in insulin action, sympathetic neural outflow and renal ** sodium ** excretion.	sodium	leptin	SUBSTRATE
role of tyrosine phosphorylation in ## leptin ## activation of atp-sensitive ** k+ ** channels in the rat insulinoma cell line cri-g1.	k+	leptin	ACTIVATOR
** serotonin ** receptor 5-ht5a in rat hippocampus decrease by ## leptin ## treatment.	serotonin	leptin	INDIRECT-DOWNREGULATOR
the augmentation of reactive oxygen species levels in ## leptin ## treated cardiomyocytes was reversed by 0.1-10μmol/l ** gw6471 ** (40%, 52% and 58%).	gw6471	leptin	DOWNREGULATOR

Fig. 12 Evidence sentences for interactions

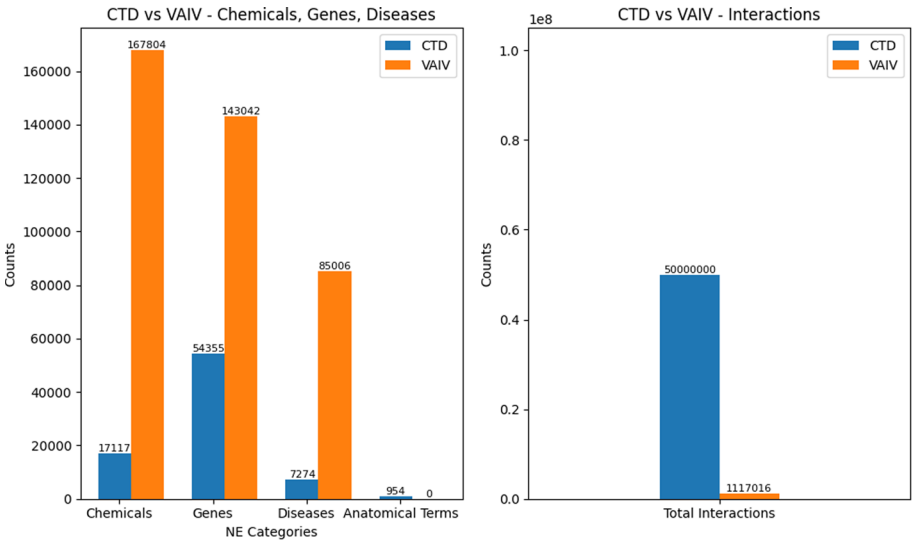


Fig. 13 Resource comparisons with CTD

extensive information through natural language queries, even at the level of relation and interaction.

We also compared our system with other text mining systems like DrugCentral 2023 [30], DrugBank [22], DigSeE [31], and Drugs.com [32]. Figure 14 shows the results from searching for 'leptin' in DrugCentral 2023 [30], which identified only two related drugs, 'metreleptin' and 'setmelanotide'. The results provided by DrugCentral are limited. In our system, when querying how each drug is related to leptin, information on 'metreleptin' was provided from TTD's target function documents and a related paper on 'setmelanotide' was also found, as shown in Fig. 15. In the case of DrugBank, no interaction for leptin was found.

Figure 16 shows the results of DiGSeE (disease gene search engine with evidence sentences) [31] regarding the association between 'leptin (LEP)' and 'insulin resistance,' which corresponds to frequent interacting pair in our system. It retrieved only one relevant document. DiGSeE identifies biological events such as gene expression, regulation,

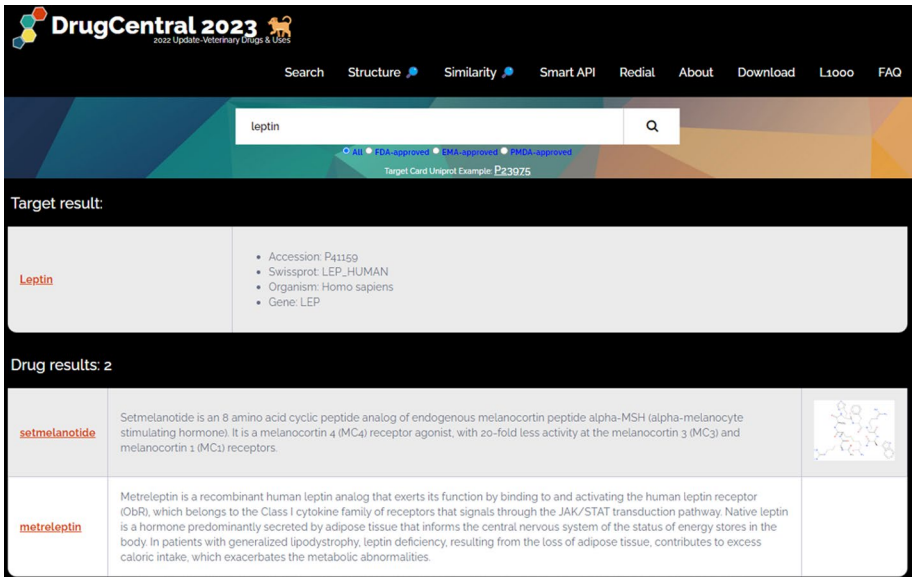
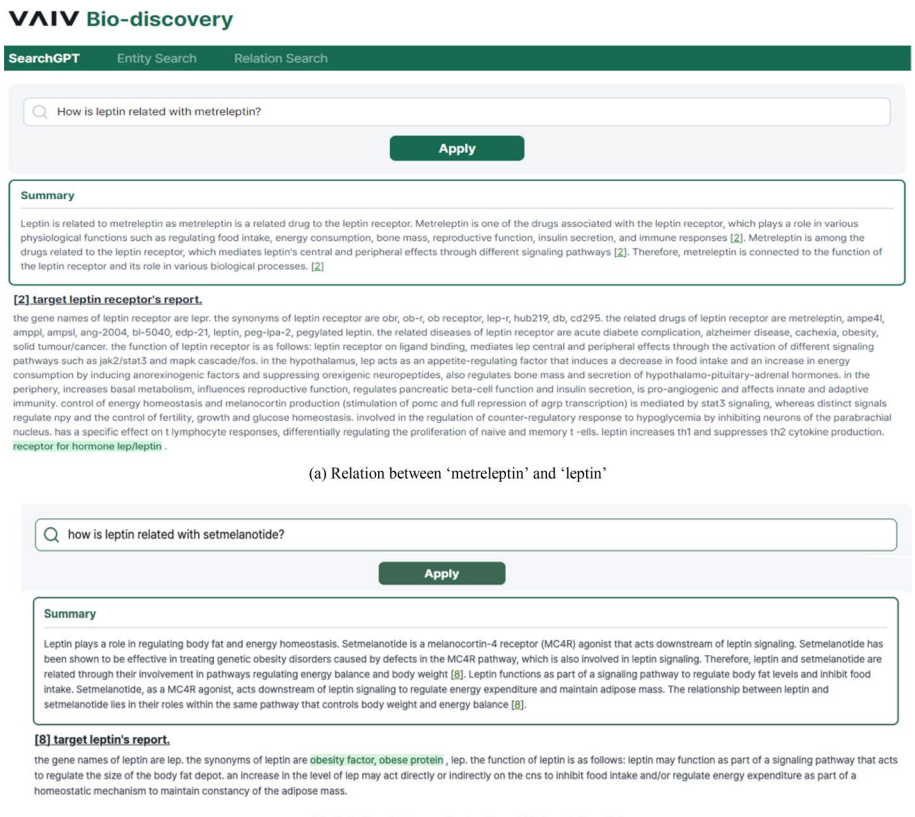


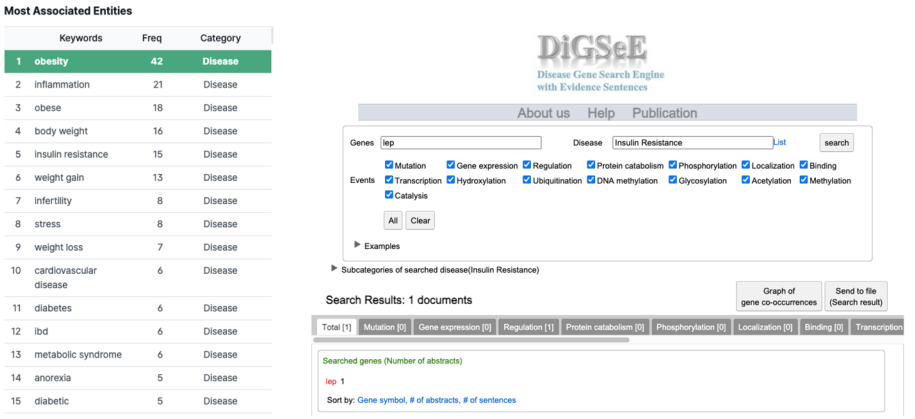
Fig. 14 Drugs related to 'leptin' in DrugCentral 2023. This figure shows the search results from DrugCentral 2023 (https://drugcentral.org/?q=leptin&approval=). Permission to use the screenshot has been granted by the DrugCentral team



(a) Relation between 'metreleptin' and 'leptin'

(b) Relation between 'setmelanotide' and 'leptin'

Fig. 15 Relations related to 'leptin' in our system



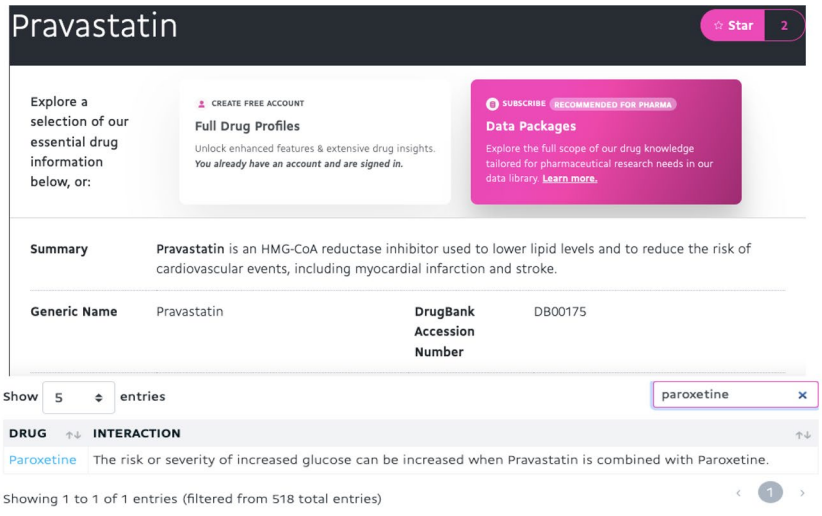
(a) Diseases associated with 'leptin' (our system) **(b)** The association between 'Insulin Resistance' disease and 'leptin' (DiGSeE)

(c) Relevant document (DiGSeE)

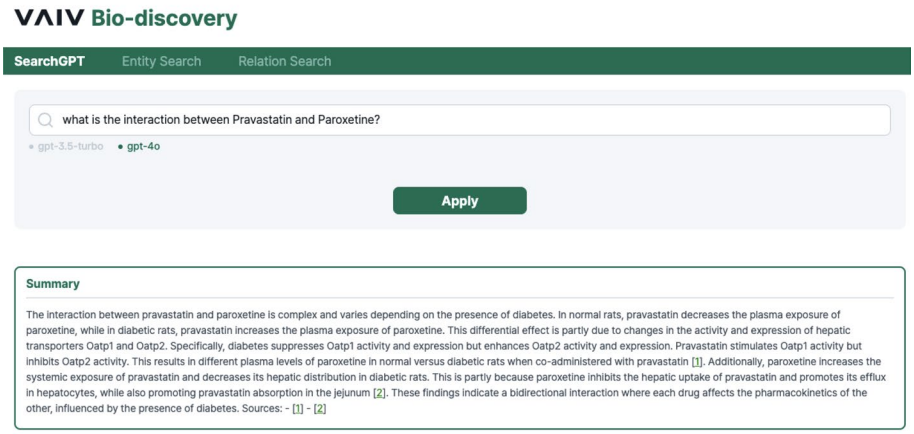
Fig. 16 CDR comparison**. b** and **c** This figures show the search results and retrieved document from DiGSeE (http://210.107.182.61/geneSearch/ Gene Query = lep and Disease Query = Insulin Resistance).Permission to use the screenshot has been granted by the DiGSeE team

phosphorylation, localization, and protein catabolism in the development of diseases to understand the associations between diseases and genes. The performance seems to require further improvement, as it primarily utilized the Turku event extraction system [33] to locate biological events, which achieved an F-measure of 52.86% (precision 58.13% and recall 48.46%).

Furthermore, to investigate how well our system summarizes in response to questions, we compared interaction descriptions from DrugBank [22], as shown in Fig. 17. As shown in the example, the information on drug interactions provided by DrugBank is concise, often limited to a single sentence and mainly related to increases or decreases in interactions. The description patterns are typically phrases like "**A** may decrease the excretion rate of **B**, which could result in a higher serum level," or "The metabolism of **A** can be increased when combined with **B**." While these descriptions are concise, they may need to elaborate on the complex mechanisms of actual drug interactions, requiring



(a) DrugBank



(b) Our system

Fig. 17 Comparison of interaction description/summarizaiton between 'Pravastatin' and 'Paroxetine'. (a) The figure shows the search results from DRUGBANK online (https://go.drugbank.com/drugs/DB00175 Interaction Drug = paroxetine).Permission to use the screenshot has been granted by the DRUGBANK team

more detailed information, especially for research or scientific analysis. Additionally, we investigated the interaction between 'pravastatin' and 'paroxetine' using the interaction checker on Drugs.com [32]. The description is likely intended as medication advice for patients, as follows: Taking pravastatin with paroxetine may increase blood glucose levels, especially in patients with diabetes. Consult your doctor about your medication use. In contrast, our system's summarized answer further explains the mechanism of how the two drugs interact, detailing the effects on their activity.

Finally, to evaluate RAG-based summarization, we used the BioASQ Task B datasets [34], excluding the cases where our system answered 'no papers exist with information that matches your question.' We used 258 questions from the Task11B-GoldenEnriched dataset (330 questions) of BioASQ Task B on Biomedical Semantic QA without adding the PubMed articles for the task. The task uses benchmark datasets containing development and test questions in English, along with gold standard (reference) answers constructed by a team of biomedical experts. Participants have to respond with relevant

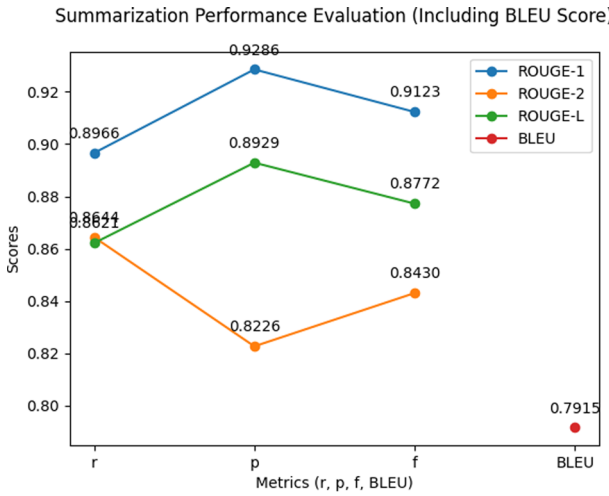


Fig. 18 Summarization Performance (our system)

concepts, articles, snippets, and RDF triples from designated resources, as well as exact and 'ideal' answers. We utilized ideal answers as references and our system's answers as candidates.

To assess the text summarization quality for questions, two widely used metrics, BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) were adopted. The BLEU score measures the similarity between a generated text and a reference text based on the precision of n-grams. A higher BLEU score indicates a higher degree of similarity, reflecting that the generated summary accurately captures the content of the reference summary. ROUGE compares the overlap between the generated summary and the reference summary. ROUGE-1, ROUGE-2, and ROUGE-L measure the overlap of unigrams, bigrams and the longest common subsequences, respectively. Our system demonstrates high QA performance with a ROUGE-1 score of 0.912 (F-score) and a BLEU score of 0.795 using only the 2023 PubMed baseline, as shown in Fig. 18. This means that we can, to some extent, prevent the potential harm of drug associations, which can arise from incorrect interpretations of the summarizations, since the information is based solely on the given publications.

Discussion and further improvements

The 2023 annual baseline of 219,317 and 6,924 daily update abstracts from PubMed might not provide comprehensive, high-quality information that we need. In addition, consistently updating with the most current and relevant information is to provide distinctive features that enable scientists to discover new and novel findings. PubMed's annual baselines contain many older papers, indicating that a significant portion of information may already be well-known to users. In order to incorporate the latest relevant studies, we integrated PubMed daily update files.

However, we found that half of these were already included in the annual baseline. The volume of new information is also quite small. To address this issue, we intend to

further expand our dataset by utilizing arXiv[6] which is a free distribution service and an open-access archive containing nearly 2.4 million scholarly articles in the various fields. Additionally, publicly accessible abstracts will be collected using the PubMed platform's 'E-utilities' API.

We will enhance neural search by incorporating filters like citation counts, journal prestige and impact factor to assess the retrieved papers, which provide indications of their relevance, significance, and recency. For future development, additional named entities such as cell lines, cell types, species, biological process and body parts should be considered as well as different interaction types especially in CDR. Our system's expansion will include extracting other important interactions such as Gene–Gene Interactions, Protein–Protein Interactions, Gene-Disease Associations, Protein-Disease Associations or Drug-Metabolite Interactions. The performance of relation extraction depends on the improvement of learning algorithms but, more crucially, on the availability of well-constructed and sufficient training data. Thus, it is essential to secure high-quality training data in future research efforts. We also plan to visualize the interactions related to entities as an additional functionality. Further research on NE tagging and relation extraction is also required for more accurate relation search.

Conclusions

In this work, we introduced a novel biomedical search system that incorporates biomedical entity and relation extraction. It provides an efficient way to find biomedical entities and relations associated with a specific entity from scientific literature. In addition, the system gives an answer for a natural language query through neural search and summarization by RAG (retrieval augmented generation) and LLM. This enables researchers or curators to quickly grasp the research findings of interest within a large collection of research papers.

We performed emerging technologies such as transformer-based pretrained deep learning for relation extraction, neural search, LLM model and RAG for language generation that enhances the performance and capabilities of biomedical text mining systems. We also emphasize the significance of biomedical text mining and sophisticated search techniques in discovering valuable information from the vast amount of unstructured text data generated in biomedical research. This research contributes to the rapidly evolving biomedical field by introducing a new service to access relevant knowledge.

Abbreviations

NLP	Natural language processing
NER	Named entity recognition
RE	Relation extraction
LLL	Large language model
BLURB	Biomedical language understanding & reasoning benchmark
RAG	Retrieval augmented generation
QA	Question-answering
MeSH	Medical subject headings
DDI	Drug-drug interactions
CPR	Chemical compounds and protein/gene relations
CDR	Chemical compounds/drug and diseases relations
CDT	Comparative toxicogenomics database

[6] https://arxiv.org/

TTD Therapeutic target database

Acknowledgements
Not applicable.

Author contributions
SK analyzed PubMed abstracts regarding named entities and implemented T5slim_dec for Drug-Drug Interaction (DDI), Chemical/Drug-Protein Relation (CPR), Chemical/Drug-Disease Relation (CDR) extraction, including natural language processing and text-mining techniques. JY designed and implemented a database indexing/search model, user interface, and web service including summarization of answers to user's natural language queries. Both SK and JY were major contributors in writing the manuscript. All authors read and approved the final manuscript.

Funding
This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2020R1I1A1A01073125).

Availability of data and materials
The data supporting the findings of this study is available at https://bio.vaiv.kr.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 30 January 2024 Accepted: 16 August 2024
Published online: 21 August 2024

References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaizer L. Attention Is All You Need. NIPS 2017. Proceedings of Advances in Neural Information Processing Systems. 2017 Dec 4–9;Long Beach, CA;USA. https://arxiv.org/abs/1706.03762
2. OpenAI. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774
3. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agrawal S, Herbert-Voss A, Krueger G, Henighan T, Child R. Language Models are Few-Shot Learners. In: Proceedings of 34th Conference on Neural Information Processing Systems. 2020 Dec 6–12;Vancouver; Canada
4. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR. 2020;21(140):1–67.
5. Lee J, Yoon W, Kim S, Kim D, Kim S, So C, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
6. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2021;3(1–2):1–23.
7. Yasunaga M, Leskovec J, Liang P. LinkBERT: Pretraining Language Models with Document Links. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022 May 22–27; Dublin, Ireland. p. 8003–16.
8. Phan LN, Anibal JT, Tran H, Chanana S, Bahadıro E, Peltekian A, Altan-Bonnet G. SciFive: a text-to-text transformer model for biomedical literature. 2021. https://arxiv.org/pdf/2106.03598.pdf. Access 30 Jan 2024.
9. Kim SH, Yoon JT, Kwon OY. Biomedical relation extraction using dependency graph and decoder-enhanced transformer model. Bioengineering (Basel). 2023;10(5):586. https://doi.org/10.3390/bioengineering10050586.
10. Sarrouti M, Tao C, Randriamihaja MY. Comparing Encoder-Only and Encoder-Decoder Transformers for Relation Extraction from Biomedical Texts: An Empirical Study on Ten Benchmark Datasets, In: Proceedings of the BioNLP 2022 workshop. 2022 May 26; Dublin, Ireland. p. 376–82.
11. Dodge J, Sap M, Marasović A, Agnew W, Ilharco G, Groeneveld D, Mitchell M, Gardner M. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021 Nov 7–11; Punta Cana, Dominican Republic. p.1286–305.
12. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the NAACL-HLT 2019. 2019 June 2–7; Minneapolis, USA. p.4171–86.
13. Chen Q, Sun H, Liu H, Jiang Y, Ran T, Jin X, Xiao X, Lin Z, Chen H, Niu Z. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics. 2023;39(9):btad557.
14. Liu P, Yuan W, Fu J. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. ACM Comput Surv. 2023;55:1–35.
15. Nakamura TA, Calais PH, Reis DC, Lemos AP. An anatomy for neural search engines. J Inf Sci. 2018;480:339–3534.

16. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W, Rocktäschel T, Riedel S, Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020 Dec 6–12. No. 793. p. 9459–74.
17. Davis AP, Wiegers TC, Wiegers J, Wyatt B, Johnson RJ, Sciaky D, Barkalow F, Strong M, Planchart A, Mattingly CJ. CTD Tetramers: a new online tool that computationally links curated chemicals, genes, phenotypes, and diseases to inform molecular mechanisms for environmental health. Toxicol Sci. 2023;195(2):155–68.
18. Davis AP, Wiegers TC, Johnson RJ, Sciaky D, Wiegers J, Mattingly CJ. Comparative toxicogenomics database (CTD): update 2023. Nucleic Acids Res. 2023;51(D1):D1257–62.
19. Zhou Y, Zhang Y, Zhao D, Yu X, Shen X, Zhou Y, Wang S, Qiu Y, Chen Y, Zhu F. TTD: Therapeutic Target Database describing target drugaability information. Nucleic Acids Res. 2024;52(D1):D1465–77.
20. Casero Á. Named entity recognition and normalization in biomedical literature: a practical case in SARS-CoV-2 literature. 2021. https://oa.upm.es/67933/. Accessed 30 Jan 2024
21. Legrand J, Gogdemir R, Bousquet C, Dalleau K, Devignes MD, Digan W, Lee C, Ndiaye NC, Petitpain N, Ringot P, Smaïl-Tabbone M, Toussaint Y, Coulet A. PGxCorpus, a manually annotated corpus for pharmacogenomics. Sci Data. 2020. https://doi.org/10.1038/s41597-019-0342-9.
22. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;D1(46):D1074–82.
23. Coudert E, Gehant S, Castro E, Pozzato M, Baratin D, Neto T, Sigrist CJ, Redaschi N, Bridge A. The UniProt Consortium, Annotation of biologically relevant ligands in UniProtKB using ChEBI. Bioinformatics. 2023;39(1):btac793.
24. Krallinger M. Overview of the Chemical-Protein relation extraction track. In: Proceedings of the BioCreative VI workshop. 2017 Oct 20; Bethesda, Maryland, USA. p.141–146
25. Segura-Bedmar I, Mart´ınez P, Herrero-Zazo M. SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In: Proceedings of Seventh International Workshop on Semantic Evaluation. 2013 June 14–15;Atlanta, Georgia, USA. p. 341–350
26. Miranda A, Farrokh M, Luoma J, Sampo P, and Alfonso A, Martin K. Overview of DrugProt BioCreative VII track: quality evaluation and large scale text mining of drug-gene/protein relations. Proceedings of the seventh BioCreative challenge evaluation workshop, 2021 Nov 8–10.
27. Wei CH, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, Wiegers TC, Lu Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. Database. 2016;2016:baw032.
28. Apache Hbase Team. Apache HBase™ Reference Guide. https://hbase.apache.org/book.html. Accessed 30 Jan 2024
29. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. https://ar5iv.org/abs/1907.11692. Accessed 30 Jan 2024
30. Avram S, Wilson TB, Curpan R, Halip L, Borota A, Bora A, Bologa CG, Holmes J, Knockel J, Yang JJ, Oprea TI. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. Nucleic Acids Res. 2023;51(D1):D1276–87. https://doi.org/10.1093/nar/gkac1085.
31. Kim J, So S, Lee HJ, Park JC, Kim JJ, Lee H. DigSee: disease gene search engine with evidence sentences (version cancer). Nucleic Acids Res. 2013;41:510–7. https://doi.org/10.1093/nar/gkt531.
32. http:// https://www.drugs.com/drug_interactions.html, Accessed 30 Jan 2024
33. Björne J, Heimonen J, Ginter F, Airola A, Pahikkala T, Salakoski T. Extracting complex biological events with rich graph-based featuresets. 2009 In: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Boulder, CO, USA, pp. 10–18
34. Nentidis A, Katsimpras G, Krithara A, Paliouras G. BioASQ-QA: a manually curated corpus for biomedical question answering. Sci Data. 2023;10:170. https://doi.org/10.1038/s41597-023-02068-4.

Publisher's Note