Received: 1 February 2025 | Revised: 12 May 2025 | Accepted: 30 May 2025 | Published online: 19 June 2025

Medinformatics
2025, Vol. 2(4) 241–255
DOI: 10.47852/bonviewMEDIN52025348

REVIEW

# Literature-Based Discovery (LBD): Towards Hypothesis Generation and Knowledge Discovery in Biomedical Text Mining

**Balu Bhasuran[1,2], Gurusamy Murugesan[3] and Jeyakumar Natarajan[1,3,*]**

*[1]DRDO-BU Center for Life Sciences, Bharathiar University, India*

*[2]School of Information, Florida State University, USA*

*[3]Department of Bioinformatics, Bharathiar University, India*

**Abstract:** Biomedical knowledge is growing at an astounding pace with a majority of this knowledge represented as scientific publications. Text mining tools and methods represent automatic approaches for extracting hidden patterns and trends from this semi-structured and unstructured data. In biomedical text mining, literature-based discovery (LBD) is the process of automatically discovering novel associations between medical terms otherwise mentioned in disjoint literature sets. LBD approaches have proven to successfully reduce the discovery time of potential associations that are hidden in the vast amount of scientific literature. The process focuses on creating concept profiles for medical terms such as a disease or symptom and connecting them with a drug and treatment based on the statistical significance of the shared profiles. This knowledge discovery approach introduced in 1989 remains a core task in text mining. Currently, the ABC principle-based two approaches namely open discovery and closed discovery are mostly explored in the LBD process. This review starts with a general introduction about text mining followed by biomedical text mining followed by a brief introduction of the core ABC principle and its associated two approaches open discovery and closed discovery in the LBD process. This review discusses the deep learning applications in LBD by reviewing the role of transformer models and neural networks-based LBD models and their future aspects. Additionally, the potential of Large Language Models in enriching the LBD process is discussed with challenges and solutions and finally reviews the key biomedical discoveries generated through LBD approaches in biomedicine and concludes with the current limitations and future directions of LBD.

**Keywords:** biomedical text mining, literature-based discovery, ABC principle, open and closed discovery, novel associations, concept profiles

## 1. Introduction

Key findings and insights from scientific research and clinical investigations frequently appear as unstructured text in publications and clinical records [1]. With the ongoing advancements in biomedical research, the volume of published literature has experienced rapid growth in recent years [2]. Consequently, scientists and clinical researchers face considerable difficulties in remaining up-to-date and uncovering hidden insights from the massive corpus of millions of biomedical publications [3]. This extensive body of unstructured data brings challenges related to data collection, management, exploration, and the discovery of new knowledge. A promising, comprehensive solution to these issues is biomedical text mining (BTM) [4].

According to Hearst [5], text mining (TM) can be defined as “the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources”. TM is the process of generating high-quality information in the form of novel, relevant, and interesting patterns, trends, facts, or hypotheses by sifting through a large volume of unstructured data [6–9]. The process of TM pipeline consists of Information Retrieval (IR), Information Extraction (IE), and Knowledge Discovery and Hypothesis Generation [10–14]. In the context of TM, IR is the process of finding relevant natural language text from a set of literature-based databases. Normally, IR is performed as a query-based or document-based search for retrieving abstract or full text from digital libraries or databases [15–17]. IE can be defined as the automatic process of extracting structured information from semi-structured and/or unstructured machine-readable text [18]. The sole purpose of automated TM is the discovery of new knowledge and the generation of new ideas or hypotheses from literature by Zeng et al. [19].

BTM is concerned with the extraction of information regarding biological entities and their relationships, such as genes and proteins, diseases, drugs, cell type, miRNA, phenotypes, or even more broadly biological events and pathways from the scientific text [20–24]. Furthermore, the extracted information has been used for hypothesis generation, knowledge discovery, annotation of specialized databases, tools and manual curation of biological

[*] **Corresponding author:** Jeyakumar Natarajan, DRDO-BU Center for Life Sciences, Bharathiar University and Department of Bioinformatics, Bharathiar University, India. Email: n.jeyakumar@buc.edu.in

databases such as infer novel relationships: fish oil and Raynaud disease, magnesium deficiency and migraine, creation of databases CTD, OMIM, DisGeNET, STRING, building sophisticated web servers PubTator, mirCancer, PolySearch, DISEASES, PKDE4J, and formation of discovery platforms such as BEST, DigSee, Beegle, and Implicitome [25–33]. Thus, BTM has become an integral part of many resources serving a wide audience of researchers and scientists [34–41]. In general, the BTM pipeline encompasses the three core steps of traditional TM: IR, IE, and Knowledge Discovery from Text (KDT).

Schematic flowchart for identifying, screening, and including relevant studies for this review is shown in Figure 1. In this review, we have collected all the relevant articles from PubMed search engine using words and subwords representing LBD such as "knowledge discovery" and "literature based discovery". The selected literatures were screened based on relevancy and topic interest which resulted in a total of 156 articles. With other conditions such as out of scope or in sufficient details resulted a total of 84 articles to be included in this review.

Within the KDT framework, the primary objective of TM is to generate hypotheses with a high degree of reliability by traversing and linking numerous biomedical concepts drawn from separate bodies of literature [42, 43]. Discovery platforms and literature-wide analysis studies (LWAS) that aim to address these issues are collectively termed Literature-Based Discovery (LBD) in TM [44]. In BTM, the Knowledge Discovery process is performed as a novel connection between medical terms or biological network analysis and prioritization methods. Disease-specific case studies, drug searches for cancers by integrating pathways and molecules, gene prioritization, and global disease network generation are some major examples of these kinds in BTM [45]. Knowledge Discovery is sometimes referred to as hypothesis generation. Hypothesis generation is the process of generating unknown facts by utilizing information discovered with the use of IR and IE. Generating hypotheses in the biomedical field is a significant task to infer unknown biomedical facts that can be used to guide the design of experiments in the future or explain existing experimental results [46]. This, in turn, helps to determine new

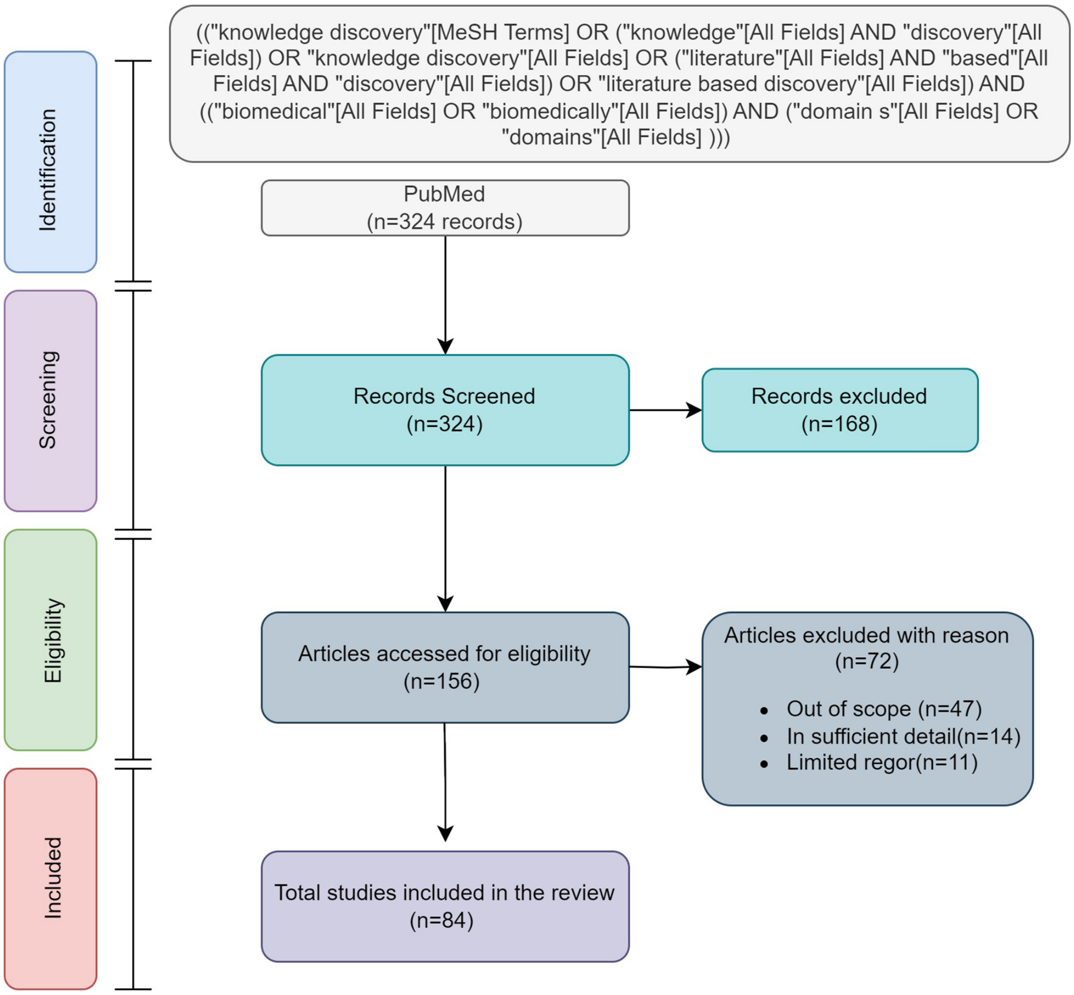
Identification
Screening
Eligibility
Included
(("knowledge discovery"[MeSH Terms] OR ("knowledge"[All Fields] AND "discovery"[All Fields]) OR "knowledge discovery"[All Fields] OR ("literature"[All Fields] AND "based"[All Fields] AND "discovery"[All Fields]) OR "literature based discovery"[All Fields]) AND (("biomedical"[All Fields] OR "biomedically"[All Fields]) AND ("domain s"[All Fields] OR "domains"[All Fields] )))
PubMed (n=324 records)
Records Screened (n=324)
Records excluded (n=168)
Articles accessed for eligibility (n=156)
Articles excluded with reason (n=72)
• Out of scope (n=47)
• In sufficient detail(n=14)
• Limited regor(n=11)
Total studies included in the review (n=84)

**Figure 1. PRISMA style flowchart for the literature selection process in this review. A total of 324 articles were retrieved from PubMed using the specified query for literature-based discovery (LBD) in the biomedical domain. After initial screening, 168 records were excluded. Of the remaining 156 articles assessed for eligibility, 72 were excluded based on the following predefined criteria: (A) Out of scope—articles not focused on LBD or not within the biomedical context ($n = 47$); (B) Insufficient detail—lacking methodological description, experimental setup, or results ($n = 14$); and (C) Limited rigor—studies without empirical validation or use of benchmark data ($n = 11$). The final set of 84 articles met the inclusion criteria, which required peer-reviewed publications explicitly describing LBD methods (co-occurrence, semantic, deep learning, or LLM-based), applied within biomedical research between 1986 and 2023.**

drug targets or novel interactions between biomedical concepts that have not been proved before [47].

Our review addresses a critical gap in the literature by providing a comprehensive, end-to-end synthesis of LBD in BTM, which is currently fragmented across co-occurrence methods, semantic approaches, graph-based models, and more recently, deep learning and Large Language Models (LLM)-based paradigms. While previous reviews have primarily focused on classical LBD techniques or isolated case studies, our manuscript uniquely integrates:

1) Foundational LBD principles (e.g., ABC model) with
2) Modern neural architectures (e.g., AGATHA, LINE, CNNs) and
3) Emerging LLM applications (e.g., GPT-4 and RAG frameworks)

This is the first review, to our knowledge, that holistically traces the evolution from Swanson's foundational models to LLM-augmented hypothesis generation systems, contextualizing each advancement in light of biomedical literature growth and computational scalability.

We also emphasize the urgency and significance of this review given the exponential rise in biomedical publications and the corresponding need for scalable, AI-enabled tools to derive insights and generate hypotheses. We further argue that the fusion of transformer-based models and knowledge graphs (KGs), as explored in our review, will define the next generation of LBD systems, making our study both forward-looking and practically relevant.

One of the well-known approaches to this task is proposed by Swanson [48, 49] using the ABC principle to link disjoint literature sets for biomedical knowledge discovery [50]. The ABC principle states that if concept A and concept B were associated directly in one set of literature, while concept B and concept C were in direct relation to an independent disjoint set, then the union of these literature sets allows a new possible inference relation between concepts A and C linking via the concept B.

Kastrin and Hristovski [51] did the first inclusive scientometric overview of the LBD study covering 35 years (1986–2020) using 409 documents from six bibliographic databases. The overview found Rindflesch TC, Kostoff RN, Hristovski D, Smalheiser NR, and Swanson DR as the top five authors in LBD based on several publications. The study also generated a Co-authorship network and document co-citation network in this domain and top journals publish the studies in the LBD domain. We recommend this study for a better understanding of LBD in biomedicine, its origin, and evolution [51].

Concept profiles are the major component in the LBD study. Generating a concept profile involves systematically gathering and organizing information related to a specific concept. Begin by clearly defining the concept and conducting a thorough search of academic databases to collect relevant literature, using specific keywords, synonyms, and related terms to ensure comprehensive coverage. Extract key terms, phrases, and themes from the literature, focusing on definitions, attributes, functions, interactions, and significant findings. Identify related concepts frequently associated with the primary concept and analyze the context and connections to determine interactions, focusing on cause-effect relationships and correlations. Details of the concept profiles are included in the Literature-Based Discovery Methods section. For LBD studies using biomedical text, the following sources can be useful for concept recognition, curation, and normalization.

## 2. Auxiliary Knowledge Sources

### 2.1. Unified medical language system

Unified medical language system (UMLS) enables semantic understanding and interoperability among different software applications and systems by combining widely used dictionaries in the biomedical field. UMLS contains three knowledge sources namely (i) Metathesaurus (ii) Semantic Network (iii) SPECIALIST Lexicon.

1) Metathesaurus: It is the main component of the UMLS, and it is organized by combining various biological concepts (such as Gene, Protein, Disease names, etc.). UMLS utilizes a metathesaurus to connect the alternative names of the same concepts from various sources of dictionaries. Metathesaurus not only links the same concepts from various sources but is also used to identify relationships among different concepts.
2) Semantic network: The biological concepts described in the UMLS Metathesaurus are grouped into subject categories called semantic types. For example: The concept of breast cancer belongs to the semantic type ["Disease or Syndrome"] and magnesium is categorized as ["chemical"]. UMLS also contains the relation between these semantic types called "semantic relations".
3) SPECIALIST lexicon: The SPECIALIST Lexicon contains the information (word usage) used by the natural language processing (NLP) processing. Each entry in this lexicon includes the morphological, syntactic, and orthographic information for each word or term.

### 2.2. Medical subject headings (MeSH)

MeSH was introduced by NLM for indexing and retrieval of PubMed articles, MeSH terms provide abstract or summarized biological concepts used in the paper. MESH terms are classified into three sub-types: (i) Descriptors: denote the main concepts of the article described for example, if an article explores the role of magnesium deficiency in neurological disorders, it would be indexed under the descriptors "Magnesium Deficiency" and "Neurological Disorders." Descriptors are standalone terms compared to other terms. (ii) Qualifiers: mainly useful if it is used in conjunction with descriptors. (iii) Supplementary Concept Records (SCR): SCR index named entities associated with the article such as gene, disease name, chemical, etc. Apart from the above three sub-types, MESH also contains a code called MeSH tree code which is arranged hierarchically. Thus, the MESH concept provides an effective way of searching for articles on specific biomedical subjects.

### 2.3. SemMedDB

Semantic relations are important for TM tasks such as knowledge discovery and hypothesis generation. SemMedDB [52] is the repository of semantic relations extracted from PubMed articles titles and abstracts using the rule-based system called SemRep. SemMedDB Contains the predictions of (subject-predicate-object) triples from the PubMed articles. SemMedDB uses UMLS Metathesaurus for concept extraction and relation extraction, it uses the Semantic Network concept. The Semantic MEDLINE Database (SemMedDB) indexes semantic predications triples (subject-predicate-object) extracted by the semantic interpreter SemRep from PubMed citations.

#### *2.3.1. SemRep tool*

SemRep is a program that extracts three-part clauses, called semantic predicates, from sentences contained in biomedical text. The predicate consists of a subject argument, an object argument, and the relationship between them. This is a UMLS-based program that uses UMLS metathesaurus concepts and their associative relationships to extract predicates. SemRep is available

**Table 1. Literature sources and NLP tools in text mining for LBD**

| Type | Name | Web-Link | Type | Current status |
|---|---|---|---|---|
| Literature Sources | MEDLINE | https://www.nlm.nih.gov/medline/index.html | Online DB | Working |
| | Scopus | https://www.elsevier.com/en-in/solutions/scopus | Online DB | Working |
| | Science Direct | https://www.sciencedirect.com/ | Online DB | Working |
| | Europe PMC | https://europepmc.org/ | Online DB | Working |
| | bioRxiv | https://www.biorxiv.org/ | Online DB | Working |
| NLP Tools | SemRep | https://semrep.nlm.nih.gov/ | Standalone/Downloadable | Working |
| | MetaMap | https://metamap.nlm.nih.gov/ | Standalone/Downloadable | Working |
| | cTAKES | https://ctakes.apache.org/ | Standalone/Downloadable | Working |

as a standalone program on the Linux platform and can be run interactively or in batch mode using the SKR scheduler.

The current version of SemMedDB provides approximately 96.3 million predictions from SemRep using 29.1 million citations from the MEDLINE database. This PubMed scale MySQL database provides information about the PubMed citation, One-to-many relationships of the concept with UMLS metathesaurus information, and links between predictions and between a prediction and a sentence.

Literature sources and NLP tools in TM for LBD are given in Tables 1 and 2 provide curated knowledge sources for LBD in biomedicine below.

## 3. LBD Methods

This section briefly discusses the various terms and concepts explored in LBD such as the ABC principle, Concept profile, and open and closed discovery process followed by key applications and discoveries in the BTM domain.

### 3.1. ABC principle

Using BTM integrated with prior knowledge (already known biomedical associations), intuition, and experience scientific discoveries are made in biomedicine. TM plays an important role by assisting this process using automatically discovering interesting novel hypotheses. In this regard, TM researchers adapted the widely explored ABC principle for hypothesis generation and knowledge discovery [49]. The ABC principle can be stated as, If concept A and concept B are directly associated in one body of literature, while concept B and concept C are similarly linked in a separate, disjoint body of literature, then merging these two sets enables a new possible inference of a connection between concepts A and C via concept B [53]. This approach enables the establishment of a new link between Concept A and Concept C via shared Concept B. The concepts can be any relevant biological entity such as gene, protein, drug, disease, cell line, miRNA, antibody, peptide, or common medical terms from repositories like MeSH [54].

The two core aspects of the ABC principle that are explored widely in LBD studies are the identification of shared(related) interesting(promising) concepts and the exploration of those relationships based on certain particulars (biologically relevant). For example, the starting point of the search can be a disease X [55]. The first step of the approach will be identifying related medical concepts to X such as a particular biomarker gene Y1 or drug Y2. In the second step, the relationship is further explored in such a way that if the mutation in the Y1 gene that causes disease X, or drug Y2 is likely to treat disease X or aggravate X [56–58]. The approach aims to derive some kind of novelty in the relationship and discover more Y concepts (chemical, miRNA, or medically relevant reactions) thereby creating more new possible hypotheses between X and Y [59–61]. According to the way the concepts are searched, the approaches are classified as open discovery and closed discovery. An open discovery process aims for hypothesis generation by navigating through connected concepts at different levels [56, 62–65]. A closed discovery process starts with known concepts at both ends A and C

**Table 2. Curated knowledge sources for LBD in biomedicine**

| Curated knowledge source | Web-Link | Type | Current status |
|---|---|---|---|
| **OMIM** (Online Mendelian Inheritance in Man) (*Hamosh et al., 2005*) | http://www.ncbi.nlm.nih.gov/omim | Online DB | Working |
| **CTD** (Comparative Toxicogenomics Database) (*Davis et al., 2018*) | http://ctd.mdibl.org | Online DB | Working |
| **STRING** (*Szklarczyk et al., 2017*) | http://string-db.org/ | Online DB | Working |
| **DisGeNET** (*Pinero et al., 2017*) | http://www.disgenet.org | Online DB | Working |
| **PharmGKB** (Pharmacogenomics Knowledgebase) (*Thorn et al., 2013*) | http://www.pharmgkb.org | Online DB | Working |
| **UniProt** (*The UniProt Consortium, 2018*) | http://www.uniprot.org/ | Online DB | Working |
| **MEDIC** (merged disease vocabulary) (*Davis et al., 2012*) | http://ctd.mdibl.org/voc.go?type=disease | Online DB | Working |
| **DO** (Disease Ontology) (*Kibbe et al., 2015*) | http://www.disease-ontology.org | Online Tool | Working |
| **UMLS** (Unified Medical Language System) (*Olivier Bodenreider, 2004*) | http://umlsks.nlm.nih.gov | Standalone/ Downloadable | Working |
| **MeSH** (Medical Subject Headings) (*CE Lipscomb, 2000*) | https://www.nlm.nih.gov/MeSH/ | Online Tool/ Standalone | Working |
| **SNOMED CT** (Systematized Nomenclature of Medicine-Clinical Terms) (*Elkin et al., 2006*) | http://www.snomed.org/ | Online DB | Working |

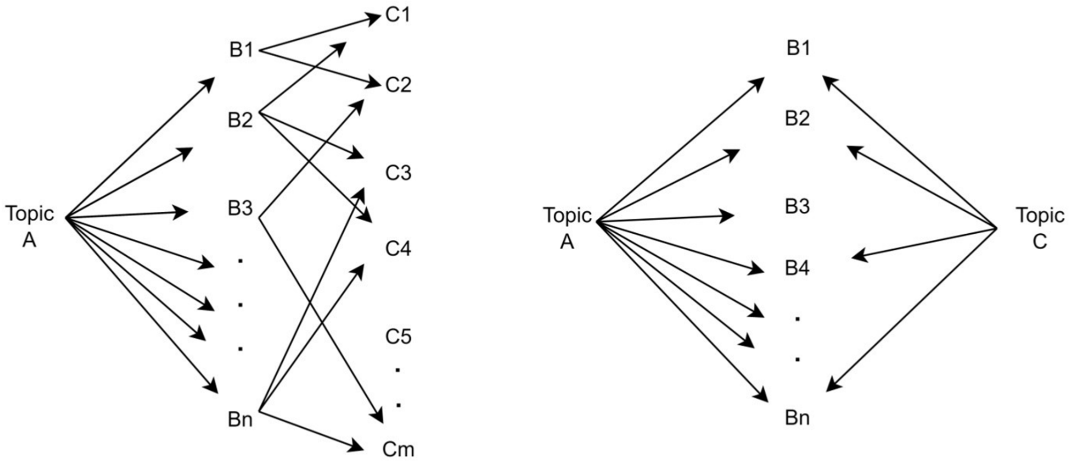

**Figure 2. Open and closed discovery approaches in literature-based discovery**

respectively. In this process, the approach searches for B terms that can support the claim that the A–C association is a relevant one [66–70]. A schematic representation of open and closed discovery approaches in LBD is depicted in Figure 2 [79].

## 3.2. Concept profile

The concept profile of a biological entity represents a set of terms that are related to the entity either through curated known association or through a co-occurrence mentioned in a biologically relevant context [71]. Consider a topic such as Alzheimer's disease (AD), which is an irreversible, progressive brain disorder. The profile for this topic distilled from a suitable text collection could identify, for example, terms representing the genes, proteins, symptoms, drug treatments, other diseases, and population groups associated with the disease, i.e., "statistically related" to it. In the majority of the cases biologically co-occurrence implies semantic association [72–75]. One way to create a concept profile is to apply MeSH metadata on MEDLINE databases using dictionary matching or automatic concept identifiers such as MetaMap or cTAKES. A concept profile can be represented as,

$$\text{Profile}(\mathbf{T}_i) = \left\{w_{i,1} \times m_1, \cdots, w_{i,j} \times m_j, \cdots, w_{i,n} \times m_n\right\} \quad (1)$$

where $j = 1, 2, ..., n$ and $m_j$ denotes the $j$-th MeSH term and $w_{i,j}$ is the weight representing the association between concept **T** and the MeSH term $m_j$. There are a total of n MeSH terms in the vocabulary.

## 3.3. Open discovery

An open discovery process aims for hypothesis generation by navigating through connected concepts at different levels. Initially, there is only the starting concept that can be a scientific problem or research question and the end of the discovery is not defined. For example, Swanson's [48, 49] initial study was to find a new treatment for Raynaud's disease. The discovery approach uses disease C as the initial concept searches for interesting clues (B), typically treating drugs, molecular pathways, or physiological processes that play a role in the disease under scrutiny. Next, the approach finds A-terms, typically substances/drugs or pathways, that act on the selected Bs. The major challenge of open discovery support tools is to contain the vast amount of possibilities identified in the initial search. Finally, a hypothesis can be formulated like the substance A can be used for the treatment of disease C. Since search space is expanded into multiple levels due to connected concepts, a better understanding of the problem with domain knowledge is necessary for open discovery [76–78].

## 3.4. Closed discovery

A closed discovery process starts with known concepts at both ends A and C respectively. In this process, the approach searches for B terms that can support the claim that the A–C association is a relevant one. For the sample example as discussed above the closed discovery approach starts from both disease C and substances/drugs A, the approach searches for common intermediate B terms. The more pathways or physiological processes between A and C in the search results, the more likely this hypothesis is a valid one [79–81]. Due to the simplicity and better search paradigm, most of the LBD approaches are focused on closed discovery. Since both concepts are already known, the approaches simply search for B terms between them. Various association hypotheses such as gene-disease and drug-disease are generated using a closed approach [82–87]. Novel potentially relevant interesting and spurious biological link identification using open and closed discovery approaches is depicted in Figure 3.

## 4. Applications in BTM

Because of the sheer volume of biomedical research literature, scientists face significant challenges when trying to sift through all relevant articles on a particular disease, gene, chemical, or miRNA to formulate research hypotheses or uncover novel connections. One of the early landmark studies in LWAS was conducted by Swanson DR, who pioneered a hidden relationship model by examining disjoint sets of literature. This approach led him to propose several innovative hypotheses—such as the links between magnesium and migraine, fish oil and Raynaud's syndrome, and somatomedin C and arginine [48, 49, 52]. None of these connections had been predicted or reported before, but they were subsequently confirmed, marking the beginning of a new era in BTM. Another well-known TM-based knowledge discovery system developed by Smalheiser et al. [87] named Arrowsmith uses B-term phrases and title words connecting the articles with a two-node approach-based searching [88]. Following this root and adapting the famous ABC principle, Hristovski et al. [64] released BITOLA, a MEDLINE database-based meaningful relation generator using user-given MeSH terms as pivot concepts. The web server expects the user to give a meaningful concept and incorporates external

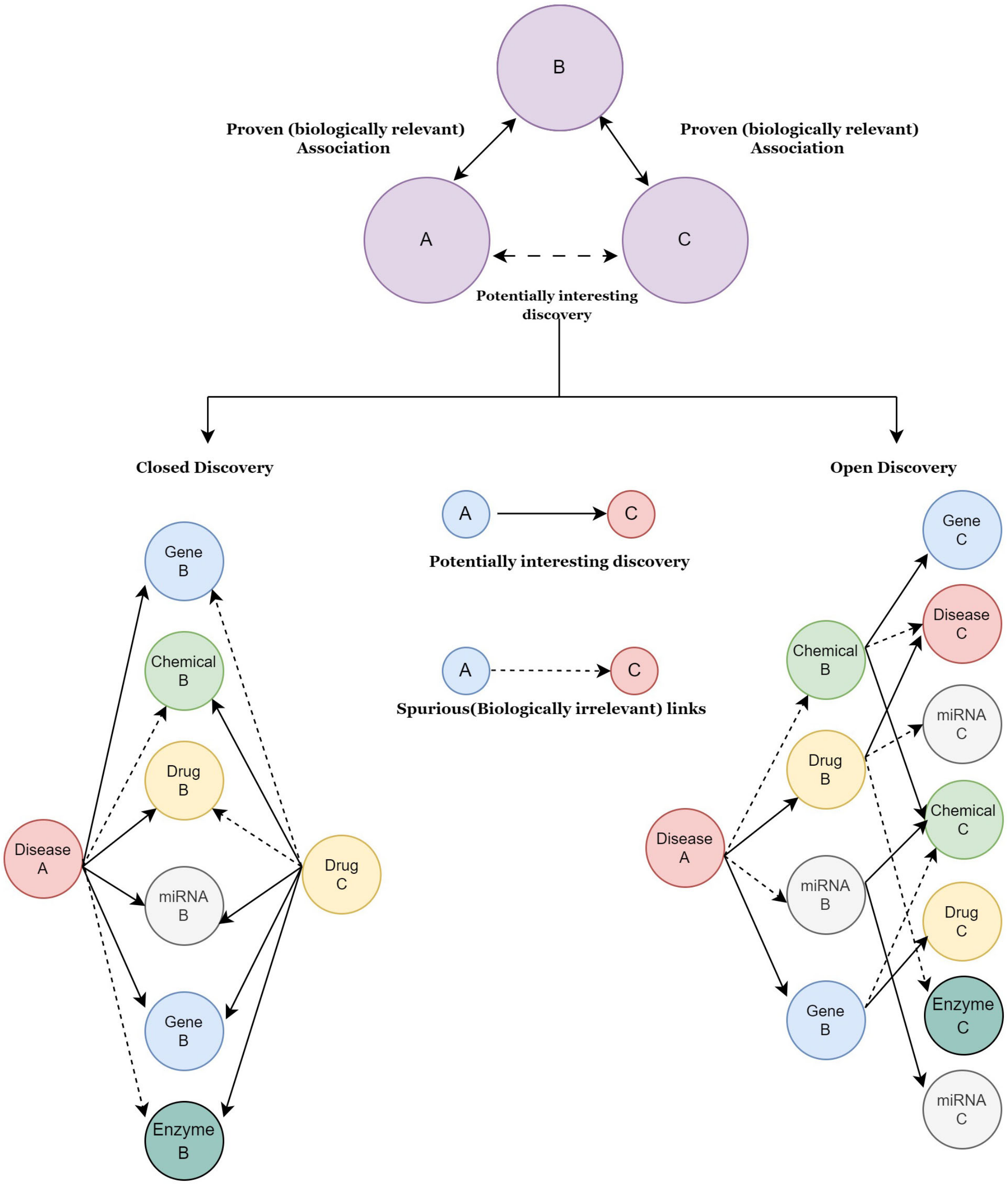
B
Proven (biologically relevant) Association
Proven (biologically relevant) Association
A
C
Potentially interesting discovery
Closed Discovery
Open Discovery
A
C
Potentially interesting discovery
A
C
Spurious(Biologically irrelevant) links
Gene B
Chemical B
Drug B
Disease A
miRNA B
Drug C
Gene B
Enzyme B
Gene C
Disease C
Chemical B
miRNA C
Drug B
Chemical C
Disease A
miRNA B
Drug C
Gene B
Enzyme C
miRNA C

**Figure 3. Novel potentially relevant interesting and spurious biological link identification using open and closed discovery approaches. Through closed discovery, Disease A and Drug C are connected through gene B and miRNA B whereas other discoveries are not relevant. Through open discovery, Disease A to Drug C is connected via Gene B, Disease A to Disease C is connected via Drug B, and other connections are not relevant.**

knowledge sources such as a chromosomal location for performance improvement [64]. Another major real-time discovery tool FACTA+ created by Tsuruoka et al. [88] is based on concept co-occurrence at the abstract level integrating hidden association generation, bio-molecular events, and network visualization [89]. Fleuren and Alkema [17] developed CoPub 5.0, an integrative framework with co-occurrence and keyword-based searching, ABC principle-based hidden connection, and Cytoscape software-based network construction. CoPub 5.0 has three search modes namely term search (to retrieve abstract and keyword relation extraction for a particular term), pair search (analyze the new relation or known relation), and set of terms (relation between multiple terms) to answer biological questions. Figure 4 shows various approaches and examples in LBD systems such as co-occurrence bases, semantic relation-based, graph-based, and hybrid approaches. Table 3 represents details of knowledge discovery tools using BTM sources. Table 4 shows the major LBDs in biomedicine using the ABC principle.

Recent work by Tropmann-Frick and Schreier [89] discussed the various drug repurposing approaches for COVID-19 using LBD showing the potential and immediate applications of the field. They used three LBD systems Arrowsmith, BITOLA, and SemBT for the search of repurposable drugs for COVID-19 using the ABC principle. With a closed discovery approach using Arrowsmith, they used COVID-19 and the drug “remdesivir” as A and B concepts, for open discovery they used “molecular mechanisms of pharmacological action” as the target concept.

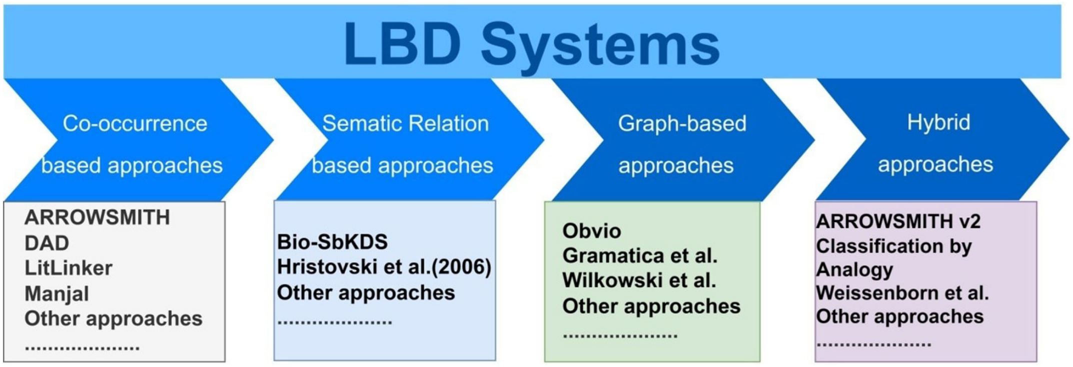

**Figure 4. Various approaches and examples in LBD systems**

Using BITOLA they used SARS and "chloroquine" in closed discovery and SARS and "lactate dehydrogenase" as source and target concepts, respectively. The study generated rank frequency, rank coefficient, frequency (AB, BC), novel discovery status, confidence values, etc., to statistical insights into the discovery [89].

One improvement to the ABC principle is discussed by Baek et al. [90], in their study of plausible new hypothesis generation from PubMed. They discussed two aspects namely context surrounding and clinical validation. The proposed solution emphasized multiple B terms in metabolite-related hypotheses with diverse biological types. The study found that lactosylceramide and arterial stiffness are associated with the involvement of a potential pathway connecting the entities and nitric oxide, malondialdehyde, and they clinically validated the generated hypothesis [90]. The same author further expanded the new ABC principle to context-based and context-assignment-based ABC models by using four biological context elements: cell, drug, disease, and organism. This study showed that there is a 50–70% improvement in precision for identifying an association between APOE–MAPT and FUS–TARDBP by comparing the co-occurrence-based ABC model with the context-based ABC models [91].

## 5. Deep Learning LBD Models

### 5.1. Transformer models

Contextual word embedding and transfer learning methods provided immense momentum and new dimensions to NLP. The biomedical domain adapted this momentum very quickly through various models pre-trained from weights such as BioBERT [92] or generated from scratch such as ELECTRAMed [93]. The baseline for these models is the Bidirectional Encoder Representations from Transformers (BERT) architecture, a bidirectional self-attention model that uses encoder layers for two tasks, masked language modeling and next sentence prediction. Lee et al. [92] proposed a domain-specific language model BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) by further training the weights of BERT (English Wikipedia and Books Corpus) using PubMed abstracts and PMC full-text articles. The BioBERT model base and large versions were shown to outperform general BERT models and other biomedical models in three tasks, namely Named Entity Recognition in 9 biomedical datasets, Relation Extraction in 3 biomedical datasets, and Question Answering in 3 biomedical datasets [92]. The wider success of BioBERT enabled the researchers to develop new transfer learning models using biomedical and clinical literature resulting in PubMedBERT [94], ClinicalBERT [95], MT-clinical BERT [96], Umlsbert [97], ELECTRAMed [93], BioMegatron [98], etc. Most of these models are pre-trained with various combinations of scientific literature data PubMed and PMC, clinical data MIMIC-III, and biological databases MeSH and UMLS Metathesaurus.

Even though these models are trained with entire biomedical literature (29 M) knowledge and are performing well in entity recognition, linking, and summarization tasks, a well-designed LBD task is not yet widely formulated using biomedical transfer learning models. Very recently, Sybrandt et al. [99] proposed AGATHA, a graph-based transformer model for hypothesis generation. This deep learning-based system using SciBERT tested using a temporal holdout set used a data-driven ranking criteria for generating new biomedical connections. The study constructed large-scale semantic graphs containing over 10 billion edges, representing sentences, entities, n-grams, lemmas, and terms derived from UMLS and MeSH, along with predictors from SemRep. Term pairs were ranked, and the generated hypotheses were validated using methods such as Heuristic-Based Ranking, Subdomain Recommendation, Edge2Vec Comparison, and Ablation Studies [99]. These new approaches demonstrate that neural word embeddings, such as BERT—which capture sentence context and perform well across multiple prediction tasks—hold significant potential for literature-based discovery (LBD). Deep learning-based discovery models like AGATHA which is exploiting the association of contextual vector representations and graph neural networks (NNs) are opening a new dimension for researchers in biomedicine.

Very recently, Millikin et al. [100], introduced Serial KinderMiner (SKiM) available as an open-source tool and web interface which identifies ABC linkages for LBD discoveries and additionally reported the results for drug repurposing and a case study using Cancer literature. The authors further supplemented the SKiM tool with a KG and transformers to interpret the discoveries. As step 1, the proposed algorithm used case-insensitive matching of the user-given terms to get the indexed PubMed abstracts and extracted the co-occurred terms. In step 2, a KG was generated from 34 million PubMed abstracts using entities and relation identified by a fine-tuned PubMedBERT using annotated data. Finally, an ABC relation is provided only if it is present in both co-occurrence extraction and in the KG [100].

### 5.2. NNs

Crichton et al proposed four graph-based, NN methods using Large-scale Information Network Embedding (LINE) in open and closed discovery and compared the performance with the LION LBD system in the context of cancer case discoveries and a

**Table 3. Detailed representation of knowledge discovery tools using biomedical text mining**

| Discovery tool | Web-Link | System description | Type | Current status |
|---|---|---|---|---|
| DigSee (*Kim et al., 2017*) | http://gcancer.org/digsee | Direct(explicit) gene-disease associations from genes involved in the bio-molecular events with sentence scoring | Online Tool | Inoperative |
| LION LBD (*Pyysalo et al., 2018*) | http://lbd.lionproject.net | Implicit and Explicit associations generation using mapped ontology and concept graph with a special emphasis on Cancer | Online Tool | Working |
| The Implicitome (*Hettne et al., 2016*) | http://knowledge.bio | Indirect(implicit) gene-disease associations using concept profiles using the ABC principle and association score | Online Tool | Inoperative |
| Textpresso Central (*Muller et al., 2018*) | http://www.textpresso.org/tpc | In-depth search and annotation tool with customization and integration option | Online Tool | Inoperative |
| Beegle (*ElShal et al., 2015*) | http://beegle.esat.kuleuven.be/ | Implicit and explicit associations identified through co-occurrence and concept profile, integrated with a prioritization tool | Online Tool | Inoperative |
| CoPub 5.0 (*Fleuren et al., 2011*) | http://www.copub.org | Integrative framework with co-occurrence and keyword-based searching, ABC principle-based hidden connection | Online Tool | Inoperative |
| GS2D (*Miguel A. Andrade-Navarro and Jean Fred Fontaine, 2016*) | http://cbdm.uni-mainz.de/geneset2diseases | Direct(explicit) gene-disease associations with co-occurrence statistics and disease enrichment analysis | Online Tool | Available |
| FACTA+ (*Tsuruoka et al., 2011*) | http://refine1-nactem.mc.man.ac.uk/facta/ | Concept co-occurrence at the abstract level integrating hidden association generation, bio-molecular events, and network visualization | Online Tool | Inoperative |
| Anni 2.0 (*Jelier et al., 2008*) | http://www.biosemantics.org/anni | Implicit and Explicit associations with co-occurrence and ontologies | Online Tool | Inoperative |
| Arrowsmith (*Smalheiser et al., 2009*) | http://arrowsmith.psych.uic.edu | B-term phrases and title words connecting the articles with a two-node approach-based searching | Online Tool | Working |
| FACTA (*Tsuruoka et al., 2008*) | http://www.nactem.ac.uk/software/facta/ | Direct(explicit) associations with co-occurrence statistics and point-wise mutual information | Online Tool | Working |
| PolySearch (*Cheng et al., 2008*) | http://wishart.biology.ualberta.ca/polysearch | Large number of dictionaries and bag-of-words for direct(explicit) associations | Online Tool | Working |
| DISEASES (*Pletscher-Frankild et al., 2015*) | http://diseases.jensenlab.org/ | Direct(explicit) associations integrated with cancer mutation data and manually curated databases | Online Tool | Working |
| PolySearch2 (*Liu et al., 2015*) | http://polysearch.ca | Update of PolySearch with tightness measure based on word position | Online Tool | Working |
| Anni (*Jelier et al., 2007*) | http://www.biosemantics.org/Anni | Concept profile weighting using likelihood ration | Online Tool | Inoperative |
| BITOLA (*Hristovski et al., 2005*) | http://www.mf.uni-lj.si/bitola/ | User-given MeSH term as pivot concepts with external knowledge sources such as chromosomal location | Online Tool | Working |
| iTextMine (*Ren et al., 2018*) | http://research.bioinformatics.udel.edu/itextmine | Automated workflow with parallel processing for explicit associations | Online Tool | Working |
| DEXTER (*Gupta et al.,2018*) | http://biotm.cis.udel.edu/DEXTER | Disease expressions extraction with co-occurrence and argument filtering | Online Tool | Working |
| MELODI (*Elsworth et al., 2018*) | www.melodi.biocompute.org.uk | Graph-based database for mechanistic pathways identification | Online DB | Working |

**Table 4. Important literature-based discoveries in biomedicine using the ABC principle**

| Medical entity 1 (A) | Medical entity 2 (C) |
|---|---|
| Migraine | Magnesium |
| Raynaud disease | Fish Oil |
| Indomethacin | Alzheimer's Disease |
| Estrogen | Alzheimer's Disease |
| Calcium-Independent Phospholipase A2 | Schizophrenia |
| Magnesium deficiency | Neurologic |
| Thalidomide | Chronic Hepatitis C |
| Testosterone | Sleep |
| Somatomedin C | Arginine |
| Chlorpromazine | Cardiac Hypertrophy |
| Diethylhexyl (DEHP) | Sepsis |
| Sleep | Depression |

time-slicing based approach with post-cut-off publication year bases evaluation sets [101]. Baseline models were generated using 8 co-occurrence metrics namely Co-occurrence count, Document count, Jaccard Index, Symmetric conditional probability, Normalized point-wise mutual information, Chi-squared ($\chi 2$), Student's *t*-test (*t*-test), and log-likelihood ratio. The neural methods are proposed as link prediction where node embeddings are created using LINE along with Jaccard Index-based weighted edges [101].

For closed discovery, Multi-Layer Perceptron (MLP) architecture is trained as a classifier with the link between A and C being taken into account from the graph. The first model Closed discovery 1 (CD-1) generated a score for every A-B or B-C link as the second model Closed discovery 2 (CD-2) A-B-C embeddings link as the single input trained the model to assign a prediction score for the entire association A-B-C. The authors claimed that this approach gives more flexibility in terms of the length of the association (the number of B entities involved in the association). The Open Discovery 1 (OD-1) followed the same pipeline as CD-1 with a difference in using the accumulator function for ranking based on prediction scores. The final model Open Discovery 1 (OD-2) used a convolutional neural network (CNN) with a single vector input and prediction output pipeline. This model allowed the removal of aggregator and accumulator functions along with the merging of many paths from A-Bs to the same C concept. A link is given as a single-dimension input to CNN and generates an A-C link prediction score. To maintain a consistent input window size, the authors used elementwise summation and applied zero padding to fill any gaps. This CNN model used a ReLU activation, max pooling, and Softplus activation functions [101]. This final model also gives insights that we can use the other deep learning or transformer architectures such as BERT for LBS where a token length of 512 or fixed is expected as input. This approach also sheds light on the use of graph NNs in LBD.

Preiss [102] investigated a novel approach to drug repurposing by analyzing word evolution in biomedical literature. Traditional LBD connects knowledge pairs from separate publications, often leading to an overgeneration of potential repurposing candidates. Applications of NNs to LBD circumvent the first and last problems by utilizing text directly, eliminating the need for separate extraction of knowledge pairs. The authors propose further exploration of NNs, specifically by using word embeddings to detect changes in a drug's context before it is repurposed and by evaluating the accuracy of a model based on time series word embeddings to predict a drug's suitability for repurposing. This study proposes using word embeddings, which capture the context of words in chronological publication intervals, to detect changes in drug usage. By constructing time series word embeddings from MEDLINE abstracts and annotating repurposed drugs, the researchers trained deep learning models to predict repurposing potential. The study generates bi-monthly word embeddings from MEDLINE abstracts to create time series for each drug, focusing on changes in context over time. Two labeling methods, based on UMLS relations and SemRep extracted triples, were used to annotate instances of drug repurposing. Deep learning models were then trained using these time series and annotations, with 5-fold cross-validation performed to optimize hyperparameters and evaluate accuracy. The results show a prediction accuracy of 65% using UMLS labels and 81% using SemRep labels, demonstrating the method's effectiveness. The author claimed that the approach offers a scalable and data-driven alternative to traditional LBD methods, potentially expediting the identification of candidate drugs for repurposing [102]. Cuffy and McInnes [103] employed a NN architecture for LBD, representing terms as concepts through PubTator and embedding them using the LINE algorithm. The model receives A and C concept embeddings as input and generates probability distributions for all B terms, identifying explicit and implicit relationships. Three methods—concept averaging, concatenation, and the Hadamard product—are used to combine A and C embeddings, with feature scaling applied to enhance model performance. The model's output is represented in two ways: a reduced output with a subset of unique concepts and a full output encompassing all concepts in the dataset. Reduced output improves computational efficiency and model generalization, making it ideal for hypothesis testing, while the full output captures a wider range of relationships for hypothesis generation. The model is trained with binary cross-entropy loss and evaluated by ranking predicted B concepts against known B concepts. This approach significantly reduces the need for domain-specific knowledge and manual effort in hypothesis generation and testing [103].

Pu et al. [28] presented a graph embedding-based link prediction method for LBD in the context of AD. The researchers collected an AD-specific corpus and constructed a KG from it, annotating 16,452 papers published between 1977 and 2021 with relevant AD-specific concepts and relationships. They applied graph embedding techniques to predict new knowledge by identifying potential links between nodes in the graph. The study evaluates the impact of different link prediction models and time-sliced evaluation methodologies on the effectiveness of LBD. Results showed that the Structural Deep Network Embedding (SDNE) model consistently performed best in predicting links as knowledge evolved over 20 years. The study highlights the importance of considering varying prediction window lengths, as this can significantly impact the evaluation and interpretation of LBD models. The approach demonstrates potential for scalable knowledge discovery in AD and can be generalized to other diseases [29]. Wang et al. [104] introduced Contextual Literature-Based Discovery (C-LBD), a new approach for generating scientific hypotheses by integrating background contexts and seed terms into the hypothesis generation process. This method contrasts with traditional LBD, which typically focuses on predicting pairwise relations between discrete concepts without considering the broader context. The researchers propose a modeling framework that retrieves "inspirations" from past scientific papers to ground the generated hypotheses in relevant contexts, enhancing their relevance and novelty. The framework includes two tasks: idea-sentence generation and idea-node prediction, utilizing various retrieval modules such as semantic similarity, KGs, and citation networks. The study evaluates the performance of different models, including GPT-4, showing that while advanced models like GPT-4 can generate contextually relevant ideas, they often lack the technical depth and novelty found in human-generated scientific ideas. The research highlights the challenges and potential of using AI to assist in scientific discovery, demonstrating improvements over traditional LBD methods but also recognizing the need for further advancements in generating high-quality, innovative hypotheses [104]. To summarize the deep learning-based LBD approaches, we can use a transformer model like BERT for hypothesis generation by encoding sentences, entities, and their n-gram, to which semantic predictions and UMLS and MeSH data can be infused. It is also possible to train open and closed LBD systems using MLP and CNN-based deep learning architecture by formulating the ABC problem as link prediction using KGs where entities are represented with a graph NN and A-B-C links are given as sequence input with padding and the classifiers are generated a prediction score for A-C links. These approaches will pave the path to the future direction of LBD.

## 6. LLMs

Generative AI models, commonly known as LLMs [105] such as OpenAI GPT-4 (https://openai.com/research/gpt-4) or Google Med-PaLM [106], are one of the future directions of NLP and biomedical informatics where one multi-modal can perform various IE downstream tasks. There is ample support for the claim that LLMs can outperform existing state-of-the-art models in various NLP tasks, especially in biomedical informatics. Here we discuss some of the future directions of LBD using LLMs.

Nedbaylo and Dimitar [107] explored using ChatGPT for LBD to generate research hypotheses by identifying hidden connections within scientific literature. This study uses GPT-3.5 and GPT-4 models to simulate the discovery of relationships between medical concepts through prompt engineering. The researchers employed a bifurcated approach to prompt engineering, separating disease characteristics and potential interventions to reduce bias. The method involved using two distinct chat windows to reduce bias and enhance the creative potential of the LLMs. In the first window, the model generated a list of disease characteristics without revealing the disease name, aiming to identify general physiological and pathological traits. In the second window, based on these characteristics, the model proposed potential interventions, also without knowing the specific disease, to ensure the connections were made independently. The prompts were meticulously designed to be clear and specific, leveraging the LLM's inherent knowledge base to simulate the discovery of potential relationships, reminiscent of traditional LBD methods. The researchers used both zero-shot and few-shot learning techniques, evaluating the model's ability to generate novel connections and validating these outputs against existing scientific literature using the ScholarAI plugin. Results showed that while the models often defaulted to established medical knowledge, they occasionally produced novel hypotheses, such as the use of adaptogens like Rhodiola rosea for migraine management. However, the black-box nature of LLMs and the difficulty in steering them towards innovative connections present significant challenges. The major one was since GPT-4's training data are not publicly disclosed, there is no option to determine if the predictions are actually connecting two disjoint literature sets. With LLMs, it is challenging to know if the model is filling in gaps or relying on existing literature in its training data. Another major challenge was that without references for its outputs, tracing the LLM's "thought process" is difficult, making validation heavily dependent on empirical testing and clinical trials, which may not always be feasible. The authors also noted that even with a few-shot approach, prompting LLMs to generate novel hypotheses remains challenging, as the models tend to produce well-established or known information. Even with these challenges, the study underscores the potential of LLMs in enriching the LBD process but highlights the need for further research and advanced prompt engineering techniques to optimize their application. Studies like this open new dimensions in LDB using LLM.

One solution to this can be the use of Retrieval-augmented generation (RAG) [108]. Recently, LLMs have been queried with knowledge source-based prompts using RAG to improve the quality of the response along with necessary references. Generating discovery with reference allows researchers to verify the discovery process back to the root of LBD, the concept profiles. With additional prompting strategies like the chain of thought (CoT) [109], researchers can control the LLM discovery in the intermediate steps. Another approach will focus on prompting LLMs to search through the vast amount of text, find anomalies or less reported results compared to highly published association studies as well as validation of the novel discovery consistency with existing knowledge. In the future, we may see LBD systems deployed using AI agents where multiple LLMs work together for LBD.

## 7. Limitations and Future Directions of LBD

LBD has demonstrated significant potential in accelerating the identification of hidden associations within vast bodies of scientific literature. Since its introduction in 1989, it has remained a foundational technique in BTM. However, several limitations continue to hinder its full potential, particularly in the creation and use of underlying knowledge bases. Current approaches often rely on incomplete or inconsistently structured data. A key area for improvement lies in integrating advanced NLP tools with curated biomedical databases and bioinformatics resources. By enriching these knowledge repositories through precise concept extraction, semantic normalization, and multi-source integration, researchers can enhance both the quality and relevance of discovered associations.

Despite progress in NLP and TM, many LBD systems still fail to generate biologically or clinically valid concept linkages. This limitation is especially critical in biomedical domains where inaccurate associations can mislead downstream research. To address this, future systems should incorporate advances from KG networks, which provide structured and explainable representations of relationships between concepts, and social media mining, which can offer real-world contextual signals often missing in formal literature. Additionally, many existing systems do not adequately account for contextual relevance—often disregarding concept features distributed across multiple contexts or de-emphasizing medically important traits that fall outside the primary research focus. Other promising future directions include enhancing logic and reasoning capabilities for hypothesis generation, reducing dependence on manual curation and domain expertise, and developing fully automated LBD pipelines. Interactive and dynamic visualizations of discovered concepts and their interrelations could further improve user interpretability and engagement. More recently, NN and transformer-based architectures have been proposed for LBD tasks, bringing new opportunities for semantic generalization and automated inference.

A transformative shift in LBD research is now underway with the emergence of LLMs such as GPT-4, Gemini, and Llama. These

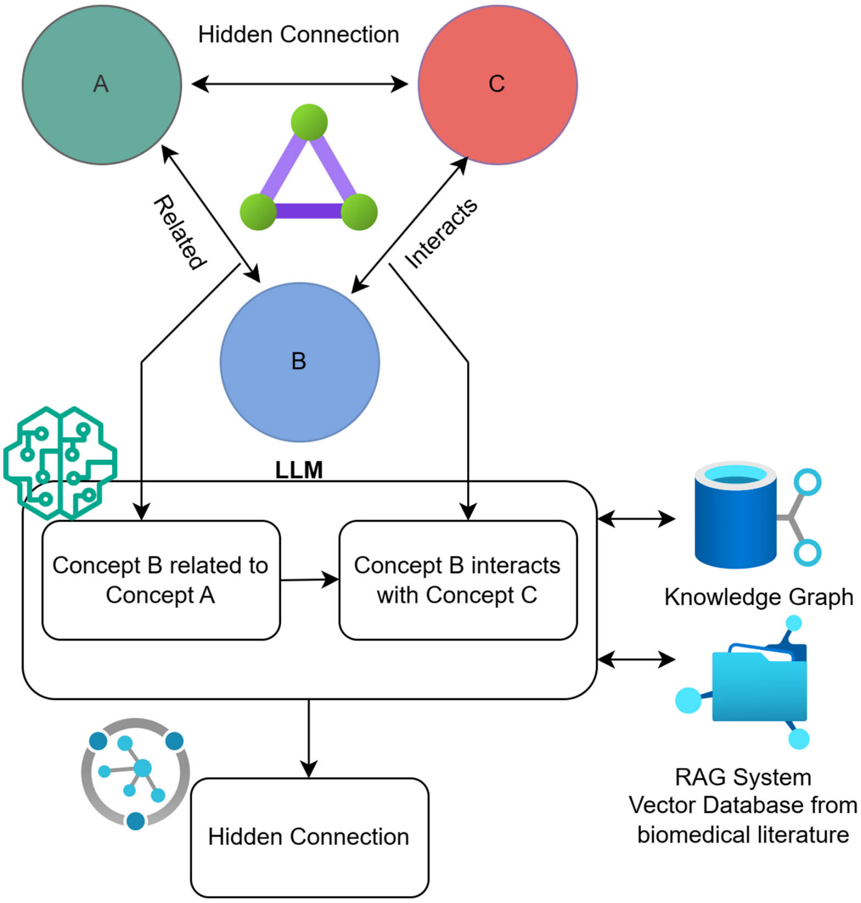

**Figure 5. Conceptual framework of literature-based discovery (LBD) using Large Language Models (LLMs). This illustration follows the classical ABC model where Concept A (e.g., disease) and Concept C (e.g., treatment) are indirectly linked through a shared Concept B (e.g., physiological trait or pathological feature). The LLM autonomously generates two steps: (A) identifying ConceptBs related to ConceptA and (B) determining interventions (ConceptCs) that interact with ConceptBs. This bidirectional discovery process is supported by retrieval-augmented generation (RAG) systems and knowledge graphs to validate or enrich the proposed relationships. The ultimate goal is to reveal a novel or "hidden connection" between A and C through structured reasoning.**

models, when integrated with LBD frameworks, can simulate the reasoning needed to identify latent connections between medical concepts. Figure 5 illustrates how LLMs operationalize the classical ABC model: first identifying intermediate features (Concept B) related to a known concept (Concept A), then independently generating candidate treatments or interventions (Concept C) that interact with Concept B. This enables the model to hypothesize indirect or hidden connections between A and C, potentially revealing novel therapeutic strategies. Critically, the integration of RAG systems and KGs significantly strengthens this LLM-driven discovery framework. While LLMs can generate plausible connections from pre-trained knowledge, they may lack access to real-time or domain-specific information. RAG systems address this gap by retrieving up-to-date biomedical literature from vectorized databases during inference, enhancing relevance and recency. In parallel, KGs provide structured validation of proposed relationships—linking drugs, diseases, symptoms, and pathways with logical consistency. Together, RAG and KG modules transform LLM-based LBD from a black-box heuristic process into a semi-transparent, evidence-grounded reasoning pipeline. This convergence of technologies marks a new era in hypothesis generation—one that blends generative intelligence with verifiable biomedical evidence.

## Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

## Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## Author Contribution Statement

**Balu Bhasuran:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Gurusamy Murugesan:** Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jeyakumar Natarajan:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration.

## References

[1] Nadif, M., & Role, F. (2021). Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, *22*(2), 1592–1603. https://doi.org/10.1093/bib/bbab016

[2] Zhao, S., Su, C., Lu, Z., & Wang, F. (2021). Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, *22*(3), bbaa057. https://doi.org/10.1093/bib/bbaa057

[3] Kononova, O., He, T., Huo, H., Trewartha, A., Olivetti, E. A., & Ceder, G. (2021). Opportunities and challenges of text mining in materials research. *iScience*, *24*(3), 102155. https://doi.org/10.1016/j.isci.2021.102155

[4] Wang, M., Wang, M., Yu, F., Yang, Y., Walker, J., & Mostafa, J. (2021). A systematic review of automatic text summarization for biomedical literature and EHRs. *Journal of the American Medical Informatics Association*, *28*(10), 2287–2297. https://doi.org/10.1093/jamia/ocab143

[5] Hearst, M. A. (1999). Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 3–10. https://doi.org/10.3115/1034678.1034679

[6] Cheerkoot-Jalim, S., & Khedo, K. K. (2021). A systematic review of text mining approaches applied to various application areas in the biomedical domain. *Journal of Knowledge Management*, *25*(3), 642–668. https://doi.org/10.1108/JKM-09-2019-0524

[7] Trieu, H. L., Miwa, M., & Ananiadou, S. (2022). BioVAE: A pre-trained latent variable language model for biomedical text mining. *Bioinformatics*, *38*(3), 872–874. https://doi.org/10.1093/bioinformatics/btab702

[8] Leaman, R., Wei, C. H., Allot, A., & Lu, Z. (2020). Ten tips for a text-mining-ready article: How to improve automated discoverability and interpretability. *Plos Biology*, *18*(6), e3000716. https://doi.org/10.1371/journal.pbio.3000716

[9] Song, B., Li, F., Liu, Y., & Zeng, X. (2021). Deep learning methods for biomedical named entity recognition: A survey and qualitative comparison. *Briefings in Bioinformatics*, *22*(6), bbab282. https://doi.org/10.1093/bib/bbab282

[10] Bhasuran, B., Murugesan, G., Abdulkadhar, S., & Natarajan, J. (2016). Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *Journal of Biomedical Informatics*, *64*, 1–9. https://doi.org/10.1016/j.jbi.2016.09.009

[11] Bhasuran, B. (2022). BioBERT and similar approaches for relation extraction. In K. Raja (Ed.), *Biomedical text mining* (pp. 221–235). Humana Press. https://doi.org/10.1007/978-1-0716-2305-3_12

[12] Tóth, B., Berek, L., Gulácsi, L., Péntek, M., & Zrubka, Z. (2024). Automation of systematic reviews of biomedical literature: A scoping review of studies indexed in PubMed. *Systematic Reviews*, *13*(1), 174. https://doi.org/10.1186/s13643-024-02592-3

[13] Jin, Q., Leaman, R., & Lu, Z. (2024). PubMed and beyond: Biomedical literature search in the age of artificial intelligence. *EBioMedicine*, *100*, 104988. https://doi.org/10.1016/j.ebiom.2024.104988

[14] Cesario, E., Comito, C., & Zumpano, E. (2024). A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. *Neurocomputing*, *568*, 127079. https://doi.org/10.1016/j.neucom.2023.127079

[15] Sivarajkumar, S., Mohammad, H. A., Oniani, D., Roberts, K., Hersh, W., Liu, H., . . . , & Wang, Y. (2024). Clinical information retrieval: A literature review. *Journal of Healthcare Informatics Research*, *8*(2), 313–352. https://doi.org/10.1007/s41666-024-00159-4

[16] Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., . . . , & Lu, Z. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. *Briefings in Bioinformatics*, *25*(1), bbad493. https://doi.org/10.1093/bib/bbad493

[17] Fleuren, W. W. M., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, *74*, 97–106. https://doi.org/10.1016/j.ymeth.2015.01.015

[18] Tucciarone, I., Secci, G., Contiero, B., & Parisi, G. (2024). Sustainable aquaculture over the last 30 years: An analysis of the scientific literature by the text mining approach. *Reviews in Aquaculture*, *16*(4), 2064–2076. https://doi.org/10.1111/raq.12950

[19] Zeng, H., Luo, C., Jin, B., Sarwar, S. M., Wei, T., & Zamani, H. (2024). Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*, 1441–1452. https://doi.org/10.1145/3589334.3645477

[20] Zhu, R., Tu, X., & Huang, J. X. (2021). Utilizing BERT for Biomedical and Clinical Text Mining. In K. C. Lee, S. S. Roy, P. Samui, & V. Kumar (Eds.), *Data analytics in biomedical engineering and healthcare* (pp. 73–103). Academic Press. https://doi.org/10.1016/B978-0-12-819314-3.00005-7

[21] Bhasuran, B., & Natarajan, J. (2023). DisGeReExT: A knowledge discovery system for exploration of disease–gene associations through large-scale literature-wide analysis study. *Knowledge and Information Systems*, *65*(8), 3463–3487. https://doi.org/10.1007/s10115-023-01862-1

[22] Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, *23*(6), bbac409. https://doi.org/10.1093/bib/bbac409

[23] Bhasuran, B., Subramanian, D., & Natarajan, J. (2018). Text mining and network analysis to find functional associations of genes in high altitude diseases. *Computational Biology and Chemistry*, *75*, 101–110. https://doi.org/10.1016/j.compbiolchem.2018.05.002

[24] Bhasuran, B., & Natarajan, J. (2018). Automatic extraction of gene-disease associations from literature using joint ensemble learning. *PloS One*, *13*(7), e0200699. https://doi.org/10.1371/journal.pone.0200699

[25] Zong, H., Wu, R., Cha, J., Feng, W., Wu, E., Li, J., . . . , & Shen, B. (2024). Advancing Chinese biomedical text mining with community challenges. *Journal of Biomedical Informatics*, *157*, 104716. https://doi.org/10.1016/j.jbi.2024.104716

[26] Bonner, S., Barrett, I. P., Ye, C., Swiers, R., Engkvist, O., Bender, A., . . . , & Hamilton, W. L. (2022). A review of biomedical datasets relating to drug discovery: A knowledge graph perspective. *Briefings in Bioinformatics*, *23*(6), bbac404. https://doi.org/10.1093/bib/bbac404

[27] Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., . . . , & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. *Cell*, *187*(22), 6125–6151. https://doi.org/10.1016/j.cell.2024.09.022

[28] Pu, Y., Beck, D., & Verspoor, K. (2023). Graph embedding-based link prediction for literature-based discovery in Alzheimer's Disease. *Journal of Biomedical Informatics*, *145*, 104464. https://doi.org/10.1016/j.jbi.2023.104464

[29] Wysocki, O., Wysocka, M., Carvalho, D. S., Bogatu, A., Gusicuma, D., Delmas, M., . . . , & Freitas, A. (2024). An LLM-based knowledge synthesis and scientific reasoning

framework for biomedical discovery. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 3*, 355–364. https://doi.org/10.18653/v1/2024.acl-demos.34

[30] Daowd, A., Barrett, M., Abidi, S., & Abidi, S. S. R. (2021). A framework to build a causal knowledge graph for chronic diseases and cancers by discovering semantic associations from biomedical literature. In *2021 IEEE 9th International Conference on Healthcare Informatics*, 13–22. https://doi.org/10.1109/ICHI52183.2021.00016

[31] Tandra, G., Yoone, A., Mathew, R., Wang, M., Hales, C. M., & Mitchell, C. S. (2023). Literature-based discovery predicts antihistamines are a promising repurposed adjuvant therapy for Parkinson's disease. *International Journal of Molecular Sciences*, *24*(15), 12339. https://doi.org/10.3390/ijms241512339

[32] Liu, Z., Roberts, R. A., Lal-Nag, M., Chen, X., Huang, R., & Tong, W. (2021). AI-based language models powering drug discovery and development. *Drug Discovery Today*, *26*(11), 2593–2607. https://doi.org/10.1016/j.drudis.2021.06.009

[33] Zhang, T., Leng, J., & Liu, Y. (2020). Deep learning for drug–drug interaction extraction from the literature: A review. *Briefings in Bioinformatics*, *21*(5), 1609–1627. https://doi.org/10.1093/bib/bbz087

[34] Alachram, H., Chereda, H., Beißbarth, T., Wingender, E., & Stegmaier, P. (2021). Text mining-based word representations for biomedical data analysis and protein-protein interaction networks in machine learning tasks. *PloS One*, *16*(10), e0258623. https://doi.org/10.1371/journal.pone.0258623

[35] Macnee, M., Pérez-Palma, E., Schumacher-Bass, S., Dalton, J., Leu, C., Blankenberg, D., & Lal, D. (2021). SimText: A text mining framework for interactive analysis and visualization of similarities among biomedical entities. *Bioinformatics*, *37*(22), 4285–4287. https://doi.org/10.1093/bioinformatics/btab365

[36] Wang, L. L., & Lo, K. (2021). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. *Briefings in Bioinformatics*, *22*(2), 781–799. https://doi.org/10.1093/bib/bbaa296

[37] Manoharan, S., & Iyyappan, O. R. (2022). A hybrid protocol for finding novel gene targets for various diseases using microarray expression data analysis and text mining. In K. Raja (Ed.), *Biomedical text mining* (pp. 41–70). Humana Press. https://doi.org/10.1007/978-1-0716-2305-3_3

[38] Reddy, S., Bhaskar, R., Padmanabhan, S., Verspoor, K., Mamillapalli, C., Lahoti, R., . . . , & Sinha, S. (2021). Use and validation of text mining and cluster algorithms to derive insights from corona virus disease-2019 (COVID-19) medical literature. *Computer Methods and Programs in Biomedicine Update*, *1*, 100010. https://doi.org/10.1016/j.cmpbup.2021.100010

[39] González-Márquez, R., Schmidt, L., Schmidt, B. M., Berens, P., & Kobak, D. (2024). The landscape of biomedical research. *Patterns*, *5*(6), 100968. https://doi.org/10.1016/j.patter.2024.100968

[40] Li, P. H., Chen, T. F., Yu, J. Y., Shih, S. H., Su, C. H., Lin, Y. H., . . . , & Huang, J. H. (2022). pubmedKB: An interactive web server for exploring biomedical entity relations in the biomedical literature. *Nucleic Acids Research*, *50*(W1), W616–W622. https://doi.org/10.1093/nar/gkac310

[41] Jung, H., & Lee, B. G. (2020). Research trends in text mining: Semantic network and main path analysis of selected journals. *Expert Systems with Applications*, *162*, 113851. https://doi.org/10.1016/j.eswa.2020.113851

[42] Weber, L., Münchmeyer, J., Rocktäschel, T., Habibi, M., & Leser, U. (2020). HUNER: Improving biomedical NER with pretraining. *Bioinformatics*, *36*(1), 295–302. https://doi.org/10.1093/bioinformatics/btz528

[43] Yu, W., Zhu, C., Li, Z., Hu, Z., Wang, Q., Ji, H., & Jiang, M. (2022). A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, *54*(11s), 227. https://doi.org/10.1145/3512467

[44] Gupta, M., & Agrawal, P. (2022). Compression of deep learning models for text: A survey. *ACM Transactions on Knowledge Discovery from Data*, *16*(4), 61. https://doi.org/10.1145/3487045

[45] Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., & Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, *9*, 156043–156070. https://doi.org/10.1109/ACCESS.2021.3129786

[46] Alkan, B. B., Karakuş, L., & Direkci, B. (2023). Knowledge discovery from the texts of Nobel Prize winners in literature: Sentiment analysis and latent Dirichlet allocation. *Scientometrics*, *128*(9), 5311–5334. https://doi.org/10.1007/s11192-023-04783-6

[47] Lahav, D., Saad Falcon, J., Kuehl, B., Johnson, S., Parasa, S., Shomron, N., . . . , & Hope, T. (2022). A search engine for discovery of scientific challenges and directions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 11982–11990. https://doi.org/10.1609/aaai.v36i11.21456

[48] Swanson, D. R. (1988). Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, *31*(4), 526–557. https://doi.org/10.1353/pbm.1988.0009

[49] Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, *30*(1), 7–18. https://doi.org/10.1353/pbm.1986.0087

[50] Gallai, V., Sarchielli, P., Coata, G., Firenze, C., Morucci, P., & Abbritti, G. (1992). Serum and salivary magnesium levels in migraine. Results in a group of juvenile patients. *Headache: The Journal of Head and Face Pain*, *32*(3), 132–135. https://doi.org/10.1111/j.1526-4610.1992.hed3203132.x

[51] Kastrin, A., & Hristovski, D. (2021). Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020). *Scientometrics*, *126*(2), 1415–1451. https://doi.org/10.1007/s11192-020-03811-z

[52] Kilicoglu, H., Shin, D., Fiszman, M., Rosemblat, G., & Rindflesch, T. C. (2012). SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics*, *28*(23), 3158–3160. https://doi.org/10.1093/bioinformatics/bts591

[53] Swanson, D. R. (1990). Somatomedin C and arginine: Implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine*, *33*(2), 157–186. https://doi.org/10.1353/pbm.1990.0031

[54] Swanson, D. R., & Smalheiser, N. R. (1997). An interactive system for finding complementary literatures: A stimulus to scientific discovery. *Artificial Intelligence*, *91*(2), 183–203. https://doi.org/10.1016/S0004-3702(97)00008-8

[55] Gopalakrishnan, V., Jha, K., Jin, W., & Zhang, A. (2019). A survey on literature based discovery approaches in biomedical domain. *Journal of Biomedical Informatics*, *93*, 103141. https://doi.org/10.1016/j.jbi.2019.103141

[56] Gopalakrishnan, V., Jha, K., Xun, G., Ngo, H. Q., & Zhang, A. (2018). Towards self-learning based hypotheses generation in biomedical text domain. *Bioinformatics*, *34*(12), 2103–2115. https://doi.org/10.1093/bioinformatics/btx837

[57] Henry, S., & McInnes, B. T. (2019). Indirect association and ranking hypotheses for literature based discovery. *BMC Bioinformatics*, *20*(1), 425. https://doi.org/10.1186/s12859-019-2989-9

[58] Henry, S., & McInnes, B. T. (2017). Literature based discovery: Models, methods, and trends. *Journal of Biomedical Informatics*, *74*, 20–32. https://doi.org/10.1016/j.jbi.2017.08.011

[59] Kostoff, R. N., & Patel, U. (2015). Literature-related discovery and innovation: Chronic kidney disease. *Technological Forecasting and Social Change*, *91*, 341–351. https://doi.org/10.1016/j.techfore.2014.09.013

[60] Henry, S., Wijesinghe, D. S., Myers, A., & McInnes, B. T. (2021). Using literature based discovery to gain insights into the metabolomic processes of cardiac arrest. *Frontiers in Research Metrics and Analytics*, *6*, 644728. https://doi.org/10.3389/frma.2021.644728

[61] Hettne, K. M., Thompson, M., van Haagen, H. H. H. B. M., van der Horst, E., Kaliyaperumal, R., Mina, E., . . . , & Schultes, E. A. (2016). The implicitome: A resource for rationalizing gene-disease associations. *PLoS One*, *11*(2), e0149621. https://doi.org/10.1371/journal.pone.0149621

[62] Muratov, E. N., Amaro, R., Andrade, C. H., Brown, N., Ekins, S., Fourches, D., . . . , & Tropsha, A. (2021). A critical overview of computational approaches employed for COVID-19 drug discovery. *Chemical Society Reviews*, *50*(16), 9121–9151. https://doi.org/10.1039/D0CS01065K

[63] Elbattah, M., Arnaud, É., Gignon, M., & Dequen, G. (2021). The role of text analytics in healthcare: A review of recent developments and applications. In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies*, *5*, 825–832. https://doi.org/10.5220/0010414508250832

[64] Hristovski, D., Peterlin, B., Mitchell, J. A., & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *International Journal of Medical Informatics*, *74*(2–4), 289–298. https://doi.org/10.1016/j.ijmedinf.2004.04.024

[65] Lavrač, N., Martinc, M., Pollak, S., Pompe Novak, M., & Cestnik, B. (2020). Bisociative literature-based discovery: Lessons learned and new word embedding approach. *New Generation Computing*, *38*(4), 773–800. https://doi.org/10.1007/s00354-020-00108-w

[66] Li, L., Zheng, X., Zhou, Q., Villanueva, N., Nian, W., Liu, X., & Huan, T. (2020). Metabolomics-based discovery of molecular signatures for triple negative breast cancer in Asian female population. *Scientific Reports*, *10*(1), 370. https://doi.org/10.1038/s41598-019-57068-5

[67] Pratt, W., & Yetisgen-Yildiz, M. (2003). LitLinker: Capturing connections across the biomedical literature. In *Proceedings of the 2nd International Conference on Knowledge Capture*, 105–112. https://doi.org/10.1145/945645.945662

[68] Preiss, J., & Stevenson, M. (2017). Quantifying and filtering knowledge generated by literature based discovery. *BMC Bioinformatics*, *18*(S7), 249. https://doi.org/10.1186/s12859-017-1641-9

[69] Pyysalo, S., Baker, S., Ali, I., Haselwimmer, S., Shah, T., Young, A., . . . , & Korhonen, A. (2019). LION LBD: A literature-based discovery system for cancer biology. *Bioinformatics*, *35*(9), 1553–1561. https://doi.org/10.1093/bioinformatics/bty845

[70] Rather, N. N., Patel, C. O., & Khan, S. A. (2017). Using deep learning towards biomedical knowledge discovery. *International Journal of Mathematical Sciences and Computing*, *3*(2), 1–10. https://doi.org/10.5815/ijmsc.2017.02.01

[71] Sang, S., Yang, Z., Liu, X., Wang, L., Zhang, Y., Lin, H., . . . , & Zhang, Y. (2018). A knowledge graph based bidirectional recurrent neural network method for literature-based discovery. In *2018 IEEE International Conference on Bioinformatics and Biomedicine*, 751–752. https://doi.org/10.1109/BIBM.2018.8621423

[72] Grissa, D., Junge, A., Oprea, T. I., & Jensen, L. J. (2022). Diseases 2.0: A weekly updated database of disease–gene associations from text mining and data integration. *Database*, *2022*, baac019. https://doi.org/10.1093/database/baac019

[73] Chen, Q., Allot, A., & Lu, Z. (2021). LitCovid: An open database of COVID-19 literature. *Nucleic Acids Research*, *49*(D1), D1534–D1540. https://doi.org/10.1093/nar/gkaa952

[74] Venugopal, V., Sahoo, S., Zaki, M., Agarwal, M., Gosvami, N. N., & Krishnan, N. A. (2021). Looking through glass: Knowledge discovery from materials science literature using natural language processing. *Patterns*, *2*(7), 100290. https://doi.org/10.1016/j.patter.2021.100290

[75] Zhang, L., Carter, R. A., Qian, X., Yang, S., Rujimora, J., & Wen, S. (2022). Academia's responses to crisis: A bibliometric analysis of literature on online learning in higher education during COVID-19. *British Journal of Educational Technology*, *53*(3), 620–646. https://doi.org/10.1111/bjet.13191

[76] Srinivasan, P., & Libbus, B. (2004). Mining MEDLINE for implicit links between dietary substances and diseases. *Bioinformatics*, *20*(suppl_1), i290–i296. https://doi.org/10.1093/bioinformatics/bth914

[77] Srinivasan, P. (2004). Text mining: Generating hypotheses from MEDLINE. *Journal of the American Society for Information Science and Technology*, *55*(5), 396–413. https://doi.org/10.1002/asi.10389

[78] Thilakaratne, M., Falkner, K., & Atapattu, T. (2019). A systematic review on literature-based discovery: General overview, methodology, & statistical analysis. *ACM Computing Surveys*, *52*(6), 129. https://doi.org/10.1145/3365756

[79] Weeber, M., Vos, R., Klein, H., de Jong-van den Berg, L. T. W., Aronson, A. R., & Molema, G. (2003). Generating hypotheses by discovering implicit associations in the literature: A case report of a search for new potential therapeutic uses for thalidomide. *Journal of the American Medical Informatics Association*, *10*(3), 252–259. https://doi.org/10.1197/jamia.M1158

[80] Weeber, M., Klein, H., de Jong-van den Berg, L. T., & Vos, R. (2001). Using concepts in literature-based discovery: Simulating Swanson's Raynaud–fish oil and migraine–magnesium discoveries. *Journal of the American Society for Information Science and Technology*, *52*(7), 548–557. https://doi.org/10.1002/asi.1104

[81] Weissenborn, D., Schroeder, M., & Tsatsaronis, G. (2015). Discovering relations between indirectly connected biomedical concepts. *Journal of Biomedical Semantics*, *6*(1), 28. https://doi.org/10.1186/s13326-015-0021-5

[82] Workman, T. E., Fiszman, M., Cairelli, M. J., Nahl, D., & Rindflesch, T. C. (2016). Spark, an application based on serendipitous knowledge discovery. *Journal of Biomedical Informatics*, *60*, 23–37. https://doi.org/10.1016/j.jbi.2015.12.014

[83] Xie, Q., Yang, K. M., Heo, G. E., & Song, M. (2020). Literature based discovery of alternative TCM medicine for adverse reactions to depression drugs. *BMC Bioinformatics*, *21*(S5), 405. https://doi.org/10.1186/s12859-020-03735-8

[84] Yang, H. T., Ju, J. H., Wong, Y. T., Shmulevich, I., & Chiang, J. H. (2017). Literature-based discovery of new candidates for drug repurposing. *Briefings in Bioinformatics*, *18*(3), 488–497. https://doi.org/10.1093/bib/bbw030

[85] Yetisgen-Yildiz, M., & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *Journal of Biomedical Informatics*, *39*(6), 600–611. https://doi.org/10.1016/j.jbi.2005.11.010

[86] Zhou, H., Yu, H., & Hu, R. (2017). Topic discovery and evolution in scientific literature based on content and citations. *Frontiers of Information Technology & Electronic Engineering*, *18*(10), 1511–1524. https://doi.org/10.1631/FITEE.1601125

[87] Smalheiser, N. R., Torvik, V. I., & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*, *94*(2), 190–197. https://doi.org/10.1016/j.cmpb.2008.12.006

[88] Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., & Ananiadou, S. (2011). Discovering and visualizing indirect associations between biomedical concepts. *Bioinformatics*, *27*(13), i111–i119. https://doi.org/10.1093/bioinformatics/btr214

[89] Tropmann-Frick, M., & Schreier, T. (2022). Towards drug repurposing for COVID-19 treatment using literature-based discovery. In M. Tropmann-Frick, H. Jaakkola, B. Thalheim, Y. Kiyoki, & N. Yoshida (Eds.), *Information modelling and knowledge bases XXXIII* (pp. 215–232). IOS Press. https://doi.org/10.3233/FAIA210488

[90] Baek, S. H., Lee, D., Kim, M., Lee, J. H., & Song, M. (2017). Enriching plausible new hypothesis generation in PubMed. *PloS One*, *12*(7), e0180539. https://doi.org/10.1371/journal.pone.0180539

[91] Kim, Y. H., & Song, M. (2019). A context-based ABC model for literature-based discovery. *PloS One*, *14*(4), e0215313. https://doi.org/10.1371/journal.pone.0215313

[92] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

[93] Miolo, G., Mantoan, G., & Orsenigo, C. (2021). ELECTRAMed: A new pre-trained language representation model for biomedical NLP. *arXiv Preprint: 2104.09585*. https://doi.org/10.48550/arXiv.2104.09585

[94] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., . . . , & Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 2. https://doi.org/10.1145/3458754

[95] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W. H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–28. https://doi.org/10.18653/v1/W19-1909

[96] Mulyar, A., Uzuner, O., & McInnes, B. (2021). MT-clinical BERT: Scaling clinical information extraction with multitask learning. *Journal of the American Medical Informatics Association*, *28*(10), 2108–2115. https://doi.org/10.1093/jamia/ocab126

[97] Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., & Wong, A. (2020). UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1744–1753. https://doi.org/10.18653/v1/2021.naacl-main.139

[98] Shin, H. C., Zhang, Y., Bakhturina, E., Puri, R., Patwary, M., Shoeybi, M., & Mani, R. (2020). BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 4700–4706. https://doi.org/10.18653/v1/2020.emnlp-main.379

[99] Sybrandt, J., Tyagin, I., Shtutman, M., & Safro, I. (2020). AGATHA: Automatic graph mining and transformer based hypothesis generation approach. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2757–2764. https://doi.org/10.1145/3340531.3412684

[100] Millikin, R. J., Raja, K., Steill, J., Lock, C., Tu, X., Ross, I., . . . , & Stewart, R. (2023). Serial KinderMiner (SKiM) discovers and annotates biomedical knowledge using co-occurrence and transformer models. *BMC Bioinformatics*, *24*(1), 412. https://doi.org/10.1186/s12859-023-05539-y

[101] Crichton, G., Baker, S., Guo, Y., & Korhonen, A. (2020). Neural networks for open and closed literature-based discovery. *PLoS One*, *15*(5), e0232891. https://doi.org/10.1371/journal.pone.0232891

[102] Preiss, J. (2024). Using word evolution to predict drug repurposing. *BMC Medical Informatics and Decision Making*, *24*(S2), 114. https://doi.org/10.1186/s12911-024-02496-1

[103] Cuffy, C., & McInnes, B. T. (2023). Exploring a deep learning neural architecture for closed literature-based discovery. *Journal of Biomedical Informatics*, *143*, 104362. https://doi.org/10.1016/j.jbi.2023.104362

[104] Wang, Q., Downey, D., Ji, H., & Hope, T. (2023). Learning to generate novel scientific directions with contextualized literature-based discovery. *arXiv Preprint: 2305.14259*. https://doi.org/10.48550/arXiv.2305.14259

[105] Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., & Ting, D. S. W. (2023). Large language models in medicine. *Nature Medicine*, *29*(8), 1930–1940. https://doi.org/10.1038/s41591-023-02448-8

[106] Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P. C., . . . , & Natarajan, V. (2024). Towards generalist biomedical AI. *NEJM AI*, *1*(3), AIoa2300138. https://doi.org/10.1056/AIoa2300138

[107] Nedbaylo, A., & Hristovski, D. (2024). Implementing literature-based discovery (LBD) with ChatGPT. In *2024 47th MIPRO ICT and Electronics Convention*, 120–125. https://doi.org/10.1109/MIPRO60963.2024.10569439

[108] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., . . . , & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 793.

[109] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., . . . , & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 1800.

**How to Cite:** Bhasuran, B., Murugesan, G., & Natarajan, J. (2025). Literature-Based Discovery (LBD): Towards Hypothesis Generation and Knowledge Discovery in Biomedical Text Mining. *Medinformatics*, *2*(4), 241–255. https://doi.org/10.47852/bonviewMEDIN52025348