# Augmented Knowledge Graphs for Literature-Based Discovery (AKG-LBD): A Novel Framework to Enhance Semantics-Based LBD For Biomedical Knowledge Discovery

by

Ali Daowd

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2023

Dalhousie University is located in Mi'kma'ki, the ancestral and unceded territory of the Mi'kmaq. We are all Treaty people.

# Table of Contents

# List of Tables

# List of Figures

# Abstract

The biomedical literature is expanding exponentially, generating a vast amount of knowledge that frequently goes unnoticed. Consequently, there is an urgent need to develop methods to mine knowledge from published literature to facilitate the automated discovery of hidden biomedical knowledge. Literature-Based Discovery (LBD) is a novel paradigm that aims to uncover new knowledge from the literature via transitive inference. Advances in text mining and knowledge extraction methods have enabled semantics-based LBD, which extracts knowledge in the form of *subject-predicate-object* semantic triples represented in a Knowledge Graph (KG). The *subject* and *object* are normalized biomedical concepts, and the *predicate* denotes the semantic relation between them.

Semantics-based LBD has not seen large scale adoption due to several challenges. Firstly, knowledge extraction methods result in incomplete knowledge extraction due to missing semantic relations. Secondly, extracted biomedical entities are represented by granular and ambiguous representations, leading to a large discovery search space. Thirdly, the over-generation of spurious discoveries as output obscures meaningful discoveries. This dissertation investigates semantics-based methods and KG representation learning to develop novel solutions addressing the fundamental challenges in semantic-based LBD. Specifically, we address the challenges by: (i) incorporating state-of-the-art knowledge extraction to acquire semantic-based knowledge from the literature; (ii) utilizing concept disambiguation and semantic alignment techniques to resolve ambiguity and granularity of concept representations; (iii) leveraging a multi-step Knowledge Graph Completion (KGC) methodology to augment the literature-based KG by predicting missing relations using KG embeddings; and (iv) presenting a knowledge filtering and ranking approach based on the principles of information theory to prioritize interesting discoveries. The outcome of this dissertation is the novel *Augmented Knowledge Graphs for LBD* (*AKG-LBD*) framework that enhances traditional semantics-based LBD frameworks. The *AKG-LBD* framework is assessed by replicating biomedical discoveries published in peer-reviewed journals. The results indicate that *AKG-LBD* can discover meaningful knowledge with high precision relative to baseline approaches. The main implication of this dissertation is that KGC methods, combined with semantic alignment, enhances the performance of semantics-based LBD by generating augmented literature-based KGs. Additionally, the knowledge filtering and ranking methods are capable of prioritizing interesting knowledge which facilitates the exploration of meaningful biomedical discoveries.

# List Of Abbreviations Used

| | |
|---|---|
| LBD | Literature-Based Discovery |
| AKG-LBD | Augmented Knowledge Graphs for Literature-Based Discovery |
| KG | Knowledge Graph |
| KGC | Knowledge Graph Completion |
| KGE | Knowledge Graph Embeddings |
| KB | Knowledge Base |
| UMLS | Unified Medical Language System |
| MeSH | Medical Subject Headings |
| GO | Gene Ontology |
| PRO | Protein Ontology |
| NLP | Natural Language Processing |
| IC | Information Content |
| LTC | Linking Term Count |
| OR | Odds Ratio |
| $X^2$ | Pearson's Chi-square |
| LLR | Log-Likelihood Ratio |
| COF | Co-Occurrence Frequency |
| mAP | Mean Average Precision |
| AP@K | Average Precision at K |
| AR@K | Average Recall at K |
| RR | Relative Rank |
| ARR | Average Relative Rank |

# Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Syed Sibte Raza Abidi, for his unwavering support, guidance, and mentorship throughout this long and difficult journey. I am forever indebted to him for providing me the opportunity and supportive environment to grow academically and pursue my goals. This thesis would not be possible without his mentorship and support.

I extend my sincere thanks and appreciation to my co-supervisor, Dr. Samina Abidi, for her endless support and guidance. I am fortunate to have Dr. Samina Abidi as a mentor and co-supervisor.

I would like to thank my supervisory committee members, Dr. Samuel Stewart and Dr. Christian Blouin, for their support and taking the time to read this thesis. My sincere thanks to Dr. Samuel Stewart for his valuable advice and guidance throughout this journey.

I would like to extend my appreciation to Dr. Jim Warren for kindly accepting to be my thesis external examiner.

Thanks to Dr. William Van Woensel for his insightful comments and challenging questions during my time with the NICHE research group.

A special thanks to my colleagues, Asil Naqvi and Jaber Rad, for their help and encouragement throughout the years. I am fortunate to have them as friends who helped make my time here memorable.

Lastly, I would like to thank my parents and brother for the endless love. I am fortunate to have a compassionate and understanding family who have always supported me in all my pursuits. This thesis would not be possible without them by my side.

# Chapter 1 Introduction

Finding and reviewing relevant scholarly literature is one of the key aspects of scientific knowledge synthesis and discovery. With the recent exponential growth in the volume of published literature, the research community is struggling to stay up-to-date with the latest developments and knowledge published in the literature. It is estimated that the doubling time of biomedical knowledge has increased from 3.5 years in 2010 to just 73 days in 2020 (Densen, 2011). Furthermore, the substantial increase in the number of specialized scientific journals, and articles per journal, has led to fragmentation of evidence-based knowledge (Jaradeh et al., 2019; Landhuis, 2016). Consequently, researchers tend to deal with fragments of incomplete knowledge rather than complete and complementary knowledge, and therefore many valuable implicit connections that exist between disparate bodies of knowledge remain undiscovered.

Literature-Based Discovery (LBD) has emerged as a novel computational approach to discover and synthesize knowledge by connecting and reasoning over disparate bodies of literature, to uncover implicit connections which have not been explicitly stated or connected (Henry & McInnes, 2017). LBD is premised on Don Swanson's ABC theory that if two unrelated articles explicitly specify an association between biomedical concepts A-B (in one article) and B-C (in another article), then it is indicative of an A-C association – which was not explicitly specified in any of the explored articles (Swanson, 1986b). Swanson's theory was used to propose dietary fish oil (A) as a treatment for Raynaud's disease (C) due to their shared association with blood viscosity and platelet aggregation (B) (Swanson, 1986a). This LBD-driven hypothesis was later confirmed in a clinical study by DiGiacomo et al. (DiGiacomo et al., 1989). Subsequently, Swanson continued to utilize LBD methods to propose other discoveries on migraines and magnesium (Swanson, 1988), Alzheimer's disease and estrogens (Smalheiser & Swanson, 1996), and somatomedins and arginine (Swanson, 1990).

Over the years, LBD has assisted in countless discoveries and hypotheses generation in the biomedical field, including drug development and repurposing (R. Zhang et al., 2021), and adverse drug event prediction (Bougiatiotis et al., 2020). For example, Zhang et al. applied LBD to discover novel treatments for prostate cancer using drug-gene, gene-cancer, and

gene-gene associations extracted from literature (R. Zhang et al., 2014). The authors found three drugs already used for prostate cancer therapy and 18 candidate drugs which have not been previously identified as prostate cancer medications. In another recent study, a graph-based LBD approach was introduced to develop a novel drug repurposing framework for COVID-19 (R. Zhang et al., 2021). This study resulted in identifying five novel medications (i.e., paclitaxel, SB 203580, alpha 2-antiplasmin, metoclopramide, and oxymatrine) for COVID-19. Similarly, LBD was applied to propose new treatments for cataracts (Kostoff, 2008), Parkinson's disease (X. Zhang & Che, 2021), and multiple sclerosis (Kostoff, Briggs, et al., 2008). LBD have been applied to areas outside biomedicine, such as climate change (Marsi et al., 2017) and discovering novel water purification techniques (Kostoff, Solka, et al., 2008), however, the vast majority of LBD research remains within the biomedical domain.

Applicability of LBD methods in biomedicine is due to its potential to connect silos of knowledge that describe underlying biological interactions related to diseases. Such knowledge can provide a better understanding of underlying pathological mechanisms, thus allowing researchers to deduce previously unknown knowledge (Deftereos et al., 2011). As such, the LBD principles resemble the domain of systems medicine, which seeks to understand pathological disease mechanisms from a holistic perspective incorporating biochemical, physiological, and environmental interactions (Saqi et al., 2016).

## 1.1 The Literature-Based Discovery (LBD) Process:

LBD is a relatively mature field which has incorporated various computational methods and techniques to uncover novel discoveries from disparate literature-based knowledge instances. A high-level overview of the LBD process is depicted in Figure 1.1.

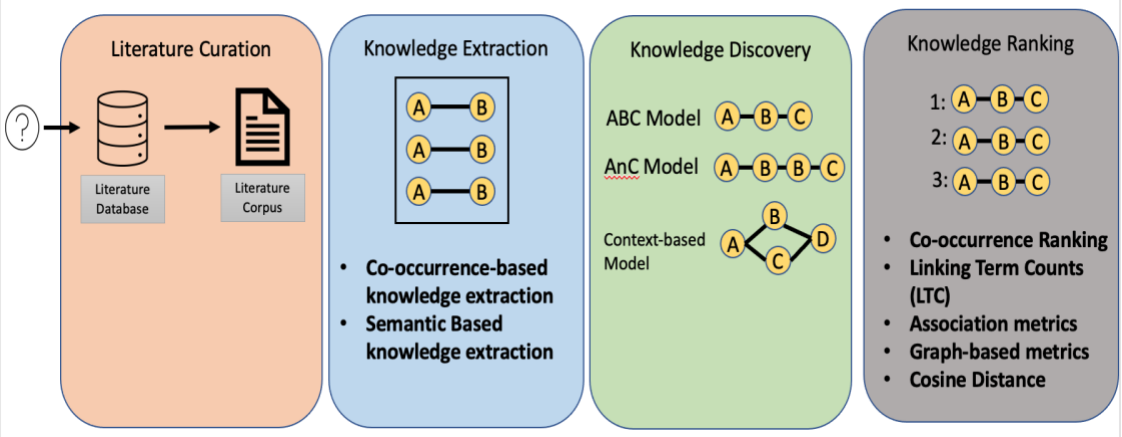


*Figure 1.1: Schematic of the LBD Process*

The LBD process is generally initiated by defining a discovery task and literature relevant to the discovery domain (Thilakaratne et al., 2019), to identify and extract literature-based knowledge. As various methods have been employed in LBD to extract literature-based knowledge, it is important to establish a clear definition of what constitutes literature-based knowledge. Conceptually, knowledge in scientific literature is asserted as statements (i.e., sentences) describing a set of concepts and how they relate to one another. A concept is universally defined as meaning expressed by different domain-specific terms and/or phrases. However, the definition of what constitutes a relationship between concepts can vary across different LBD approaches. Broadly, LBD can be categorized based on the employed approach for knowledge extraction: co-occurrence and semantic-based LBD. The semantic-based approach characterizes a relationship between concepts based on verbal clauses in sentences that denote some notion of associative or causal predicate (Cairelli et al., 2013). This approach applies Natural Language Processing (NLP) and relation extraction methods on literature corpora, thereby representing instances of literature-based knowledge as *subject-predicate-object* semantic triples (Kilicoglu et al., 2020). The subject and object concepts are normalized to concepts from standardized terminologies, such as the Unified Medical Language System (UMLS) or Medical Subject Headings (MeSH), while the predicate is an annotated ontological relation. As such, the semantic-based approach provides meaningful and expressive semantic relations denoting functional relationships between biomedical concepts.

In contrast, the co-occurrence based approach is based on the assumption that the co-occurrence of two biomedical concepts in a sentence implies a form of association between

them (Swanson & Smalheiser, 1997). This approach utilizes rudimentary text mining methods to identify normalized ontological concepts from literature corpora, and represents instances of literature-based knowledge as a co-occurrence network. While co-occurrence methods have the advantage of simplicity, the underlying assumption of association is inherently weak as co-occurrences do not adequately capture the semantics in text, which is a crucial factor in the discovery process to elucidate causal, mechanistic, or associative relations between concepts (Hristovski et al., 2006). Hence, co-occurrence-based knowledge instances can only be interpreted as associations and not as domain-specific relations, which complicates the interpretation of the generated knowledge. Conversely, semantic-based methods seek to exploit the semantics of text to capture meaningful knowledge and extract annotated semantic relations linking concepts (Ahlers et al., 2007).

The extraction of literature-based knowledge is typically followed by the discovery task, whereby a well-defined discovery model is applied to logically connect disparate instances (spanning multiple unrelated documents) of knowledge to infer novel *discovery paths,* characterized by a chain of associations or relations between a source and target biomedical concepts, established through one or more intermediate concepts which expound the indirect relation/association between the source and target. Most contemporary LBD works utilize discovery models influenced by Swanson's ABC model (Swanson, 1986a). The popular ABC model has two variants: open-discovery and closed-discovery (Thilakaratne et al., 2019). In open-discovery, the discovery task starts with a predefined concept of interest (referred to as A) and seeks to identify implicit connections with an unknown target concept (referred to as C). In closed-discovery, the source (A) and target (C) concepts are predefined, and the aim is to find meaningful intermediate (B) concepts that may explain the indirect association or relation between (A) and (C). Extensions of the ABC model include: the AnC model where n represents multiple intermediates (i.e., B1, B2, B3…Bn) (Wilkowski et al., 2011), context-based ABC model (Kim & Song, 2019), and context-assignment ABC model (Kim & Song, 2019). Regardless of the applied discovery model, the output consists of discovery paths which are explored to generate novel knowledge relevant to the predefined discovery task.

The final task in the LBD process is to assign ranking scores for the output discovery paths according to some notion of statistical association, semantic relatedness, or interestingness-

based measure. This task is necessary to sort the LBD output, thereby ensuring that the most meaningful and novel discovery paths are prioritized over nosy and trivial paths. Common ranking metrics include: co-occurrence frequencies (Heo et al., 2019), linking term counts (Yetisgen-Yildiz & Pratt, 2009), association-based measures (e.g., Pearson's Chi-square, odds ratio, and log-likelihood ratio) (Henry & McInnes, 2019), graph-based measures (e.g., Jaccard index, common neighbours, centrality, and PageRank) (Kastrin et al., 2016), and nearest neighbour analysis (e.g., cosine or Euclidean distance) (Henry & McInnes, 2019).

## 1.2 Challenges of Semantic Literature-Based Knowledge Discovery:

Despite its promise, semantic LBD remains confined within research settings and has not seen wide-scale adoption among the scientific community. This can be attributed to several challenges, such as incomplete literature-based knowledge extraction (Henry & McInnes, 2017), representation and normalization of biomedical entities in text (H.-T. Yang et al., 2017), ranking output discoveries (Rastegar-Mojarad, Elayavilli, Li, Prasad, et al., 2015; Sebastian et al., 2017), differentiating noise from meaningful discoveries (Raja et al., 2020), and evaluating the generated discoveries (Gopalakrishnan et al., 2019). These challenges have a significant impact on the knowledge discovery process and adoption of LBD by the scientific community (Phang et al., 2022), and addressing these challenges is important for the performance of LBD.

In this thesis, the following research challenges are pursued:

1. **Ambiguity and granularity of biomedical concept representations**: Existing semantic-based knowledge extraction tools (i.e., semantic parsers) entail the mapping of biomedical terms in text to standardized concepts from medical terminological resources, such as the Unified Modeling Language System (UMLS), and representing them by unique concept identifiers (Demner-Fushman et al., 2017). This process creates several challenges in LBD owing to the ambiguity and granularity of biomedical concept representations in medical terminological systems. For instance, biomedical semantic parsers, such as SemRep, fail to disambiguate genes and proteins terms in text to standardized concepts. This

limitation results in mapping genes/proteins in text to multiple concepts on the basis of shared aliases and exact string matching (Kilicoglu et al., 2020). For example, given the following sentence: "*TTF1 gene-expression in human proliferating thyroid-diseases*", the TTF1 gene term is mapped to two distinct gene concepts in UMLS: C1384616 (thyroid transcription factor 1) and C1421218 (transcription termination factor 1), as these genes share the same alias (i.e., TTF1). This is particularly problematic when analyzing cancer literature, as genes and proteins play crucial yet different roles in the development and treatment of cancers, thus failing to distinguish between genes/proteins with identical aliases leads to ambiguity and imprecision in the knowledge extraction process and downstream discovery tasks.

Secondly, the reliance on comprehensive terminological resources for knowledge extraction creates highly granular representation of biomedical concepts. For example, generic and brand drugs are represented uniquely in UMLS and assigned distinct concept identifiers. From the LBD perspective, such granular representation of concepts can result in semantic triples that convey the same underlying knowledge but are represented differently. This in turn leads to a large discovery search space, as there are more unique triples to consider for knowledge discovery (Vlietstra et al., 2017). Ideally, a condensed representation of concepts merges fine-grained concepts into higher-level concepts, thereby requiring users to inspect fewer knowledge instances without compromising the knowledge domain coverage. Hence, a significant challenge in LBD is ensuring that terminological resources knowledge extraction provide sufficient coverage of biomedical sub-domains (e.g., genetics, clinical medicine, molecular biology, etc.) while maintaining the right level of granularity, such that fine-grained concepts are merged and represented as atomic generalized concepts (Pyysalo et al., 2019).

2. **Incomplete extraction of semantic-based knowledge**: Semantics-based LBD employ methods grounded in information retrieval and Natural Language Processing (NLP) to extract knowledge from literature (Henry & McInnes, 2017). These methods have limitations when dealing with complex corpora such as biomedical literature. Most LBD frameworks use domain-specific semantic

parsers, such as SemRep, to identify and extract knowledge from unstructured text in the form of *subject-predicate-object* semantic triples. Domain-specific semantic parsers are reported to identify and extract meaningful knowledge with high precision, however, they are also prone to low recall, thereby resulting in an incomplete knowledge extraction (Kilicoglu et al., 2020). Since LBD is premised on the principle of overlapping assertional knowledge, if explicit relations between concepts are not extracted, then the hidden implicit relations cannot be discovered. In other words, if the A-B association is missing, then the A-C implicit association (via the intermediate B) will never be found. As such, working with incomplete literature-based knowledge remains a significant challenge in LBD (Henry & McInnes, 2017).

3. **Ranking output discoveries:** LBD methods are prone to generating many discovery outputs which makes the task of reviewing all of them complex and, at times, impractical (Henry & McInnes, 2019). Knowledge ranking is an essential component of the LBD process to facilitate pruning of output discoveries to achieve a manageable set of meaningful discoveries. Co-occurrence frequencies tend to favour knowledge discoveries consisting of frequently co-occurring concepts (Thilakaratne et al., 2019). Association measures, such as Chi-square and log-likelihood ratio, are expectation-based (i.e., not null-invariant) statistical measures which are strongly influenced by the total number of null co-occurrences (Henry et al., 2019). Individually, these ranking metrics are capable of prioritising relevant knowledge to some extent, however, this does not necessarily imply novelty and/or interestingness, which are important characteristics in knowledge discovery. Hence, prior research has emphasised (Sebastian et al., 2017) on the need to develop LBD ranking techniques which rely on multiple properties of knowledge instances to prioritise interesting and novel knowledge discoveries.

## 1.3 Research Objectives and Contributions:

The overarching objective of this research work is to investigate and develop novel solutions to the aforementioned challenges faced by semantic based LBD systems. Accordingly, the following objectives are pursued in this thesis:

1. To examine the integration of multiple biomedical-specific knowledge extraction tools for the acquisition of representative and precise semantic-based knowledge from the literature;
2. To investigate semantic consolidation techniques and condensed biomedical terminologies for the consolidation of granular concepts acquired from the literature;
3. To explore and evaluate novel representation learning methods that address limitations of incomplete knowledge extraction from the literature;
4. To assess and compare knowledge ranking measures that prioritize interesting and meaningful discoveries generated by the semantics-based LBD process;
5. To validate and compare the performance of semantics-based LBD in real-world knowledge discovery tasks targeting molecular oncology and drug repurposing.

***Objective #1*** is pursued by combining two well-established biomedical knowledge extraction tools, namely SemRep and PubTator, to acquire semantics-based knowledge in the form of *subject-predicate-object* triples. SemRep is utilized as the primary semantic parser for semantic triple extraction, and PubTator is employed to resolve ambiguous gene/protein concept representations. ***Objective #2*** is pursued by leveraging condensed biomedical terminologies, in addition to semantic alignment and mapping techniques to create consolidated representations of biomedical concepts that encompass more granular concepts.

***Objective #3*** is investigated by exploring the incorporation of state-of-the-art knowledge representation techniques—i.e. Knowledge Graphs (KG), to represent literature-based knowledge in the form of *subject-relation-object* triples, with *subject* and *object* concepts represented as nodes and the semantic relation between them as directed edges. We leverage biomedical knowledge bases and ontologies to integrate curated knowledge with the literature-based KG. Subsequently, we investigate the application of novel KG representation techniques – i.e., KG embeddings – to predict missing edges between existing nodes using Knowledge Graph Completion (KGC) methods. This multi-step knowledge integration and completion approach is intended to address the limitations of incomplete knowledge extraction from the literature.

***Objective #4*** *and* ***objective #5*** are pursued by leveraging the *ABC* discovery paradigm. Explicitly, we premise our knowledge discovery and ranking approach on the following assumptions: (1) an implicit relationship between *A* and *C* may potentially exist if an explicit semantic relationship is lacking; and (2) the interestingness of *ABC* discoveries can be estimated using information theory-centric measures, such that the lower the probability of encountering novel knowledge instances, the more interesting the *ABC* discovery path. We posit that focusing on infrequent knowledge instances increases the likelihood that the implicit relation between them is novel and yet to be explored.

In terms of research outcome, our research has led to the development of ***A****ugmented* ***K****nowledge* ***G****raphs for* ***LBD*** (*AKG-LBD*) framework that extends traditional semantics-based LBD approaches by (i) resolving the limitations of ambiguous knowledge extraction; (ii) introducing a concept consolidation component to consolidate fine-grained concepts into higher-level representations; (iii) augmenting literature-based knowledge via a multi-step knowledge integration and completion methodology that leverages knowledge graph representation learning and curated biomedical knowledge; and (iv) presenting a novel ranking approach that prioritizes *ABC* knowledge discoveries based on notions of *interestingness* and *rarity*.

This dissertation makes the following contributions to the field semantic LBD:

1. Integration of two well-regarded biomedical knowledge extraction tools – i.e., SemRep and PubTator - to enhance the accuracy and representativeness of semantic-based knowledge extracted from the literature;
2. Novel semantic consolidation methods that leverage biomedical terminologies to consolidate fine-grained concepts, thereby reducing the discovery search space without compromising coverage of the knowledge domain;
3. Novel LBD methodology centred on Knowledge Graphs (KGs) and knowledge representation learning techniques to address challenges of incomplete knowledge extraction in semantic LBD via knowledge graph completion;
4. Knowledge ranking metrics based on information theory to prioritize novel and interesting *ABC*-based knowledge discovery paths.

To summarize, Table 1.1 provides an overview of the fundamental challenges in semantic-based LBD, the pursued research objectives to address the challenges, and the contributions made towards enhancing semantic-based LBD.

*Table 1.1: Summary of research challenges, objectives, and contributions*

| Semantic-based LBD challenges | Research objectives | Research contributions |
|---|---|---|
| Ambiguity and granularity of biomedical concept representations | To examine the integration of multiple biomedical-specific knowledge extraction tools for the acquisition of representative and precise semantic-based knowledge from the literature | Integration of two well-regarded biomedical knowledge extraction tools – i.e., SemRep and PubTator - to enhance the accuracy and representativeness of semantic-based knowledge extracted from the literature |
| | To investigate semantic consolidation techniques and specialized biomedical terminologies for the consolidation of semantically similar concepts acquired from the literature | Novel semantic consolidation methods that leverage biomedical terminologies to consolidate fine-grained concepts, thereby reducing the discovery search space without compromising coverage of the knowledge domain |
| Incomplete extraction of semantic-based knowledge | To explore and evaluate novel representation learning methods that address limitations of incomplete knowledge extraction from the literature | Novel LBD methodology centred on Knowledge Graphs (KGs) and knowledge representation learning techniques to address challenges of incomplete knowledge extraction in semantic LBD via knowledge graph completion |
| Ranking output discoveries | To assess and compare knowledge ranking measures that prioritize interesting and meaningful discoveries generated by the semantic-LBD process | Knowledge ranking metrics based on information theory to prioritize novel and interesting ABC-based knowledge discovery paths |

## 1.4 Thesis outline:

This section outlines the chapters of this dissertation with a brief description of their content.

**Chapter 2**: provides a review of existing LBD literature to describe the motivations, underlying principles, and main application areas of LBD. Next, LBD is described as a knowledge discovery framework consisting of a set of interconnected components for literature curation, knowledge extraction, knowledge representation, knowledge discovery, and ranking output discoveries. Using this framework, we provide a review of existing methods and approaches utilized in LBD. This chapter concludes with a detailed taxonomy of LBD approaches and a discussion on outstanding methodological challenges and limitations in current LBD frameworks.

**Chapter 3**: introduces the research design and methodological approach to achieve the outlined objectives. Additionally, the chapter conceptualizes the development of the *AKG-LBD* framework by describing its components and underlying theoretic principles for semantic-based LBD. Next, this chapter describes the primary literature-based sources, knowledge extraction tools, and domain-specific knowledge resources used throughout this research. Finally, we present the evaluation scheme to assess the performance of the AKG-LBD framework in replicating real-world biomedical discoveries compared to traditional LBD frameworks.

**Chapter 4**: describes the implementation of AKG-LBD for the discovery of novel knowledge in the field of cancer genomics and drug repurposing. This chapter outlines the methods and techniques employed to implement the framework components for: *literature curation, semantic-based knowledge extraction, semantic consolidation, literature-based knowledge representation, knowledge integration and completion, and knowledge discovery and ranking.*

**Chapter 5**: presents the results of implementing the AKG-LBD framework using biomedical literature focused on cancers. Next, this chapter presents the evaluation results for the performance of the AKG-LBD framework on real-world knowledge discovery tasks and discusses the findings. Finally, we compare the discovery output of AKG-LBD with other well-established LBD systems.

**Chapter 6**: concludes this dissertation, discusses limitations and challenges, and provides recommendations for future work.

# Chapter 2 Literature Review

This chapter provides the necessary background information on Literature-Based Discovery (LBD), its motivation, main application areas, and reviews current approaches for LBD. **Section 2.1** introduces the underlying principles of LBD and its relevance for knowledge synthesis and discovery in the biomedical domain. **Section 2.2** defines the main components of LBD frameworks in terms of literature curation, knowledge extraction and representation, discovery models, and ranking. **Section 2.3** presents a review of existing formalized LBD systems. **Section 2.4** concludes this chapter by introducing a detailed taxonomy of LBD frameworks and summarizing the gaps in literature.

## 2.1 Literature-Based Discovery:

Literature-Based Discovery (LBD) is a data-driven methodology to discover and synthesize knowledge from published literature by connecting and reasoning over disconnected fragments of knowledge to uncover implicit associations between them. The principles of LBD are premised on Don Swanson's "Undiscovered Public Knowledge" which describes an intuitive syllogism to identify potentially new knowledge via transitive reasoning (Swanson, 1986b). Swanson proposed the *ABC* theory for knowledge discovery; which states that given two concepts *A* and *C* found in disjointed literature fragments, if concept *A* is associated with a concept *B* in one fragment, and the same concept *B* is associated with concept *C* in another fragment, then there is an implicit association between concepts *A* and *C* which is yet to be explored (as depicted in Figure 2.1).

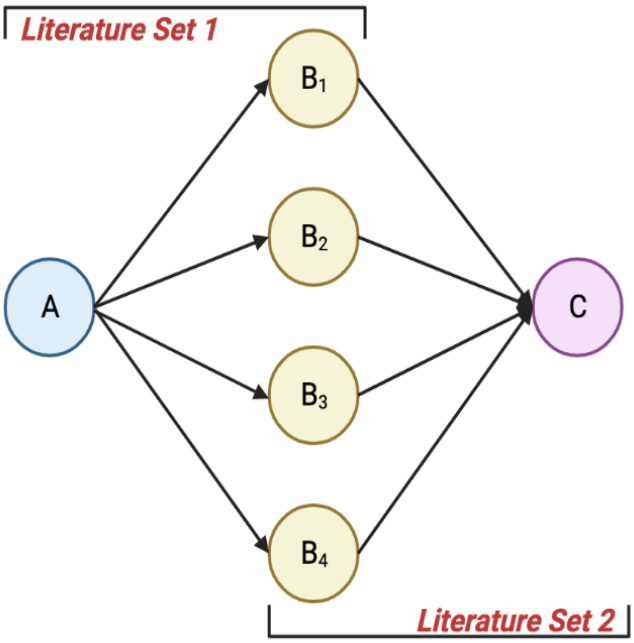


*Figure 2.1: Schematic Overview of ABC Theory*

Swanson applied the *ABC* theory to investigate the unknown association between fish oil (concept *A*) and Raynaud's syndrome (concept *C*) based on their shared connections with intermediary physiological processes (concept *B*) (Swanson, 1986a). This work revealed that one fragment of literature described that fish oil is associated with lowering blood viscosity, reduction of platelet aggregation, and inhibition of vasoconstriction. Concomitantly, the other fragment of the literature revealed that a reduction in blood viscosity and platelet aggregation, together with inhibition of vasoconstriction contribute to the prevention of Raynaud's syndrome. Hence, by applying principles of transitive reasoning, Swanson proposed that "dietary fish oil might ameliorate or prevent Raynaud's syndrome". This LBD-driven hypothesis was later confirmed in an independent clinical in 1989 (DiGiacomo et al., 1989). Crucially, Don Swanson's ABC theory reveals distinctive features of published literature which can be leveraged to discover implicit scientific knowledge that otherwise would remain unknown. The first feature is that knowledge in published literature is *complementary* – i.e., two literature fragments are complementary if new knowledge can be inferred when the fragments are contextualized together, but new knowledge would not be inferred if each fragment is explored in isolation (Swanson, 1986b). The second feature is *non-interaction*; complementary literature are often mutually isolated as none or few articles in one fragment may cite articles in other fragments despite

existing logical scientific connections between them (Swanson, 1986b). Considering the fish oil-Raynaud's syndrome discovery, it was observed that the knowledge within fragment A (i.e., articles on dietary fish oil) complimented findings in fragment C (i.e., articles on Raynaud's syndrome), such that when contextualized simultaneously, new complementary knowledge was inferred. Further, Swanson reported that out of 3,000 articles in fragment A and 1,000 articles in fragment C, only 4 articles had cross-fragment citations (Swanson, 1986a). Swanson suggests that considering complementary but non-interactive fragments of literature can reveal 'undiscovered public knowledge' by identifying implicit connections between them which are not apparent when articles within each fragment are explored in isolation.

### 2.1.1 Applications of LBD in Biomedicine:

Since its inception in 1986, LBD has led to many discoveries and hypotheses in various scientific fields, including biomedical sciences (Zhang et al., 2014, 2021), environmental sciences (Marsi et al., 2017), industrial engineering (Vicente-Gomila, 2014), and crime sciences (Schroeder et al., 2007). However, most LBD applications remain within the clinical and biomedical domains due to its potential in describing basic biological interactions and yield a better understanding of underlying pathological mechanisms, thus, allowing researchers to deduce previously unknown knowledge (Deftereos et al., 2011). In recent years, LBD has been used to propose new treatments for neurological diseases and cancers via drug repurposing (Nian et al., 2022; H.-T. Yang et al., 2017), discovering underlying mechanisms of metabolic diseases (Cairelli et al., 2013), and identifying novel adverse drug events (Hristovski et al., 2016). The following sections provide a review of the most promising LBD applications in the biomedical domain.

**Drug Discovery:**

Traditional drug development is an expensive, time-consuming, and high-risk process. On average, a single drug can take 10-15 years to be approved for clinical use and costs over $1 billion (Sun et al., 2022). LBD can help in reducing the time and cost needed to develop new drugs by discovering hidden knowledge in large-scale literature to provide a better understanding of the underlying physiological interactions necessary for drug development. For example, Zhang et al. applied LBD to discover novel drugs treating

prostate cancer using drug-gene, gene-cancer, and gene-gene semantic triples extracted from the literature to uncover implicit drug -disease associations via a shared gene. Yang et al. leveraged disease-gene and gene-drug associations extracted from the literature to discover new therapeutic indications using drug-target similarity metrics (H.-T. Yang et al., 2017). In another study, knowledge extracted from biomedical literature was integrated with empirical evidence from DNA microarray data to support hypothesis generation for drug discovery (Hristovski et al., 2010). These studies leverage knowledge extracted from the literature to describe interactions between drugs, genes, proteins, and diseases to uncover indirect connections between a drug and a target disease. Understanding these underlying interactions can help with discovering new indications for existing drugs (i.e., drug repurposing), and additionally lead to the development of new drugs (H.-T. Yang et al., 2017).

**Discovering Disease Correlations:**

Comorbidity - defined as the presence of two or more chronic diseases in an individual - is a prevalent phenomenon which complicates healthcare and causes significant limitations in patients' quality of life. Understanding the underlying mechanisms by which one chronic disease may cause the onset of other diseases is necessary for the prevention and/or treatment of comorbidity (Vos et al., 2014). However, investigating all possible co-existing disease pairs is a difficult task given the large number of epidemiological and clinical data (Vos et al., 2014). Further, population-level epidemiological studies do not offer plausible explanations about the underlying physiological and/or pathological mechanisms of comorbid diseases. LBD presents an efficient approach to investigate and explain the relationships between comorbid diseases by extracting shared physiological, genomic, or proteomic biomarkers from the literature (Biswas et al., 2021). Chen et al. introduced a LBD framework to explore relationships between Chronic Obstructive Pulmonary Disease (COPD) and other comorbidities based on shared genomic profiles extracted from biomedical literature (G. Chen et al., 2019). The study discovered several novel comorbidities associated with COPD: *acute lung injury, pulmonary sarcoidosis, Bird Fancier's Lung*, *Eosinophilic Granuloma,* and *Pulmonary Veno-Occlusive Disease*. In another study, the BITOLA LBD system was utilized to discover the underlying genetic interactions linking myocardial infarction and depression (Dai et al., 2019). Rindflesch et

al. leveraged LBD to explain the indirect relationship between inflammatory bowel disease and epilepsy via shared biomarkers (Rindflesch et al., 2018). The study was informed by statistical correlations extracted from Electronic Health Record (EHR) data describing statistically significant associations between epilepsy and inflammatory bowel disease. However, since epidemiological evidence alone cannot explain the relationship between the two diseases, the authors leveraged LBD to extract and navigate literature-based knowledge to explain the underlying mechanisms of this relationship. These studies indicate that LBD offers a novel approach to support epidemiological and clinical research by explaining the underlying mechanisms for disease correlations and interactions, which may result in the onset of comorbidity.

**Drug Interactions and Pharmacovigilance:**

The combined use of multiple drugs, referred to as polypharmacy, is a common therapeutic regimen for patient groups with chronic diseases and comorbidities. Polypharmacy may pose a serious risk for patient safety as drugs may interact and influence the pharmacologic effects of one another when administered in combination. Identifying such drug-to-drug interactions is a critical and complex task due to the lack of sufficient clinical data and knowledge (Bougiatiotis et al., 2020). LBD has emerged as a solution to automate the detection of drug interactions from heterogeneous data sources, including biomedical literature and structured databases. For example, Bougiatiotis et al. applied a graph-based LBD approach to automate the discovery of interacting drug pairs using semantic knowledge extracted from scientific literature combined with structured knowledge from Gene Ontology, DrugBank, and Disease Ontology (Bougiatiotis et al., 2020). Explicitly, this study leveraged semantic relations denoting drug interactions to encode path-based features as numeric vectors, which were subsequently utilized to predict novel interactions between drug pairs. Similarly, Kastrin et al. characterized the discovery of novel drug interactions as a link prediction problem on a large-scale network using information extracted from literature-based sources and curated biomedical knowledge bases (Kastrin et al., 2018). LBD has also been applied to outline the underlying mechanisms of known drug interactions based on shared biomarkers which are in turn associated with observed adverse effects. Hristovski et al. utilized the formalized SemBT LBD system to generate semantic knowledge paths consisting of a *drug* as the starting concept and an *adverse effect*

as the end concept, with the aim of identifying *genes or proteins* as intermediates which may interpret the link between the drug and corresponding adverse effect (Hristovski et al., 2016).

Scientific literature contains a wealth of hidden knowledge on drug interactions and associations with adverse effects. LBD methods leverage such knowledge sources to automate the discovery of novel drug interactions to support pharmacovigilance studies. Additionally, LBD methods have the advantage of providing evidence-based interpretations of the discovered knowledge directly from the literature, which can help with manual reviews to validate discoveries (Wilkowski et al., 2011).

## 2.2 LBD Framework Components:

Knowledge discovery is considered a secondary process as it relies on the output of external methods and techniques grounded in information retrieval, natural language processing, semantic knowledge extraction, and knowledge representation (Thilakaratne et al., 2019). Regardless of the methods and techniques used, a typical LBD framework takes a set of literature as input, extracts and represents knowledge from unstructured literature text, applies a pre-defined knowledge discovery model, and ranks the output discoveries according to some metric which prioritizes interesting discoveries (Henry & McInnes, 2017). Overtime, LBD shifted from a largely manual process to adopting computational techniques to automate knowledge discovery. Hence, there is a wide range of existing LBD methods and techniques which primarily differ in terms of how literature-based knowledge is acquired, represented, and subsequently utilized to infer implicit connections between disparate fragments of the literature. To provide a comprehensive understanding of LBD, we break LBD into a set of distinct components common to most existing frameworks, and discuss the methods and techniques constituting the following components: (i) *literature-curation*, (ii) *knowledge extraction,* (iii) *knowledge representation*, (iv) *knowledge discovery*, and (vi) *knowledge ranking*.

### 2.2.1 Literature Curation Component:

The first component of any LBD framework is concerned with curating a baseline literature corpus as input for knowledge acquisition, representation, and subsequent discovery. This

component has significant influence on knowledge discovery as the extent to which any given LBD system can infer novel and plausible discoveries is dependent on the content and quality of the input literature (Thilakaratne et al., 2020). Literature curation in LBD consists of two main tasks: (i) defining a literature search strategy; and (ii) determining the sections of articles to be used for knowledge extraction.

**Literature search strategy in LBD:**

This task is primarily concerned with defining a comprehensive literature search strategy that can capture literature fragments of interest for downstream knowledge acquisition and representation. When examining existing LBD frameworks (Kim & Song, 2019; Zhang et al., 2014, 2021), we find that there are two types of literature search strategies used: *generalized*, and *specific.* The generalized literature search approach deals with complete literature databases, such as MEDLINE and IEEE Xplore Digital Library, to capture as many literature fragments as possible. In this approach, the literature search is not constrained to specific topics, rather the search strategy involves determining the most suitable literature databases to extract scientific articles published within a specified date range. Examples of a generalized literature search approach in LBD include works by Kastrin et al. who utilized the full MEDLINE database to extract biomedical articles published between 1945 and 2010 (Kastrin et al., 2016). In another study, Huang et al. utilized the Chinese periodical database of literature to extract all articles published between 1989 and 2009 (Huang et al., 2012). Formalized LBD systems, such as LION and MELODI-PRESTO, also use generalized literature search strategies to generate the baseline corpus of biomedical literature (Elsworth et al., 2018; Pyysalo et al., 2019).

Conversely, specific literature search strategies in LBD target specific topics in a scientific domain, which are often determined based on the discovery task of interest. For example, Kim et al. applied LBD to discover interactions between neurodegenerative associated proteins using a subset of biomedical articles focused on neurodegenerative diseases as the input (Kim & Song, 2019). Likewise, Zhang et al. sought to develop LBD methods to repurpose existing drugs for COVID-19 using a literature search strategy focused on coronaviruses (Zhang et al., 2021). It is also worth noting that early LBD studies used the specific literature search strategy to define the baseline literature corpus (Weeber et al., 2001).

Generalized and specific literature search approaches in LBD mostly differ with respect to several factors which may influence subsequent components of the LBD framework. A specific search approach tends to identify literature related to the context of discovery and, therefore, provides the benefit of potentially discovering knowledge from complementary articles within specified fragments of the literature. Conversely, the generalized approach may utilize all existing articles in a given database, which provides the advantage of potentially discovering knowledge from non-interacting fragments of the literature. Further, the generalized approach often results in retrieving a copious amount of articles which may increase the computational cost of the knowledge extraction process, whereas the specific search approach retrieves fewer articles for knowledge extraction and is associated with lower computational costs for knowledge extraction.

**Determining sections of articles for LBD:**

This task has important implications on LBD as it determines the size of the input literature corpus, in addition to the quality of knowledge that can be extracted. Low quality input will have a direct impact on the LBD output and ultimately the potential to make meaningful implicit discoveries (Thilakaratne et al., 2020). Further, the LBD literature curation component requires finding the right balance between representative coverage of the current state of knowledge and a manageable baseline corpus, as the subsequent knowledge extraction tasks can generate varying amounts of literature-based knowledge depending on which sections of scientific articles are being used as input.

Scientific articles typically have a standard format and style consisting of the following main sections (Katz, 2009):

- **Title**: encapsulates the main research topic and may describe the most essential findings of the study.
- **Abstract**: provides a condensed and focused summary of the full study. It is often consistent with the main text and, therefore, can be considered a standalone document that sufficiently represents the full study.
- **Main body**: provides a detailed description of the research topic, methodology, and results. The main text is often organized according to the IMRAD structure - i.e., *Introduction*, *Materials and methods*, *Results*, and *Discussion*).

Different sections of scientific articles have been used as input for knowledge extraction in LBD, including: *titles only, titles and abstracts, full-text* (i.e., title, abstract, and main body), and *keywords* (e.g., MeSH descriptors). Early LBD systems, such as Arrowsmith (Swanson & Smalheiser, 1997), relied on the use of *title only* as LBD input. This approach has the benefit of providing a narrow coverage of the literature which requires low computational resources to extract and represent literature-based knowledge. However, a *title only* LBD input results in missing a considerable amount of meaningful knowledge found in other sections of a given article, as titles are not necessarily descriptive nor representative of the scientific content (Tullu, 2019). With advances in text mining and natural language processing techniques, succeeding LBD studies leveraged *titles and abstracts* as input for knowledge extraction. The combination of *titles* and *abstracts* provides a constrained yet faithful representation of the literature, as the content of abstracts reflect the most important findings in a given scientific article (Moreau et al., 2021). However, this comes at the expense of increasing computational costs required to process the baseline corpus for knowledge extraction when compared to *titles only* as the input (Thilakaratne et al., 2020). Du et al. argue that inclusion of *abstracts* may introduce some noise during knowledge extraction compared to *titles only* and, therefore, proposed using subsections of abstracts by only considering conclusive abstract sentences as input (Du & Li, 2020). Nonetheless, *titles and abstracts* remain the most commonly used input in the majority of recent LBD studies (Thilakaratne et al., 2020).

Few LBD studies have considered using *full-text* of scientific publications as LBD input (Lever et al., 2018). This is mainly due to the significant computational costs associated with processing large-scale corpora, in addition to the potential of introducing an overwhelming amount of noise for knowledge extraction (K. B. Cohen et al., 2010). Further, some literature databases restrict access to full text scientific articles, thereby limiting the number of articles which can be used as LBD input (Thilakaratne et al., 2020). Finally, *keywords* as LBD input has gained some traction among recent LBD studies in biomedicine by using Medical Subject Heading (MeSH) descriptors as surrogates for abstracts and full-text articles (Kastrin et al., 2016). MeSH descriptors are standardized biomedical concepts and used to manually index biomedical publications in PubMed in a process known as MeSH indexing; whereby qualified human indexers read full articles,

identify the important topics and, accordingly, assign MeSH concepts to articles. Hence, this manually-driven indexing process provides a complete concept-based representation of a given article and is deemed a suitable alternative to other LBD input types. However, the use of MeSH descriptors as LBD input limits the applicability of novel text mining and NLP methods for knowledge acquisition, as the input simply consists of standardized keywords as opposed to a literature-based corpus.

Thilakarante et al. investigated the use of common LBD input types (i.e., *titles only, abstracts and titles,* and *keywords*) from a cost-benefit perspective; whereby *cost* was defined in terms of the knowledge acquired from different input types and *benefit* was defined in terms of the number of novel knowledge discoveries via LBD (Thilakaratne et al., 2020). Further, the authors applied a time-slicing approach to replicate the well-known Swanson discoveries. The study revealed that *title and abstract* provided the best cost-benefit compromise, followed by *keywords* and *titles*.

### 2.2.2 Knowledge Extraction Component:

Knowledge extraction is a critical component in the LBD framework which ultimately determines the extent to which implicit connections between disparate literature fragments can be uncovered. Knowledge in LBD can be defined as a set of concepts (entities) extracted from unstructured literature corpora and some notion of association or relation between those concepts. This suggests that knowledge extraction involves two distinct but interrelated tasks: (i) identifying and extracting biomedical entities in text; and (ii) establishing associations or relations between concepts. Commonly used computational methods to achieve these tasks can be grouped into the following categories: *term-based*, *concept-based*, and *semantic-based* methods. The following sections provide a detailed discussion on these high-level categories and outline their strengths and weaknesses in the context of LBD.

**Term-Based Knowledge Extraction:**

Term-based knowledge extraction is one of the first methods used in pioneering biomedical LBD systems. This method operates on surface forms of raw words to identify and extract biomedical terms in unstructured text (Lindsay & Gordon, 1999). Given a literature-based corpus as input, a term-based knowledge extraction method aims to identify all n-gram

terms and phrases while excluding stop or noise words from a predefined list (Gordon & Lindsay, 1996). Hence, this knowledge extraction method is largely a manual process requiring continuous refining of extracted n-grams. Associations between extracted terms are established using n-gram co-occurrences, on the premise that frequently, or rarely, co-occurring terms are likely to have logical biomedical associations (Henry & McInnes, 2017).

Term-based knowledge extraction for LBD has many limitations. Firstly, these methods do not rely on standardized biomedical terminologies and, as a result, fail to capture the variability and ambiguity of biomedical literature (Henry & McInnes, 2017). For example, these methods fail to identify synonymous terms referring to a single concept – e.g., *Vasopressin* and *Argipressin*– without some form of manual intervention. Preiss et al. demonstrated that such lexical ambiguities have adverse effects on the performance of LBD, and that word sense disambiguation is a critical to address the lexical ambiguities in biomedical literature (Preiss & Stevenson, 2016). Secondly, the underlying assumption of associations based on n-gram co-occurrences is inherently flawed as co-occurrence frequency distributions do not capture the semantics in text, which is a crucial factor in the discovery process to elucidate causal or mechanistic relations between biomedical concepts. Even with high co-occurrence frequencies, the captured co-occurring concepts can only be interpreted as associations and not as relations, which further complicates the interpretation of generated knowledge. Hence, despite the success of term-based LBD in early implementations, it has been replaced by more advanced methods which identify and extract knowledge in the form of standardized biomedical concepts.

**Concept-Based Knowledge Extraction:**

Concept-based knowledge extraction in LBD emerged to deal with the ambiguity and complexity of biomedical literature. These methods leverage novel text mining and natural language processing techniques to automate the extraction and normalization of biomedical terms and phrases in text by mapping them to concepts in controlled biomedical vocabularies, such as UMLS (Aronson & Lang, 2010). Hence, these methods streamline the automatic identification and normalization of terms and phrases in biomedical text, while dealing with synonymy by collapsing terms/phrases into atomic concept-based representations. This also provides the benefit of reducing the number of unique concepts

extracted from text (Gopalakrishnan et al., 2019). One of the first formalized LBD systems to use such methods is the *DAD* LBD system which utilized MetaMap to extract and map biomedical terms and phrases in the literature to standardized UMLS concepts (Weeber et al., 2000). Similarly, Gabetta et al. used text mining methods to identify and extract UMLS concepts from biomedical literature as the knowledge extraction component in a LBD framework designed to discover genes associated with heart disease (Gabetta et al., 2013). Virtually all associations in concept-based knowledge extraction methods are established in a similar manner to term-based methods - i.e., using n-gram co-occurrences. Co-occurrence associations are typically established based on a predefined context window size. For example, given the following concept normalized sentence: "Metformin (UMLS:C0025598) use during docetaxel (UMLS:C0246415) chemotherapy (UMLS:C3665472) did not significantly improve (UMLS:C0184511) prostate cancer (UMLS:C0600139)", a window size of 3 will result in the following co-occurrence associations for *Metformin (UMLS:C0025598)*: {C0025598 - C0246415}, {C0025598 - C3665472}, {C0025598 - C0184511}, and {C0025598 - C0600139}, while a smaller window size of 2 will result in fewer co-occurrences: {C0025598 - C0246415}, and {C0025598 - C3665472}. Larger window sizes can result in numerous co-occurrences, some of which may be too noisy or uninteresting for LBD, while smaller window sizes can result in fewer co-occurrences and potentially miss many novel associations (Henry et al., 2019).

Concept-based knowledge extraction resolves some of the limitations in term-based methods, such as addressing lexical ambiguity, but there are several limitations with regards to how associations between concepts are established. The co-occurrence of concepts in a sentence does not necessarily indicate the existence of a biomedically relevant association, which may result in generating false positive literature-based knowledge and, therefore, inaccurate LBD output. Additionally, the co-occurrence approach is known to generate a huge volume of concept associations, due to the flexible notion of what constitutes an association, resulting in a large discovery space and LBD output. This makes the task of reviewing LBD output a laborious and impractical task. Another important limitation is that co-occurrence associations do not provide any insights into the semantics of a given association. For instance, an association between a co-occurring drug and protein

does not indicate whether the underlying mechanism of association is inhibition or activation of the protein, which is a significant source of ambiguity. Providing some insight into the underlying biomedical interactions is an important factor for knowledge discovery (Ahlers et al., 2007). Interestingly, Swanson's renowned *dietary fish oil - Raynaud's disease* hypothesis was entirely based on recognizing the underlying mechanistic associations between the target concepts - i.e., (i) fish oil *lowers* blood viscosity and platelet activity, and (ii) lower blood viscosity and platelet activity *prevent* Raynaud's disease (Swanson, 1986a). Hence, LBD output should provide additional insights into the nature of associations between concepts constituting knowledge discoveries.

**Semantic-Based Knowledge Extraction:**

A semantic-based knowledge extraction in LBD constitutes leveraging domain-specific semantic parsers in an effort to extract relational knowledge from unstructured text by exploiting the contextual syntactic and semantic features of sentences. Given a sentence as input, semantic parsers extract knowledge in the form of *subject-predicate-object* semantic triples, whereby the *subject* and *object* are concepts normalized to controlled vocabularies, and the *predicate* is a semantic relation between them (Luo et al., 2017). Biomedical semantic parsers, such as SemRep (Kilicoglu et al., 2020) and PKDE4J (Song et al., 2015), consist of two interdependent modules for entity extraction and relation extraction. The entity extraction module deals with the task of extracting and normalizing terms in text to standardized biomedical concepts, such as UMLS. While the relation extraction module is responsible for extracting semantic relations by exploiting lexical, syntactic and semantic features of the input sentence (Kilicoglu et al., 2020). For example, given the following sentence from a PubMed abstract (PMID: 32151063): "Tamoxifen (TAM) is a hydrophobic anticancer agent and a selective estrogen modulator (SERM), approved by the FDA for hormone therapy of BC.", SemRep will extract the following semantic triples: [Tamoxifen – ISA – Antineoplastic Agent], [Tamoxifen – ISA – Selective Estrogen Receptor Modulator], and [Tamoxifen – Treats – Breast cancer]. Unlike co-occurrence associations, semantic literature-based knowledge characterizes the underlying biomedical interactions between concepts as *predicates* (Gopalakrishnan et al., 2019).

Hristovski et al. implemented one of the earliest semantic-based LBD frameworks by utilizing SemRep and BioMedLee to capture different types of relational knowledge from

biomedical literature; SemRep was used to capture clinical knowledge related to treatment of diseases, while BioMedLee was used to capture genotypic and phenotypic knowledge (Hristovski et al., 2006). This semantic-based knowledge extraction facilitated leveraging patterns of semantic relations to uncover novel therapeutic mechanisms to treat Huntington's disease. The authors note that such relational patterns cannot be inferred with co-occurrence associations alone, since the underlying relation between concepts is unknown. Similar knowledge extraction approaches can be found in other studies (Ahlers et al., 2007; Cairelli et al., 2015; Cameron et al., 2013; Du & Li, 2020; Zhang et al., 2014). Semantic literature-based knowledge provides many advantages to the LBD process. Biomedical semantic parsers, such as SemRep and PKDE4J, combine the entity extraction and relation extraction into a single pipeline, thereby bypassing the need to use external methods and techniques to establish relations between concepts. Additionally, these domain-specific semantic parsers are reported to extract semantic relations with high precision; one of the most commonly used semantic parsers in LBD, SemRep, is reported to achieve 73%-96% precision for the task of relation extraction (Kilicoglu et al., 2020). Further, a wide range of semantic relations can be extracted by biomedical semantic parsers relating to clinical medicine (e.g., TREATS, DIAGNOSES), molecular interactions (e.g., INTERACTS_WITH, INHIBITS, STIMULATES), disease etiology (e.g., CAUSES, ASSOCIATED_WITH, PREDISPOSES), and pharmacogenomics (e.g., AFFECTS, AUGMENTS, DISRUPTS), in addition to hierarchical, spatial and temporal relations (e.g., IS_A, PRECEDES, LOCATION_OF). Lastly, the extraction of meaningful semantic relations facilitates robust knowledge filtering techniques to eliminate non-informative knowledge instances, as opposed to statistical-based filtering techniques employed for co-occurrence associations which are prone to eliminating potentially novel knowledge (Thilakaratne et al., 2019).

Despite the success of semantic-based knowledge extraction in integrating semantics into LBD frameworks, there are few fundamental shortcomings which adversely impact the knowledge discovery process. Firstly, biomedical semantic parsers suffer from the problem of low recall which results in missing potentially meaningful knowledge (i.e., semantic triples). SemRep is known to achieve recall rates between 55% and 70% (Kilicoglu et al., 2020). Incomplete knowledge extraction is attributed to the performance of the entity

extraction modules. Generally, biomedical semantic parsers rely on dictionary-based text mining pipelines, such as MetaMap, to identify and extract concepts in text. Limitations of dictionary-based text mining methods are well documented and discussed in (Demner-Fushman et al., 2017). Of note is that dictionary-based text approaches fail to detect out-of-dictionary and, therefore, are unable to detect new terminology (Demner-Fushman et al., 2017; Song et al., 2015).

Another fundamental weakness is related to the complex task of relation extraction. Generally, relation extraction for biomedical corpora requires addressing challenges such as coreference resolution, detecting relations in long-distance arguments and identifying implicit relations beyond sentence boundaries (i.e., when there is no explicit textual evidence of a relation) (Drury et al., 2022). While efforts have been made to resolve some challenges such as coreference resolution (Kilicoglu et al., 2016, 2020; Miwa et al., 2012), detecting semantic relations between concepts across sentence boundaries remains unresolved. Studies on biomedical corpora have shown that a vast majority of biomedical relations go beyond boundaries of clausal sentences (Kilicoglu et al., 2020; Rastegar-Mojarad, Elayavilli, Li, & Liu, 2015). As such, failure to address these challenges can aggravate the problem of incomplete knowledge extraction.

Regardless of the limitations, a semantic-based knowledge extraction approach provides many benefits to LBD owing to the integration of semantics into the knowledge discovery process, while limiting the discovery search space to precise and meaningful literature-based knowledge. Prior research has shown that utilizing approaches which extract knowledge from the literature in high volumes will result in a LBD process that outputs numerous potential discoveries. This makes the task of reviewing LBD outputs complex and impractical. Hence, an ideal knowledge extraction approach for LBD should provide a reasonable trade-off between acquiring a tractable amount of high quality knowledge (i.e., containing few false positives) and ensuring that the acquired knowledge is sufficiently complete for knowledge discovery.

Preiss et al. compared the impact of employing concept- and semantic-based knowledge extraction methods on LBD, with a focus on the scale of output discoveries (Preiss et al., 2015). The authors used titles and abstracts of MEDLINE articles as input for the knowledge extraction component. MetaMap was used to extract concept-based knowledge,

and several generic and biomedical-specific semantic parsers were used to extract semantic-based knowledge from the same input. Consequently, concept- and semantic-based approaches were evaluated within a LBD framework to determine the most suitable ones for knowledge discovery. The study revealed that the concept-based approach extracted a large volume of knowledge from the literature, which also translated into generating numerous discovery outputs, ranging from approximately 700 million to 14 billion potential discoveries. Conversely, semantic-based approaches extracted fewer knowledge instances and, as a result, the discovery outputs were several orders of magnitudes fewer than the former approach. Interestingly, LBD via the semantic-based approach replicated a larger proportion of the evaluation discoveries while maintaining a tractable amount of output discoveries. This indicated that the semantic-based approach improved performance of LBD without sacrificing coverage of knowledge within the source literature. It is also worth noting that knowledge extraction using the biomedical-specific semantic parser SemRep resulted in a better LBD performance compared to the generic ReVerb and Stanford parsers. This suggests that domain-specific semantic parsers are better suited for domain-specific knowledge discovery.

### 2.2.3 Knowledge Representation Component:

Raw instances of literature-based knowledge comprise of many latent lexical, semantic, and topological features which can be leveraged to facilitate the discovery process. For example, lexical statistics can be used to represent term- and concept-based associations in terms of co-occurrence frequencies to emphasize features of importance (high co-occurrence frequency) or rarity (low co-occurrence frequency), while distributional semantics methods can be used to represent features of semantic similarity and relatedness. Hence, transforming the raw instances of knowledge into representations that capture latent features is an important task in the LBD framework. This section reviews the common knowledge representation methods utilized in LBD to extract latent features of raw knowledge instances.

**Statistical-Based Knowledge Representation:**

Statistical-based knowledge representation methods in LBD rely on lexical features to represent knowledge instances in terms of direct co-occurrence frequency distributions

across the source literature corpora. These methods are commonly used to represent co-occurrence-based associations capturing features of importance or rarity for a given instance of literature-based knowledge. In the context of LBD, the importance of knowledge instances is often characterized based on the frequency distributions of co-occurring terms or concepts in the literature. The implication is that high co-occurrence frequencies suggest a meaningful association between terms or concepts and, therefore, are good candidates for knowledge discovery. Conversely, co-occurring frequencies can also be used to represent rarity of knowledge instances, based on the premise that rarely co-occurring terms or concepts are less researched and, therefore, more interesting for knowledge discovery. Commonly used statistical methods include concept frequency (Gordon & Lindsay, 1996), document frequency (Ittipanuvat et al., 2014), relative concept frequency (Lindsay & Gordon, 1999), term frequency-inverse document frequency (TF-IDF) (Srinivasan & Libbus, 2004), and mutual information (Wren, 2004).

**Distributional Semantics Knowledge Representation:**

Distributional semantics consist of novel computational techniques to generate vector-based representations of knowledge instances based on patterns of co-occurrences found in the input literature. A range of distributional semantic techniques have been proposed to semantically represent knowledge instances in LBD, including latent semantic indexing (LSI) (Gordon & Dumais, 1998), singular value decomposition (SVD) (Henry et al., 2018), reflective random indexing (RRI) (T. Cohen et al., 2010), and word embeddings (Heo et al., 2019). These techniques are primarily premised on the distributional hypothesis; which states that terms or concepts that occur in similar contexts exhibit similar vector representations and, therefore, are semantically related.

Exploiting lexical and semantic similarity features of literature-based knowledge facilitates knowledge discovery by identifying closely related pairs of non-interacting terms or concepts. Gordon et al. utilize LSI to represent term co-occurrences within the literature as dense vectors and then compute the cosine similarity of non-co-occurring terms following Swanson's ABC framework for knowledge discovery (Gordon & Dumais, 1998). The benefits of this approach lies in reducing the number of discovery outputs to a focused subset of potential discoveries with strong semantic similarity between the source (A) and target (C) terms. More recently, word embedding methods have been utilized to generate

rich semantic-based vector representations of terms and/or concepts by iterating over the input corpus of text to encode semantic and contextual information surrounding target words and/or concepts (Q. Chen et al., 2020; Henry et al., 2018; Heo et al., 2019). Such vector-based representations are amenable to various vector operations, including projection, addition, and subtraction (Tshitoyan et al., 2019).

Distributional semantics provide several benefits for LBD frameworks. Representing knowledge instances as vectors makes them amenable to various machine learning methods for knowledge discovery and ranking, such as nearest neighbour and clustering analysis. Additionally, distributional semantic methods are capable of capturing global and local semantic features which are not readily captured with lexical statistics. Prediction-based methods, such as word embeddings, can efficiently represent domain knowledge from large-scale literature without the need to add knowledge from external sources (Tshitoyan et al., 2019). However, word embeddings are 'black box' methods which generate vector representations that are not readily interpretable (Mikolov et al., 2013). Additionally, applications of distributional semantics in LBD are mostly confined to concept- and term-based knowledge due to the nature of these methods in encoding knowledge directly from the source literature. Relational knowledge (i.e., semantic triples) require graph-based embedding methods which can encode the topological and semantic features of *subject*/*object* concepts in addition to the relations (*predicates*) between them.

**Graph-Based Knowledge Representation:**

Literature-based knowledge can be inherently represented and visualized as co-occurrence networks or multi-relational Knowledge Graphs (KGs). Co-occurrence networks represent concepts as nodes and associations between concepts as undirected edges, indicating that concept pairs co-occur in a sentence or a document. KGs are utilized to represent *subject-predicate-object* semantic triples, with the *subject* and *object* concepts represented as nodes, and *predicates* as multi-relational directed edges. In general, graph-based knowledge exhibits a range of topological and semantic features which can be leveraged in LBD to facilitate knowledge discovery. The use of graph-based analysis is commonly used to analyze complex biological networks and, accordingly, has been adapted for LBD to analyze co-occurrence networks and KGs. Examples of graph-based analysis in LBD include degree centrality (Goodwin et al., 2012), betweenness centrality, closeness

centrality, Eigenvector centrality (Özgür et al., 2010), and personalized PageRank (Petric et al., 2014). These metrics evaluate the topological features of nodes via associated edges to define their relative importance in a graph (Naderi Yeganeh et al., 2020; Özgür et al., 2010). Additionally, proximity-based measures have been utilized to compute association scores for edges based on the graph's topology. Common neighbours, preferential attachment, and Jaccard similarity have been applied on literature-based co-occurrence networks to predict links between a pair of nodes based on their shared connections (Kastrin et al., 2016).

More recently, advances in graph representation learning methods led to novel embedding techniques (i.e., graph embeddings) capable of encoding latent semantic and topological features of nodes and multi-relational edges as low-dimensional vectors (Mohamed et al., 2021). Traditional graph theoretic measures are only effective in extracting prominent topological and similarity features of a graph, however graph embeddings adapt through a learning process to encode optimal topological and semantic features of a given graph. Examples of these techniques include random walk embeddings (*Node2Vec, DeepWalk, LINE*), tensor decomposition embeddings (*DistMult, ComplEx, SimplE*), geometric embeddings (*TransE, RotatE*), and deep learning embeddings (*ConvE, ConvKB, ConvR*) (Mohamed et al., 2021). Random walk embeddings are typically applied for co-occurrence networks, since these techniques view relations between nodes as undirected edges and do not consider their semantic attributes. Geometric, decomposition, and deep learning embeddings are multi-relational embeddings which can encode vector-based representations for nodes and edges separately and, therefore, can be applied to large-scale KGs constructed from literature-based semantic triples. Graph embeddings in LBD are typically used in downstream knowledge discovery tasks, such as link prediction (Kastrin et al., 2016) and entity prediction (Zhang et al., 2021). Graph embeddings have also been used in vector-based operations to rank LBD outputs (Crichton et al., 2020).

Overall, graphs provide many advantages for LBD owing to their versatility as knowledge representation and visualization techniques. Literature-based knowledge, whether extracted as co-occurrence associations or semantic triples, can be intuitively organized and visualized as graphs which can facilitate novel knowledge discovery methods, such as discovery browsing, discovery patterns, and path finding algorithms (Baek et al., 2017;

Wilkowski et al., 2011; Zhang et al., 2021). Further, application of graph theoretic methods on literature-based graphs can generate meaningful knowledge attributes, which are not readily apparent when considering raw instances of knowledge. Finally, graphs constructed from literature-based knowledge can be easily extended by integrating knowledge from external knowledge bases and ontologies (Mohamed et al., 2021), which are typically organized and represented as graphs or triple stores.

### 2.2.4 Knowledge Discovery Component:

In LBD frameworks, knowledge discovery is characterized by the component that takes literature-based knowledge as input and applies user-defined discovery models with the aim of identifying implicit or indirect connections between non-interacting knowledge instances. Current knowledge discovery models used in LBD range from the traditional *ABC* model to prediction-based models that can predict future links between non-interacting knowledge entities. The choice of a discovery model is typically influenced by the type of knowledge extracted from the literature (i.e., term-, concept-, or semantic-based) and how this knowledge is represented. For example, discovery models for link prediction are applicable to relational-based knowledge represented as a graph or co-occurrence network. Hence, the knowledge discovery component can be considered as a secondary process that relies on the output of preceding LBD components. In this section, we provide a detailed review of current discovery models used in LBD frameworks.

**ABC-Based Discovery Models:**

Traditional discovery models in LBD are based on Swanson's *ABC* theory to identify associations between non-interacting concepts via transitive reasoning. Explicitly, the *ABC* theory states that given two non-interacting concepts *A* and *C*, if *A* has a direct association with concept *B*, and *B* has a direct association with *C*, then there is an implicit association between *A* and *C*. This theory was formalized as the open- and closed-based *ABC* model for knowledge discovery (Henry & McInnes, 2017). The open-discovery variant starts with a single predefined source concept (*A*) and aims to find intermediate concepts (*B*) which are used to identify the target concept (*C*). Conversely, the closed-discovery variant starts with predefined *A* and *C* concepts and seeks to identify meaningful *B* concepts which are used as bridges to link the two predefined concepts (i.e., *A* and *C*). In essence, the *ABC*

model aims to discover and expound the indirect association between *A* and *C* via an intermediate concept *B*. This knowledge discovery model has contributed to many biomedical discoveries and used to expound the indirect association between dietary fish oil and Raynaud's syndrome, somatomedin and arginine (Swanson & Smalheiser, 1997), and magnesium and migraine (Swanson, 1988).

The *ABC* paradigm remains one of the most prevalent knowledge discovery models used in LBD frameworks due to its intuitive syllogism in discovering hidden knowledge associations between a source and target concepts (i.e., *A* and *C*). Further, it is adaptable to nearly all forms and representations of literature-based knowledge, including term-, concept-, and relational-based knowledge (Baker & Hemminger, 2010). However, despite its prevalent use, the *ABC* paradigm tends to generate a large number of candidate discoveries, making it difficult to assess or interpret manually without effective filtering or ranking procedures (Smalheiser, 2012).

Wilkowski et al. expanded on the underlying principles of the *ABC* theory to introduce the *AnC* discovery paradigm, where *n* is a chain of multiple intermediate concepts -- i.e., $A$-$B_1$-$B_2$-$B_3$...$B_n$-$C$ (Wilkowski et al., 2011). The *AnC* model is typically applied to literature-based knowledge represented as graphs since it leverages graph theoretic measures, such as degree centrality, to iteratively generate 'discovery sub-graphs'. Specifically, the discovery process begins with a seed concept (*A*) to create the first sub-graph, then the most influential nodes ($B_1$) are identified and selected, using degree centrality measures, as new seed concepts to expand the initial sub-graph. This iterative discovery approach tends to provide broader insights into the underlying associations between the source (*A*) and target (*C*) concepts due to the presence of multiple bridging intermediate concepts. However, the *AnC* paradigm requires active intervention/input from users to limit the sub-graph expansion task.

Another graph-based discovery model is the *discovery patterns* approach by Hristovski et al. which leverages semantic triples represented as knowledge graphs (Hristovski et al., 2006). Discovery patterns expand on the *ABC* theory to introduce graph traversal techniques that search a given KG based on a set of semantic constraints to identify indirect relations between a source concept (*A*) and a target concept (*B*). These semantic constraints refer to a sequence of semantic types of nodes and relations. To illustrate this notion,

Hristovski et al. proposed the *maybe_treats* discovery pattern for drug repurposing: if a disease (concept *A*) is related to a pathological function (concept *B*) via a causal relation, and a drug (concept *C*) is related to the same pathological function (concept *B*) via an inhibitory relation, then the drug (concept *C*) *maybe_treats* the disease (concept *A*) (Hristovski et al., 2006). In another recent study, similar discovery patterns were proposed to discover novel treatments for COVID-19 using the following constraints: (Drug A) - INHIBITS/INTERACTS_WITH - (Biological Function B) - AFFECTS/PREDISPOSES/CAUSES - (COVID-19) (Zhang et al., 2021). The benefits of this approach lie in providing fully interpretable discovery paths which may explain the underlying mechanistic relations between a source and a target concept. Further, since this approach relies on strict graph traversal constraints, the number of generated candidate discovery paths are limited and can be easily assessed and reviewed by domain experts. However, some degree of domain knowledge is necessary to define a logical pattern of nodes and edges.

Overall, ABC-based models offer an intuitive approach to knowledge discovery and have contributed to real-world discoveries in the biomedical domain. Virtually all formalized biomedical LBD systems - including LION LBD (Pyysalo et al., 2019), MELODI-PRESTO (Elsworth & Gaunt, 2021), and Arrowsmith (Smalheiser & Swanson, 1996) - utilize ABC-based models to uncover hidden associations in literature-based knowledge. Further, the underlying notion of overlapping assertional knowledge tends to generate fully interpretable discovery paths which can be easily assessed and reviewed by domain experts to validate their significance. However, the extent to which ABC-based models can generate significant discoveries depends on the completeness of knowledge extracted from the literature; if the A-B association is missing due to incomplete knowledge extraction, then the A-C implicit association (via the intermediate B) will not be discovered.

**Prediction-Based Discovery Models:**

Prediction-based discovery models incorporate data-driven methods leveraging lexical, topological, and semantic features to predict links between disparate knowledge entities. Discovery models in this category are distinguishable from traditional approaches due to the utilization techniques and paradigms which go beyond the *ABC* theory of overlapping assertional knowledge, such as link prediction and entity prediction.

Several LBD studies have characterized knowledge discovery as a link prediction task, where the goal is to predict pairs of entities that can potentially be linked in the future based on the current state of knowledge (Crichton et al., 2018; Eronen & Toivonen, 2012; Kastrin et al., 2014, 2016; Zhang et al., 2021). These studies leverage graph-based representations to extract inherent topological and semantic features, which are in turn utilized for supervised and unsupervised link prediction. For example, Kastrin et al. represented literature-based knowledge as a large-scale co-occurrence network and utilized unsupervised and supervised link prediction methods on a set of unlinked nodes to predict the probability that a given pair of nodes may establish a link in the future (Kastrin et al., 2016). For unsupervised link prediction, Adamic-Adar, Common Neighbours, and Jaccard Index proximity measures were utilized based on the assumption that nodes with similar neighbours are more likely to establish a direct link in the future. In the supervised link prediction (learning-based) setting, feature vectors were created by combining proximity measures and used as input for decision trees, k-nearest neighbours, logistic regression, naïve Bayes, and random forest to classify whether a link can potentially exist between unlinked nodes. This study revealed that supervised learning-based approaches outperformed unsupervised link prediction for knowledge discovery.

Recently, neural network models have been utilized as link prediction methods for knowledge discovery. Crichton et al. utilized vector-based representations generated from network embedding algorithms (DeepWalk, LINE, and Node2Vec) as input for neural network models to predict future links between nodes as an approach for knowledge discovery (Crichton et al., 2018). The authors concluded that neural networks have great potential for knowledge discovery as they perform better than baseline approaches, such as Jaccard Index, in predicting links between nodes with no or few common neighbours.

In addition to link prediction, knowledge discovery can also be characterized as a task of entity prediction by applying KG embedding algorithms on literature-based KGs. Zhang et al. represented *subject-predicate-object* semantic triples extracted from biomedical literature as a KG and applied various KG embedding algorithms to generate vector-based representations of nodes and semantic edges (Zhang et al., 2021). Consequently, entity prediction was performed as a ranking task, whereby embeddings for a given *predicate* and *object* are used as input in a prediction function which generates high scores for plausible

combinations of *subject-predicate-object* triples. This approach was used to discover novel treatments for COVID-19 by using embeddings for *treats* and *COVID-19* as inputs for the prediction function.

Prediction-based methods for knowledge discovery have the potential to uncover novel links between disparate knowledge entities and can help overcome limitations of ABC theory-based models (Kastrin & Hristovski, 2021). Recent advances in representation-based learning methods can facilitate the development of novel knowledge discovery models that are capable of predicting previously unknown knowledge with high accuracy (Jiang et al., 2020; Mohamed et al., 2021). However, these approaches are derived from black-box algorithms which are not inherently interpretable (Rudin, 2019). Further, applications of prediction-based discovery models are limited to large-scale literature-based knowledge and, therefore, require significant computational resources to develop and train highly accurate models.

### 2.2.5 Filtering and Ranking Component:

The number of discoveries generated by the aforementioned discovery models can be too many to review and assess manually by domain experts. Hence, limiting the number of output LBD to a small subset of meaningful and interesting knowledge is an important component in LBD frameworks. Limiting the output of the knowledge discovery component can be done through knowledge filtering and knowledge ranking. Commonly used techniques for knowledge filtering and ranking are discussed in the sections below.

**Filtering Techniques:**

Filtering techniques in LBD are focused on identifying and eliminating uninformative, spurious, or uninteresting concepts or relations constituting output discovery paths. For example, concepts such as *neoplasms, pharmacologic substance,* or *pathologic processes* are recurring broad concepts in biomedical literature which may not provide any useful information for knowledge discovery. Similarly, relational-based knowledge may include generalized hierarchical or spatio-temporal semantic relations which do not relay any useful insights into the underlying mechanistic association between concepts.

Knowledge filtering techniques typically leverage hierarchical and/or semantic type information from biomedical ontologies and standardized vocabularies, such as the UMLS

or MeSH. For example, Qian et al. utilized MeSH to identify and eliminate concepts in the first and second levels of the MeSH tree hierarchy as they were deemed too broad for knowledge discovery (Qian et al., 2012). Hristovski et al., leveraged the UMLS Semantic Network, which classifies biomedical concepts into several semantic types and high-level groups, to restrict the output of a ABC-based knowledge discovery model to concepts with semantic groups *Physiology* while eliminating all concepts with other semantic groups (Hristovski et al., 2006). Similar approaches are also used to filter uninformative or uninteresting semantic relations given a specified discovery task. For example, LBD focused on drug repurposing may focus on retaining semantic relations denoting therapeutic and substance interactions (e.g., *TREATS, PREVENTS, INHIBITS, STIMULATES)* (Zhang et al., 2021).

**Ranking Techniques:**

Ranking refers to the task of sorting output discoveries to prioritize interesting and meaningful knowledge over uninteresting and uninformative ones. This task is necessary to limit the number of output of ABC-based discovery models exhibiting the small world phenomenon (Henry & McInnes, 2017). In this context, the small world phenomenon occurs when intermediate knowledge entities (*B*) are linked to many target entities (*C*), thereby generating many *ABC* discovery paths.

Ranking measures in LBD utilize features of knowledge instances extracted during the knowledge representation phase to sort output of discovery models based on lexical, semantics, or graph-based topological features. For example, literature-based knowledge represented in terms of co-occurrence frequencies are utilized to rank *ABC* discovery paths based on the popularity of the *A*-*B* and *B*-*C* associations (Pyysalo et al., 2019). Vector-based representations are utilized to compute scores for ABC discovery paths using distance metrics to estimate the relatedness between the A and C entities (Heo et al., 2019). Similarly, topological features of graph-based knowledge representations can be utilized to compute proximity-based measures for A and C nodes/entities (Kastrin et al., 2016). Hence, ranking measures are defined in parallel to knowledge representation methods - i.e., statistical-based, distributional semantics-based, and graph-based ranking techniques.

Statistical-based ranking relies on lexical features to rank discovery paths based on popularity, rarity, strength of association, or linking term count (Henry & McInnes, 2019).

Measures of popularity are premised on the notion that high co-occurrence frequencies suggest meaningful associations between entities extracted from the literature and, therefore, are prioritized over rarely co-occurring entities which are assumed to be spurious or insignificant associations. Conversely, measures of rarity assume that low co-occurrence frequencies represent novel and interesting associations for knowledge discovery. Association measures utilize co-occurrence frequency distributions to quantify discovery paths based on the likelihood that the A-B and B-C associations co-occur together more often than expected. Examples of association measures include: odds ratio, mutual information, chi-square, log-likelihood ratio, and dice coefficient (Henry et al., 2019). Lastly, linking term count quantifies discovery paths based on the number of unique intermediate entities between the source and target entities. A study comparing the performance of multiple statistical-based metrics found that linking term count provides the best ranking approach for LBD (Yetisgen-Yildiz & Pratt, 2009). Overall, statistical-based ranking is relatively easy to compute and rationalize for LBD given their reliance on basic lexical statistics. However, these metrics can only quantify direct co-occurring associations between knowledge entities - i.e., A-B and B-C. As such, ranking scores for ABC or AnC discovery paths are obtained by summing or averaging direct associations.

Distributional semantics ranking measures utilize vector-based representations of knowledge entities to compute distance-based metrics quantifying the association between entities constituting a given discovery path (Henry & McInnes, 2017). Examples of commonly used distance metrics include cosine distance (Gopalakrishnan et al., 2018), euclidean distance (van der Eijk et al., 2004), and information flow (Bruza et al., 2006). These measures can be utilized to quantify the association between directly linked entities (e.g., A-B and B-C) and indirectly linked entities (e.g., A-C).

Graph-based ranking measures leverage graph characteristics to compute ranking scores for output discovery paths. Centrality measures, such as degree centrality, betweenness centrality, and closeness centrality, are used as indicators to quantify the significance of a knowledge entity in a graph (Özgür et al., 2010; Wilkowski et al., 2011). More recently, proximity measures have been adapted to rank graph-based discovery paths based on the notion of common neighbouring nodes (Kastrin et al., 2016). Explicitly, this ranking measure is based on the number of shared links between the source (A) and target (C)

nodes, such that if the source and target have many links to other nodes in common, they are considered to have strong implicit association.

Overall, there is no consensus on the best approach for ranking in LBD. Few studies have compared the performance of statistical-based ranking measures and reported varying results (Pyysalo et al., 2019; Yetisgen-Yildiz & Pratt, 2009). Recent LBD studies proposed to combine multiple metrics to rank output discoveries. For example, Gopalakrishnan et al. combined co-occurrence frequency and degree centrality metrics to rank the output of a graph-based LBD framework (Gopalakrishnan et al., 2018), while Sybrandt et al. proposed to combine distributional semantics and graph-based ranking metrics (Sybrandt & Safro, 2018). A combined approach to knowledge ranking can provide more flexibility by leveraging several features/properties of literature-based knowledge to prioritize novel and meaningful discoveries (Thilakaratne et al., 2019).

## 2.3 Review of LBD Systems:

In this section, we review existing formalized LBD systems that are publicly available. While there exist several LBD frameworks that have not been formalized as web-based systems, this review primarily focuses on systems that are accessible, with the aim to perform a comparative analysis within the context of this thesis. Table 2.1 presents a concise classification of the methodologies employed by these systems.

*Table 2.1: Classification of formalized LBD systems*

| **LBD system** | **Components** | | | |
|---|---|---|---|---|
| | *Literature curation* | *Knowledge extraction* | *Knowledge discovery* | *Filtering and ranking* |
| Arrowsmith | Titles only | Term-based co-occurrence associations | Closed-based ABC discovery | Semantic type filtering and probabilistic ranking based on a regression model |
| BITOLA | Titles and abstracts | Concept-based co-occurrence associations | Open- and closed-based ABC discovery | Association rule mining to filter and rank generated discoveries |
| SemBT | Titles and abstracts | Semantic-based knowledge in the form of *subject-predicate-object* triples | Open- and closed-based ABC discovery | Semantic triple counts to rank discoveries |

| LBD system | Components | | | |
|---|---|---|---|---|
| | *Literature curation* | *Knowledge extraction* | *Knowledge discovery* | *Filtering and ranking* |
| MELODI | Titles and abstracts | Semantic-based knowledge in the form of *subject-predicate-object* triples | Open- and closed-based ABC discovery | Semantic triple counts, odds ratio, and including Fisher's exact test |
| LION-LBD | Titles and abstracts | Concept-based co-occurrence associations | Open- and closed-based ABC discovery | Co-occurrence frequency and statistical association measures |

**Arrowsmith:**

Arrowsmith is one of the first co-occurrence LBD systems that uses the ABC discovery to uncover links between disparate sets of biomedical literature (Smalheiser, 2005; Swanson & Smalheiser, 1997). The system requires users to identify source and target terms to initiate the LBD process. Subsequently, titles of biomedical publications which include the source or target term are retrieved from MEDLINE. A term-based approach is used to extract literature-based knowledge, whereby all n-gram co-occurrences of terms that are found in the titles are extracted. Generic and common terms are eliminated, based on a predefined stoplist, and the remaining terms are mapped to UMLS via MetaMap to assign each term a standardized semantic type. Additional filters allows users to retain terms which belong to one or more of desired semantic types. ABC discoveries are generated on the basis of a probabilistic regression model to estimate the probability of relevance for each intermediate B term (Torvik & Smalheiser, 2007). Finally, the system generates a ranked list of B terms based on a cohesion score.

Arrowsmith is a pioneering LBD system that has influenced the development of many subsequent systems. It uses a rudimentary but effective approach to uncover co-occurrence based knowledge from the literature via a user-friendly interface. However, Arrowsmith has several limitations which can be attributed to the limited computational resources at the time of system development. The knowledge extraction approach achieves high recall but low precision since all n-gram co-occurrences are considered valid associations (Henry & McInnes, 2017). Additionally, this approach does not consider synonymous terms as concepts/entities which introduces further ambiguity and imprecision into LBD. Another

source of ambiguity is the co-occurrence associations which do not indicate the nature of the relationship between A-B and B-C terms. Lastly, Arrowsmith uses titles of biomedical articles as input for knowledge discovery, which means that it is potentially excluding huge amount of valuable knowledge found in abstracts.

**BITOLA:**

Developed by Hristovski et al., BITOLA is an interactive LBD system that specializes in the discovery of disease-gene associations (Hristovski et al., 2005). It extracts concept-based knowledge from biomedical literature using the co-occurrence approach, whereby the co-occurrence of biomedical concepts in a scientific article is considered a meaningful association. Additionally, the system integrates biomedical knowledge from curated genomic databases (i.e., HUGO and LocusLink) related to the chromosomal locations for diseases and genes. To extract associations, BITOLA uses association rule-mining to identify relevant ABC discoveries as follows: given a predefined source concept (A), relevant intermediate (B) terms are identified using association rule A → B, then target (C) concepts co-occurring with the intermediate (B) concepts are also identified using association rule B → C. Then using the background curated knowledge, it eliminates target concepts whose chromosomal location does not correspond to the location of the start concept (A). The final list of ABC discoveries are presented using a ranking metric that combines the confidence of A→B and B→C associations. BITOLA supports open and closed-based discovery, however, it restricts users to discovery tasks involving a gene as the source concept and a disease as the target concept.

The limitations of BITOLA are attributed to its co-occurrence approach for knowledge extraction, which results in extracting spurious associations (Hristovski et al., 2006). Additionally, the lack of semantic relations between concepts means that users are required to manually review articles to gain an understanding of the nature of interaction.

Hristovski et al, also developed the semantic-based version of BITOLA, called SemBT, by leveraging the SemRep knowledge extraction tool to extract *subject-predicate-object* triples from the literature (Hristovski et al., 2010). SemBT is also a specialized LBD system that focuses on tasks related to gene-disease associations. However, rather than using association rule mining to identify A-B and B-C associations, SemBT relies on meaningful

semantic knowledge. Further, it uses raw counts of semantic triples to rank generated discoveries.

**MELODI PRESTO:**

MELODI is a semantic-based LBD system that aims to discover hidden biomedical knowledge by identifying intersecting semantic triples extracted from biomedical literature (Elsworth et al., 2018). The system uses SemRep to extract *subject-predicate-object* triples from abstracts and titles published in PubMed. The extracted triples subsequently undergo an enrichment step which involves comparing the frequency of *subject* and *object* concepts within a subset of literature with global frequencies across all abstract/titles in PubMed to eliminate common and spurious semantic triples. The enrichment step also prioritizes triples which occur frequently in independent articles, thereby reducing the effect of triples which occur frequently in single articles. The strength of semantic triples is represented using statistical and raw co-occurrence measures including Fisher's exact test, Odds ratio, and triple frequency.

A previous version of the system represented the generated ABC discoveries using property graphs, however, the most recent version, called MELODI PRESTO, represents the knowledge in a tabular format and provides users with an Application Programming Interface (API) to interact with the system (Elsworth & Gaunt, 2021).

As a semantic-based LBD system, MELODI has the potential to uncover meaningful knowledge discoveries from the literature and to provide mechanistic insights on the relationship between non-interacting biomedical concepts. However, the system relies on SemRep to extract knowledge, which means that the extracted knowledge is possibly incomplete due to low recall of SemRep (Kilicoglu et al., 2020).

**LION-LBD:**

LION-LBD is a recently developed LBD system leveraging text mining tools to extract biomedical concepts from the literature and representing them as co-occurrence associations (Pyysalo et al., 2019). It integrates PubTator to extract and normalize biomedical concepts, in addition to a dedicated NLP classifier that classifies sentences in the literature into one of the 37 categories of the hallmarks of cancers taxonomy. Thus, in addition to detecting genes/proteins, chemicals, diseases, and mutations, the system detects references to the biological interactions which lead to cancers. It uses n-gram co-

occurrence frequencies to represent associations between biomedical concepts, and implements open and closed ABC discovery models. Generated discoveries are ranked using raw frequency counts in addition to multiple statistical metrics such as Chi-square, t-test, log-likelihood ratio, and Jaccard coefficient. It provides an intuitive web-based application and uses a graph-based interface to represent ABC discovery paths.

LION-LBD is a novel LBD system that leverages state-of-the-art text mining tools and NLP methods to extract knowledge focused on cancer biology from the literature. As a co-occurrence based system, it achieves high recall in knowledge extraction but the co-occurrence associations do not provide insights into the mechanistic interactions between concepts (Henry & McInnes, 2017). Nevertheless, LION-LBD has demonstrated that it is capable of uncovering novel cancer discoveries from the literature with high precision (Pyysalo et al., 2019).

## 2.4 Conclusion:

This chapter presented a comprehensive review of the underlying principles, motivations, and components of LBD as a knowledge discovery framework. LBD, as conceived by Don Swanson's early works, is based on the notion that scientific knowledge is complementary but resides in segregated silos of non-interacting literature. Additionally, the accelerated rate of scientific publications, especially in biomedicine, suggests that the current state of knowledge is well-advanced but since it is dispersed across several publications, it remains untapped. Hence, logically connecting these silos of knowledge can uncover hidden links between pieces of disparate literature-based knowledge.

Contemporary approaches to LBD rely on external computational methods and techniques to extract, represent, and analyze literature-based knowledge. In this regard, LBD can be described as a framework consisting of the following components: literature curation, knowledge extraction, knowledge representation, knowledge discovery, and ranking generated discoveries. Each component is characterized by several approaches, and the output from one component is used as input for the next component. Figure 2.2 depicts a taxonomy of the common approaches utilized by each component of the LBD framework.


*Figure 2.2: Taxonomy of LBD Framework Components*

The literature curation component defines the input data sources used for extraction of literature-based knowledge. This component involves defining a literature search strategy and determining which article sections to utilize for knowledge extraction (e.g., abstracts and titles, full-text, MeSH descriptors). The output of this component consists of literature-based corpora or a set of MeSH descriptors in addition to associated metadata, such as article publication dates.

The knowledge extraction component relies on various text mining and relation extraction methods to extract knowledge from literature text. In this setting, literature-based knowledge is characterized by a set of concepts and some notion of association or relationship between them. Contemporary LBD frameworks utilize two forms of literature-based knowledge: concept-based and relational-based knowledge. Concept-based knowledge is extracted by identifying and normalizing domain-specific terms in literature text, and associations between normalized concepts are established based on their co-occurrence in a sentence. Relational-based knowledge consists of *subject-predicate-object* triples extracted by domain-specific semantic parsers, whereby the *subject* and *object*

entities are normalized biomedical concepts and *predicate* is a semantic relationship between them. The output of this component is a large set of co-occurrence associations or semantic triples extracted from literature text.

Knowledge representation is the component that represents the extracted literature-based knowledge in a computable format, such as co-occurrence networks or large-scale knowledge graphs. Knowledge representation facilitates the extraction of latent features which can be utilized for knowledge discovery, filtering, and ranking.

Knowledge discovery component applies a used-defined discovery model to literature-based knowledge to uncover implicit links and associations between disparate knowledge entities. Several discovery models have been proposed, including the ABC-based and prediction-based models. The output of this component is a set of potential discovery paths denoting potential implicit associations.

Lastly, the knowledge filtering and ranking component eliminates non-interesting discoveries and returns a ranked list of discovery paths. This component is often accompanied by various visualization techniques to assist domain experts in reviewing potential discovery paths.

Overall, LBD is considered a relatively mature field with many novel approaches being introduced in recent years addressing major challenges, such as over-generation of discoveries (Henry, 2019), link prediction for open- and closed-based discovery (Crichton et al., 2020), entity prediction-based discovery (Zhang et al., 2021), and improving extraction of concept-based knowledge from literature text (Crichton et al., 2017). However, few methodological challenges remain unaddressed. First, commonly used biomedical semantic parsers for relational-based knowledge extraction suffer from the problem of low recall, which results in incomplete extraction of literature-based knowledge. Considering the ABC model of discovery, if relations between A and B or B and C are missing, then the implicit relation between A and C will not be discovered. As such, limitations in relational-based knowledge extraction may prohibit the wide-scale adoption of LBD among the scientific community. Recent LBD studies emphasize the need for novel approaches to address such challenges by automatically inferring new knowledge from existing knowledge or by leveraging manually curated knowledge from biomedical knowledge bases. Secondly, normalization of biomedical terms in literature text results in

mapping closely related terms to multiple concept representations from standardized biomedical terminologies. This can lead to a highly granular concept-based normalization of biomedical terms, thereby significantly increasing the number of unique concepts and, in turn, literature-based knowledge instances. Vlietstra et al. called for investigating mechanisms to associate and collapse closely related concepts with one another, thereby condensing the number of literature-based knowledge instances without compromising the coverage of domain knowledge (Vlietstra et al., 2017). Such mechanisms can also reduce the discovery search space, thereby requiring LBD users to review fewer potential discoveries (Vlietstra et al., 2017). We posit that addressing these challenges can improve the adoption of LBD as an effective knowledge discovery methodology in biomedicine.

# Chapter 3 Research Approach and Design

In the previous chapter, we reviewed existing literature to understand LBD as a multi-component framework that processes scientific literature to extract literature-based knowledge by implicitly connecting disparate knowledge instances to discover previously unknown knowledge. Our review revealed that conventional semantic-based LBD frameworks rely on domain-specific semantic parsers to extract knowledge in the form of semantic triples, however these framework do not address the challenges posed by incomplete knowledge extraction and the granular representation of biomedical concepts. Further, filtering and ranking methods applied to LBD outputs are primarily focused on prioritizing knowledge discoveries that are either statistically significant or consist of frequently co-occurring knowledge instances in the literature, thereby dismissing novel knowledge instances which usually occur rarely (i.e. with a lot frequency for it to be detected by co-occurrence methods) in the literature.

Motivated by the findings of the literature review, the overarching objective of this dissertation is to explore novel knowledge graph-centric solutions to tackle the limitations of semantic-based LBD. Our focus is to (i) leverage state-of-the-art knowledge extraction tools and techniques to extract semantics-based knowledge that resolves the ambiguity of biomedical concepts and faithfully represents the current state of knowledge in the literature; (ii) consolidate the representation of granular biomedical concepts by leveraging condensed terminologies with semantic alignment and mapping techniques; (iii) address the incompleteness of literature-based knowledge using a multi-step knowledge integration and knowledge graph completion approach to predict missing semantic relations; and (iv) develop LBD filtering and ranking measures premised on information theory to prioritize meaningful discoveries that are not captured by traditional statistical association measures. This chapter presents our methodological approach to achieve the outlined objectives. While doing so, we conceptualize the development of the ***A**ugmented **K**nowledge **G**raphs for **LBD*** (AKG-LBD) framework that leverages the strengths of knowledge graphs, semantic consolidation techniques, and knowledge integration and completion methods. This chapter is organized as follows. **Section 3.1** outlines the solution approach to address challenges and limitations of traditional LBD frameworks. **Section 3.2** introduces the

AKG-LBD framework which extends traditional LBD frameworks by encapsulating the proposed solution approaches outlined in the previous section. **Section 3.3** describes the data and tools used throughout this research, including the literature sources and semantic tools used to extract and represent knowledge from the literature. **Section 3.4** outlines the evaluation framework used to assess the performance of the AKG-LBD framework. Finally, **Section 3.5** concludes this chapter by summarizing the main components of AKG-LBD and setting the stage for the next chapter, which describes the implementation and evaluation of the framework.

## 3.1 Addressing Challenges of Traditional LBD Frameworks

This section presents our approach to address the limitations of traditional semantic-based LBD frameworks as noted in the LBD literature. In sections 3.1.1 and 3.1.2, we discuss the challenges pertaining to the extraction of semantics-based knowledge from the literature and its consequence on LBD. Then in section 3.1.3, we discuss the challenges in filtering and ranking *ABC* discovery paths generated by LBD.

### 3.1.1 Ambiguity and Granularity of Biomedical Concept Representations:

Semantic-based knowledge extraction (i.e., semantic parsing) is the process of extracting structured knowledge in the form of *subject-predicate-object* semantic triples from unstructured textual sources, such as the literature (Milošević & Thielemann, 2023). The *subject* and *object* are standardized biomedical concepts, and the *predicate* is a semantic relation between them. For example, given the following sentence in (Day et al., 2020): "Tamoxifen (TAM) is a hydrophobic anticancer agent and a selective estrogen modulator (SERM), approved by the FDA for hormone therapy of BC.", the semantic parsing process extracts the following semantic triples:

- C0039286: Tamoxifen – ISA – C0732611: Selective Oestrogen Receptor Modulator
- C0039286: Tamoxifen – TREATS – C0006142: Breast Cancer (BC)

Semantic parsing involves identifying biomedical terms in text and then disambiguating them to standardized concepts using comprehensive biomedical terminologies, such as the UMLS (Kilicoglu et al., 2020). In the biomedical domain, disambiguation of biomedical

terms present a significant challenge, because terms used in the literature can have multiple meanings depending on the surrounding context and, thus, may be represented as different concepts based on the context (Preiss & Stevenson, 2016). This is particularly problematic when dealing with gene or protein terms, as genes and proteins are often referenced in the literature by non-unique aliases (i.e., short name), such that the same alias may refer to different genes/proteins. For example, the following sentences taken from published articles use the "TTF1" alias non-uniquely to refer to two different genes:

- **Sentence 1**: "mTOR Inhibition Promotes TTF1-Dependent Redifferentiation and Restores Iodine Uptake in Thyroid Carcinoma Cell Lines"
- **Sentence 2**: "TTF1 mediates the transcription of ribosomal RNA"

In the first sentence, "TTF1" refers to the *Thyroid Transcription Factor 1* gene which regulates the expression of thyroid-specific genes, while in the second sentence "TTF1" refers to the *Transcription Termination Factor 1* gene which encodes a termination factor that mediates RNA transcription. However, the use of the "TTF1" alias in the literature is ambiguous, as it is a shared alias of two distinct genes that have different functions. In semantics-based knowledge extraction, the challenge arises in determining the most plausible concept from a biomedical terminology given an ambiguous gene/protein alias in the literature. Without applying disambiguation, gene/protein aliases in the literature may be represented by multiple concepts, resulting in ambiguous concept-based representations. Current biomedical semantic parsers, such as SemRep, do not disambiguate gene/protein aliases in the literature (Preiss & Stevenson, 2016). Instead, rudimentary string matching techniques are used to match a gene/protein alias to corresponding concepts in UMLS or the NCBI Gene database. For instance in the case of "TTF1", the gene term matches aliases of two distinct concepts and is represented the following UMLS concepts: C1421218 (Transcription Termination Factor 1 gene) and C1384616 (Thyroid Transcription Factor 1 gene). In the context of LBD, this creates a significant source of ambiguity and imprecision, as it can be difficult to determine the true concept without referring back to the source article. Preiss et al. highlighted the significance of concept disambiguation during knowledge extraction for LBD, emphasizing that the performance of LBD is sensitive to the precision of concept disambiguation (Preiss & Stevenson, 2016). The authors suggested

that precise concept disambiguation has the potential to enhance the quality of LBD output by minimizing the generation of noisy knowledge.

Further, the reliance on comprehensive and highly granular terminological resources, such as the UMLS, for knowledge extraction presents additional challenges for LBD. The UMLS is a compendium of many biomedical terminologies, such as MeSH, SNOMED CT, ICD10, HGCN, and Gene Ontology among many others. Hence, UMLS is a highly granular terminological resource due its wide coverage of concepts and biomedical domains. As a result, semantically equivalent biomedical entities are often represented as distinct concepts. For example, the breast cancer targeting drug Fulvestrant is represented as three distinct concepts in UMLS: C0935916 (Fulvestrant) denoting the generic name, C0701491 (Faslodex) denoting the trade name, and C0123085 (ZM-182780) denoting the research code. In drug nomenclature, these entities are used interchangeably to refer to the same drug, indicating their semantic equivalence. Less granular terminologies, such as MeSH or NCI, represent Fulvestrant as a single concept that encompasses all its naming variations. Similarly, UMLS represents the benign enlargement of the prostate as C1704272 (Benign Prostatic Hyperplasia) and C0005001 (Benign Prostatic Hypertrophy), whereas MeSH represents the disease as a single concept (D011470) that encompass both synonyms. In genomics, orthologous genes in different species are represented as distinct concepts despite being associated with similar biological phenomena and exhibiting similar functions across species (e.g., association of TP53 in mice and humans with tumor suppression) (Gabaldón & Koonin, 2013). The granularity of concepts is also reflected in the semantic-based knowledge extraction process. For example, the following semantic triples, describing the relationship between Fulvestrant and cancer cell growth, convey the same underlying semantics but are represented differently:

- C0701491: Faslodex – INHIBITS – C0007595: Cell Growth
- C0935916 Fulvestrant – INHIBITS – C0007595: Cell Growth

In the context of LBD, the presence of such equivalent semantic triples result in considerable increase in the number of unique triples, which in turn increases the discovery search space (Vlietstra et al., 2017). Prior research has demonstrated that utilizing condensed terminologies is better suited for downstream predictive tasks, such as relation

prediction, as it leads to a reduction in the dimensionality of the represented data (Rasmy et al., 2020).

We address the challenges of ambiguity and granularity of concept representations as follows. Our approach to resolving ambiguous concepts involves utilizing PubTator; a cutting-edge biomedical text mining tool. PubTator employs a stand-alone concept disambiguation module based on a convolutional neural network model that is capable of identifying and representing ambiguous gene/protein terms in the literature as standardized biomedical concepts, taking into account the surrounding semantic and syntactic contexts (Wei et al., 2019). PubTator's disambiguation module achieves an accuracy of 85%, surpassing the 55% accuracy achieved by rule-based tools (Kilicoglu et al., 2020; Wei et al., 2019). Our aim is to incorporate PubTator as part of the semantic-based knowledge extraction process to resolve ambiguous gene/protein concepts found in the extracted semantic triples.

Our approach to addressing the granularity of concept representations involves leveraging semantic alignment and mapping techniques to map granular concepts into higher-level, condensed representations without compromising the coverage of the domain being investigated (i.e., cancers). We utilize and integrate condensed biomedical terminologies to represent genes, proteins, drugs, chemicals, and diseases concepts found in semantic triples extracted from the literature. Our aim is to merge (i) semantically equivalent concepts, such as *Faslodex* and *Fulvestrant*, into unified concept representations; and (ii) fine-grained concepts, such as orthologous genes/proteins, into higher-level representations that encompass granular concepts.

### 3.1.2 Incomplete Extraction of Semantic Knowledge From the Literature:

Incomplete knowledge extraction is a significant challenge in semantic-based LBD due to the limitations of domain-specific semantic parsers, which suffer from the problem of low recall due to their inability to recognize key semantic relations between concepts in the literature corpus (Henry & McInnes, 2017; Kilicoglu et al., 2020). This limitation is attributed to the complexity of resolving co-references, detecting relations between distant arguments, and recognizing implicit relations that extend beyond sentence boundaries and lack of explicit textual evidence (Drury et al., 2022). For LBD, such limitations have a

significant negative impact on the accuracy and completeness of the discovery process (Henry & McInnes, 2017), because if the semantic relation between biomedical concepts A and B is missing, then the implicit and indirect association between concepts A and C will not be discovered thus limiting knowledge discovery.

To address the incomplete knowledge extraction challenge, our approach uses Knowledge Graph (KG) to represent and reason over the semantic knowledge extracted from the literature. KGs are semantics preserving representations of real-world knowledge, and in our work KGs are employed to extend the knowledge coverage for knowledge discovery in two steps:

**Step 1—Knowledge integration**: KG can facilitate the integration of heterogeneous knowledge resources to augment the knowledge represented within a KG (Ji et al., 2022). To extend the knowledge coverage of the *baseline* literature-derived KG (which typically suffers from incomplete knowledge), we integrate manually curated biomedical knowledge to generate an *integrated KG* with an extended knowledge coverage. This integration step relies on Knowledge Bases (KBs) which are populated with semantics-based knowledge, in the form of *subject-predicate-object* triples, curated directly from biomedical literature by expert biocurators. This ensures that the knowledge constituting the *integrated KG* come from the same source (i.e., biomedical literature). Further, we only consider the subset of curated knowledge within biomedical KBs which are missing from the *baseline* KG. Hence, the *integrated KG* augments the *baseline* KG with knowledge that was previously missing. We do point out that this does not fully address the problem of knowledge incompleteness, as the manual curation process to populate biomedical KBs is time-intensive and does not guarantee that the latest scientific findings are incorporated into knowledge bases.

**Step 2—Knowledge Graph Completion (KGC)**: To further extend the knowledge coverage of the *integrated KG*, we employ KGC methods that augment the KG by predicting previously missing knowledge instances (Z. Chen et al., 2020). KGC can be categorized as the task of predicting missing concepts (entity prediction) or predicting missing relations (relation prediction). In entity prediction, the goal is to predict missing head (subject) or tail (object) entities in a given incomplete triple - i.e., (*head, relation, ?*) or (*?, relation, tail*). While the goal in relation prediction is to predict the missing

relationship in a given incomplete triple - i.e., (*head, ?, tail*). In essence, given any two elements in an incomplete triple, KGC aims to predict the missing third element.

In our prior research, we have demonstrated that KGC methods can address the limitations of incomplete knowledge extraction from the literature by predicting missing semantic relations to facilitate semantics-based LBD for discovery tasks, such as drug repurposing (Daowd et al., 2022). Informed by our previous research, we characterize KGC as a task of *relation prediction* in a closed-world KG (i.e., the *integrated KG*), by inferring missing relations between pre-existing concepts – i.e. given an incomplete triple (*subject*, ?, *object*) that is missing the predicate, the goal of KGC is to predict the most plausible *predicate* using the KG semantics and topography in order to generate a complete (*subject, predicate, object*) triple to further augment the *integrated KG*.

It may be noted that predicting relations between all possible combinations of *subject* and *object* concept is computationally expensive and will result in a dense KG which can be difficult to navigate for knowledge discovery. Hence, it is necessary to have an informed relation prediction approach that pre-defines a subset of concepts that have some form of implicit and logical association between them, according to literature-based sources, but are not linked by a semantic relation in the biomedical KG. As such, we propose an informed relation prediction approach that leverages Medical Subject Heading (MeSH) descriptors as a literature-based knowledge resource to identify logical and implicit biomedical associations which are missing from the *integrated KG*. MeSH is the controlled biomedical vocabulary used by the National Library of Medicine to manually index articles in MEDLINE and PubMed (Baumann, 2016). The indexing process involves qualified human indexers to read full-text articles (scientific publications) to identify the main topics and concepts, and accordingly to assign relevant MeSH descriptors to an article to provide a complete concept-based representation of the article's main scientific content. We operate under the assumption that when two MeSH descriptors are assigned to the same article, it implies the presence of an implicit association between the two descriptors. This assumption serves as the basis for informed relation prediction by generating a subset of incomplete (*subject*, ?, *object*) triples, whereby the *subject* and *object* are MeSH descriptors which are assigned to the same article (i.e., co-occurring concepts).

### 3.1.3 Ranking and Filtering of Discovery Outputs:

Filtering and ranking discovery outputs is an important to prioritize meaningful and novel discoveries from the literature (Thilakaratne et al., 2019). Existing methods to quantify the quality of generated discoveries are prone to biases and limitations, such as overemphasizing frequently co-occurring knowledge instances or relying on expectation-based statistics that are affected by the number of null co-occurrences (Henry & McInnes, 2017; Thilakaratne et al., 2019). Such approaches tend to miss discovery paths characterized by rarely co-occurring knowledge instances, although such less frequent/co-occurring paths can lead to novel and interesting discoveries. According to the principles of information theory, more knowledge can be gained from learning about rare and unexpected occurrences than common ones (C. Chen & Song, 2017). This notion can be directly applied to LBD, as uncommon and infrequently co-occurring knowledge have the potential to reveal novel and interesting connections between disparate knowledge instances (Sebastian et al., 2017). Thus, we argue that it is important to consider discovery paths comprising rare knowledge instances (i.e., *subject-predicate-object* triples) in the ranking process.

Building on well-established metrics in LBD and the principles of information theory, we propose a knowledge filtering and ranking approach to account for discovery paths composed of unique and rarely occurring knowledge instances. In line with *ABC*-based discovery models, we investigate information theory-centric measures with established LBD metrics, such as linking term counts and concept specificity, to quantify the information content in *ABC* discovery paths to prioritize novel and meaningful *ABC* discovery paths, which can be applied to open- and closed-based discovery paradigms. Our approach utilizes the following metrics in a multi-step filtering and ranking process:

**Concept specificity**: refers to the degree to which a standardized concept precisely describes a biomedical entity based on its position within a hierarchical vocabulary, and is determined by its distance from the root concept (Gopalakrishnan et al., 2018). Intuitively, a concept is more specific if it is further away from the root, and less specific if it is closer to the root. We apply this metric to assign *specificity scores* to each concept constituting a given *ABC* discovery path. Subsequently, we define a *specificity score threshold* to

eliminate discovery paths consisting of generic concepts while retaining paths consisting of specific biomedical concepts.

**Linking Term Count (LTC):** refers to the number of unique intermediate (*B*) concepts that link the source (*A*) and target (*C*) concepts in a *ABC* discovery path (Henry & McInnes, 2017; Yetisgen-Yildiz & Pratt, 2009). LTC is a well-established co-occurrence frequency-based metric in LBD to quantify indirect associations between the source (*A*) and target (*C*) concepts, such that the strength of an indirect association increases as the number of intermediate linking terms increases. Previous research has shown that LTC is a simple yet effective metric in filtering and ranking LBD output, and has outperformed metrics such as minimum weight association, linking set association, and cosine distance (Henry & McInnes, 2019). In our work, we utilize LTC as a filtering mechanism to eliminate weak indirect associations between the source and target concepts.

**Triple count:** refers to the occurrence of a triple in the literature. Although commonly used in LBD as a ranking mechanism (Thilakaratne et al., 2019), in this framework, triple count serves as a filtering mechanism to eliminate noisy triples that occur less frequently than a predefined threshold. The underlying idea is that by removing spurious and noisy triples, the remaining ones are more likely to be representative of literature-based knowledge.

**Information Content (IC):** in information theory, IC refers to the amount of information gained from learning about or observing an event (C. Chen & Song, 2017). As shown in Figure 3.1, IC is low as the probability of encountering an event increases (i.e., common event), and high as the probability decreases (i.e., novel or rare event). This is attributed to the fact that we are less likely to encounter a rare event, which may result in new knowledge that challenges our cognitive and belief structures. This notion is commonly applied to quantify interestingness or degree of surprisal associated with an event (C. Chen & Song, 2017). Similarly, IC can be utilized in LBD to quantify how much knowledge can be gained from encountering a discovery path based on co-occurrence information derived from the literature. In this context, IC is defined as the negative log of the probability of observing a discovery path consisting of *A*-*B* and *B*-*C* knowledge instances:

$$IC(ABC) = -log_2 p(ABC)$$

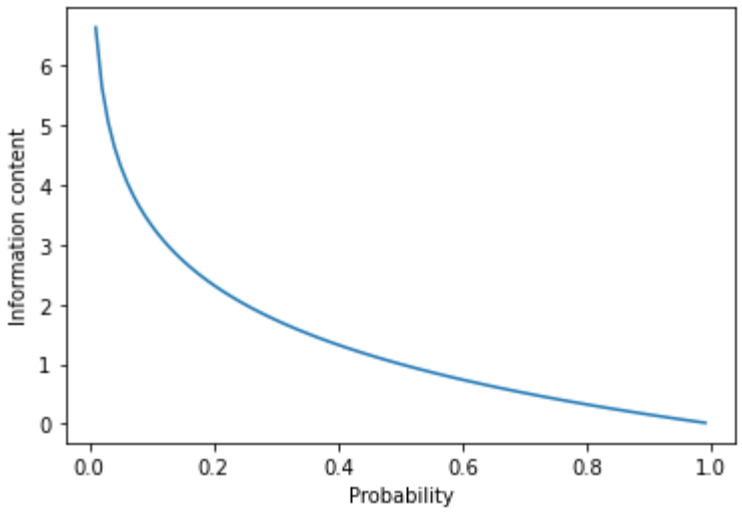


*Figure 3.1: Relationship between probability and information content*

In our work, we utilize IC as the primary metric to rank *ABC* discovery paths generated from open- and closed-based LBD paradigms. We posit that this ranking approach can help to prioritize discovery paths that may be missed by traditional filtering and ranking methods which are typically biased towards statistically significant or frequently occurring *A*-*B* and *B*-*C* associations.

## 3.2 Augmented Knowledge Graphs for LBD (AKG-LBD) Framework:

The previous section outlined the challenges and limitations of traditional semantic-based LBD frameworks and proposed novel solutions to address those challenges. In this section, we describe how the proposed solutions are employed to extend traditional LBD frameworks by incorporating components targeting semantic consolidation, knowledge completion and integration, and ranking the generated *ABC* discovery paths.

Our research has led to the development of the ***A**ugmented **K**nowledge **G**raphs for **LBD*** **(AKG-LBD)** framework as a knowledge graph (KG) centric approach that aims to progressively generate augmented and more complete KGs to provide semantics-driven LBD. AKG-LBD builds upon traditional LBD frameworks consisting of the input literature curation, knowledge extraction, knowledge representation, and discovery components, and additionally incorporates two novel components targeting (i) concept-based semantic consolidation; and (ii) knowledge integration and completion.

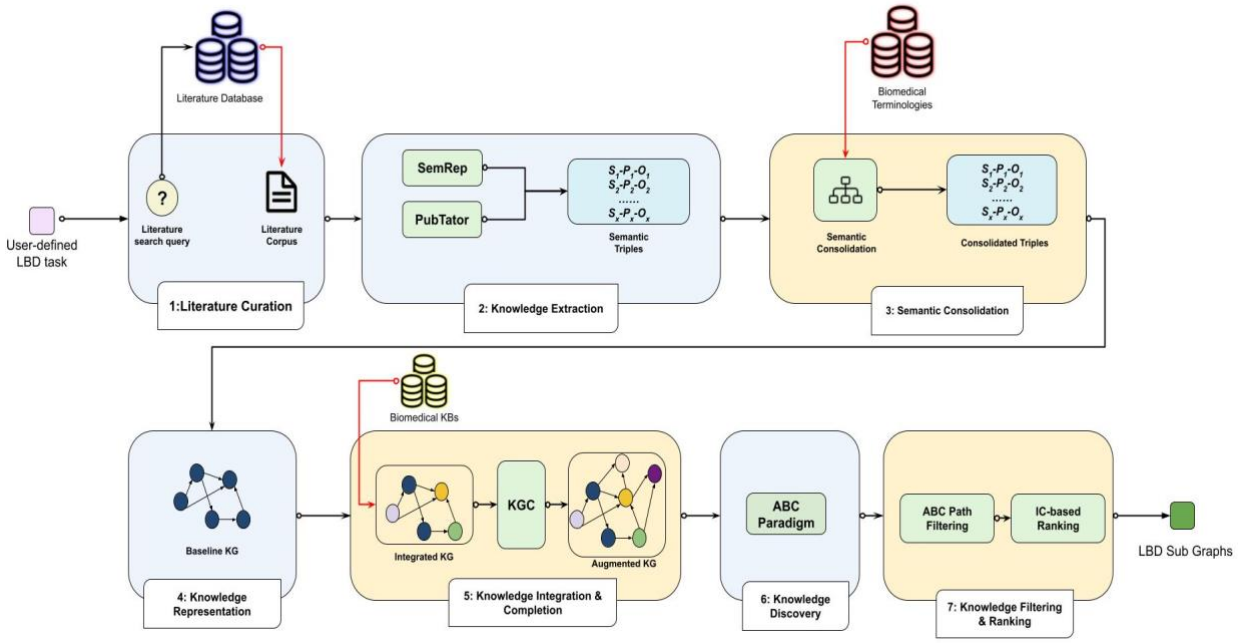


*Figure 3.2: Schematic of the AKG-LBD Framework. Components 3 and 5 are extensions to traditional semantic-based LBD*

In addition to introducing new components, AKG-LBD integrates novel methods to filter and rank graph-based knowledge discovery paths by adapting information theory-centric metrics. A schematic representation of the AKG-LBD framework is illustrated in Figure 3.2. The components comprising AKG-LBD are described in the following sub-section:

1. **Input (literature curation) component**

AKG-LBD's input component is purposed to query biomedical literature databases to identify a corpus of articles relevant to the specific biomedical domain being queried for knowledge discovery. Structured search queries are defined by users using a combination of title and abstract keywords derived from biomedical terminologies (Thilakaratne et al., 2020). Specifically, users are required to input domain-specific biomedical terms and/or phrases which are mapped to their corresponding concepts using biomedical terminologies, such as UMLS or MeSH. Additionally, users can define a date range to specifically search for articles published between specific dates. In response to the query, each identified article is represented by its title, abstract, publication date, and a unique article identifier. The output of the literature curation component is a structured corpus of articles that are used by the subsequent knowledge extraction component to extract semantics-based knowledge for LBD.

2. **Knowledge Extraction Component**

The knowledge extraction component leverages SemRep, as a biomedical semantic parser, and PubTator, as a biomedical text mining tool, to extract semantic-based knowledge in the form of *subject-predicate-object* triples from the literature-based corpus. For example, given the following sentence in (Day et al., 2020): "Tamoxifen (TAM) is a hydrophobic anticancer agent and a selective estrogen modulator (SERM), approved by the FDA for hormone therapy of BC.", the knowledge extraction component would generate the semantic triples:

- C0039286: Tamoxifen (Pharmacologic Substance) – ISA – C0732611: Selective Oestrogen Receptor Modulator (Pharmacologic Substance)
- C0039286: Tamoxifen (Pharmacologic Substance) – TREATS – C0006142: Breast Cancer (Neoplastic Process)

The *subject* and *object* entities are standardized UMLS concepts and are represented in the form of *Concept Identifier*: *Concept Name (Semantic Type)*. The *Concept Identifier* is a unique UMLS identifier, the *Concept Name* is the concept's preferred ontological name, and *Semantic Type* is a standardized classification of the concept's type based on the UMLS Semantic Network. The *predicate* denotes a semantic relationship between the *subject* and *object* based on verbal constructs expressed in the sentence.

The extracted semantic triples undergo a filtering process based on the concept semantic type and predicate (i.e., hierarchical, causal, therapeutic), whereby only relevant semantic triples are retained for the downstream knowledge discovery task. To maintain the provenance of the extracted knowledge, the semantic triples are mapped back to the sentences in the article from which they were identified—the triples are tagged with the unique identifier of their source article. The output of this component consists of semantic-based knowledge (i.e., *subject-predicate-object* triples), which are subsequently used as input by the next component to consolidate semantically similar *subject* and *object* concepts into unified representations.

**3. Semantic Consolidation Component:**

The semantic consolidation component addresses the challenge of concept granularity and is responsible for consolidating semantically equivalent and fine-grained concepts (i.e., *subjects* and *objects*) into unified concept representations. Our rationale is that the presence of granular concepts in semantic triples results in distinct knowledge instances that convey

the same underlying knowledge, thus unnecessarily increasing the discovery search space for LBD. Hence, to resolve this issue, this component leverages semantic alignment and mapping techniques to map UMLS concepts (found in the semantic triples) to corresponding concepts in condensed biomedical terminologies that encompass fine-grained and semantically equivalent concepts, whilst ensuring that the scope of the domain knowledge remains intact.

**4. Knowledge Representation Component**

Knowledge Graphs (KGs) provide semantically-rich representations of semantics- and literature-based knowledge. In our prior research we explored the use of KGs to represent semantics-based causal associations for chronic diseases and cancers from the literature to facilitate knowledge discovery (Daowd et al., 2021b, 2021a). In this work, we build on our previous research to construct large-scale literature-based KGs using the consolidated *subject-predicate-object* triples. The *subject* and *object* concepts are represented as nodes, and *predicates* are represented as labelled directed edges denoting the semantic relationship between the nodes. Nodes are associated with attributes denoting a concept's name, unique terminological identifier, semantic type, and semantic group. Similarly, edges are associated with attributes denoting the predicate type and the unique article identifier from which the relationship was extracted. The output of this component is the *baseline literature-based Knowledge Graph* (KG), which represents the available (incomplete) biomedical knowledge. It is utilized by the next component to augment the knowledge contained within the KG.

**5. Knowledge Integration and Completion Component**

The knowledge integration and completion component is a novel component in the AKG-LBD framework responsible for extending the biomedical knowledge coverage of the *baseline literature-based KG*. This component implements two complementary tasks: (i) the integration of biomedical knowledge from curated Knowledge Bases (KBs) to the *baseline KG* to supplement knowledge that was initially missing to yield an a more complete *integrated literature-based K*G. Knowledge integration is restricted to KBs curated directly from biomedical literature, thereby ensuring a unified source of knowledge to supplement the *baseline KG*; and (ii) KGC to predict semantic relations in incomplete triples (*s, ?, o*). KGC is implemented as novel graph-based representation learning methods

to encode KG nodes and edges into low-dimensional vectors, which in turn can be used for relation prediction. The outcome of this task yields an *augmented literature-based KG* addressing the limitations of semantic parsers with respect to incomplete knowledge extraction. The *augmented KG* is utilized for knowledge discovery, where graph traversal techniques and novel ranking metrics can be applied to uncover hidden and novel associations between disparate knowledge instances.

6. **Knowledge Discovery and Ranking Component:**

The knowledge discovery and ranking component applies graph traversal techniques and novel discovery path ranking metrics to discover hidden and novel associations between disparate knowledge instances within the *augmented KG*. For knowledge discovery, open and closed ABC-based discovery models are applied to the *augmented K*G—the underlying assumption is that the source (A) and target (C) concepts do not co-occur in the literature and, therefore, are not linked by a semantic relation (i.e., predicate) in the *augmented KG*. The graph traversal method implemented aims to infer indirect associations between the source and target concepts through an intermediate entity (B) which provides significant information gain by its association with the source and target entities. The knowledge discovery task tends to generate a large number of potential discoveries which are filtered based on interestingness and novelty using principles of information theory that prioritizing *A-B* and *B-C* associations that yield the greatest amount of information content. The top-ranking ABC discovery paths are aggregated as a sub-graph and serve as the discovered knowledge by AKG-LBD. The discovered knowledge's sub-graph can be interrogated, explored and visualized using an interactive (property) graph visualization interface.

## 3.3 Data and Material:

The purpose of this section is to describe the relevant datasets and material used to extract knowledge from the literature and to subsequently generate literature-based KG for LBD.

### 3.3.1 Literature Dataset:

The main source of literature used in this research is extracted from the MEDLINE repository of biomedical literature. The MEDLINE repository is maintained by the National Library of Medicine (NLM) compiling scientific articles from over 5,200

specialized journals and publications covering domains of clinical medicine, molecular biology, oncology, biochemistry, biomedical sciences, and environmental health (Lu et al., 2015). The MEDLINE database provides the title, abstract, and publication dates for more than 25 million biomedical articles. Additionally, one of the distinctive characteristics of MEDLINE is that articles are indexed with NLM's Medical Subject Heading (MeSH) descriptors which can facilitate the extraction of specific fragments of the literature related to a particular discovery task, such as drug repurposing or molecular oncology.

In LBD research, the title and abstract sections of scientific articles are the most commonly used sections as input for knowledge extraction - these sections provide a faithful summary of an article's content. Other common input types include MeSH descriptors, however, recent research has demonstrated that titles and abstracts provide the most optimal input for knowledge extraction in LBD compared to other input types as measured by their information richness (Thilakaratne et al., 2020). Full-text articles are rarely used as input due to the substantial computational resources required for extracting semantic-based knowledge from them (Thilakaratne et al., 2019). Moreover, metadata, such as article publication dates and unique identifiers, are incorporated to ensure the traceability and provenance of the retrieved articles.

Based on the above findings, we consider the following data in the MEDLINE database for this research: *title*, *abstract*, *article publication date*, *unique article identifier (PMID),* and *MeSH descriptors*. We utilize article titles and abstracts to generate the baseline biomedical corpora as input for knowledge extraction. Publication dates and PMIDs are used to keep track of the retrieved articles and to generate time-sliced views of biomedical literature. Lastly, we utilize MeSH descriptors as an additional knowledge resource for Knowledge Graph Completion.

### 3.3.2 Tools To Extract Literature-Based Knowledge:

Literature-based knowledge is extracted using domain-specific semantic parsers and text mining tools. In the following sections, we provide an overview of the tools used to extract semantic knowledge from the literature in the form of (*subject, predicate, object*) triples.

#### 3.3.2.1 SemRep:

SemRep is a biomedical-specific semantic parser developed by the National Library of Medicine (NLM) to extract semantic-based knowledge from biomedical corpora (Kilicoglu et al., 2020). SemRep uses rule-based Natural Language Processing (NLP) methods and combines syntactic and semantic techniques which leverage structured biomedical knowledge in the Unified Modeling Language System (UMLS). Using sentences as input, SemRep identifies and extracts semantic relations in the form of (*subject, predicate, object*) triples. The *subject* and *object* entities extracted by SemRep are primarily UMLS concepts represented by a Concept Unique Identifier (CUI), concept name, and concept Semantic Type (ST). The *predicate* is a relation type derived from the UMLS Semantic Network. A wide range of semantic relation types are extracted by SemRep relating to clinical medicine (e.g., treats, diagnoses), substance interactions (e.g., interacts with, inhibits, stimulates), disease etiology (causes, predisposes), pharmacogenomics (e.g., affects, disrupts), in addition to hierarchical, spatial, and temporal relations (e.g., is a, precedes, location of).

SemRep is reported to achieve high precision rates for the extraction of gene-disease relations (76%), pharmacogenomic relations (73%), gene-function relations (65%), and substance interactions (59%) (Kilicoglu et al., 2020). Overall, the performance of SemRep as a semantic parser for biomedical text yields 69% precision rates. However, SemRep suffers from low recall rates (42%) due to limitations in extracting implicit relations across sentence boundaries (Kilicoglu et al., 2020). Another limitation is the lack of entity disambiguation component for genes/proteins, as terms denoting genes/proteins are mapped to standardized concepts through exact string matching. This approach can be problematic for LBD, as genes/proteins in biomedical literature are commonly referenced by short-form abbreviated symbols which can be shared by many different genes. For example, TTF1 is the short-form symbol (i.e., alias) for thyroid transcription factor 1 gene and is also the symbol for transcription termination factor 1 gene. Given the following sentence “TTF1 may represent a therapeutic target for the treatment, prevention, and control of obesity” (Park et al., 2022), SemRep maps TTF1 to two distinct gene concepts in UMLS - i.e., TTF1 (C1421218 | C1384616), thus generating ambiguous semantic triples with several possible UMLS concepts for the *subject* or *object*.

Despite these limitations, SemRep remains widely used in LBD studies to extract semantic-based knowledge from biomedical literature, and has also been applied to clinical notes and gray literature (Kilicoglu et al., 2020). In this study, we utilize SemRep as a semantic parser to extract (*subject, predicate, object*) triples from biomedical literature to generate the *baseline KG*. We aim to address limitations of ambiguous gene mappings by combining the output of SemRep with advanced text mining tools for entity disambiguation.

#### 3.3.2.2 PubTator Central:

PubTator Central (PTC) is a biomedical text mining system leveraging advanced methods for named entity recognition and entity disambiguation to generate normalized annotations for gene/protein, disease, and chemical concepts in biomedical corpora (Wei et al., 2019). PTC incorporates an entity disambiguation module to accurately identify and disambiguate biomedical terms in text. Entity disambiguation is an important aspect of biomedical text mining responsible for assigning accurate normalized concepts to ambiguous terms. PTC addresses this challenge by leveraging convolutional neural network models capable of identifying and annotating ambiguous terms with correct biomedical concepts by considering the surrounding semantic and syntactic contexts. The accuracy of PTC's entity disambiguation module is reported to out-perform traditional rule-based methods; PTC yields 85% accuracy compared to 55% in rule-based systems. The disambiguation module is available as a stand-alone, open-source download (https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/tmTools/download/BioConceptDisambiguation.zip). We utilize PTC's entity disambiguation module in this research to overcome limitations of SemRep in assigning gene *subjects* or *objects* multiple UMLS concepts.

### 3.3.3 Curated Knowledge Bases:

There is a wealth of information in biomedical knowledge bases (KBs) that can be harnessed to supplement knowledge extracted from the literature. Biomedical KBs, such as the Comparative Toxicogenomics Database (CTD) (Davis et al., 2019) and Gene Ontology (GO) (The Gene Ontology Consortium, 2019) capture biomedical knowledge describing associations between genes-diseases, chemicals-diseases, and genes-molecular functions. These associations are manually curated by professional curators who are tasked

with reading scientific literature to identify associations between biomedical concepts and represent them using controlled terminologies such as MeSH and NCBI Gene Entrez. Additionally, the manual curation process involves annotating relations between concepts to semantically describe the nature of associations. For example in CTD, associations between chemicals and diseases are annotated as "marker/mechanism (M)" to denote causal correlations (e.g., *increased exposure to chemical X correlates with disease Y*) or "therapeutic (T)" to denote potential therapeutic correlations (e.g., *chemical X is used to treat disease Y*). Similarly, GO associations between genes and molecular functions are annotated as "*enables*" to indicate a gene's role in a particular molecular function. Hence, knowledge contained within such KBs can be readily represented as (*subject, predicate, object*) triples and integrated with biomedical KGs. In our work, we leverage structured knowledge in CTD and GO to extract curated biomedical associations involving chemicals-diseases, genes-diseases, genes-genes, and genes-molecular functions.

## 3.4 AKG-LBD Evaluation Framework:

The success of LBD hinges on its capability to generate impactful and novel knowledge discoveries (Henry & McInnes, 2017). However, determining what constitutes a discovery in the biomedical domain is a challenging task since the 'discovered' knowledge has not yet been published in existing scientific literature. Therefore, the efficacy of LBD systems has to be evaluated in a formal (retrospective) setting, by replicating discoveries published in peer-reviewed journals or clinical trials using a technique known as time-slicing (Henry & McInnes, 2017; Thilakaratne et al., 2019). In this setting, a simulated discovery environment is created by masking literature published at a predefined cut-off-date corresponding to the date of the target discovery publication. Literature published prior to the cut-off-date is used as input for LBD, while literature published after the cut-off-date is used to validate whether the targeted discovery can be replicated by the LBD system.

To assess the performance of AKG-LBD, we apply a discovery replication-based evaluation using time-slicing techniques to assess whether the discovered knowledge (i.e. output of the LBD system) correspond to prospective real-world biomedical knowledge published in peer-reviewed journals. Our evaluations will consider the following aspects: (i) the impact of knowledge integration and completion components on LBD output; and

(ii) the performance of ranking metrics on discovery replication tasks. Further, we plan to compare the output of AKG-LBD to the following formalized LBD systems: Arrowsmith (Swanson & Smalheiser, 1997), BITOLA (Hristovski et al., 2005), MELODI-PRESTO (Elsworth & Gaunt, 2021), and LION-LBD (Pyysalo et al., 2019). The following sections outline the discovery replication evaluation settings using cancer and drug test cases.

### 3.4.1 Targeted Cancer Discovery Test Cases:

Five real-world cancer discoveries are adapted as test cases to evaluate AKG-LBD in discovering novel biomedical knowledge. These test cases were curated from high quality peer-reviewed articles by cancer biologists as described in (Pyysalo et al., 2019). Specifically, biomedical articles covering cancer biology published between 2006 and 2017 were surveyed to identify novel cancer-related discoveries that can be described as a causal chain of three concepts, in conformity with the *ABC* discovery model. Discovery candidates identified in the initial survey were further filtered according to the following conditions: A-B and B-C concepts must co-occur in at least 100 articles, while A-C concepts do not co-occur in any articles. These prerequisites ensure that a direct association between A and C concepts have not been discovered previously, thus, signifying the novelty of candidate discoveries. This process resulted in five cancer discovery paths consisting of a source concept (A), an intermediate concept (B), and a target concept (C) as outlined in Table 3.1. These curated discoveries have been previously used to evaluate co-occurrence-based LBD systems, however, to our knowledge they have not been evaluated using semantic-based LBD systems.

Evaluation of the cancer test cases will require applying time-slice techniques using the date of publication outlined in Table 3.1 as the cut-off-date to create the input literature corpus for each test case. For comparison purposes, time-sliced literature sets are utilized for knowledge discovery by applying our AKG-LBD framework and a traditional semantic-based LBD framework (i.e., without knowledge integration and completion components). The rationale is to assess whether our knowledge integration and completion methods have any significant impact on LBD outputs in terms of replicating previously published discoveries. We report the number of replicated test case discoveries as

evaluation metrics, in addition to their Relative Ranks (RR). RR measures the rank of the valid discovery path relative to the rank of the top scoring path.

$$RR = \frac{Ranking\ Score_{valid}}{Rank\ Score_{Max}}$$

Where, $Ranking\ Score_{valid}$ is the ranking score of the valid discovery path, $Ranking\ Score_{Max}$ is the ranking score of the top ranked discovery path.

*Table 3.1: Cancer discovery test cases*

| **Concept *A*** | **Concept *B*** | **Concept *C*** | **Date of Publication** |
|---|---|---|---|
| Nuclear factor erythroid 2-related factor 2 (NRF2) | Reactive oxygen species | Pancreatic cancer | 2011 |
| Interleukin-17A (IL-17) | Mitogen-activated protein kinase 14 (p38a) | Dual specificity protein phosphatase 1 (MKP-1) | 2015 |
| Neurogenic locus notch homolog protein 1 (NOTCH1) | Cellular Senescence | CCAAT/enhancer-binding protein beta (C/EBPb) | 2016 |
| Nuclear factor NF-kappa-B p105 subunit (NFKB) | Apoptosis regulator Bcl-2 (BCL2) | Adenoma | 2016 |
| Stromal cell-derived factor 1 (CXCL12) | Cellular Senescence | Thyroid cancer | 2017 |

### 3.4.2 Drug Repurposing Case Study:

In this evaluation setting we assess the capability of AKG-LBD in repurposing existing drugs for new therapeutic indications targeting cancers. Time-slicing is used to divide literature covering cancers into two sets: a pre-cut-off set representing literature published before a specified cut-off-date and a post-cut-off set representing literature published after the cut-off-date. The pre-cut-off literature is used as input for LBD, while the post-cut-off literature is used to create a *silver* standard dataset to validate output discoveries. As such, this evaluation setting does not consider replicating targeted discoveries as is the case in

the cancer test cases, rather the objective is to uncover all of the literature-based knowledge reported in the post-cut-off silver standard dataset.

The silver standard is created by identifying a subset of knowledge (i.e., semantic triples) present in the post-cut-off literature but absent from the pre-cut-off literature set. We selected 2015 as the cut off date to split the literature dataset into a pre-cutoff discovery set and a post-cutoff validation set (i.e., silver standard). Subsequently, we use the validation set to identify a subset of semantic triples consisting of a chemical substance as the *subject* entity, a neoplastic disease as the *object* entity, and *TREATS* as the predicate - i.e., *Chemical Substance-TREATS-Cancer*. Semantic triples which occur in less than 3 articles are eliminated to ensure that the silver standard consists of relevant and well-researched knowledge. Further, to avoid overlap, we make sure that no semantic triples in the validation set also exist in the discovery set. Further, we filter the *subject* and *object* entities using the concept specificity as a metric, which measures the distance from the root concept, to ensure that the knowledge is specific. The final silver standard dataset consisted of 377 *Chemical-TREATS-Cancer* triples and 187 candidate drugs for repurposing.

The discovery task is initiated by pre-defining a source concept (i.e., *Chemical*) from the silver standard dataset, determine the semantic type of the intermediate (i.e., *Genes*) and target concepts (i.e., *Cancer*), and define the semantic relationships between the source, intermediate, and target concepts. Specifically, we consider two patterns of semantic relationships (i.e., predicates):

- **Pattern 1**: *Drug* - [INHIBITS, AFFECTS, DISRUPTS] - *Gene* - [CAUSES, PREDISPOSES, ASSOCIATED_WITH] - *Cancer*
- **Pattern 2**: *Drug* - [INTERACTS_WITH, STIMULATES] - *Gene* - [PREVENTS, DISRUPTS] - *Cancer*

The first pattern considers a drug that has an inhibitory effect on a gene, and the gene is involved in causal relationship with the target cancer. The second pattern considers a drug that has a stimulatory effect on a gene, which in turn may prevent or disrupt the onset of a target cancer. Finally, a candidate discovery path is considered a valid path if these conditions are satisfied:

1. *Drug-TREATS-Cancer* relationship is present in the silver standard (i.e., post-cut-off literature)

2. The *Drug* and *Gene* have an established interaction in CTD
3. The *Gene* and *Cancer* have an established association in CTD

We report the following metrics to evaluate the performance of the AKG-LBD framework for the drug repurposing task:

**Average Precision at K (AP@K):** measures the number of valid (i.e., true positive) *ABC* paths at the top *K* ranks relative to total number of paths at *K*, averaged over the number of discovery queries (i.e., source concepts). Given a ranked list of *ABC* discovery paths relevant to a pre-defined source (*A*) concept, we calculate the precision within the top *K* ranks and then compute the average for all source concepts. Formally, AP@K is defined as:

$$AP@K = \frac{1}{Q}\sum \frac{Valid(ABC) \mid rank(ABC) \leq K}{N(ABC) \mid rank(ABC) \leq K}$$

Where, $Valid(ABC) \mid rank(ABC) \leq K$ is the total number of valid *ABC* discovery paths with ranks less than or equal to *K*, $N(ABC) \mid rank(ABC) \leq K$is the total number of *ABC* discovery paths with ranks less than or equal to *K*, and *Q* is the total number of source concepts (i.e., queries).

**Average Recall (of valid paths) at K (AR@K):** measures the number of valid *ABC* paths at the top *K* ranks relative to the total number of valid paths, averaged over the number of discovery queries (i.e., source concepts). Given a ranked list of *ABC* discovery paths relevant to a pre-defined source (*A*) concept, we calculate the recall within the top *K* ranks and then compute the average recall for all source concepts. Formally, AR@K is defined as:

$$AR@K = \frac{1}{Q}\sum \frac{Valid(ABC) \mid rank(ABC) \leq K}{N_{valid}}$$

Where, $Valid(ABC) \mid rank(ABC) \leq K$ is the total number of valid *ABC* discovery paths with ranks less than or equal to *K*, $N_{valid}$ is the total number of valid *ABC* discovery paths, and *Q* is the total number of source concepts (i.e., queries)

**Mean Average Precision (mAP):** measures the average precision at each valid path for multiple discovery queries, thus provides a single measure of the framework's performance

across all queries. Given a ranked list of *ABC* discovery paths relevant to a pre-defined source (*A*) concept, we calculate the precision at each valid path (i.e., true positive), and take the average of these values to calculate the average precision (AP). Then, mAP is the arithmetic mean of AP over the total number of discovery queries. Formally, mAP is defined as:

$$mAP = \frac{1}{Q}\sum AP(q)$$

Where, $AP(q)$ is the average precision of valid paths for discovery query $q$, $Q$ is the total number of discovery queries.

**Average Relative Rank (ARR):** measures the rank of a valid *ABC* discovery path relative to the top ranked path for the discovery query, averaged over the total number of queries. Formally, ARR is defined as:

$$ARR = \frac{1}{Q}\sum \frac{Ranking\ Score_{valid}}{Ranking\ Score_{Max}}$$

Where, $Ranking\ Score_{valid}$ is the ranking score of the valid discovery path, $Ranking\ Score_{Max}$ is the maximum ranking score (i.e., score of top ranking path) of a discovery path, and $Q$ is the total number of discovery queries.

## 3.5 Summary:

In this chapter we presented the underlying research approach for developing a semantic-based LBD framework—i.e. AKG-LBD—to address the challenges faced by traditional LBD frameworks. AKG-LBD extends traditional frameworks by incorporating components targeting concept-based semantic consolidation, knowledge completion and integration, and ranking graph-based discovery paths. The input (literature curation) component queries biomedical literature databases to identify a corpus of articles relevant to a specific biomedical domain. The knowledge extraction component extracts semantic-based knowledge from the retrieved corpus of articles. The semantic consolidation component consolidates semantically similar or closely associated concepts into unified/atomic concept representations. The knowledge representation component represents the consolidated subject-predicate-object triples as a Knowledge Graph (KG), which is incomplete biomedical knowledge. The knowledge integration and completion

component extends the baseline literature-based KG with meaningful and heterogenous biomedical knowledge.

Our approach to LBD is evaluated in two discovery replication settings using time-slicing techniques: (i) replicating recent discoveries in cancer biology published in peer-reviewed journals; and (ii) repurposing existing drugs for new cancer indications. The former approach assesses the effectiveness of AKG-LBD in confirmed previously established discoveries, while the latter approach evaluates the potential of AKG-LBD to forecast future knowledge occurrences in the literature.

In the next chapter, we present the implementation of AKG-LBD as a framework for discovering knowledge related to cancer and detail the techniques and methods utilized in each phase. While the implementation of AKG-LBD centers on two targeted discovery task (i.e., cancer discovery and drug repurposing), we posit that the components of the framework can be employed for knowledge discovery across a wide range of biomedical domains.

# Chapter 4 Methods

This chapter describes the implementation of the AKG-LBD framework for the discovery of new knowledge in the field of oncology. Explicitly, this chapter outlines the methods employed to implement all components constituting AKG-LBD described in the previous chapter, from the initial curation of the input literature corpus to the filtering and ranking of the LBD output (i.e., discoveries). Further, we detail the methods and techniques used to implement the novel components of semantic consolidation, knowledge integration, and Knowledge Graph Completion (KGC). While the methods presented here are specific to a specific biomedical domain (i.e., cancers), we posit that our proposed methods can be adapted to a wide range of discovery tasks across the biomedical domain.

This chapter is structured into the following sections. **Section 4.1** describes the input component, which involves the literature search strategy and the resulting corpus of literature used as the basis for knowledge discovery. **Section 4.2** outlines the knowledge extraction component utilizing biomedical-specific semantic parsers and text mining tools to extract semantic-based knowledge from the literature corpus. **Section 4.3** presents the methods employed for the semantic consolidation of literature-based knowledge. **Section 4.4** describes the component tasked with representing semantic-based knowledge as a large-scale Knowledge Graph (KG). **Section 4.5** presents the implementation of the knowledge integration and completion component, describing the methods and techniques used to augment the baseline literature-based KG with meaningful and accurate biomedical knowledge. **Section 4.6** describes the knowledge discovery and ranking component, which is tasked with generating and prioritizing interesting implicit connections between disparate knowledge instances. Finally, **section 4.7** concludes this chapter by summarizing the novel contributions presented and discussing the potential applications in the broader biomedical domain.

## 4.1 Input Corpus Curation For Literature-Based Knowledge Discovery:

The objective of this component is to query the MEDLINE repository via PubMed, as the primary biomedical literature database, to extract a literature corpus that covers the topics related to the intended discovery task.

PubMed can be queried using specific words/phrases mentioned in articles (i.e., title, abstract, or full-text) or using the controlled MeSH terminology, which is used for indexing each article in the database via a set of standardized biomedical concepts that describe the article's content. Our approach to query PubMed uses concepts in the MeSH terminology, which also eliminates the need to address different spelling variations or synonyms when specifying a biomedical concept in a query. For instance, the standardized MeSH concept "Breast Neoplasms" encompasses various spelling variations and synonyms, such as "breast tumor", "cancer of the breast", "breast carcinoma", "mammary cancer", and "malignant neoplasm of the breast". Using such controlled representation of concepts to query PubMed improves the efficiency and precision of the literature search. In contrast, using non-standardized words/phrases requires specifying all spelling variations and synonyms to ensure the retrieval of all relevant articles.

Using MeSH concepts also provides the advantage of retrieving articles focusing on the desired topic, rather than mentioning it in passing (Baumann, 2016). Further, the MeSH terminology is organized in a hierarchical structure, so including a generalized (high-level) concept in a query will also include its sub-types (i.e., child concepts), which are more specific. Such comprehensive querying is not possible using words/phrases.

Accordingly, we developed a structured query formulation method that leverages the MeSH terminology and its hierarchical structure to map user-defined terms to standard biomedical concepts, which are used to perform a literature search on PubMed. Figure 4.1 outlines the schema for our query formulation method.

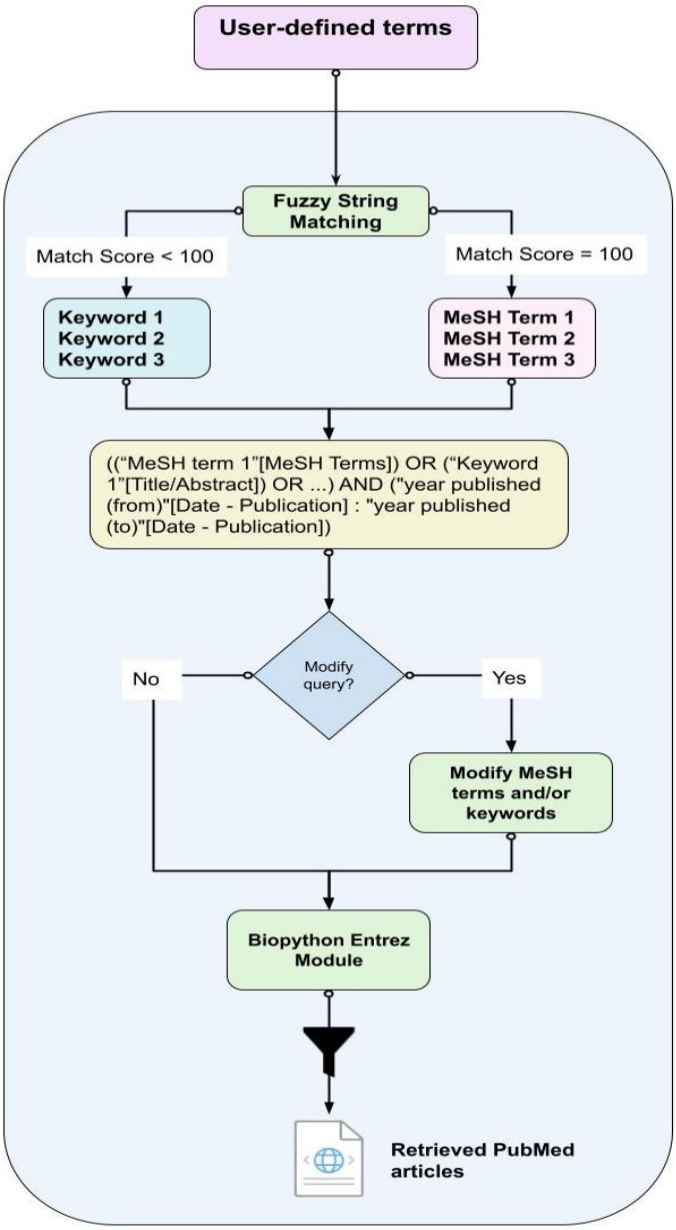


*Figure 4.1: PubMed query formulation method*

Using a Command-Line Interface (CLI), users input a set of terms and specify date ranges for article publications to initiate the search query formulation. String matching techniques are used to map the input (user-defined) terms to standard MeSH concepts via corresponding concept names (i.e., MeSH entry terms). We utilize a fuzzy string matching algorithm implemented in the open-source fuzzywuzzy Python package (https://github.com/seatgeek/fuzzywuzzy) for comparing two strings using the Levenshtein distance (also referred to as edit distance). The algorithm calculates the minimum number

of edits required to change the input (user-defined) term to a corresponding MeSH concept name and, accordingly, assigns a matching score ranging from 0-100, with 100 indicating a perfect match. In the case of a perfect match, the matched MeSH concept name and associated synonyms are retained for the query as *MeSH terms*, and in the case of no match user-defined terms are utilized instead as a *title or abstract keywords*. Consequently, MeSH concept names, user-defined terms, and specified date ranges are used to formulate the PubMed search query based on the following template:

- ((“**MeSH term 1”**[*MeSH Terms*]) OR (“**Keyword 1”**[*Title/Abstract*]) OR ...) AND ("**year published (from)**"[*Date - Publication*] : "**year published (to)**"[*Date - Publication*])

The template can include several MeSH terms and title/abstract keywords based on the input provided and the number of matched concept names. Moreover, the boolean “OR” operator is used to create a comprehensive query to retrieve as many relevant articles as possible. The resulting query is presented back to the user to make any essential modifications before performing the PubMed search task. The finalized query is then used to retrieve PubMed articles via Biopython’s Entrez module (Cock et al., 2009), which provides APIs to access and download PubMed articles in structured XML formats. For each article, we retain its title, abstract, and unique PubMed identifier (PMID). The retrieved articles undergo a filtering process to eliminate articles without titles or abstracts. The final set of articles—i.e. the literature corpus—are stored in plain text format for the subsequent knowledge extraction component.

### 4.1.1 PubMed Query Formulation to Retrieve Cancer Literature:

We employed the query formulation methods to retrieve from PubMed a corpus of literature covering cancer research. We restricted the PubMed search to articles published between 1975 and 2021. The query formulation was based on a set of 28 terms related to cancer and its associated hallmarks (i.e., hallmarks of cancer) (Knijnenburg et al., 2015). The full list of terms are as follows:

- Cancer, neoplasm, tumor, carcinogenesis, tumorigenesis, cell proliferation, cell growth, cell cycle, cell division, apoptosis, autophagic cell death, autophagy, regulated cell death, cellular senescence, angiogenesis, cell differentiation, cell

adhesion, cell migration, taxis, DNA repair, inflammation, glycolysis, immunity, telomere, actin, cell movement, epithelial-mesenchymal transition, and endocytosis.

The input list was converted to lowercase strings before passing it to the fuzzy string matching algorithm to identify corresponding MeSH terms and synonyms. All input terms were mapped to standard MeSH terms via the string matching task. However, in the query modification phase, we sought to include all input terms as title or abstract keywords, in addition to their corresponding matched MeSH terms and synonyms, to ensure a comprehensive and exhaustive search strategy. The query used to retrieve PubMed articles was finalized as follows:

**("Neoplasms"[MeSH] OR "cancer"[Title/Abstract] OR "neoplasm"[Title/Abstract] OR "carcinogenesis"[Title/Abstract] OR "Carcinogenesis"[MeSH] OR "tumorigenesis"[Title/Abstract] OR "tumor"[Title/Abstract] OR "cell proliferation"[Title/Abstract] OR "Cell Proliferation"[MeSH] OR "cell growth"[Title/Abstract] OR "Cell Cycle"[MeSH] OR "cell cycle"[Title/Abstract] OR "Cell Division"[MeSH] OR "cell division"[Title/Abstract] OR "signaling pathway"[Title/Abstract] OR "Apoptosis"[MeSH] OR "apoptosis"[Title/Abstract] OR "caspase"[Title/Abstract] OR "Autophagic Cell Death"[MeSH] OR "autophagic cell death"[Title/Abstract] OR "autophagy"[Title/Abstract] OR "Regulated Cell Death"[MeSH] OR "regulated cell death"[Title/Abstract] OR "programmed cell death"[Title/Abstract] OR "Cellular Senescence"[MeSH] OR "cellular senescence"[Title/Abstract] OR "telomere"[Title/Abstract] OR "Telomere"[MeSH] OR "angiogenesis"[Title/Abstract] OR "Cell Differentiation"[MeSH] OR "cell differentiation"[Title/Abstract] OR "Blood Circulation"[MeSH] OR "blood circulation"[Title/Abstract] OR "Cell Adhesion"[MeSH] OR "cell adhesion"[Title/Abstract] OR "Epithelial-Mesenchymal Transition"[MeSH] OR "Cell Movement"[MeSH] OR "cell migration"[Title/Abstract] OR "cell projection"[Title/Abstract] OR "actin"[Title/Abstract] OR "Taxis Response"[MeSH] OR "taxis"[Title/Abstract] OR "Cell Polarity"[MeSH] OR "cell polarity"[Title/Abstract] OR "DNA Repair"[MeSH] OR "DNA repair"[Title/Abstract] OR "Inflammation"[MeSH] OR**

**"inflammation"[Title/Abstract] OR "Glycolysis"[MeSH] OR "glycolysis"[Title/Abstract] OR "Immunity"[MeSH] OR "Endocytosis"[MeSH]) AND ("year published (from)"[1975] : "year published (to)"[2021])**

The resulting modified query was used to retrieve cancer literature from PubMed via Biopython's Entrez module (Cock et al., 2009). We retrieved 5,549,328 articles and after excluding articles without a full title or abstract, we were left with a literature corpus of 5,531,702 articles that were stored as ASCII-formatted plain text containing the title, abstract, and PMID of each article.

## 4.2 Knowledge Extraction from Biomedical Literature:

The corpus of literature extracted by the previous component serves as the input to the semantic-based knowledge extraction component. The knowledge extraction component combines biomedical specific semantic parsing and text mining tools to extract knowledge in the form of *subject-predicate-object* triples from the literature. The knowledge extraction was executed in two phases: (i) semantic knowledge extraction (using biomedical-specific semantic parser); and (ii) disambiguation of biomedical concepts. The output of these phases was combined to generate the final semantic-based knowledge.

### 4.2.1 Semantic-Based Knowledge Extraction:

We investigated several biomedical-specific semantic parsers used for LBD, namely SemRep (Kilicoglu et al., 2020), PKDE4J (Song et al., 2015), BioMedLEE (L. Chen & Friedman, 2004), and BELMiner (Ravikumar et al., 2017). Among these tools, SemRep was the most suitable semantic parser for our purposes based on the following factors. Firstly, SemRep is widely adopted in LBD research which attests to its reliability and relevance within the field. Secondly, SemRep is a broad coverage semantic parsers due to its capability in identifying a wide range of biomedical concepts and semantic relations from the literature. Furthermore, SemRep is publicly accessible and maintained on a regular basis, which further enhances its practicality for LBD. In contrast, BELMiner is not publicly accessible and has limited adoption in LBD research. BioMedLEE is focused on biomolecular knowledge extraction which limits its applicability as a comprehensive tool for our purposes, as it fails to encompass a broader range of biomedical concepts and

relations, such as drug-disease associations. PKDE4J is a recently developed tool compared to SemRep and is capable of extracting a wide range of semantic relations, however, it was not accessible at the time of our investigation. Based on these factors, SemRep emerged as the most practical choice.

Accordingly, we employed SemRep (batch mode) as a broad coverage biomedical semantic parser to extract semantic-based knowledge—i.e. *subject-predicate-object* semantic predications—from the literature corpus. The *subject* and *object* were noted as standardized ontological concepts with specific semantic types, while the *predicate* as a semantic relation type in an extended version of the UMLS Semantic Network. A detailed description of the full SemRep pipeline can be found in (Kilicoglu et al., 2020); however, for completeness, a summarized description of the pipeline is presented here.

The first processing step in SemRep is the pre-linguistic analysis to split and tokenize sentences and detect biomedical acronyms/abbreviations using MetaMap. Next, lexical and syntactic analysis of tokens is performed to capture lemmas, part-of-speech tags, and multi-word expressions. Subsequently, shallow parsing analysis is applied to identify simple noun phrases. The next step is referential analysis, which involves identifying and mapping biomedical terms in text to standardized UMLS concepts. To achieve this, MetaMap maps noun terms to UMLS concepts, including their unique concept identifiers (CUIs), preferred names, and semantic types. However, due to the limited coverage of gene/protein concepts in UMLS, ABGene is used to identify and map gene/protein terms in text to NCBI Gene identifiers. It should be noted that ABGene does not perform disambiguation on gene/protein terms, but instead uses exact string matching to map these terms to corresponding NCBI Gene concepts. The final step was relational analysis to generate three types of predicates: hypernymic (i.e., IS_A), comparative (e.g., HIGHER_THAN), and associative (e.g., CAUSES, ASSOCIATED_WITH) predications. Hypernymic and comparative predicates were generated through specialized techniques described in (Kilicoglu et al., 2020). Associative predicates were generated through uniform trigger detection and argument identification mechanisms. A set of *indicator rules* were applied to identify lexical elements (i.e., verbs, relational nouns, prepositions, and adjectives) that correspond to particular predicate types. SemRep relies on the SemRep relational ontology to specify associative predicate types that determine the underlying relation/association

between a pair of concepts and their respective semantic types. Overall, the ontology includes 25 associative predicates denoting therapeutic, substance interaction, etiological, and pharmacogenomic relations.

The raw output of the SemRep knowledge extraction was the following elements: *text*, *entity*, and *relation*. The *text* element includes the sentence from which the triple was extracted from and the PubMed Identifier (PMID) of the article. The *entity* element describes attributes of the mapped subject or object concept, including the concept's unique ontological identifier, ontological name, semantic type, and the corresponding term mention in text. We note that SemRep normalizes terms extracted by MetaMap to UMLS concepts, and gene/protein terms extracted by ABGene to NCBI Gene concepts. Moreover, SemRep assigns each concept a semantic type (e.g., disease, organic chemical, molecular function) based on the classification of biomedical concepts in the UMLS Semantic Network. Lastly, the *relation* element outlines the *subject* and *object* concepts (i.e., entities) constituting a semantic predication, in addition to the *predicate* which denotes the underlying association between concepts. An example of the unprocessed pipe-delimited output is provided in Appendix A.

We utilized the *relation* element to create *subject-predicate-object* triples, and retained the *text* and *entity* elements to maintain the semantic types of *subjects* and *objects*, preferred names, CUIs, originating sentences from which the triples were extracted, the subject and object terms in text, and the PMID of the source article. An example of the processed (tabular) output of SemRep is provided in Table 4.1.

As a final processing step, semantic triples were filtered based on the predicate type and semantic types of *subject* and *object* concepts. Specifically, we retained 13 predicates which denote functional and associative relations and excluded comparative (e.g., LESS_THAN, HIGHER_THAN) and hypernymic (IS_A) relations, as these are deemed irrelevant for cancer related knowledge discovery. Table 4.2 presents the 13 retained predicates.

We leveraged the UMLS Semantic Network to categorize the assigned semantic types into high-level semantic groups. For example, the semantic types of "Disease or Syndrome", "Neoplastic Process", and "Pathologic Function" were aggregated into a high-level semantic group called "Disorders". In our work, we targeted the following semantic groups

*Table 4. 1: Processed SemRep output*

| PMID | Subject ID | Subject Name | Subject Semantic Type | Predicate | Object ID | Object Name | Object Semantic Type | Sentence |
|---|---|---|---|---|---|---|---|---|
| 29871641 | C0005740 | bleomycin | Amino Acid, Peptide, or Protein | CAUSES | C0034069 | Pulmonary Fibrosis | Disease or Syndrome | Parthenolide attenuated bleomycin-induced pulmonary fibrosis via the NF-κB/Snail signaling pathway |
| 28059095 | 6121\|7405\|8626\|10970 | RPE65\|UVRAG\|TP63\|CKAP4 | Gene or Genome | ASSOCIATED_WITH | C0017525 | Giant Cell Tumors | Neoplastic Process | P63 expression in giant cell tumors of the bone seems to be associated with H3F3 gene mutations |
| 32608212 | C0072652 | pyranoprofen | Organic Chemical | DISRUPTS | C0162638 | Apoptosis | Cell Function | Our study aimed to test the anti-apoptosis effects of pranoprofen (PF), a specific prostaglandin E2 (PGE2) inhibitor, on human CHs |

which relate to cancer research: *chemicals and drugs*, *genes and molecular sequences, disorders,* and *physiology*. Table 4.3 presents the full list of semantic groups and corresponding semantic types utilized in this research.

*Table 4.1: UMLS Semantic Network predicates*

| **Predicate** | **Description** |
|---|---|
| *AFFECTS* | Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity |
| *ASSOCIATED_WITH* | Has a relationship with/to |
| *INTERACTS_WITH* | Substance interaction |
| *CAUSES* | Brings about a condition or an effect. Implied here is that an agent, such as for example, a pharmacologic substance or an organism, has brought about the effect |
| *STIMULATES* | Increases or facilitates the action or function of (substance interaction) |
| *INHIBITS* | Decreases, limits, or blocks the action or function of (substance interaction) |
| *TREATS* | Applies a remedy with the object of effecting a cure or managing a condition |
| *DISRUPTS* | Alters or influences an already existing condition, state, or situation. Produces a negative effect on |
| *AUGMENTS* | Expands or stimulates a process |
| *PREDISPOSES* | To be a risk to a disorder, pathology, or condition |
| *PRODUCES* | Brings forth, generates or creates |
| *PREVENTS* | Stops, hinders or eliminates an action or condition |
| *COMPLICATES* | Causes to become more severe or complex |

*Table 4.2: Concept semantic groups and corresponding semantic types*

| **Semantic Group** | **Semantic Type** |
|---|---|
| *Chemicals & Drugs* | Antibiotic |
| | Biologically Active Substance |
| | Clinical Drug |
| | Element, Ion, or Isotope |
| | Enzyme |

| | |
|---|---|
| | Hormone |
| | Hazardous or Poisonous Substance |
| | Immunologic Factor |
| | Inorganic Chemical |
| | Organic Chemical |
| | Pharmacologic Substance |
| | Receptor |
| | Vitamin |
| *Disorders* | Acquired Abnormality |
| | Anatomical Abnormality |
| | Cell or Molecular Dysfunction |
| | Congenital Abnormality |
| | Disease or Syndrome |
| | Finding |
| | Injury or Poisoning |
| | Neoplastic Process |
| | Pathologic Function |
| | Sign or Symptom |
| *Genes & Molecular Sequences* | Gene or Genome |
| | Molecular Sequence |
| | Nucleotide Sequence |
| *Physiology* | Cell Function |
| | Clinical Attribute |
| | Genetic Function |
| | Molecular Function |
| | Organism Attribute |
| | Organism Function |
| | Organ or Tissue Function |

| | |
|---|---|
| | Physiologic Function |

### 4.2.2 Disambiguation of Gene and Protein Concepts:

While SemRep is a well-regarded semantic parser with applications in LBD, clinical guideline development, question-answering systems, and clinical decision support, it has certain limitations in disambiguating gene and protein terms in literature text and mapping to NCBI Gene identifiers, thus resulting in a high rate of false-positive concept mappings (Kilicoglu et al., 2020). SemRep fails to resolve ambiguities in the short-form (abbreviated) name of genes/proteins found in the literature, as gene names are used non-uniquely and the same abbreviated name can refer to different genes. For instance, *NAP1* is the abbreviated name of 5 genes: nucleosome assembly protein 1 like 1, NCK associated protein 1, napsin A aspartic peptidase, and acyl-CoA thioesterase 8.

For semantic-based knowledge extraction, this limitation results in triples where the *subject* or *object* are ambiguous as they are represented by multiple gene/protein concepts (i.e., UMLS or NCBI Gene). Table 4.4 provides examples of semantic triples consisting of ambiguous *subject* or *object* entities which are mapped to multiple NCBI or UMLS concept identifiers. For instance, SemRep represents the gene term "NRF2" as two distinct gene concepts: 2551 (GABPA) and 4780 (NFE2L2), which have different biological functions. In such cases, determining the true concept becomes challenging without referring to the source article. The significance of this limitation becomes more apparent when extracting knowledge from cancer literature, as genes/proteins play crucial roles in the development of cancer. In high-stake discovery tasks, such as drug repurposing, it is imperative that the extracted semantic triples consist of unambiguous subject or object that are normalized to a single concept.

To overcome this limitation, this phase of the knowledge extraction component investigates methods to autonomously disambiguate gene and protein concepts when analyzing literature texts. Our approach is to combine SemRep with a biomedical text mining tool—i.e. PubTator—to perform concept disambiguation, utilizing the syntax and semantics of the target gene/protein term and its neighbouring terms to achieve a context-driven outcome. The intent is to resolve ambiguous gene and protein concepts found in semantic triples by disambiguating them to a single concept. To the best of our knowledge,

*Table 4.3: Examples of semantic triples consisting of ambiguous gene or protein concepts*

| PMID | Subject ID | Subject Name | Predicate | Object ID | Object Name | Sentence |
|---|---|---|---|---|---|---|
| 28866133 | 1215\|6566\|27349\|28985 | CMA1\|SLC16A1\|MCAT\|MCTS1 | AUGMENTS | C1159978 | osteoclast differentiation | ***MCT1*** expression is significantly upregulated during osteoclast differentiation |
| 29483950 | 3737\|9261 | KCNA2\|MAPKAPK2 | ASSOCIATED_WITH | C0152018 | Esophageal carcinoma | Expression of ***MK2*** and ETV1 are prognostic factors in patients, with esophageal adenocarcinoma |
| 28059095 | 6121\|7405\|8626\|10970 | RPE65\|UVRAG\|TP63\|CKAP4 | ASSOCIATED_WITH | C0017525 | Giant Cell Tumors | ***P63*** expression in giant cell tumors of the bone seems to be associated with H3F3 gene mutations |
| 28125038 | C0003818 | Arsenic | STIMULATES | 2551\|4780 | GABPA\|NFE2L2 | ***Nrf2*** is required for basal and arsenic-induced p62 up-regulation |
| 29805671 | C0297674\|23545\|27239\|170589 | cyclin A2\|ATP6V0A2\|GPR162\|GPHA2 | ASSOCIATED_WITH | C0001418 | Adenocarcinoma | Significant correlations were detected between LINC00968, miR-9-3p and ***CCNA2*** in lung adenocarcinoma |
| 29761522 | C0334227 | Tumor cells, malignant | INTERACTS_WITH | C0030190\|5054\|5055 | Plasminogen Activator Inhibitor 1\|SERPINE1\|SERPINB2 | Extracellular vesicles (EV) shed from cancer cells may contribute to the regulation of TF and PAI-1 |
| 30881493 | C0259275\|2069\|2099 | BRCA1 Protein\|EREG\|ESR1 | ASSOCIATED_WITH | C0006142 | Malignant neoplasm of breast | Fanconi anemia group D2 protein (FANCD2) and ***breast cancer type 1 susceptibility protein (BRCA1)***, within the FA/BRCA pathway, are involved in the regulation of DNA damage repair, which is associated with breast cancer (BC) progression |

this is the first LBD framework that combines the output of these two established biomedical tools to ensure the accuracy and representativeness of the extracted knowledge. We investigated a number of concept recognition and annotation tools, and decided to use PubTator (Wei et al., 2019) to process the literature corpus to annotate disambiguate gene and protein concepts. PubTator employs a novel convolutional neural network based method to address the concept disambiguation task, which involves identifying the most plausible concept by analyzing the syntax and semantics of the target term and its neighboring terms.

In this phase of knowledge extraction, a subset of the literature corpus—i.e. articles that include gene or protein terms in their titles or abstracts, and is identified based on the output of SemRep--was used as input to PubTator. Specifically, we searched for semantic triples where the *subject* or *object* is a gene/protein (based on the semantic group) and is represented by multiple concepts, which we consider as an ambiguous triple. These triples can be easily identified, as the *subject/object* concepts consist of multiple CUIs and/or NCBI Gene identifiers separated by the pipe character "|" - e.g., (C0030190|5054|5055). Next, we retrieve the titles and abstracts associated with ambiguous semantic triples and utilize them as input into PubTator for concept disambiguation. It should be noted that the output of this process does not consist of semantic triples. Instead, it provides disambiguated annotations of gene/protein terms in literature text and represents them as NCBI Gene concepts. Table 4.5 provides examples of the output of the concept disambiguation process, which presents precise representations of biomedical entities mentioned in sentences. For instance, mention of the NRF2 gene in the following sentence: "*Nrf2* is required for basal and arsenic-induced p62 up-regulation" is disambiguated to NFE2L2 (NFE2 like bZIP transcription factor 2) and accordingly represented using the NCBI Gene identifier: 4780.

*Table 4.4: Output of concept disambiguation*

| **PMID** | **Sentence** | **Gene/Protein Term** | **Gene/Protein Concept Name** | **Gene/Protein Concept Identifier** |
|---|---|---|---|---|
| 28866133 | MCT1 expression is significantly upregulated during osteoclast differentiation | MCT1 | Solute carrier family 16 member 1 (SLC16A1) | 6566 |
| 29483950 | Expression of MK2 and ETV1 are prognostic factors in patients, with esophageal adenocarcinoma | MK2 | MAPK activated protein kinase 2 (MAPKAPK2) | 9261 |

| PMID | Sentence | Gene/Protein Term | Gene/Protein Concept Name | Gene/Protein Concept Identifier |
|---|---|---|---|---|
| 28059095 | P63 expression in giant cell tumors of the bone seems to be associated with H3F3 gene mutations | P63 | Tumor protein p63 (TP63) | 8626 |
| 28125038 | Nrf2 is required for basal and arsenic-induced p62 up-regulation | Nrf2 | NFE2 like bZIP transcription factor 2 (NFE2L2) | 4780 |
| 29805671 | Significant correlations were detected between LINC00968, miR-9-3p and CCNA2 in lung adenocarcinoma | CCNA2 | Cyclin A2 (CCNA2) | 890 |
| 29761522 | Extracellular vesicles (EV) shed from cancer cells may contribute to the regulation of TF and PAI-1 | PAI-1 | Serpin family E member 1 (SERPINE1) | 5054 |
| 30881493 | Fanconi anemia group D2 protein (FANCD2) and breast cancer type 1 susceptibility protein (BRCA1), within the FA/BRCA pathway, are involved in the regulation of DNA damage repair, which is associated with breast cancer (BC) progression | BRCA1 | Breast cancer type 1 susceptibility protein (BRCA1) | 672 |

Table 4.6 provides a comparison of the ambiguous concepts extracted by SemRep with the disambiguated concepts extracted by PubTator. We note that PubTator utilizes convolutional neural network classifiers to disambiguate gene/protein terms to the most likely concept, achieving an accuracy of 85.2%. In contrast, SemRep uses simple string matching techniques to match gene/protein terms in text to one or more standardized concepts, which contributes to the ambiguous gene/protein concepts in the extracted semantic triples.

*Table 4. 2: Comparison of ambiguous and corresponding disambiguated concepts*

| PMID | Sentence | Gene/Protein Term | Ambiguous Concepts (SemRep) | Disambiguated Concept (PubTator) |
|---|---|---|---|---|
| 28866133 | MCT1 expression is significantly upregulated during osteoclast differentiation | MCT1 | CMA1\|SLC16A1\|MCAT\|MCTS1 | SLC16A1 |
| 29483950 | Expression of MK2 and ETV1 are prognostic factors in patients, with esophageal adenocarcinoma | MK2 | KCNA2\|MAPKAPK2 | MAPKAPK2 |

| PMID | Sentence | Gene/Protein Term | Ambiguous Concepts (SemRep) | Disambiguated Concept (PubTator) |
|---|---|---|---|---|
| 28059095 | P63 expression in giant cell tumors of the bone seems to be associated with H3F3 gene mutations | P63 | RPE65\|UVRAG\|TP63\|CKAP4 | TP63 |
| 28125038 | Nrf2 is required for basal and arsenic-induced p62 up-regulation | Nrf2 | GABPA\|NFE2L2 | NFE2L2 |
| 29805671 | Significant correlations were detected between LINC00968, miR-9-3p and CCNA2 in lung adenocarcinoma | CCNA2 | cyclin A2\|ATP6V0A2\|GPR162\|GPHA2 | Cyclin A2 |
| 29761522 | Extracellular vesicles (EV) shed from cancer cells may contribute to the regulation of TF and PAI-1 | PAI-1 | Plasminogen Activator Inhibitor 1\|SERPINE1\|SERPINB2 | SERPINE1 |
| 30881493 | Fanconi anemia group D2 protein (FANCD2) and breast cancer type 1 susceptibility protein (BRCA1), within the FA/BRCA pathway, are involved in the regulation of DNA damage repair, which is associated with breast cancer (BC) progression | BRCA1 | BRCA1 Protein\|EREG\|ESR1 | BRCA1 |

To resolve ambiguous gene/protein concepts in semantic triples, we replace the ambiguous subject/object with the corresponding disambiguated concept from the output of PubTator. This is accomplished by aligning the ambiguous gene/protein concepts with the disambiguated concepts on the basis of matching PMIDs and gene/concept terms within sentences. For example, consider the following semantic triple which has an ambiguous *object* mapped to multiple concepts:

- **Ambiguous semantic triple**: C0003818 (Arsenic) – STIMULATES - 2551|4780 (GABPA|NFE2L2)

To resolve the ambiguous *object*, we first determined the PMID associated with the triple and then identify the gene/protein term corresponding to the ambiguous *object* (i.e., short-form name of gene/protein in sentence):

- **PMID**: 28125038
- **Object term**: Nrf2
- **Ambiguous object concept**: 2551|4780 (GABPA|NFE2L2)

Next, using the output of PubTator, we identify the matching PMID, gene/protein term, and the corresponding disambiguated concept:

- **PMID**: 28125038
- **Gene/protein term**: Nrf2
- **Disambiguated concept**: 4780 (NFE2L2)

Accordingly, we replace the ambiguous *object* with its corresponding disambiguated concept extracted by PubTator, resulting in the following semantic triple:

- **Disambiguated semantic triple**: C0003818 (Arsenic) – STIMULATES - 4780 (NFE2L2)

We note that in some cases PubTator failed to annotate gene/protein terms in sentences with disambiguated concept representations. In such cases, we cannot resolve ambiguous *subject/object* concepts in semantic triples, as there is no corresponding disambiguated concept. We argue that unresolved ambiguous semantic triples are detrimental to the knowledge discovery process due to the presence of imprecise *subject* and *object* concepts (Preiss & Stevenson, 2016). Hence, unresolved ambiguous *subjects* or *objects* were eliminated, and by extension the corresponding semantic triple is also eliminated. This elimination did not result in a significant loss of knowledge, as the eliminated triples represented less than 0.5% of all extracted triples.

The outcome of this component consists of a set of semantic triples where the *subject* or *object* is a disambiguated gene/protein represented by a precise concept. These triples are merged with the remaining triples extracted in section 4.2.1.

## 4.3 Consolidation of Granular Concepts in Semantic Triples:

The granularity of UMLS concepts in semantic triples presents a challenge for LBD due the presence of semantically equivalent triples that convey the same underlying knowledge but are represented with distinct concepts (Vlietstra et al., 2017). This leads to an increase in the number of unique knowledge instances, thereby expanding the search space for discovery. To resolve this challenge, the semantic consolidation component aims to merge fine-grained concepts into higher-level/generalized concepts without compromising the semantics (i.e. intended meaning) of the acquired literature-based knowledge. The intent is to reduce the number of unique fine-grained concepts that constitute the semantic triples, to achieve a smaller set of (generalized) concepts and, as a result, a reduced search space for LBD. We utilize widely recognized terminological resources for semantic consolidation, such as MeSH, GO, and PRO, to ensure that the semantic triples are compatible with external biomedical knowledge bases. The idea is to facilitate the integration of knowledge from external sources for downstream knowledge completion tasks.

The semantic concept consolidation component leverages semantic alignment techniques to align hierarchical structures of targeted terminologies with UMLS, then map concepts to corresponding concept representations in target terminologies. Our aim is to create consolidated and standardized representations of concepts, where multiple fine-grained concepts are merged into broader and more encompassing concepts in target terminologies. To initiate semantic consolidation, we identified target terminologies for each semantic group which served as the basis for semantic consolidation. Several biomedical terminologies were explored based on the following criteria: (a) concepts in the target terminology merge two or more concepts into a single concept; and (b) the target terminology is a used as a standard for representing biomedical knowledge in external knowledge bases. Based on this criteria, we selected the following resources for semantic consolidation:

I. **Medical Subject Heading (MeSH)** (National Library of Medicine, 2023): MeSH is a widely used resource for representing biomedical knowledge, encompassing approximately 30,000 concepts organized in 16 hierarchical categories. On average, MeSH concepts subsume two corresponding UMLS concepts. MeSH is

used as the target terminological resource to merge concepts in the "*chemicals and drugs*" and "*disorders*" semantic groups into higher-level representations.

II. **Gene Ontology (GO)** (The Gene Ontology Consortium, 2019): GO is a well-known biomedical ontology consisting of 50,000 concepts covering physiological, biological, and molecular entities organized in a hierarchical structre. On average, GO concepts subsume two corresponding UMLS concepts. GO is used to merge concepts in the "*physiology*" semantic group into higher-level representations.

III. **Protein Ontology (PRO)** (Natale et al., 2011): PRO is a specialized ontology covering gene and protein concepts organized in a hierarchical structure. PRO includes high-level concepts which abstract over orthologous genes and proteins and, thus, subsume more than two corresponding NCBI Gene/UMLS concepts. We utilize PRO as a terminology to map concepts in the "*genes and molecular sequences*" semantic group into higher-level representations.

As a first step towards consolidation of concepts, we aggregated the subject and object concepts in semantic triples into high-level semantic groups: *chemicals and drugs*, *genes and molecular sequences*, *disorders*, and *physiology*. Then for concepts in the *chemicals and drugs, disorders,* and *physiology semantic groups*, we constructed hierarchical networks that provide hierarchical paths between a given UMLS concept and the highest-level concept within the semantic group. These hierarchical networks are created using the UMLS Metathesaurus resources, which provide hierarchical relationships between concepts via the Computable Hierarchies (MRHIER.RRF) and Related Concepts (MRREL.RRF) files. Hierarchical networks were constructed using the outgoing CHD (has a child relationship) and RN (has a narrower relationship) relations. Figure 4.2 presents a section of the hierarchical network for the *disorders* semantic group, showing the hierarchical path between "Invasive Ductal Breast Carcinoma" and "Neoplasms".

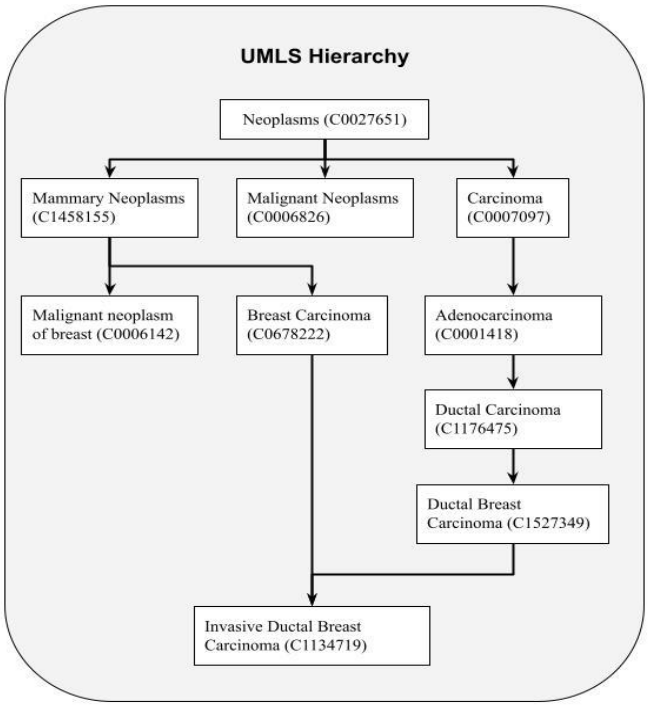


*Figure 4.2: Excerpt of the constructed UMLS hierarchy*

Concepts in the *genes and molecular sequences* semantic group, which are represented using NCBI Gene, cannot be organized hierarchically as the NCBI Gene database does not provide a hierarchy of concepts. Instead, we organized these concepts into different orthologous sub-groups, which denote genes in different species that retain the same genetic function (e.g., association of the TP53 gene in mice and humans with tumor suppression). Specifically, we used the *gene orthologs* file (https://ftp.ncbi.nih.gov/gene/DATA/gene_orthologs.gz) which is provided by NCBI Gene to describe associations between orthologous genes in different species. This grouping of gene concepts is necessary, as our aim is to standardize and merge concepts in each sub-group to higher-level concepts that abstract over associated orthologs.

The semantic consolidation of concepts in the *chemicals and drugs*, *disorders*, and *physiology* semantic groups commenced by extracting a subset of concepts from MeSH and GO hierarchies that correspond to these semantic groups. From MeSH, we extracted all child concepts within the *Diseases* and *Chemicals and Drugs* root categories. From GO, we extracted all child concepts of *Biological Process* (GO:0008150) and *Molecular Function* (GO:0003674). Subsequently, we constructed hierarchical networks of MeSH and GO concepts, then aligned them with the UMLS hierarchies. The alignment is

accomplished using the MRCONSO resource, which is part of the UMLS Metathesaurus that provides one-to-many links between a given UMLS concept and concepts from other terminologies, such as MeSH and GO. This ensures that one or more UMLS concepts are mapped to a corresponding concept from MeSH or GO.

Figure 4.3 illustrates an example of the concepts alignment and mapping to consolidate one or more UMLS concepts to corresponding MeSH targets. In this example, the UMLS hierarchy consists of 10 unique concepts, with "Neoplasm" as the root concept and "Invasive Ductal Breast Carcinoma" as a leaf concept. The MeSH hierarchy consists of 9 unique concepts, with "Neoplasm" as the root concept and "Carcinoma, Ductal, Breast" as a leaf concept. Through the mapping process, multiple granular UMLS concepts were consolidated to a single MeSH concept. We note that UMLS concepts “Invasive Ductal Carcinoma” and “Ductal Breast Carcinoma” were consolidated into a single MeSH concept which subsumed both UMLS concepts. Similarly, “Mammary Neoplasms”, “Malignant neoplasm of breast”, and “Breast Carcinoma” were consolidated into a single MeSH concept (“Breast Neoplasms”).

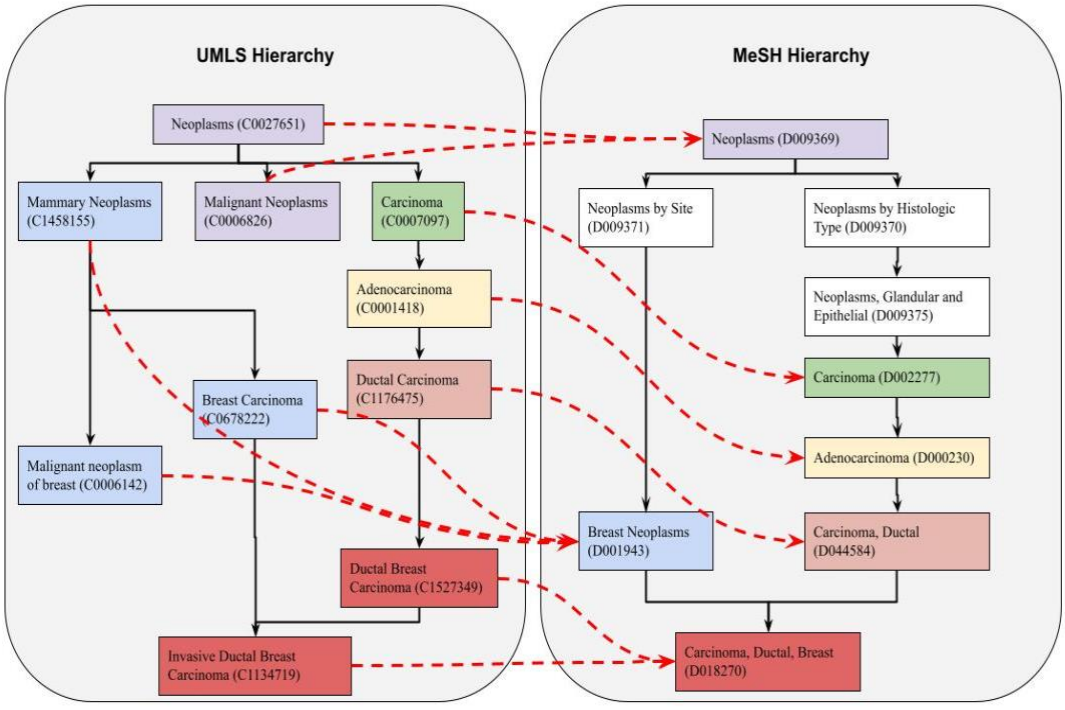


*Figure 4.3: Consolidation of concepts via alignment and mapping to MeSH*

For the consolidation of concepts in the *genes and molecular sequences* semantic group, we opted to directly map these concepts to PRO using UniProt as an intermediary resource. Explicitly, UniProt is a large biomedical knowledge base which includes information about protein sequences, functions, structures, interactions, and encoding genes. UniProt provides cross-references to many external databases, including NCBI Gene, via the idmapping file (https://proconsortium.org/download/current/). We utilized the idmapping file to directly map NCBI Gene identifiers to corresponding UniProtKB identifiers, which are used as the intermediary link to PRO. Next, we use PRO hierarchical structure to leverage hierarchical IS_A relations between genes/proteins across different species (i.e., orthologs) to consolidate and represent them by a single gene-level concept (Natale et al., 2011). These concepts are high-level abstractions that generalize over related orthologs across different species. As an example, Figure 4.4 shows a subset of the PRO hierarchy whereby the human and mouse orthologs of the CD80 gene/protein are represented by their respective organism-level concepts - i.e., hCD80 (PR:P33681) and mCD80 (PR:Q00609). These concepts are subsumed by the CD80 gene-level concept (PR:000001438) which abstracts over the human and mouse orthologs. This task entails leveraging the PRO hierarchy to map organism-level concepts to gene-level concepts to ensure that all orthologs of the same gene/protein are consolidated and represented by a single concept.

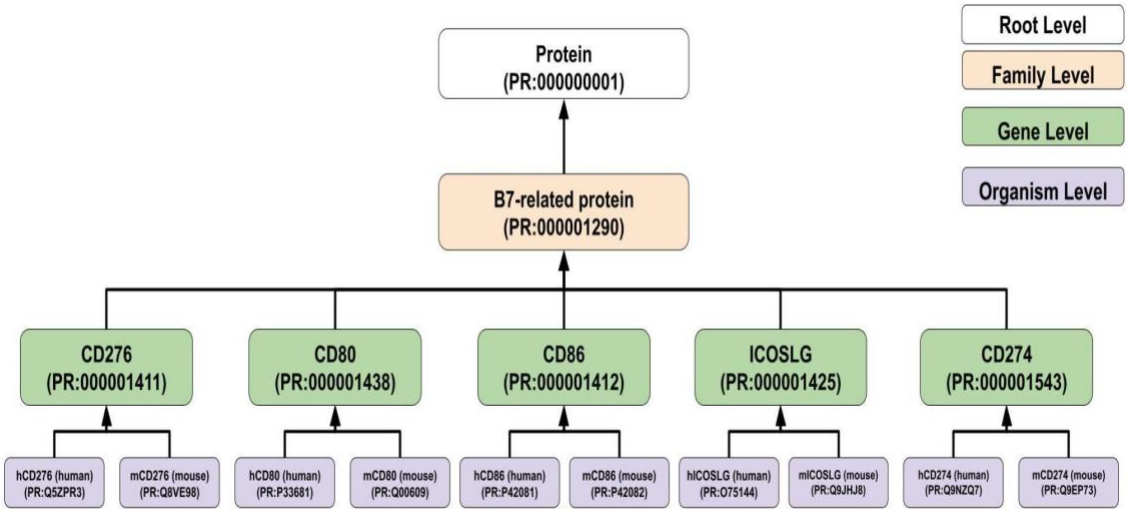


*Figure 4.4: Subset of the Protein Ontology hierarchy*

Lastly, we ensured that the provenance of consolidated concepts is maintained by creating a *concept mapping file* that holds the source UMLS or NCBI Gene concepts and their

corresponding equivalents from target terminologies (i.e., MeSH, GO, and PRO). The concept mapping file was used to replace UMLS and NCBI Gene concepts in semantic triples with corresponding concepts from MeSH, GO, and PRO.

## 4.4 Graph Based Representation of Literature-Based Knowledge:

To facilitate downstream knowledge completion and discovery, the semantic triples—i.e. *subject-predicate-object* triples—are represented as a baseline literature-based KG, where the *subjects*/*objects* are represented as nodes and *predicates* as directed semantic edges. The nodes are assigned attributes denoting the concept's unique identifier, preferred name, semantic type, and semantic group, likewise the directed semantic edges are also assigned attributes denoting the predicate type.

The *baseline KG* is created using Neo4j as it supports storing and visualization of semantic triples. We use the Cypher query language to import and represent the semantic triples as described in (Hristovski et al., 2015). Code 4.1 presents the Cypher query used to generate the *baseline KG*. The import query iterates through a Comma-Separated Values (CSV) file containing the semantic triples and associated metadata to create a graph in Neo4j. The query starts by using the `LOAD CSV` clause to read each line from the CSV file and store it in the `line` variable. The `WITH` clause is then used to pass the `line` variable to the next part of the query. The first `MERGE` clause creates a node for the *subject* in the graph with label 'Concept' and a unique identifier `'concept_id'` based on the first element in the `line` variable. If the node already exists, the query does nothing. If the node does not exist, the `ON CREATE SET` clause sets the `name`, `sem_type` (i.e., semantic type), and `sem_grp` (i.e., semantic group) attributes based on the second, third, and fourth elements of the `line` variable. The second `MERGE` clause creates a node for the *object* in the same way as the subject node. Finally, the query creates a directed edge labeled 'Relation' between the *subject* and *object* nodes using the specified predicate (i.e., semantic relation). The outcome is a baseline literature-based KG constructed from consolidated semantic triples derived from the literature. We regard the KG as baseline since given the limitations of the knowledge extraction task, the KG is incomplete due to missing nodes and relations. Moving forward, we use the *baseline KG* to progressively augment it by adding the 'missing' nodes and relations.

```
LOAD CSV FROM ’baseline_triples.csv’
  AS line
WITH line
MERGE (c1:Concept {concept_id: line[0]})
ON CREATE SET c1.name=line[1],
  c1.sem_type=line[2], c1.sem_grp=line[3]
MERGE (c2:Concept {concept_id: line[5]})
ON CREATE SET c2.name=line[6],
  c2.sem_type=line[7], c2.sem_grp=line[8]
MERGE (c1)-[r:Relation {type:
  line[4]}]->(c2)
```

*Code 4. 1: Cypher query to construct literature-based KG*

## 4.5 Knowledge Integration and Completion:

The knowledge integration and completion component presents novel methods to address the limitations of incomplete literature-based knowledge extraction by progressively augmenting the *baseline KG* across two phases to generate a more complete literature-based KG. The first phase leverages external knowledge bases to extend the *baseline KG* by integrating curated biomedical knowledge, resulting in an enhanced *integrated KG*. The second phase leverages state-of-the-art graph-based representation learning methods (i.e., knowledge graph embeddings) for Knowledge Graph Completion (KGC), whereby we predict the missing relations between biomedical concepts in the *integrated KG* to generate *augmented literature-based KG* which serves as the foundation for literature-based knowledge discovery.

The following sub-sections present our methods for the knowledge integration and knowledge completion phases.

### 4.5.1 Integration of External Knowledge Bases (KBs):

Biomedical KBs, such as the Comparative Toxicogenomics Database and Gene Ontology, capture biomedical knowledge via manual curation of scientific literature. These KBs contain rich assertional knowledge describing associations and interactions between genes-diseases, chemicals-diseases, and genes-biological processes. To extend the knowledge coverage for LBD, we used external KBs to (a) collect knowledge that is missing in the literature and (b) build a knowledge-rich biomedical KG to create high quality Knowledge Graph Embeddings (KGEs). KGEs are vector-based representations of nodes and relations

constituting a KG that can be used for downstream relation prediction tasks. KG-based representation learning techniques are typically used to map nodes and relations in a continuous vector space, such that the geometric relations between vectors capture the underlying semantics of the KG. By integrating knowledge from external KBs, our aim is to improve the vector-based representations of nodes and relations in the KG.

The biomedical KBs and extracted assertional knowledge used in this work are:

1. Comparative Toxicogenomics Database (CTD) (Davis et al., 2019) for chemical-disease, and gene-disease interactions. Files containing curated CTD knowledge were downloaded from https://ctdbase.org/downloads/ on March 15, 2022.
2. Gene Ontology (GO) annotations (The Gene Ontology Consortium, 2019) for gene-biological process and gene-molecular function associations. Files containing GO annotations were downloaded from https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz on March 15, 2022.

We extracted curated assertional knowledge about cancers to supplement the literature-based semantic triples as follows: (i) searching for MeSH identifiers of all neoplastic diseases – i.e., child concepts of Neoplasms (MESH:D009369); and (ii) extracting all associated genes, chemicals and drugs from CTD. For GO annotations, we use the set of gene extracted from CTD to identify their associations with biological processes and molecular functions.

To extend the knowledge coverage of the *baseline KG*, we integrated biomedical associations – represented as (*subject, predicate, object*) triples – extracted from structured KBs. The integration commenced by replacing KB asserted predicates with semantically equivalent predicates from the UMLS semantic network. For example, curated associations in CTD are labeled as 'marker/mechanism (M)' or 'therapeutic (T)', whereby the label 'M' indicates that a chemical or gene product has a causal association with a disease, and the label 'T' indicates that a chemical or gene product has a therapeutic role in a disease. CTD associations with the 'M' label are replaced with *predisposes*, while associations with the 'T' label are replaced with *treats*. For GO annotations, the 'qualifier' label is used to interpret the associations between genes, molecular functions, and biological processes. We particularly focus on two types of GO qualifiers: (i) *involved_in* indicating that a gene or its product have a role in a biological process; and (ii) *enables* indicating that a gene or

its product are engaged in executing a molecular function. The GO qualifier *involved_in* is replaced with *affects* and *enables* is replaced with *stimulates*. The resultant of this step is an *integrated KG* that extends the baseline literature-based knowledge (i.e. the *baseline KG*) with manually curated knowledge sourced from external sources.

We do understand that the *integrated KG* is likely still missing relations due to (a) lack of available explicit knowledge in the literature, and (b) absence of meaningful semantic relations between biomedical entities, as manually curated KBs do not guarantee that the latest scientific knowledge is incorporated due to the time-intensive process of manual curation.

### 4.5.2 Knowledge Graph Completion (KGC):

To augment the *integrated KG* with additional knowledge (i.e. missing relations), we applied KGC methods to predict semantic relations between logically associated nodes (i.e., biomedical entities). KGC is pursued by leveraging Knowledge Graph Embedding (KGE) methods to embed nodes and semantic relations in the *integrated KG* as low-dimensional vectors which encode latent semantic and structural features of the KG (Mohamed et al., 2021). The following sections present our methods for KGC. First we describe the methods to embed the *integrated KG* nodes and relations as low-dimensional embeddings. Next, we describe our approach to identify implicitly associated nodes, and present the relation prediction method using the generated KG embeddings.

Missing relations are inferred by utilizing a secondary literature-based knowledge resource, characterized by Medical Subject Heading (MeSH) descriptors assigned to cancer literature, which provides a comprehensive concept-based representation of an article's scientific content. We represent MeSH descriptors by considering their co-occurrences to create partial/incomplete semantic triples, whereby the *subject* and *object* are existing nodes in the *integrated KG* but the relation between them is missing or unknown - i.e., *subject - ? - object*. Consequently, these partial/incomplete triples are used as input for KGC, with the objective of predicting the missing relation to generate complete *subject-predicate-object* triples to augment the *integrated KG*. The output of this task results in an *augmented literature-based KG* which serves as the foundation for literature-based knowledge discovery.

#### 4.5.2.1 Embedding the Integrated KG:

To predict missing relations between concepts represented in the *integrated KG*, we leveraged KGE methods that encode the *integrated* KG concepts and relations as vectors while preserving the KG's structure and semantic information. Figure 4.5 presents the pipeline for generating node and relation vectors via KGE methods (Bonner et al., 2022; Mohamed et al., 2021).

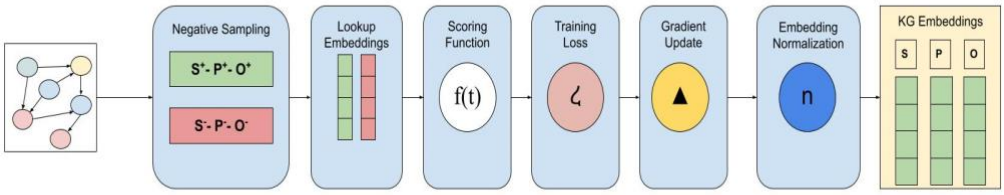


*Figure 4.5: Schematic of the training pipeline of a KGE model*

Given a set of triples as input, KGE methods first generate negative samples from positive triples using uniform random corruptions of the *subject*, *object*, or *predicate* (i.e., semantic relation). Next, the embedding lookup layer assigns positive and negative triples random embeddings from uniform or Gaussian distributions (Mohamed et al., 2021). These embeddings are then used as input for method-dependent scoring functions to generate scores for all positive and negative (corrupted) triples. Subsequently, the training loss is calculated using different variations of loss functions with the objective of maximizing scores of positive triples while minimizing scores for corrupted triples. KGE optimize the training loss with stochastic gradient descent algorithms, such as Adam and AdaGrad. Lastly, the embeddings for nodes and relations are normalized as a regularization strategy to improve their generalization. This multi-step training process is executed iteratively to update the embeddings until optimal representations of the KG's latent and semantic features are learned, such as node semantic types, local and global neighborhoods, and relation types (i.e., symmetrical, asymmetrical, one-to-one, etc.). The outcome of this process are vector-based representations of KG nodes and relations, which can be utilized for relation prediction.

This dissertation explored three KGE methods: TransE, ComplEx, and DistMult as each differs in terms of how they embed and represent KG nodes and relations. We investigated

these KGE methods because: (i) they provide competitive results for entity prediction in biomedical KGs (Nicholson & Greene, 2020); (ii) given the size of the *integrated KG*, the selected methods require less training time compared to deep learning-based models such as ConvE and ConvKB; and (iii) able to model a wide-range of relations including symmetric, antisymmetric, and many-to-many relations, which are abundant in literature-based KGs. TransE, ComplEx, and DistMult were implemented using the open source AmpliGraph 1.4.0 package (https://github.com/Accenture/AmpliGraph) for learning large-scale KG embeddings. KGE methods were initialized by inputting *subject-predicate-object* triples constituting the *integrated KG*, followed by tuning of shared hyperparameters for training. We explored a range of hyperparameter values (Table 4.6) based on previously published works (Bonner et al., 2022; Chang et al., 2020), and tuned method using a grid search on a validation set. Specifically, we split the *integrated KG* into a training KG and a validation KG (90% and 10% respectively). The embeddings generated from the training KG were used to predict the relations in the validation KG. For KGE, the combination of parameters that yielded the highest Mean Reciprocal Rank (MRR) was considered the optimal set of parameters. MRR is calculated by averaging the inverse ranks of true relations across all predictions, with values closer to 1 indicating good predictive performance. Formally, MRR is defined as:

$$MRR = \frac{1}{|P|}\sum_{p \in P}\frac{1}{p}$$

where $p$ represents the highest rank of the true predicted relation, $P$ represents all predictions. The final set of parameters for each model are presented in Table 4.7.

*Table 4.5: Hyperparameter values explored for KGE training*

| **Hyperparameter** | **Values** |
|---|---|
| Embedding dimension ($k$) | 100, 200, 300 |
| **Hyperparameter** | **Values** |
| Number of epochs | 100, 300, 500, 700 |
| Learning rate | 0.01, 0.001, 0.0001 |
| Negative samples | 10, 25 |
| Loss function | Pairwise, Multiclass Negative Log-Likelihood |

*Table 4.6: Selected hyperparameter values*

| KGE Model | Hyperparameters | | | | |
|---|---|---|---|---|---|
| | *Embedding dimension* | *Epochs* | *Learning rate* | *Negative samples* | *Loss function* |
| TransE | 300 | 300 | 0.001 | 10 | Multiclass negative log-likelihood |
| DistMult | 300 | 500 | 0. 0001 | 25 | Multiclass negative log-likelihood |
| ComplEx | 200 | 200 | 0.001 | 10 | Multiclass negative log-likelihood |

#### 4.5.2.2 Knowledge Graph Completion Using KG Embeddings:

We define KGC as the task of *relation prediction* – i.e., given an incomplete triple (*s, ?, o*), the aim is to fill the missing element within a KG with the most plausible *predicate (p)* using the information encoded in the KG embeddings. A major challenge in KGC is determining which relations between KG concepts are missing, as applying the relation prediction task on every pair of unlinked concepts/nodes requires intensive computational resources (i.e., time complexity). For any given KGE model, the time complexity of relation prediction depends on the number of dimensions, number of unique unlinked entities (i.e., concepts), and the number of relations in the KG – i.e., $O(ek + rk)$, where $e$ and $r$ represent number of entities and relations respectively, and $k$ is the number of dimensions of relation and entity embeddings. Further, when relation prediction is unconstrained, a significant number of new relations may be added, resulting in a dense KG and increased cost of knowledge discovery. Specifically, applying open- or closed-based discovery on a dense KG would result in generating a large number of *ABC* discovery paths, which would be difficult to assess and review. Thus, constraining the number of entities to a focused subset can result in a reduction of the time complexity for the relation prediction task. As such, it is necessary to pre-define a subset of *subject* and *object* concepts that have some form of implicit association between them in the literature but are not linked by a semantic relation in the KG.

Our approach to KGC is to leverage Medical Subject Heading (MeSH) concepts (National Library of Medicine, 2023) as they are used to index PubMed articles (National Library of Medicine, 2022) with complete concept-based representation of articles. When considering

the set of MeSH concepts assigned to an article, it can be argued that their co-occurrences suggest some implicit biomedical association between them. To illustrate this notion, consider the article in (Kim & Park, 2019) which describes the role of chemokine CXCL12 in promoting cellular proliferation in senescent tumor cells. Despite the absence of explicit mentions of the CXCL12 chemokine in the article's title and abstract, it can be inferred that an implicit association between CXCL12, cell proliferation, and cellular senescence exists by considering the co-occurrence of MeSH descriptors assigned to the article. We created the set of MeSH associations from PubMed articles retrieved by the literature search. Next, we classified MeSH descriptors into the following high-level semantic groups: *Chemicals*, *Diseases*, *Biological and Physiological Phenomena*, and *Proteins*. For knowledge alignment, we map entities in the *Biological and Physiological Phenomena* semantic groups to corresponding concepts in GO, and entities in the *Protein* semantic group to corresponding concepts in PRO. Subsequently, MeSH associations are aligned with semantic triples from the *integrated KG* to identify a subset of implicit biomedical associations for KGC. Identified associations are represented as incomplete (*subject, ?, object*) triples, whereby the *subject* and *object* are pre-existing entities in the KG and the goal is to predict the missing relation between them – i.e., *predicate* – to create complete (*subject, predicate, object*) triples to augment the baseline and *integrated KG*.

We pursued KGC via relation prediction by leveraging the KG embeddings and a scoring function $f\ \{t = (s, p, o)\}: \mathcal{E}, \mathcal{R}, \mathcal{E} \rightarrow \mathbb{R}$, such that when given an input incomplete triple (*s, ?, o*), all possible combinations of (*s, p, o*) triples are assigned a score proportional to the likelihood that the completed triple is true. The missing relation in the input triple is replaced with all existing *predicates* in the KG, and a score is assigned to each combination of (*s, p, o*), with high scores indicating more plausible predictions. For each incomplete input triple, we retain top *k* predicted relations and add them to the *integrated KG* for knowledge discovery tasks.

To quantify the plausibility of relation predictions, we use the following KGE-specific scoring functions.

TransE (Bordes et al., 2013) employs a distance-based scoring function as:

$$f(s, p, o) = -||s + p - o||_{L1/L2}$$

where $s, p, o$ are vectors representing the *subject*, *predicate*, and *object* respectively, and $L1/L2$ represent L1 and L2 norms respectively. Explicitly, TransE uses an additive scoring function that considers the sum of *subject* and *predicate* vectors to be approximately equal to the *object* vector for true positive triples – i.e., $s + p \approx o$.

DistMult (B. Yang et al., 2015) restricts relations as diagonal matrices to reduce the number of relation parameters. The scoring function is defined as the bi-linear dot product of the *subject, predicate,* and *object* vectors:

$$f(s, p, o) = \langle s, w_p, o \rangle$$

where $s$ and $o$ are vectors representing the *subject* and *object* entities respectively, and $w_p$ represents the *predicate* diagonal matrix.

Similarly, ComplEx (Trouillon et al., 2016) restricts relation embeddings to be diagonal matrices, but uses complex valued vectors for entities and relations. The scoring function is defined as the Hermitian dot product of the *subject, predicate,* and *object* complex vectors:

$$f(s, p, o) = Re\big(\langle s, w_p, o \rangle\big)$$

where $s, w_p, o \in \mathbb{C}^k$ and $Re(x)$ represents the real part of vector $x$.

## 4.6 Knowledge Discovery, Filtering, and Ranking:

In this section, we describe our methods for uncovering novel and meaningful discovery paths from the *supplemented KG* and, subsequently, filter and rank discoveries to create condensed sub-graphs which can be explored and visualized by domain experts. Our methods are not restricted to any particular discovery model or paradigm, and can be adapted for use with ABC, AnC, discovery patterns, and open or closed-based discovery. Nevertheless, we employ the *ABC* model as an example to demonstrate our methods.

Our methods can be described as a multi-step process that integrates the *ABC* discovery model, filtering of non-meaningful discovery paths, and the final ranking of novel discoveries.

In the first step, a single or multiple user-defined source concepts (i.e., *A*) are identified to initialize an exhaustive 2-hop KG traversal to retrieve *ABC* discovery paths. This step involves utilizing Neo4j's Cypher query language to define a matching pattern that defines

the KG traversal logic (Francis et al., 2018). Specifically, we use the Cypher query presented in Code 4.2 to initialize the KG traversal.

```
MATCH PATH = (A {name:concept name})-[R1]->(B)-[R2]->(C)
WHERE A <> C AND NOT (A)-[]-(C)
RETURN A.identifier, B.identifier, C.identifier, type(R1) AS
SemRel1, type(R2) AS SemRel2
```

*Code 4. 2: Cypher query to initiate ABC discovery*

In the above query, the KG traversal begins with a node *A* which has a specified name (represented by *concept name*) and finds all its directly related nodes (i.e., intermediate *B* nodes) through semantic relations denoted as *R1*. Alternatively, the *concept name* attribute can be replaced by *concept semantic group* or *concept semantic type* to retrieve candidate discovery paths. Then the query traverses the KG to identify all target nodes (i.e., *C*) directly related to the intermediate *B* nodes through the relation *R2*. The query then filters out any paths where the source node is the same as the target node (i.e., self-directed loops). Additionally, the query ensures that nodes A and C are not linked by a semantic relation, since one of the core assumptions of the *ABC* discovery model is that the source and target concepts are not directly related nor co-occur in the literature. Finally, the query returns paths characterized by the following variables: identifiers of the source (A), intermediate (B), and target (C) nodes, as well as the types of semantic relationships that connect them. In the next step, the preliminary discovery paths are processed by a filtering algorithm to generate the discovery subgraph. The filtering algorithm applies multiple criteria at different levels to filter *ABC* discovery paths, including concept specificity at the concept level, *A-B* and *B-C* counts at the semantic triple level, and the well-established Linking Term Counts (LTC) at the path level:

- **Concept level filtering** leverages hierarchical structures of biomedical vocabularies to define specificity of *A*, *B*, and *C* concepts (i.e., nodes) by calculating the distance from root within the vocabulary where the concept is represented. We leverage MeSH, Gene Ontology (GO), and Protein Ontology (PRO) vocabularies which organize concepts in tree-like hierarchical structures, linked by *IS_A* relations, whereby generic concepts are located near the top-level concept (root) and more specific concepts are located further down the hierarchy. As such, given a biomedical concept, the distance from root is calculated by measuring the number

of *IS_A* hierarchical relations until the top-level concept in the hierarchy (i.e., root) is reached. Intuitively, the greater the distance from root, the more specific the concept and the higher the specificity score. Concept specificity is a commonly used metric to filter and rank knowledge extracted from curated ontologies and biomedical text (Gopalakrishnan et al., 2018). In this framework, a concept is considered specific if its score is greater than a certain threshold *s*. We explore several concept specificity thresholds (*s* = 3, 4, 5, and 6).

- **Semantic triple level filtering** uses corpus-based semantic triple counts as a filtering criteria to eliminate triples which occur less than a prespecified threshold. The rationale is that triples that occur infrequently in the literature can be considered noisy and do not accurately represent literature-based knowledge. Hence, by eliminating such triples, the remaining triples that also occur infrequently but more often than the eliminated ones represent meaningful and relevant triples for LBD. We explore several thresholds for semantic triple count 3, 5, 10, and 15.
- **Path level filtering** leverages LTC which is a well-established metric in LBD that quantifies the strength of indirect associations based on the number of linking B concepts, with higher linking concepts indicating stronger indirect associations between the source (*A*) and the target (*C*) (Yetisgen-Yildiz & Pratt, 2006). LTC provides the benefit of eliminating spurious paths which have few overlapping intermediate *B* concepts. A discovery path is considered to have a strong association if the LTC score is greater than a certain threshold *t*. In this framework, we explore several thresholds of *t* = 3, 5, 10, and 15.

After the filtering is applied, the next step is concerned with ranking the retained set of discovery paths using Information Content (IC) as a metric to prioritize novel and interesting discoveries. IC is calculated as the negative log of the probability of encountering an event *x*:

$$IC(x) = -log_2\, p(x)$$

In this context, an event *x* can be defined as an *ABC* discovery path, and the probability of encountering a path can be calculated based on frequencies derived from the literature (i.e., corpus-based information content) (McInnes & Pedersen, 2015). However, IC-based metrics are designed to be biased towards low probability events. Therefore, to account for

such biases, we propose two variants of IC-based ranking metrics for LBD that also leverage LTC as a correction factor. Explicitly, we propose the $IC_{sum}$ and $IC_{path}$ ranking metrics.

$IC_{sum}$ calculates the IC of *A*-*B* and *B*-*C* triples individually, and assigns the final score to the *ABC* path by summing $IC(AB)$ and $IC(BC)$ and multiplying the IC scores by LTC of the target discovery path. Formally, $IC_{sum}$ is defined as follows:

$$IC_{sum}(ABC) = \left(-log_2\left(\frac{X_{AB}}{N}\right)\right) + \left(-log_2\left(\frac{X_{BC}}{N}\right)\right) \times LTC$$

Where $X_{AB}$ and $X_{BC}$ are the frequencies of *AB* and *BC* triples respectively, *N* is the sum of all triple frequencies in the discovery subgraph, and $LTC$ is the correction factor.

$IC_{path}$ calculates the IC of *ABC* discovery paths by determining the probability of encountering the path in the discovery subgraph. Formally, $IC_{path}$ is defined as follows:

$$IC_{path}(ABC) = -log_2\left(\frac{X_{ABC}}{N}\right) \times LTC$$

Where $X_{ABC}$ is the frequency of the *ABC* path, *N* is the sum of all path frequencies in the discovery subgraph, and $LTC$ is the correction factor.

We posit that the proposed IC-based metrics can distinguish interesting and novel discovery paths from generic ones, while also accounting for indirect *A*-*C* associations via the $LTC$ correction factor. To evaluate the effectiveness of these metrics, we compare IC-based rankings with co-occurrence frequency-based ranking and traditional association-based metrics: Log-likelihood ratio (LLR), Pearson's Chi-square ($X^2$), and Odds Ratio (OR), which are commonly used in LBD for ranking purposes (Henry & McInnes, 2019; Zhang et al., 2021). We briefly describe the baseline metrics here:

- LLR is an expectation-based statistical measure that calculates the degree to which the observed frequency values deviate from the expected values. A high LLR score indicates a concept pair is less likely to have occurred together by chance and, therefore, have strong association. We calculate LLR for *A*-*B* and *B*-*C* individually, and path scores are derived by summing the LLR scores. We utilize the Text::NSP package to calculate LLR scores (Pedersen et al., 2011).
- $X^2$ is another commonly used expectation-based measure that reflects the deviation of observed frequencies from expected frequencies. We calculate the path score by

summing the $X^2$ score for *A-B* and *B-C*. A higher score indicates that the *A-B* or *B-C* associations are more likely to be dependent, and there is less evidence in favor of the hypothesis that they are independent. The Text::NSP package is used to compute $X^2$scores (Pedersen et al., 2011).

- OR is a statistical measure of association that computes the ratio of how often concepts co-occur together divided by the total number of null co-occurrences. We calculate the path score by summing the OR score for *A-B* and *B-C*. A higher score indicates that the path consists of strongly associated concepts. The Text::NSP package is used to compute $X^2$scores (Pedersen et al., 2011).
- Co-occurrence frequency is a rudimentary ranking metric that ranks *ABC* discovery paths by summing the total number of *A-B* and *B-C* co-occurrences.

## 4.7 Summary:

This chapter introduced the methods applied in the AKG-LBD framework to enhance and improve the process of semantic-based LBD by proposing novel solutions addressing the challenges described previously (Chapter 3). This chapter is structured into 6 sections, with each section corresponding to a component in the AKG-LBD framework.

In the first section (4.1), we described our methods for acquiring biomedical literature from PubMed using a comprehensive pre-defined search query aimed at cancer literature. Specifically, we leveraged Medical Subject Heading (MeSH) descriptors, which serve as indexes to PubMed articles, and specified keywords in titles and abstracts. We introduced a query formulation algorithm using fuzzy string matching and Biopython's Entrez package to automatically formulate PubMed queries. Overall, we curated 23 terms related to cancers and their molecular and cellular hallmarks (i.e., hallmarks of cancer) and used them as input for query formulation. The output of the literature curation component consisted of titles, abstracts, and corresponding PMIDs, which were formatted as plain text to be used as input for the next component.

In section 4.2, we described the semantic-based knowledge extraction component which leverages SemRep as a semantic parser that extracts knowledge in the form of *subject-predicate-object* triples. Further, this component integrates PubTator to address limitations of SemRep in disambiguation of *Gene* and *Protein* concepts. This is a novel approach in

our framework, as it combines the output of two established biomedical knowledge extraction tools to ensure the accuracy and representativeness of literature-based knowledge.

Section 4.3 introduced the semantic consolidation component, which aims to unify the representation of fine-grained concepts into atomic concept representations. We posit that this step is necessary to (i) reduce the discovery search space without compromising the semantics of concepts; and (ii) ensure semantic compatibility with external knowledge resources and ontologies to facilitate the integration of curated biomedical knowledge.

The following section (4.4) described the representation of semantic-based knowledge (i.e., *subject-predicate-object* triples) as a large-scale graph, thereby generating the first iteration of the literature-based KG (i.e., *baseline KG*). We utilized Neo4j due to its scalability in storing large-scale graphs and its interactive visual tools which can facilitate exploration of literature-based knowledge.

Section 4.5 introduced our novel methods for progressively augmenting the baseline literature-based KG to address the problem of incomplete extraction of knowledge from the literature corpus. This component entails a two-step process: (i) leveraging curated knowledge extracted from the Gene Ontology and the Comparative Toxicogenomics Database; and (ii) using Knowledge Graph Completion (KGC) methods to predict missing semantic relations between implicitly associated *subject* and *object* concepts. The first step (i.e., knowledge integration) results in generating the *integrated KG*, which augments literature-based knowledge with high quality knowledge curated by expert biocurators. We posit that this step alone is insufficient in augmenting the *baseline KG*, as manual curation is a time-consuming process and, therefore, lags behind the current state of knowledge. Consequently, the KGC approach aims to fill this gap by predicting missing semantic relations using information-rich Knowledge Graph Embeddings (KGEs). We explore three KGE models to encode the *integrated KG* nodes and relations as vectors: TransE, DistMult, and ComplEx. As literature-based KGs contain a diverse range of relations, we selected these models for their ability to represent various relation types, including one-to-one, one-to-many, symmetrical, and asymmetrical relations. The best performing model is selected based on an evaluation methodology focusing on relation prediction. Explicitly, we adapt time-slicing and random-slicing techniques to split the *integrated KG* into training and

evaluation sets. KGE models are trained on the training set and evaluated by predicting relations in the evaluation set. The best performing model will be used for the final task of informed relation prediction to further augment the *integrated KG* with missing semantic relations. This task generates the *augmented KG*, which serves as input for knowledge discovery.

Lastly, section 4.6 describes the knowledge discovery, filtering, and ranking methods. We propose an integrated methodology that adapts to various discovery models, including *ABC*, *AnC*, discovery patterns, and closed- and open-based discovery. However, this framework utilizes the *ABC* model, as it is the most widely employed model in LBD research. The knowledge filtering algorithm is applied to different levels of *ABC* discovery paths, including concept specificity at the concept level, frequency counts at the triple level, and LTC at the path level. For ranking, we adapt IC-based metrics to propose two LBD ranking variants which also account for inherent biases by leveraging LTC as a correction factor.

# Chapter 5 Experimental Results

This chapter presents our results for the methods implemented within AKG-LBD framework. We present evaluation of AKG-LBD to replicate real-world cancer discoveries and for repurposing drugs. The main aim of the evaluation experiments is to measure the performance of AKG-LBD for knowledge discovery tasks using literature at PubMed pertaining to cancers.

The chapter is structured as follows. **Section 5.1** reports on the results of the baseline knowledge extraction from the literature and semantic consolidation of concepts. **Section 5.2** describes the characteristics and topology of the *baseline KG*. **Section 5.3** describes the results of integrating curated knowledge from external sources to create the *integrated KG*. **Section 5.4** presents the results of KGE evaluations and determining the best performing model for KGC. **Section 5.5** reports on the outcomes of KGC via relation prediction and the subsequent construction of the *augmented KG*. **Section 5.6** presents the results of knowledge discovery, filtering, and ranking for replicating real-world cancer discoveries and repurposing drugs for new cancer indications. **Section 5.7** presents a comparison of the performance of formalized LBD systems against the AKG-LBD framework. Finally, **section 5.8** summarizes the key findings of this chapter.

## 5.1 Literature-Based Knowledge Extraction and Semantic Consolidation:

Biomedical articles retrieved from PubMed constitute the literature corpus used for semantic-based knowledge extraction and subsequent construction of the baseline literature-based KG. We retrieved 5,531,702 articles covering cancers using the PubMed query formulation methodology described in Chapter 4 (section 4.1). The following sections describe the results of semantic knowledge extraction, semantic consolidation, and construction of the *baseline KG*, which serves as the baseline for LBD.

### 5.1.1 Extraction of Semantic-Based Knowledge:

Using the literature-based corpus as input to SemRep, we retrieved 38,941,970 semantic triples. The output of SemRep consisted of semantic triples (*subject-predicate-object)*

along with the semantic types of the *subject* and *object* concepts, the PMID of the article where the triple was extracted, and the corresponding sentence that the triple was derived from. Recall that *subject* and *object* concepts are represented as UMLS or NCBI Gene concepts through unique identifiers (i.e., CUIs or NCBI Gene identifiers), while *predicates* are ontological relations derived from the UMLS Semantic Network. The semantic triples were subsequently filtered based on the concept semantic types and type of *predicate* (i.e., semantic relation) to eliminate concepts and relations that are deemed uninformative for the downstream discovery task. As a result 34,262,163 triples were eliminated and 4,679,807 triples were retained, consisting of 88,909 unique *subjects*, 67,904 unique *objects*, and 13 unique *predicates* (i.e., semantic relations).

We identified a total of 20,301 ambiguous gene/protein mappings, for which we retrieved the corresponding PMIDs that were used as input into PubTator to disambiguate the ambiguous gene/protein concepts. The output of this process resulted in disambiguating 72.7% (14,759 concepts) of the ambiguous gene/protein concepts, whereas we eliminated the remaining ambiguous concepts as they were not useful for knowledge discovery tasks. After merging all the disambiguated concepts with SemRep's semantic triples, the final set of semantic triples comprised of 102,750 unique concepts, represented by UMLS CUIs or NCBI Gene identifiers, classified as 34 semantic types. We aggregated concepts into four high-level semantic groups (*Chemicals and Drugs, Disorders, Genes,* and *Physiology*) based on the semantic type classifications, for example, concepts classified as *Disease or syndrome*, *Neoplastic process*, and *Finding* are aggregated into the *Disorders* semantic group. Figure 5.1 shows the distribution of concepts extracted from the literature classified by semantic type and semantic groups.

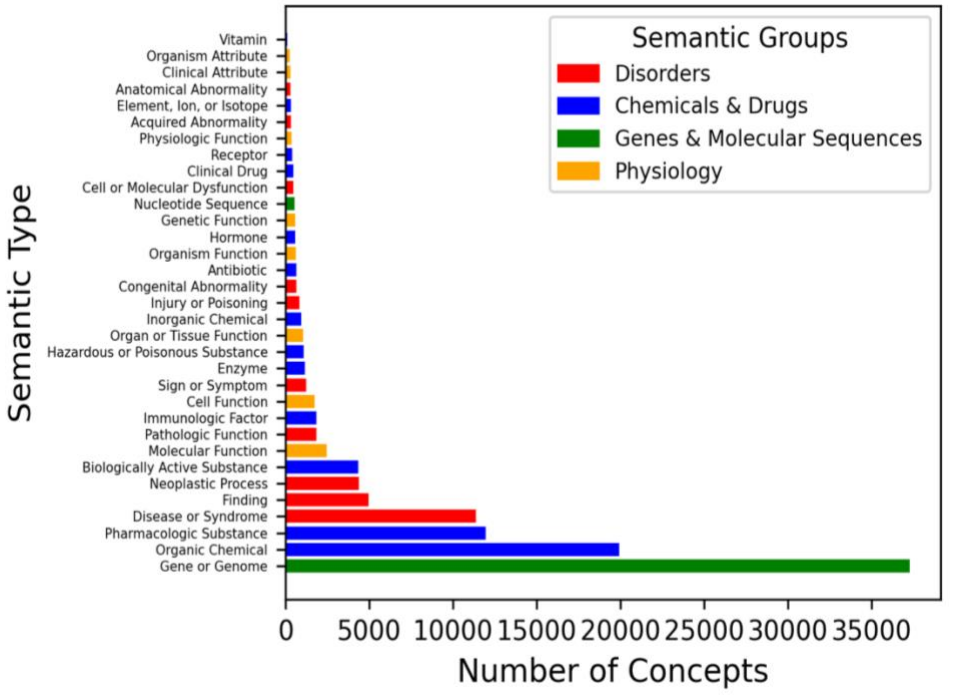


*Figure 5.1: Distribution of concepts by semantic type and high-level semantic groups*

### 5.1.2 Semantic Consolidation of Concepts:

Semantic consolidation merged granular concepts into high-level, generalized representations by leveraging biomedical terminologies, such as MeSH, Gene Ontology (GO), and Protein Ontology (PRO).

Table 5.1 shows the number of unique concepts per semantic group before and after semantic consolidation. Overall, semantic consolidation resulted in reducing the number of unique concepts by 51.78% (i.e., from 102,750 to 49,506 concepts). Concepts in the *Genes* group were impacted the most from semantic consolidation, whereby the average number of concepts subsumed by a PRO concept was 3.3. Concepts in the *Disorders* group also had a significant decrease in the number of concepts, with an average of 2.9 concepts subsumed by a MeSH concept. However, we did not observe notable consolidation of concepts in the *Chemicals and Drugs* and *Physiology* semantic groups, whereby the average number of UMLS concepts subsumed by MeSH and GO concepts was 1.3 and 1.1, respectively. These results highlight the highly granular nature of the UMLS as a knowledge resource for representing biomedical concepts in literature. The semantic consolidation task led to the merging of multiple, granular UMLS concepts into higher-level concepts in MeSH, GO, and PRO vocabularies.

*Table 5.1: Number of unique concepts before and after semantic consolidation*

| Semantic Group | Number of unique concepts before consolidation | Number of unique concepts after consolidation |
|---|---|---|
| *Chemicals and Drugs* | 34,423 | 25,286 |
| *Disorders* | 25,261 | 5,717 |
| *Genes* | 37,781 | 14,496 |
| *Physiology* | 5,285 | 4,011 |

To assess the impact of the semantic consolidation on its coverage of the domain being investigated (i.e. cancers), we examined the representation of cancer-related concepts within different semantic groups (i.e., *chemicals and drugs*, *disorders*, *physiology*, and *genes*) before and after consolidation. We used the Comparative Toxicogenomics Database (CTD) as a reference to evaluate the domain coverage, as it encompasses curated concepts covering cancers. Using CTD, we extracted concepts that represent cancer sub-types (e.g., breast cancer, colon cancer, etc.) and concepts that have known associations with the development or treatment of cancers (i.e., chemicals, genes, and physiological phenotypes). The concepts extracted from CTD – referred to as domain concepts – were aggregated into four semantic groups: *disorders*, *chemicals and drugs*, *physiology,* and *genes*. Subsequently, we evaluated the domain coverage in each semantic group by calculating the overlap between the literature-based concepts before/after consolidation and domain concepts extracted from CTD.

Table 5.2 presents the extent of the domain coverage before and after semantic consolidation across four semantic groups. We note an increase in domain coverage across all semantic groups after semantic consolidation, with the most increase in the *Genes* group (33.9% increase), followed by *Chemicals and Drugs* (4.4% increase), *Physiology* (4% increase) and *Disorders* (2% increase). These results partially validate that the semantic consolidation task did not compromise the coverage of biomedical domains across the four semantic groups, rather it improved the overall proportion of represented concepts from all the target vocabularies. This task will facilitate the integration of curated knowledge from

knowledge bases that use the same vocabularies for representing biomedical concepts and will also assist in reducing the discovery search space in LBD.

*Table 5.2: Domain coverage before and after semantic consolidation*

| **Semantic Group** | **Domain coverage before semantic consolidation (%)** | **Domain coverage after semantic consolidation (%)** |
|---|---|---|
| *Chemicals and Drugs* | 74.7% | 79.1% |
| *Disorders* | 95.8% | 97.8% |
| *Genes* | 52.4% | 86.3% |
| *Physiology* | 19.1% | 23.1% |

## 5.2 Baseline Literature-Based KG Construction:

Semantic triples (*subject-predicate-object*) were used to generate the *baseline KG* using Neo4j—a node was created for each unique *subject* or *object* concept, and directed relations between nodes were created based on the *predicates* extracted from the literature. We assigned properties to nodes signifying the concept's preferred vocabulary name, semantic type, and semantic group. Likewise, the relations in the KG were assigned properties indicating their predicate type (i.e., relation type) and relation weight, which was determined by the frequency of occurrence of the relation type between a given subject and object concept.

Table 5.3 shows that the *baseline KG* comprises over 49,000 nodes and over 1 million relations. However, the average degree centrality confirms that connections between nodes are sparse, as each node is connected to a small fraction of other nodes in the KG. This behaviour is further supported by the low graph density of 0.0006. Despite its large size, these results indicate that the *baseline KG* is incomplete due to missing relations, which further emphasizes the need to augment the KG.

*Table 5.3: Characteristics of the baseline KG*

| **Parameter** | |
|---|---|
| Nodes | 49,506 |
| Relations | 1,541,035 |
| Average Degree Centrality | 0.001 |
| Graph Density | 0.0006 |

## 5.3 Integrated KG Construction:

We augmented the *baseline KG* with missing knowledge manually curated biomedical knowledge from the Gene Ontology (GO), and Comparative Toxicogenomics Database (CTD), where we acquired 442,295 instances of curated knowledge in the form of *subject-predicate-object* triples relating to chemical-disease, gene-disease, gene-biological process, and gene-molecular function associations. To avoid duplicate triples, we eliminated approximately 5% of curated triples that overlapped with the *baseline KG*. This modest overlap in knowledge is another indication that biomedical knowledge extracted from a single source is inherently incomplete, which further underlines the need to integrate knowledge from multiple heterogeneous sources. Table 5.4 shows the breakdown of the remaining 422,233 instances of curated knowledge.

*Table 5.4: Curated knowledge extracted from biomedical KBs*

| **Curated Associations** | **Number of Acquired Triples** | **Source** |
|---|---|---|
| *Chemical-Disease* | 6,089 | CTD |
| *Gene-Disease* | 4,451 | CTD |
| *Gene-Biological Process* | 91,835 | GO |
| *Gene-Molecular Function* | 144,675 | GO |

The largest source of curated knowledge was acquired from GO, which provided a total of 236,510. CTD provided a total of 10,540 triples. This distribution is not surprising, as the

*baseline KG* had limited coverage of the *Physiology* domain, which largely comprises *molecular function* and *biological process* concepts.

Curated triples acquired from biomedical knowledge bases were merged with the *baseline KG* to generate the *integrated KG*, which now consisted of 1,788,085 unique triples, resulting in a 16% increase in the number of triples compared to the *baseline KG*. Figure 5.2 shows the distribution of the nodes and relations in the *integrated KG*. We note that knowledge integration contributed to a significant number of new nodes (i.e., concepts), particularly within the *Physiology* semantic group. With respect to relations, the integration of curated knowledge contributed to 13% of new relations in the *integrated KG*, most of which were *STIMULATES* and *AFFECTS* relation types.

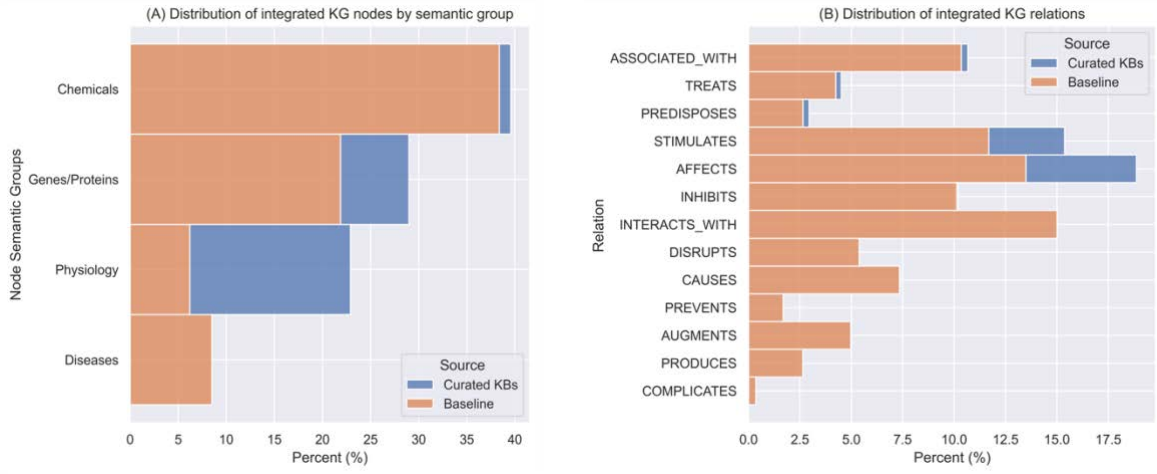


*Figure 5.2: Distribution of nodes and relation in integrated KG*

The *integrated KG* serves as input for the subsequent augmentation task which entails predicting relations between pre-existing nodes via KGC methods.

## 5.4 Evaluation of KG Embeddings for KGC:

This section presents evaluation of KG Embedding (KGE) models applied to achieve an optimal KGC performance. Using the *integrated KG* as input, we investigated three KGE models: TransE, DistMult, and ComplEx. Two forms of evaluations were conducted: a visual analysis of KGEs to assess the quality of the embeddings in distinguishing between various node types (i.e., semantic groups) and relation types, and an assessment of KGE models for KGC tasks through relation prediction.

### 5.4.1 Visual Analysis of KGEs:

We applied the Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction method to visually assess the quality of KGEs. UMAP maps node and relation embeddings into 2-dimensional spaces such that similar embeddings are placed into nearby spaces, and as such it can be used to visualize clusters within the data to uncover features captured by the KGE models. For example, we can assess how well nodes within a semantic group are clustered (local feature) and how similar semantic group clusters collocate with one another (global feature). We applied UMAP to node relation embeddings generated by all three KGE models. It may be noted that all reported results are based on tuned models, which have been trained on a training set and tuned accordingly on a validation set.

Figure 5.3 shows the 2D UMAP visualizations of node embeddings based on the different KGE models. Compared to visualizations of TransE and ComplEx, DistMult embeddings show distinct clusters of nodes based on high-level semantic groups (i.e., disease, genes/proteins, chemicals, and physiology) with relatively clear group separation, thus indicating that DistMult is capable of distinguishing between multiple types of KG nodes. This can be attributed to DistMult being a semantic matching model which exploits similarity-based functions to embed entities and relations in a KG. Furthermore, DistMult uses a multiplicative interaction function to measure the plausibility of KG triples by matching the latent semantics of entities and relations. Hence, DistMult ensures that entities and relations with similar semantics tend to have similar embeddings.

Visualization of ComplEx node embeddings shows well-defined clusters for Genes/Proteins and Physiology groups, however chemical and disease node clusters are less distinguishable compared to DistMult due to overlaps with other semantic groups.

Visualization of TransE node embeddings shows the formation of two large groups of overlapping clusters, indicating that TransE does not adequately capture the KG's local features since nodes within the same semantic groups tend to appear in separate clusters.

Overall, our visualization results confirm that the embeddings by DistMult and ComplEx capture the KG's local and global features—i.e. nodes within the same semantic groups form well-demarcated clusters, and similar semantic groups tend to collocate (e.g., chemicals and genes/proteins).

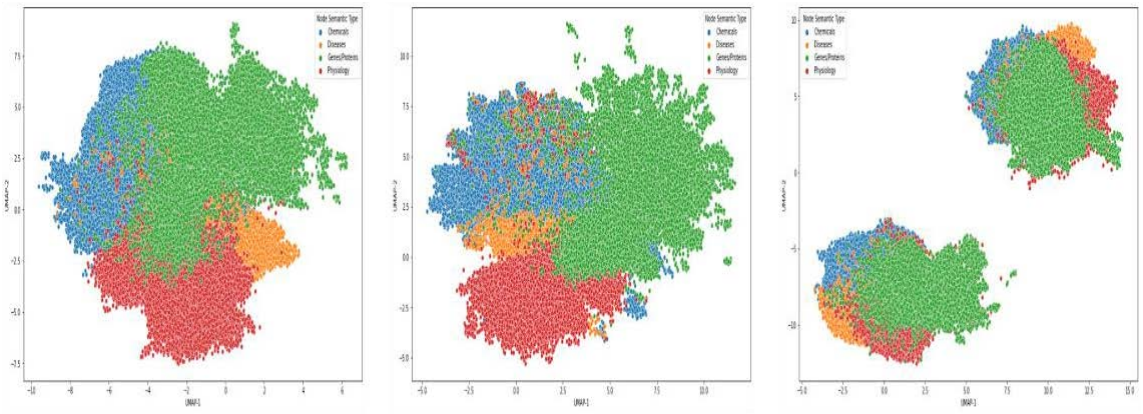

*Figure 5.3: Illustration of 2-dimensional UMAP plots for KG nodes based on different embeddings. Green represents Genes/Proteins, yellow represents Diseases, blue represent Chemicals, and red represent Physiology (i.e., GO) (From left to right: DistMult, ComplEx,*

Figure 5.4 presents 2D UMAP visualizations of relation embeddings based on the three KGE models. Visualizations of DistMult and ComplEx relation embeddings indicate that semantically similar relations are projected into nearby spaces. For example, relations denoting therapeutic (i.e., *PREVENT*S, *TREATS*), substance interactions (i.e., *INTERACTS_WITH*, *INHIBITS*, *STIMULATES*), disease etiology (i.e., *PREDISPOSES, ASSOCIATED_WITH*), and pharmacogenomics (i.e., *AFFECTS, DISRUPTS, AUGMENTS*) are projected into adjacent spaces in the DistMult and ComplEx scatterplots. This demonstrates that DistMult and ComplEx are capable of classifying semantically related relations based on their biomedical functions – i.e., pharmacogenomics, substance interaction, etc.

The visualization of TransE relation embeddings show less meaningful clusters of relations, and relations denoting therapeutic associations are further apart compared to their counterparts in DistMult and ComplEx relation embeddings. Our visualization results confirm that DistMult and ComplEx can capture relational semantics of the KG.

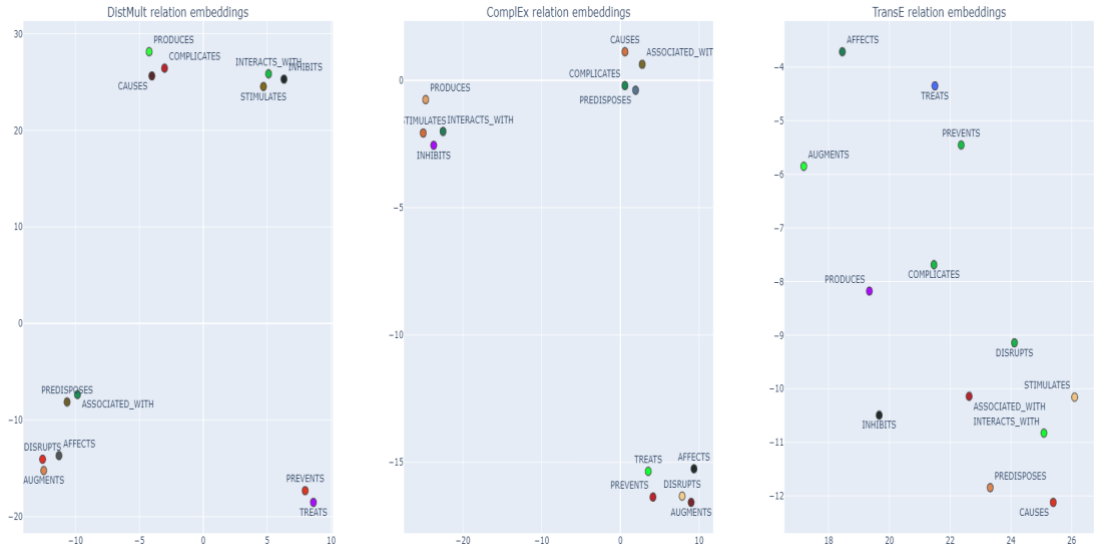


*Figure 5.4: UMAP visualization of relation embeddings*

### 5.4.2 Relation Prediction Evaluation:

We evaluated the performance of KGE models for relation prediction in terms of KGC using random- and time-slicing evaluation settings.

In the random-slicing setting, 20% of triples were randomly selected and the relations between the subject and object concepts were eliminated to form the evaluation set, while the remaining 80% of triples were retained as the training set. Since, a random split can result in train-to-test leakage, which can inflate the relation prediction results, we avoided data leakage on random split by ensuring that the subject and object concepts constituting the evaluation set are not linked by a predicate (i.e., relation) in the training set. This strategy ensured that the evaluation data is not seen by models at training time. The triples in the training set were used to train KGE models, and the output embeddings were used to predict the a priori eliminated relations in the evaluation set. Training set consisted of 1,615,467 triples and the evaluation set consisted of 403,869 incomplete triples (i.e., subject, ?, object). Figure 5.5 compares the distribution of relations in the evaluation and training sets, indicating that the evaluation set is a representative sample of the *integrated KG* relations.

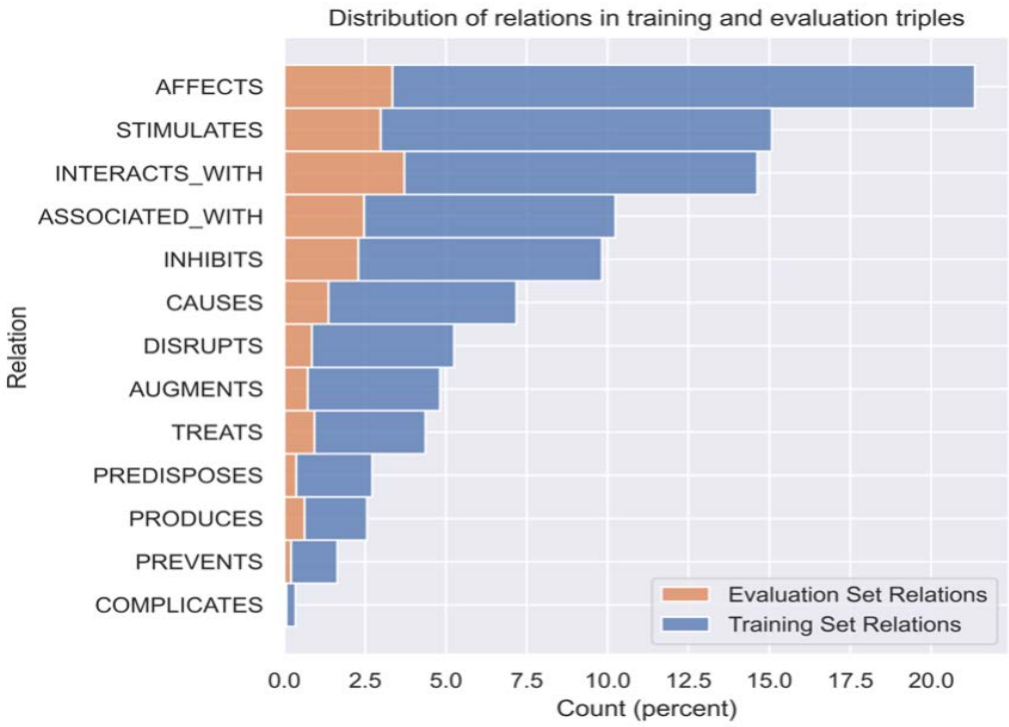


*Figure 5.5: Distribution of relations in training and evaluation triples*

For time-slicing based evaluation, we set the cutoff-date to 2015 to form the training set (pre-cutoff-date) and evaluation set (post-cutoff-date). The target was to predict relations in the evaluation set – i.e., relations that have formed after the 2015 cutoff date. The relation prediction results for TransE, DistMult and ComplEx, for both experimental settings, are presented in Table 5.5.

*Table 5.5: Relation prediction results*

| ***Model*** | **Random-slicing** | | | | | **Time-slicing** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Hits@1* | *Hits@3* | *Hits@5* | *Hits@10* | *MRR* | *Hits@1* | *Hits@3* | *Hits@5* | *Hits@10* | *MRR* |
| TransE | 0 | 0.344 | 0.509 | 0.703 | 0.234 | 0 | 0.354 | 0.519 | 0.663 | 0.232 |
| ComplEx | 0.303 | 0.701 | 0.870 | 0.981 | 0.530 | 0.302 | 0.669 | 0.818 | 0.958 | 0.515 |
| DistMult | **0.511** | **0.863** | **0.957** | **0.998** | **0.692** | **0.308** | **0.689** | **0.848** | **0.968** | **0.528** |

The evaluation results show that DistMult outperforms ComplEx and TransE across both experimental settings for relation prediction, achieving higher Hits@k and MRR scores in the random-slicing setting, thus suggesting that DistMult can effectively handle KGC tasks. In the time-slicing experiment, the difference between DistMult and ComplEx is negligible

as both models achieve similar Hits@k and MRR. The results show that TransE does not adequately capture the knowledge within the *integrated KG* as it fails to predict missing relations when Hits@k = 1. It is worth noting that literature-based KGs contain numerous one-to-many and many-to-many relations, as scientific findings tend to progress with time, hence subject and object concepts can be linked by various semantic relations depending on the biological context. This may explain the low performance of TransE, as it is not capable of embedding one-to-many and many-to-many relations, unlike DistMult and ComplEx.

To determine whether the differences in predictive performance are statistically significant, we used the Wilcoxon signed rank test (two-tailed) to compare the MRR scores of the KGE models. The Wilcoxon signed rank test is a non-parametric test used to compare model prediction performance which does not require the assumption of data normality. As per this test, if the p-value is lower than or equal to 0.05, the difference in predictive performance, based on MRR scores, is considered statistically significant. Table 5.6 shows the results of the Wilcoxon signed rank test based on the MRR scores of the following combinations of models: TransE-ComplEx, TransE-DistMult, and ComplEx-DistMult. The test suggests that there is a statistically significant difference in the performance of DistMult and ComplEx compared to TransE (p-values < 0.05) based on the MRR scores obtained from the random-slicing and time-slicing evaluations. Additionally, we observe statistically significant difference in the predictive performance between ComplEx and DistMult for random-slicing evaluations.

*Table 5.6: Results of the Wilcoxon signed-ranked test*

| **Dataset** | **P-value (MRR)** | | |
|---|---|---|---|
| | ***TransE-ComplEx*** | ***TransE-DistMult*** | ***ComplEx-DistMult*** |
| Random-slicing | < 0.001 | < 0.001 | < 0.001 |
| Time-slicing | < 0.001 | < 0.001 | 0.163 |

Our results confirm that DistMult significantly outperforms ComplEx in predicting missing relations from an incomplete KG. In the time-slicing evaluation, there was no significant difference between ComplEx and DistMult, since the p-value is greater than the significance level of 0.05, thus suggesting that the performance of ComplEx and DisMult are comparable in predicting future relations in a KG.

## 5.5 Augmented KG Construction via KGC:

As DistMult was noted to be the best performing model for KGC based on the evaluation results reported in section 5.4. it was used to train on the *integrated KG* to generate the embeddings required for KGC. The dataset of incomplete triples (i.e., *subject, ?, object*) was created using MeSH descriptors in the literature corpus—we ensured that every *subject* and *object* concept has a corresponding embedding. Table 5.7 shows the representation and number of incomplete triples used as input for relation prediction based on the semantic group of subject and object concepts

*Table 5.7: Representation and number of incomplete input triples for relation prediction*

| **Type of incomplete input triples** | **Number of triples** |
|---|---|
| Chemical, ?, Gene/Protein | 146,706 |
| Chemical, ?, Disease | 111,775 |
| Physiology, ?, Disease | 11,461 |
| Gene/Protein, ?, Gene/Protein | 831,297 |
| Gene/Protein, ?, Physiology | 134,821 |
| **Total** | 1,236,060 |

The KGC input dataset consisted of 1,236,060 incomplete (*subject, ?, object*) triples, with the goal of predicting the missing relational element (i.e., predicate). For every incomplete triple, we retained the top scoring 3 relation prediction—i.e. DistMult at Hits@3. This

resulted in a set of complete (*subject, predicate, object*) triples, whereby the *subject* and *object* represent pre-existing KG nodes and the predicate was a previously missing semantic relation. Completed triples were added to the *integrated KG* to generate the *augmented KG* which serves as the final resource for LBD.

The *augmented KG* consists of 3,024,145 unique semantic triples. Figure 5.6 shows the distribution of semantic relations in the *augmented KG* (which also includes relations from the *baseline* and *integrated KG*). We observe a notable increase in relations across all relation types as a result of the KGC task. The most commonly predicted relations were *INTERACTS_WITH*, *INHIBITS*, *STIMULATES* and *PRODUCES* given the high number of *gene-gene*, *chemical-gene* associations.

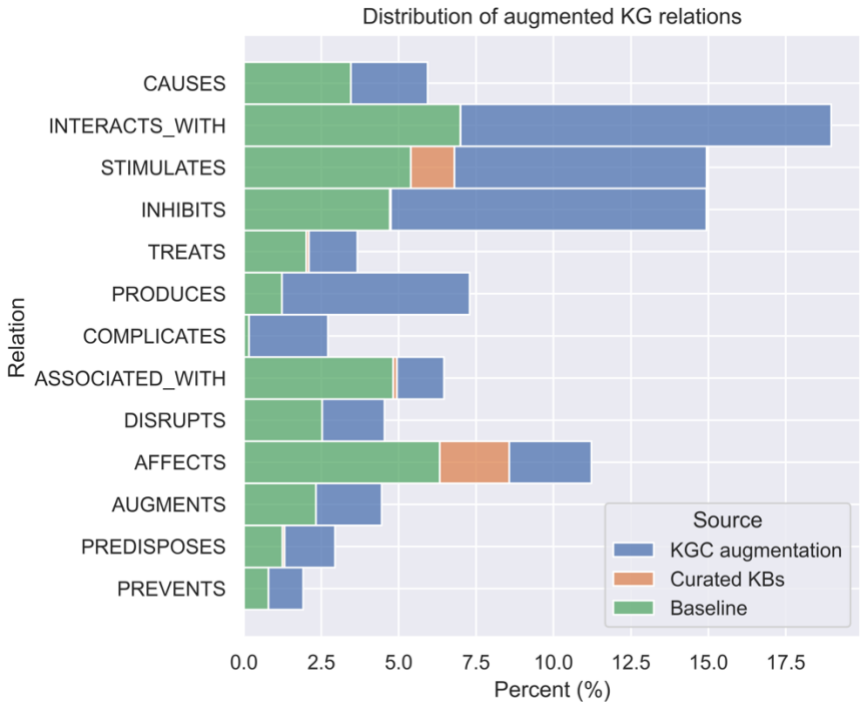


*Figure 5.6: Distribution of relations in augmented KG*

To demonstrate how the KG topology evolves throughout the multi-step knowledge completion process, we present a comparison between the *baseline, integrated*, and *augmented* KGs in Table 5.8. The comparison shows a significant increase in the number of nodes and relations in the *augmented KG* compared to the *baseline KG*. However, it is important to note that KG augmentation via KGC did not result in adding new nodes to the KG, as the task was focused on predicting relations. The density parameter indicates that the *augmented KG* is denser than both the *baseline KG* and the *integrated KG*, with a 16.7%

increase in density from the *baseline KG* to the *augmented KG*. This may indicate that the *augmented KG* is still relatively incomplete, as is the case for almost all biomedical KGs. However, we posit that a dense KG is not always beneficial, as traversing a very dense KG is significantly more costly than sparse KGs.

*Table 5.8: Comparison of the baseline, integrated, and augmented KGs*

| Parameter | Baseline KG | Integrated KG | Augmented KG |
|---|---|---|---|
| Nodes | 49,506 | 63,844 | 63,844 |
| Relations | 1,541,035 | 1,788,085 | 3,024,145 |
| Density | 0.0006 | 0.0004 | 0.0007 |

## 5.6 Literature-Based Discovery (LBD) Tasks:

In this section, we report the results of LBD task using framework for: (i) replicating cancer discoveries reported in the literature; and (ii) repurposing existing drugs for new cancer indications. We will compare the results of the different variations of our framework, each incorporating a specific KG, thus demonstrating the efficacy of our implemented methods.

### 5.6.1 Cancer Discoveries:

Cancer discoveries reported in peer-reviewed literature were replicated using the open-based *ABC* discovery model, which was applied to a time-sliced version of the baseline, integrated, and augmented KGs. Cancer discoveries were replicated using time-sliced KGs and open-based discovery to generate candidate *ABC* discovery paths. We define a replicated discovery path as any path where the source (A), intermediate (B), and target (C) concepts match the targeted discovery shown Table 5.9.

To perform LBD based discovery paths, the KGs were split at a specific point in time, corresponding to when the discovery was reported in the literature, to generate pre-cutoff KGs that serves as the literature-based KG for LBD tasks. Next, the pre-cutoff KGs were traversed to identify *ABC* discovery paths that describe an indirect relationship between a predefined source node (*A*) and a target node (*C*) through an intermediate node (*B*).

*Table 5.9: Cancer test case discoveries*

| Date of Discovery | Concept A | Concept B | Concept C |
|---|---|---|---|
| 2011 | **Name:** NRF2<br>**ID:** PR_000011170<br>**Semantic Group:** Gene | **Name**: Reactive Oxygen Species (ROS)<br>**ID**: D017382<br>**Semantic Group**: Chemical | **Name**: Pancreatic Cancer<br>**ID**: D010190<br>**Semantic Group**: Disorders |
| 2015 | **Name**: IL-17<br>**ID**: PR_000001138<br>**Semantic Group**: Gene | **Name**: p38α<br>**ID**: PR_000003107<br>**Semantic Group**: Gene | **Name**: MKP-1<br>**ID**: PR_000006736<br>**Semantic Group**: Gene |
| 2016 | **Name**: NOTCH1<br>**ID**: PR_000011331<br>**Semantic Group**: Gene | **Name**: Cellular Senescence<br>**ID**: GO_0090398<br>**Semantic Group**: Physiology | **Name**: CEBPB<br>**ID**: PR_000005308<br>**Semantic Group**: Gene |
| | **Name**: NFKB1<br>**ID**: PR_000001754<br>**Semantic Group**: Gene | **Name**: BCL2<br>**ID**: PR_000002307<br>**Semantic Group**: Gene | **Name**: Adenoma<br>**ID**: D000236<br>**Semantic Group**: Disorders |
| 2017 | **Name**: CXCL12<br>**ID**: PR_000006066<br>**Semantic Group**: Gene | **Name**: Cellular Senescence<br>**ID**: GO_0090398<br>**Semantic Group**: Physiology | **Name**: Thyroid Cancer<br>**ID**: D013964<br>**Semantic Group**: Disorders |

#### 5.6.1.1 Replication of Cancer Discoveries:

Table 5.10 presents the aggregated results of replicating the cancer discovery test cases using time-sliced versions of the baseline, integrated, and *augmented KG*s. The results indicate that the *baseline KG*s enabled the replication of only one discovery path (20% recall), the *integrated KGs* enabled the replication of two discovery paths (40% recall), and the *integrated KGs* replicated all 5 cancer discovery paths (100% recall).

The lack of replicated discoveries by the baseline and *integrated KGs* is due to the absence of numerous relations between the source, intermediate, and target concepts as a result of

incomplete knowledge extraction and/or the lack of explicit knowledge in curated knowledge bases. The improved performance of the *augmented KG* confirms the efficacy of our KG completion methods in predicting meaningful semantic relations between implicitly associated concepts. It is worth noting that the *integrated KG*, which was supplemented with curated biomedical knowledge, did not yield significant benefits to the discovery process. This further emphasizes the need for novel methods, such as KGC, to automatically infer missing knowledge in biomedical KGs.

*Table 5.10: Cancer discovery test case replication*

| **Time-sliced KG** | **Average number of semantic triples** | **Average number of ABC discovery paths** | **Recall of cancer test case discoveries (%)** |
|---|---|---|---|
| Baseline | 930,986 | 16,280 | 20% |
| Integrated | 1,133,906 | 16,441 | 40% |
| Augmented | 1,210,652 | 66,564 | 100% |

Table 5.11 presents the detailed statistics of all time-sliced KGs and the number of replicated discoveries per KG. Considering the KG sizes, we observe a significant expansion in the number of semantic triples within the *augmented KG*s compared to the integrated and *baseline KG*s. On average, the *augmented KG*s contain approximately 280,000 more semantic triples compared to the *baseline KG*s, and 76,746 more semantic triples compared to the *integrated KGs*. The increase in semantic triples is also reflected in the number of ABC discovery paths generated from the *augmented KG*s, which far surpasses the baseline and *integrated KGs*. This further underlines the significant number of knowledge instances However, we observe minimal increase in *ABC* discovery paths from the baseline to the *integrated KGs*, indicating that the curated knowledge extracted from biomedical KBs did not contribute to many knowledge instances related to the source (A) or intermediate (B) concepts. This further underlines the significant amount of relevant knowledge instances generated by KGC methods to augment the baseline and *integrated KGs*.

*Table 5.11: Replication of cancer discovery paths*

| Time-slice date | KG | Number of unique triples (KG size) | Number of Candidate *ABC* Paths | Number of Replicated Discovery Path(s) |
|---|---|---|---|---|
| 2011 | Baseline | 707,970 | 12,525 | 1 |
| | Integrated | 836,166 | 12,569 | 1 |
| | Augmented | 975,328 | 50,854 | 1 |
| 2015 | Baseline | 938,032 | 1,704 | 0 |
| | Integrated | 1,159,040 | 1,743 | 0 |
| | Augmented | 1,211,458 | 21,185 | 1 |
| 2016 | Baseline | 1,004,131 | 33,633 | 0 |
| | Integrated | 1,232,402 | 34,070 | 1 |
| | Augmented | 1,302,252 | 101,475 | 2 |
| 2017 | Baseline | 1,073,810 | 17,258 | 0 |
| | Integrated | 1,308,014 | 17,380 | 0 |
| | Augmented | 1,353,568 | 113,928 | 1 |

*Table 5. 1: Replication of cancer discovery paths*

The full set of replicated discovery paths along with their semantic relations are depicted in Figure 5.7. Semantic relations acquired by SemRep from the literature are denoted as (SemRep Relations), relations acquired from biomedical KBs are denoted as (Knowledgebase Relations), and relations predicted using KGC are denoted as (Predicted Relations). We note that, as multi-relational KGs, the discovery paths included multiple relations between the pair of A-B or B-C concepts, this is typical of literature-based KGs as scientific findings tend to progress with time or change based on the research context.

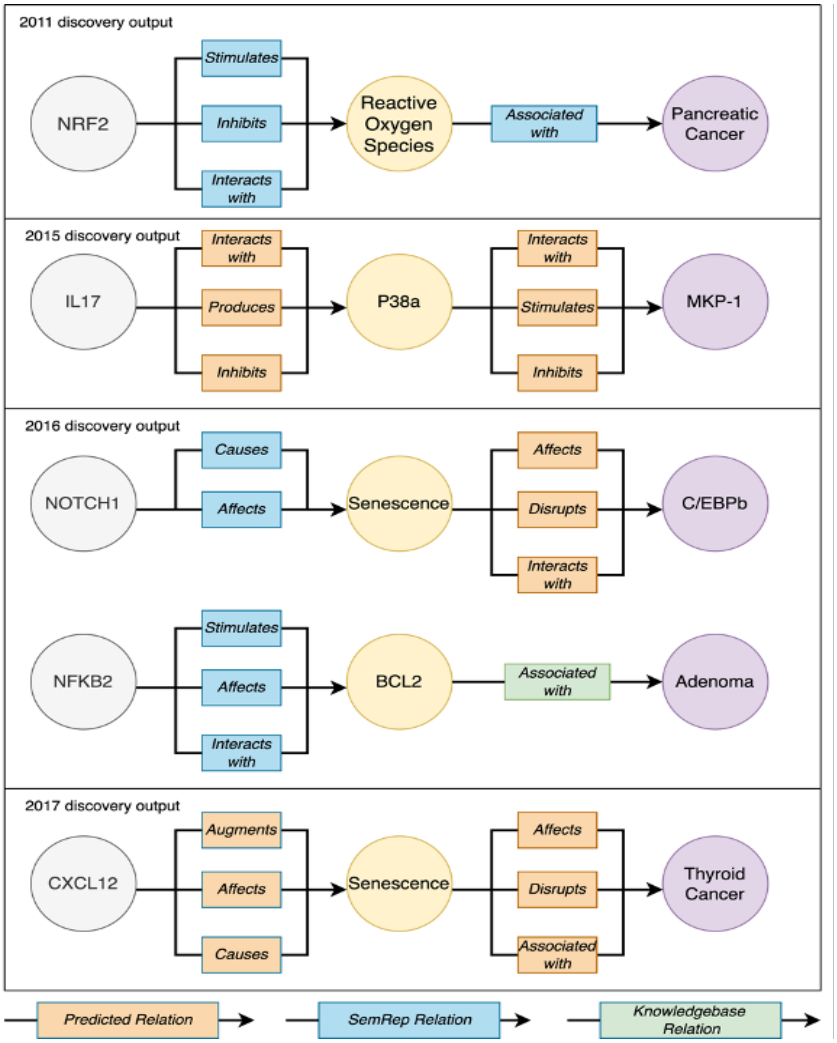


*Figure 5.7: Replicated cancer discovery paths*

We evaluated the quality of LBD outputs by cross-validating them against the original articles from which the discoveries were established. We manually screened full-text articles to identify the mechanistic relations between the biomedical concepts constituting the discoveries and cross-validated them with the retrieved discovery paths. The following subsections detail the validation of our LBD outputs by the *augmented KG* based on the reported studies in the literature.

**NRF2, ROS and pancreatic Cancer:**

According to DeNicola et al., reactive oxygen species (ROS) are highly reactive carcinogenic chemicals. Physiological stress can result in increasing ROS levels which ultimately may result in cellular and DNA damage (DeNicola et al., 2011). Under normal physiological conditions, levels of ROS are tightly controlled by the NRF2-induced

detoxification program to maintain a balanced intracellular environment. However, the authors note that in neoplastic environments NRF2 is activated suggesting that enhanced activation of NRF2 may lead to pro-carcinogenic effects. Hence, DiNicola et al. investigated the metabolic environment in pancreatic cancer cells and discovered that NRF2-induced ROS detoxification can indeed result in promoting carcinogenesis (DeNicola et al., 2011). This phenomenon is attributed to enhanced NRF2 activity which results in maintaining a balanced intracellular environment, thereby allowing the cancerous cells to grow while preventing cellular death.

Our LBD output in Figure 5.7 shows multiple distinct paths involving NRF2, ROS and pancreatic cancer. Interestingly, the semantic relations between NRF2 and ROS include inhibits, stimulates, and interacts_with. We examined the source sentences from which these triples were extracted, and found that high expression of NRF2 inhibits ROS (Li et al., 2009), while low expression of NRF2 stimulates ROS production (Singh et al., 2010). The Semantic relation between ROS and pancreatic cancer suggests a known association between them, however, this association was not in the context of an already existing cancer – rather, the association describes how high levels of ROS can promote carcinogenesis. When considering the full discovery path, it can be inferred that high levels of NRF2 can lower ROS production which is known to be associated with pancreatic cancer.

**IL17, p38a, and MKP-1:**

Gaffen et al. investigate the effect of IL-17 mediated signaling on p38a and MKP-1 in nonimmune cells (Gaffen & McGeachy, 2015). The authors suggest that the binding of IL-17 to its receptor causes the activation of p38a, which in turn activates MKP-1. The LBD output path corresponding to this discovery shows that IL-17 in linked to p38a via three predicted semantic relations, and that p38a is linked to MKP-1 also via three predicted semantic relations. The top scoring relation prediction between IL-17 and p38a was interacts_with, followed by produces, and inhibits. Similarly, the top scoring relation prediction between p38a and MKP-1 was interacts_with, followed by stimulates, and inhibits. Hence, if we consider the top scoring predictions, the LBD discovery path (IL-17 – INTERACTS_WITH – p38a – INTERACTS_WITH – MKP-1) suggests a generalized interaction between the gene products.

**NOTCH1, cellular senescence, and C/EBPb:**

A study by Hoare et al. revealed that fluctuations in NOTCH1 activity affects cellular senescence by activating TGF-B while repressing C/EBPb activity (Hoare et al., 2016). The authors suggest that at high NOTCH1 levels, cells undergo a phase of growth, while at low levels cellular senescence replaces the growth phase which is accompanied by inhibition of C/EBPb activity. Our LBD output corresponding to this discovery implies a similar hypothesis, whereby one of the paths states that (NOTCH1 – CAUSES – Senescence – DISRUPTS – C/EBPb). We note that relations between senescence and C/EBPb were predicted, as no relations were found in the literature-based semantic triples nor in the curated knowledge from biomedical KBs.

**NFKB2, BCL2, and Adenoma:**

Van der Heijden et al. investigated the role of NFKB2 and its target BCL2 in intestinal adenomas (van der Heijden et al., 2016). The authors discovered that NFKB signaling promoted the expression of anti-apoptotic factor BCL2, which in turn is associated with the outgrowth of adenomas, since high expression of BCL2 mediates the resistance of cancerous cells by preventing cellular death (i.e., apoptosis). Our LBD output corresponding to this discovery shows several discovery paths involving NFKB2, BCL2 and adenomas. Interestingly, one of the paths indeed confirms this hypothesis, which states that (NFKB2 – STIMULATES – BCL2 – ASSOCIATED_WITH – Adenomas). The other paths also illustrate some notion of substance interaction between NFKB2 and BCL2 (e.g., affects, interacts_with), however, these SemRep-based semantic relations do not contextualize the type of interaction. We found one semantic relation between BCL2 and adenomas (i.e., associated_with), which was acquired from biomedical KBs. While the association between BCL2 and various cancers, including adenomas, are well-reported in the literature, the fact that this relation was not identified by SemRep further underlines its limitations in knowledge extraction.

**CXCL12, cellular senescence, and thyroid cancer:**

CXCL12 chemokines have an important role in cell migration and cancer metastasis (Wang et al., 2006). According to Kim et al., CXCL12 chemokines are also implicated in increasing the survival of senescent cells in thyroid cancer, and ultimately promoting carcinogenesis (Kim et al., 2017). Our LBD output corresponding to this discovery

partially confirms the hypothesis that CXCL12 promotes cellular senescence, which is also known to be associated with thyroid cancer; (CXCL12 – AUGMENTS/CAUSES – Senescence – ASSOCIATED_WITH/AFFECTS – Thyroid cancer). Objectively, the source sentences from which the semantic relations between cellular senescence and thyroid cancer were extracted did not explicitly mention that senescent cells promote carcinogenesis, rather, nearly all the literature-based sources contextualize the relation between senescence and thyroid cancer as a cellular mechanism to prevent carcinogenesis.

#### 5.6.1.2 Filtering and Ranking ABC Discovery Paths to Prioritize Valid Cancer Discoveries:

LBD tends to generate numerous discovery paths which makes the task of manually reviewing and validating the output impractical. Hence, it is necessary for LBD frameworks to provide a ranked list of candidate knowledge discoveries to (a) eliminate noisy and knowledge; and (b) prioritize interesting and meaningful knowledge. This section presents our results for filtering and ranking the cancer discovery paths.

Table 5.13 presents the results of the filtering and ranking of the ABC discovery paths using the open-based discovery paradigm. The following parameters and values were explored to control the filtering process: concept specificity score (2, 3, 4, 5), Linking Term Count (LTC) (3, 5, 10, 15), and corpus-based co-occurrence frequency (3, 5, 10, 15). The values shown in Table 5.12 represent the optimal parameter values for filtering and ranking the discovery paths using two variants of the IC-based ranking metric.

The results suggest that a concept specificity score of 4 to 5 and an LTC of 10 to 15 are the most effective filtering parameters for targeted discoveries. However, determining the optimal value for the frequency parameter is inconclusive as values ranged from 5 to 15.

With respect to ranking metrics, our results indicate that $IC_{path}$ performed slightly better in ranking targeted discoveries, with an average RR of 0.82 compared to 0.79 for the $IC_{sum}$ metric. However, the difference in rankings between $IC_{path}$ and $IC_{sum}$ across all 5 discovery tasks were not statistically significant, with a p-value greater than the significance level based on the Wilcoxon signed-rank test. This suggests that the performance of $IC_{path}$ and $IC_{sum}$ are comparable in targeted LBD tasks.

*Table 5.12: Filtering and ranking of cancer discovery paths*

| Discovery Path | Filtering Parameter values | Ranking Metric | Relative Rank (RR) |
|---|---|---|---|
| NRF2 - ROS - Pancreatic cancer | Specificity > 5<br>LTC > 15<br>Triple Count > 15 | $IC_{path}$ | 0.91 |
| | | $IC_{sum}$ | 0.83 |
| IL17 - p38a - MKP-1 | Specificity > 5<br>LTC > 10<br>Triple Count > 5 | $IC_{path}$ | 0.61 |
| | | $IC_{sum}$ | 0.56 |
| NOTCH1 - Senescence - C/EBPb | Specificity > 4<br>LTC > 15<br>Triple Count > 5 | $IC_{path}$ | 0.66 |
| | | $IC_{sum}$ | 0.65 |
| NFKB2 - BCL2 - Adenoma | Specificity > 5<br>LTC > 15<br>Triple Count > 5 | $IC_{path}$ | 0.95 |
| | | $IC_{sum}$ | 0.95 |
| CXCL12 – Senescence – Thyroid cancer | Specificity > 4<br>LTC > 15<br>Triple Count > 10 | $IC_{path}$ | 0.98 |
| | | $IC_{sum}$ | 0.98 |

Table 5.13 presents the ranking results compared to baseline metrics. The performance of ranking metrics vary across the five targeted discovery tasks, which is expected since each task is applied to a unique literature-based KG using time-slicing techniques. Overall, the $IC_{path}$ resulted in an average Relative Rank (RR) of 0.82 compared to 0.65 for Odds Ratio (OR), 0.59 for Pearson's Chi-square ($X^2$), 0.48 for Log-Likelihood Ratio (LLR), and 0.17 for the Co-Occurrence Frequency (COF) metric. We observe that the $IC_{path}$ metric achieves the highest RR for 3 out of the 5 valid discovery paths, while the OR metric outperforms $IC_{path}$ in ranking 2 valid discovery paths. However, we note that $IC_{path}$ consistently achieves an RR greater than 0.6, ranging from 0.61 to 0.98. In contrast, the OR metric exhibits a wider range of RR between 0.11 and 0.98, indicating inconsistent prioritization of valid discovery paths. This can be attributed to the inherent nature of OR

as an association-based metric, which is typically influenced by the total number of null co-occurrences between the discovery path entities.

*Table 5.13: Comparison of IC-based metrics with baseline for cancer discovery path ranking*

| Discovery Path | Ranking Metric | Relative Rank (RR) |
|---|---|---|
| NRF2 - ROS - Pancreatic cancer | $IC_{path}$ | 0.91 |
| | LLR | 0.68 |
| | $X^2$ | 0.64 |
| | OR | 0.57 |
| | COF | 0.45 |
| IL17 - p38a - MKP-1 | $IC_{path}$ | 0.61 |
| | LLR | 0.56 |
| | $X^2$ | 0.89 |
| | OR | 0.96 |
| | COF | 0.02 |
| NOTCH1 - Senescence - C/EBPb | $IC_{path}$ | 0.66 |
| | LLR | 0.25 |
| | $X^2$ | 0.39 |
| | OR | 0.64 |
| | COF | 0.09 |
| NFKB2 - BCL2 - Adenoma | $IC_{path}$ | 0.95 |
| | LLR | 0.78 |

| Discovery Path | Ranking Metric | Relative Rank (RR) |
|---|---|---|
| NFKB2 - BCL2 - Adenoma | $X^2$ | 0.91 |
| | OR | 0.98 |
| | COF | 0.28 |
| CXCL12 – Senescence – Thyroid cancer | $IC_{path}$ | 0.98 |
| | LLR | 0.14 |
| | $X^2$ | 0.1 |
| | OR | 0.11 |
| | COF | 0.02 |

### 5.6.2 Drug Repurposing:

In this section, we present the outcomes of repurposing pre-existing drugs for new cancer indications using LBD on a time-sliced KG with a 2015 cutoff date. The evaluation of the association between repurposed drugs and cancer indications is based on knowledge extracted from literature published after the 2015 cutoff date. We discuss the outcomes of drug repurposing discovery in section 5.6.2.1, followed by the results of ranking discovery paths in section 5.6.2.2.

#### 5.6.2.1 Drug Repurposing Discovery:

Table 5.14 compares the performance of the discovery task on the *baseline*, *integrated*, and *augmented KG*s in terms of the number of valid discovery paths, number of repurposed drugs, and the recall of repurposed drugs relative to the silver standard. The results demonstrate that the *augmented KG* outperforms both the *baseline* and *integrated KGs* in the drug repurposing task, with a recall of 71.1% of all *Drug-Cancer* associations from the silver standard compared to a recall of 48.5% in the *baseline KG* and 48.9% *integrated KG*. These results support our hypothesis that relying solely on the integration of curated knowledge does not solve the problem of incompleteness in literature-based KGs as the

*integrated* and *baseline KG*s recall a similar number of *Drug-Cancer* associations. Further, the results indicate that the *augmented KG* generates a greater number of valid discovery paths, indicating that there were more valid intermediate genes linking a source drug and a target disease. This is a positive outcome as it implies that the *augmented KG* can generate more reliable hypotheses that can be further investigated in clinical studies.

*Table 5.14: Results of the drug repurposing discovery task without knowledge filtration*

| KG | Total Number of Discovery Paths | Number of Valid Discovery Paths | Recall of Repurposed Drugs |
|---|---|---|---|
| Baseline | 34,134 | 1,233 | 48.5% |
| Integrated | 34,587 | 1,247 | 48.9% |
| Augmented | 104,527 | 2,712 | 71.1% |

Despite the positive results, nearly 29% of drugs in the silver standard were not repurposed for new cancer indications using the *augmented KG*, which can be attributed to the strict path validity conditions. These conditions required established associations in curated knowledge bases between the source, intermediate, and target. However, if we hypothetically consider that external knowledge bases are incomplete and lag behind the current state of knowledge, relaxing the conditions for path validity to include discovery paths where only one of the source-intermediate or intermediate-target association is valid could lead to more drug repurposing. Table 5.15 presents the results of drug repurposing using relaxed path validity conditions. The number of valid paths and repurposed drugs increased across all KGs, with the *augmented KG* recalling 93.6% of all *Drug-Cancer* associations in the silver standard, while the *baseline* and *integrated KG* recalling 68.7% of associations in the silver standard. Further, we note a significant increase in the number of valid discovery paths, which means more hypotheses are generated by the relaxed validation. The relaxed conditions for path validity, where either the A-B or B-C association should be valid, are likely more practical for LBD for two reasons. Firstly, LBD is often used as a method for hypothesis generation, which means that not all source-intermediate and intermediate-target associations need to be established in curated KBs beforehand. Secondly, our validation approach was based on a single source of curated

knowledge (i.e., CTD), and it is possible that a drug-gene or gene-cancer association is established in other sources but not captured in CTD.

*Table 5.15: Results of drug repurposing discovery task using relaxed validation conditions and without knowledge filtering*

| KG | Total Number of Discovery Paths | Number of Valid Discovery Paths | Recall Repurposed Drugs |
|---|---|---|---|
| Baseline | 34,134 | 2,468 | 68.7% |
| Integrated | 34,587 | 2,496 | 68.7% |
| Augmented | 104,527 | 6,615 | 93.6% |

#### 5.6.2.2 Filtering and Ranking Drug Repurposing Discovery Paths:

Table 5.16 presents the results (i.e., valid discovery paths and recall) obtained from non-filtered discovery paths. To establish the most effective filtering thresholds for the discovery task, we examined the effects of various thresholds of concept specificity (3, 5), LTC (3, 5, 10), and co-occurrence frequency (3, 5, 10) on recall of valid discovery paths and repurposed drugs from the *augmented KG* using the relaxed validation conditions, as shown in Table 5.16. Expectedly, using lower filtering thresholds increases the recall of repurposed drugs and the number of valid paths. The combination of filtering thresholds with the lowest values yields 5,790 valid discovery paths (i.e., 87.5% of all possible valid paths) and 83.8% recall of all *Drug-Disease* associations in the silver standard. In contrast, the highest filtering thresholds resulted in the elimination of over 90,000 discovery paths, although it only generated a small fraction (16.3%) of all possible valid paths and merely 15.9% of all *Drug-Cancer* associations in the silver standard.

Using Pearson's correlation coefficient ($r$), we observe a very strong negative correlation between the *Triple Count* parameter and *Valid Discovery Paths* ($r$ = -0.91, p-value < 0.05), a weak negative correlation between *Specificity* and *Valid Discovery Paths* ($r$ =-0.37, p-value < 0.05), and a very weak negative correlation between *LTC* and *Valid Discovery Paths* ($r$ =-0.08, p-value < 0.05). Analysis of correlation between filtering parameters and recall of repurposed drugs indicates a strong negative correlation for *LTC* ($r$ = -0.65, p-value < 0.05) and *Triple Count* ($r$ = -0.61, p-value < 0.05), and a moderate negative correlation for *Specificity* ($r$ = -0.41, p-value < 0.05). Lastly, correlation of *Total Discovery Paths* (i.e., size of discovery subgraph) with filtering parameters indicates a very strong

correlation with *Triple Count* ($r$ = -0.94, p-value < 0.05), and a weak correlation with *Specificity* ($r$ = -0.27, p-value < 0.05) and *LTC* ($r$ = -0.17, p-value < 0.05). These results suggest that adjusting filtering parameters can impact the LBD process in terms of the discovery subgraph size, number of valid/meaningful discovery paths, and recall of discoveries. Additionally, the correlation of *Triple Count* with the subgraph characteristics, in terms of its size (i.e., *Total Discovery Paths*) and meaningful knowledge (i.e., *Valid Discovery Paths*), is more significant than the other parameters.

*Table 5.16: Results of drug repurposing discovery using different filtering thresholds and relaxed validation conditions*

| Filtering Parameters | | | Total Discovery Paths | Valid Discovery Paths | Recall of Repurposed Drugs |
|---|---|---|---|---|---|
| Specificity | LTC | Triple Count | | | |
| 3 | 3 | 3 | 92,572 | 5,790 | 83.8% |
| 3 | 3 | 5 | 76,340 | 4,908 | 79.6% |
| 3 | 3 | 10 | 26,876 | 2,081 | 58.4% |
| 3 | 5 | 3 | 89,254 | 5,800 | 75.6% |
| 3 | 5 | 5 | 72,491 | 4,937 | 66.8% |
| 3 | 5 | 10 | 23,859 | 1,900 | 41.1% |
| 3 | 10 | 3 | 79,950 | 5,568 | 54.1% |
| 3 | 10 | 5 | 63,112 | 4,741 | 48.0% |
| 3 | 10 | 10 | 17,979 | 1,474 | 22.0% |
| 5 | 3 | 3 | 71,329 | 4,252 | 63.1% |
| 5 | 3 | 5 | 58,548 | 3,451 | 59.7% |
| 5 | 3 | 10 | 20,920 | 1,534 | 44.6% |
| 5 | 5 | 3 | 68,568 | 4,277 | 56.5% |
| 5 | 5 | 5 | 55,674 | 3,500 | 49.6% |
| 5 | 5 | 10 | 18,638 | 1,381 | 30.2% |
| 5 | 10 | 3 | 61,301 | 4,018 | 39.3% |
| 5 | 10 | 5 | 48,274 | 3,351 | 34.2% |
| 5 | 10 | 10 | 14,010 | 1,061 | 15.9% |

Generally, the objective of an effective filtering approach for LBD is to yield a discovery subgraph which is small enough to facilitate exploration and review of generated discoveries, while retaining meaningful knowledge which constitute novel discoveries. Based on this notion and the analysis of the results in Table 5.16, we posit that filtering parameters *Specificity* = 3, *LTC* = 3, and *Triple Count* = 10 generate a condensed discovery subgraph consisting of 26,876 discovery paths in total while retaining most of the drug

repurposing discoveries (recall = 58.4%). Furthermore, if we consider querying the discovery subgraph on individual source nodes/concepts in the silver standard, the average number of discovery paths per query is reduced to 450 paths, which can be easily explored particularly when presented in a ranked order.

Table 5.17 presents the results of ranking discovery paths generated by the *augmented KG* based on the following filtering parameters *Specificity* = 3, *LTC* = 3, and *Triple Count* = 10. We compare the performance of IC-based ranking metrics $IC_{path}$ and $IC_{sum}$ against baselines: log-likelihood ratio (*LLR*), Chi-square ($X^2$), Odds Ratio (*OR*), and co-occurrence frequency (*COF*). With respect to precision, the results show that IC-based metrics outperform the baseline metrics at K = 10, 30, 50, and 100. We observe a substantial difference in the average precision between IC-based metrics and baselines. Further, the mean average precision (mAP), which measures the average precision at each valid discovery path across all source concepts/nodes, shows that IC-based metrics are effectively the better performing metrics compared to all baselines, including the traditional co-occurrence frequency.

*Table 5.17: Performance of IC-based ranking metrics compared to baseline ranking metrics using on the following filtering thresholds: specificity = 3, LTC = 3, triple count = 10*

| **Metric** | **Average Precision @ K (AP@K)** | | | | **Average Recall @ K (AR@K)** | | | | **mAP** | **RR** |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***K=10*** | ***K=30*** | ***K=50*** | ***K=100*** | ***K=10*** | ***K=30*** | ***K=50*** | ***K=100*** | | |
| $IC_{path}$ | 0.312 | 0.319 | 0.308 | 0.284 | 0.049 | 0.146 | 0.224 | 0.377 | 0.627 | 0.585 |
| $IC_{sum}$ | 0.319 | 0.332 | 0.317 | 0.291 | 0.047 | 0.141 | 0.212 | 0.358 | 0.630 | 0.587 |
| LLR | 0.232 | 0.214 | 0.218 | 0.214 | 0.045 | 0.114 | 0.178 | 0.306 | 0.465 | 0.507 |
| $X^2$ | 0.199 | 0.178 | 0.187 | 0.197 | 0.041 | 0.098 | 0.157 | 0.353 | 0.443 | 0.498 |
| OR | 0.196 | 0.173 | 0.165 | 0.167 | 0.038 | 0.093 | 0.153 | 0.276 | 0.423 | 0.478 |
| COF | 0.265 | 0.258 | 0.266 | 0.233 | 0.047 | 0.142 | 0.216 | 0.363 | 0.519 | 0.541 |

When comparing the IC-based metrics to one another, the results suggest that both metrics have similar precision when ranking drug repurposing discovery paths. This observation was also reflected in the targeted cancer discoveries, which further underlines that both IC-based metrics are well-suited for ranking LBD output.

In terms of recall, the results demonstrate that IC-based metrics consistently outperform

association-based metrics. On average, the IC-based rankings contained a higher proportion of valid discovery paths in the top 10, 20, 50, and 100 ranks compared to association-based rankings. Specifically, the top 100 ranks (i.e., K = 100), the average recall is 0.36 for $IC_{sum}$ and 0.38 for $IC_{path}$, whereas the average recall for association-based metrics ranges from 0.28 to 0.31.

We also analysed the performance of ranking metrics using low filtering parameter thresholds, as shown in table 5.18, to demonstrate that IC-based rankings are also capable of prioritizing valid discoveries in situations where the discovery sub-graph is large. Specifically, we chose the following thresholds for filtering: *Specificity* = 3, *LTC* = 3, and *Triple Count* = 3, which result in a discovery sub-graph consisting of 92,572 ABC discovery paths. The results indicate that IC-based metrics are still the better performing ranking metrics compared to baselines in terms of precision, recall, and the average relative rank of valid discovery paths. This analysis demonstrates that $IC_{path}$ and $IC_{sum}$ are scalable to large discovery sub-graph sizes and, therefore, can adapt to a wide range of LBD tasks.

*Table 5.18: Performance of IC-based ranking metrics compared to baseline ranking metrics using the following filtering thresholds: specificity = 3, LTC = 3, triple count = 3*

| **Metric** | **Average Precision @ K (AP@K)** | | | | **Average Recall @ K (AR@K)** | | | | **mAP** | **RR** |
|---|---|---|---|---|---|---|---|---|---|---|
| | ***K=10*** | ***K=30*** | ***K=50*** | ***K=100*** | ***K=10*** | ***K=30*** | ***K=50*** | ***K=100*** | | |
| $IC_{path}$ | 0.292 | 0.293 | 0.295 | 0.274 | 0.026 | 0.097 | 0.154 | 0.265 | 0.478 | 0.601 |
| $IC_{sum}$ | 0.304 | 0.294 | 0.292 | 0.251 | 0.024 | 0.083 | 0.136 | 0.232 | 0.479 | 0.603 |
| LLR | 0.212 | 0.177 | 0.176 | 0.181 | 0.026 | 0.054 | 0.078 | 0.144 | 0.358 | 0.503 |
| $X^2$ | 0.198 | 0.179 | 0.149 | 0.139 | 0.019 | 0.046 | 0.068 | 0.129 | 0.339 | 0.494 |
| OR | 0.117 | 0.112 | 0.131 | 0.129 | 0.013 | 0.035 | 0.059 | 0.119 | 0.322 | 0.476 |
| COF | 0.216 | 0.223 | 0.218 | 0.197 | 0.028 | 0.077 | 0.119 | 0.210 | 0.407 | 0.536 |

## 5.7 Comparison of AKG-LBD Output with Existing LBD Systems:

We compared the performance of AKG-LBD with other existing LBD systems, namely Arrowsmith (co-occurrence based), BITOLA (co-occurrence based), SemBT (semantic-

based), MELODI-PRESTO (semantic-based), and LION-LBD (co-occurrence based). We used the cancer discovery test cases for comparison purposes, targeting the replication of valid discovery paths (the ranking of paths by the selected LBD systems is not considered in this analysis, as not all LBD systems provide a full ranked list of discovery outputs.). A discovery path is considered replicated if the intermediate (B) concept matches the valid discovery paths in Table 5.10. We applied closed-based discovery by specifying the source and target concepts or terms. It is worth noting that the cancer discovery cases used in this analysis are adapted from the LION-LBD study, which were replicated entirely by the system (Pyysalo et al., 2019). However, we include the output of the LION-LBD system in this section for completeness.

We will to point out that the drug repurposing test cases were not used in this analysis due to (a) the size of the silver standard dataset, which consists of over 300 Drug-Cancer associations, (b) the selected LBD systems do not support the execution of discovery tasks on multiple source and target concepts simultaneously, and (c) validating the intermediate concept will be difficult, as there are multiple valid intermediates for each pair of source and target concepts.

Table 5.19 presents the results of replicating the cancer discovery test cases using the selected LBD systems when compared to AKG-LBD. A performance comparison illustrates that AKG-LBD and LION-LBD are the only systems capable of successfully replicating all the discovery test cases. Arrowsmith also showed good performance by replicating 4 out of 5 discoveries, which is expected considering the high recall of co-occurrence-based systems. However, it should be noted that Arrowsmith relies solely on information extracted from article titles. BITOLA and SemBT are specialized systems focused on gene-disease associations, and as a result these systems do have not the broader capacity to handle other discovery tasks. Despite their focus on gene-disease associations, both BITOLA and SemBT failed to replicate discovery paths involving gene-disease associations. The semantic-based MELODI-PRESTO system did not replicate any of the 5 discovery paths—this outcome could likely be affected by the limitations of the provided API, which only considers the most recent 1 million articles in the MEDLINE database.

Despite the seemingly similar performances between AKG-LBD and LION-LBD in knowledge discovery, we note that there are many differences in the respective LBD

methodology of each system. AKG-LBD is a semantic-based LBD framework which has the advantage of providing the underlying functional relationships between biomedical entities, as described in the literature. In contrast, LION-LBD provides associations between entities based on the co-occurrence of biomedical entities in a sentence or an abstract. Co-occurrence of two entities in the literature does not necessarily indicate the existence of a functional relationships between them, thus may lead to the generation of noisy knowledge discoveries. AKG-LBD employs cutting-edge semantic-based methods to extract meaningful and non-ambiguous knowledge from the literature in the form of *subject-predicate-object* triples.

Furthermore, AKG-LBD maximizes the utilization of semantic-based knowledge by employing advanced KG representation learning techniques to predict new semantic relations among biomedical entities. In contrast, LION-LBD does not utilize such predictive techniques, as the literature-based knowledge is represented as a co-occurrence network, which is not compatible with advanced KG representation learning methods.

We posit that AKG-LBD contemporizes and enhances LBD by employing novel methods that unleash the potential of semantics-based knowledge.

*Table 5.19: Results of replicating the cancer discovery test cases using various LBD systems*

| Discovery Path | Arrowsmith | BITOLA | SemBT | MELODI-PRESTO | LION-LBD | AKG-LBD |
|---|---|---|---|---|---|---|
| NRF2 - ROS - Pancreatic cancer | Replicated | Not replicated | Not replicated | Not replicated | Replicated | Replicated |
| IL17 - p38a - MKP-1 | Replicated | N/A | N/A | Not replicated | Replicated | Replicated |
| NOTCH1 - Senescence - C/EBPb | Replicated | N/A | N/A | Not replicated | Replicated | Replicated |
| NFKB2 - BCL2 - Adenoma | Not replicated | Not replicated | Not replicated | Not replicated | Replicated | Replicated |
| CXCL12 – Senescence – Thyroid cancer | Replicated | Not replicated | Not replicated | Not replicated | Replicated | Replicated |

## 5.8 Summary:

This chapter presented the results of the components of the AKG-LBD framework, with a focus on an eventual evaluation of replicating cancer discoveries and repurposing drugs for

new cancer indications. We demonstrated the step-wise enhancement of the *baseline KG* resulting in a more complete *augmented KG*.

The *baseline KG* was constructed using semantic-based knowledge extracted from the literature by combining the outputs of SemRep and PubTator. Subsequently, semantically related and/or synonymous biomedical concepts were consolidated into atomic/standardized concept representation using condensed and specialized biomedical vocabularies, such as PRO and MeSH, rather than the comprehensive UMLS vocabulary, which often represents semantically similar/related concepts under different concept identifiers. This resulted in broader coverage of the knowledge domain.

The *baseline KG* was extended by using high quality curated knowledge from biomedical KBs, resulting in an *integrated KG*. Despite supplementing the *baseline KG* with new relations and nodes, this process did not significantly enhance the density of the *baseline KG*. This was an indication that supplementing literature-based KGs with curated knowledge alone does not result in a complete biomedical KG.

The *integrated KG* was extended to yield the *augmented KG* by predicting missing relations between pre-existing nodes in the *integrated KG*. The relation prediction task relied on KGE models that encode KG nodes and relations as low-dimensional vectors, thereby making them amenable to relation prediction tasks. We presented the results of evaluating three KGE models (TransE, DistMult, and ComplEx) on relation prediction tasks. Our results indicated that the semantic matching model, DistMult, was the best performing model, due to its capability in distinguishing between different node and relation types. On the relation prediction task, DistMult achieved the highest Hits@K and MRR results compared to TransE and ComplEx. Hence, DistMult was selected to augment the *integrated KG* by predicting missing relations. The informed relation prediction task resulted in adding over 1 million new relations to the *integrated KG*, which generated the *augmented KG* with enhanced graph density compared to the baseline. Despite these efforts, we posit that the *augmented KG* is likely incomplete due to the limitations of closed-world KGC.

We evaluated the utility of the baseline, integrated, and *augmented KG*s in two LBD tasks related to replicating targeted cancer discoveries, and repurposing existing drugs for new cancer indications. In the targeted discovery task, the *baseline KG* replicated one out of

five discoveries, the *integrated KG* replicated two out of five discoveries, and the *augmented KG* replicated five out of five discoveries. Similarly, the *augmented KG* resulted in a greater number of repurposed drugs compared to the baseline and *integrated KG*s via the strict and relaxed validations. Importantly, these results indicated that the improvement in LBD performance from baseline to integrated is marginal, whereby in both evaluations one additional discovery path was replicated by the integrated compared to the *baseline KG*. This is a strong indication that integration of curated knowledge alone is not sufficient to enhance the LBD process. However, the additional step of KGC is significant towards LBD performance.

With regards to filtering and ranking discovery paths, our results indicated that there is no significant difference between the IC-based metrics, whereby both metrics perform comparably in ranking relevant and meaningful discovery paths. However, compared to the baselines, IC-based metrics showed a significant improvement in the ranking performance, given that our proposed metrics are focused, by design, on rarely occurring knowledge instances in the literature.

# Chapter 6 Conclusion and Future Work

This dissertation investigated semantics-based methods and KG representation learning techniques to develop novel solutions addressing persistent/fundamental challenges in semantic-based LBD. Our investigation resulted in the AKG-LBD framework to facilitate semantics-based knowledge discovery in the biomedical domain. We applied the AKG-LBD framework in real-world discovery tasks, resulting in the replication of biomedical discoveries published in peer-reviewed publications.

In the following sections, we provide a summary of our research motivations, discuss the enhancements the AKG-LBD framework brings to existing LBD frameworks, examine the practical implications for applying AKG-LBD on real-world discovery tasks, and conclude with a discussion on the limitations of this research and potential future work to address these limitations.

## 6.1 Research Motivations:

LBD is a promising paradigm that enables knowledge discovery from the vast and rapidly expanding biomedical literature. As the volume of publications continues to grow exponentially, keeping up with the pace of biomedical research has become increasingly challenging. In this regard, LBD offers a solution leveraging computational methods to extract and uncover hidden knowledge from the ever-growing volume of biomedical literature.

Overtime, LBD has evolved from a largely manual approach to adopting advanced knowledge extraction and representation techniques. At a high-level, contemporary LBD can be categorized as co-occurrence-based semantics-based LBD. The co-occurrence-based approach employs text mining methods to extract biomedical terms or concepts from the literature. Subsequently, associations between terms or concepts are established based on the assumption that their co-occurrence in articles implies a logical association between them. However, this assumption is inherently weak, as co-occurrence associations do not imply the existence of a mechanistic relationship between terms/concepts. Conversely, the semantic-based approach utilizes semantic parsing techniques to extract literature-based knowledge in the form of *subject-predicate-object*

triples. As such, this approach has the advantage of extracting meaningful knowledge that characterizes the underlying mechanistic relations between concepts, as described in the literature.

Despite the promise of semantic-based LBD, there are several fundamental challenges that impact its performance for knowledge discovery. We provide a summary of these challenges below:

**Challenge #1 - Granularity and ambiguity of biomedical concept representations**: The granularity of terminological resources, such as the UMLS, used for the normalization of biomedical entities in text poses a significant challenge in LBD due to the resulting semantic redundancy, where closely related entities are often represented as separate concepts. As a consequence, the knowledge discovery task will entail considering many distinct 'source' concepts which may be semantically related. , leading to an exponential expansion of the discovery search space. Additionally, the entity normalization process in semantic parsers fails to disambiguate mentions of genes or protein aliases. These challenges lead to (a) exponential expansion of the discovery search space due to the presence of granular but semantically related concepts; and (b) highly ambiguous and imprecise concept-based representations of genes and proteins.

**Challenge #2 - Incomplete extraction of semantic-based knowledge**: Biomedical semantic parsers, such as SemRep, achieve high precision in the extraction of semantic-based knowledge but also suffer from low recall (i.e., missing knowledge), which results in incomplete knowledge extraction. This is especially challenging in semantic-based LBD due to limitations in NLP methods to correctly identify implicit relations that extend beyond sentence boundaries in text. In the context of semantic-based LBD, incomplete knowledge extraction has a significant impact on knowledge discovery. Explicitly, if relations between the source-intermediate or intermediate-target are absent, then the indirect relationship between the source and target will not be discovered.

**Challenge #3 - Filtering and ranking interesting ABC-based discoveries**: LBD frameworks tend to generate a large number of candidate discoveries which makes the task of reviewing and validating discoveries a time consuming process. Most LBD frameworks use raw co-occurrence frequencies, which favour commonly occurring knowledge instances, or statistical association measures, which are affected by null co-occurrences.

Prior research has suggested that meaningful discoveries typically consist of rare and specific knowledge instances which are not captured by traditional filtering and ranking techniques.

Motivated by these challenges, this dissertation explored novel semantics-based and Knowledge Graph Completion (KGC) methods to present novel solutions addressing the fundamental challenges in semantics-based LBD.

We addressed the first challenge by extending traditional semantic LBD framework to include: (i) a semantic consolidation component to that resolves the challenge of ambiguity by leveraging condensed and specialized terminologies to represent entities denoting chemicals, diseases, genes/proteins, and biological functions; and (ii) a knowledge extraction component that integrates semantic parsers and cutting-edge text mining tools to resolve the ambiguity of gene/protein representations.

The challenge of incomplete knowledge extraction is addressed via a novel knowledge completion methodology that leverages curated biomedical knowledge, in addition to KG-based representation learning techniques to embed nodes and relations as low dimensional vectors. Subsequently, we leverage Knowledge Graph Completion (KGC) methods to predict missing semantic relations between existing nodes. Further, our approach to predicting missing relations is informed by implicit associations of MeSH descriptors found in the literature. With this approach, the literature-based knowledge is augmented progressively by supplementing it with curated knowledge first to increase coverage of the knowledge domain, then supplementing with predicted semantic relations via KGC methods.

The third challenge is resolved by introducing a knowledge filtering and ranking approach that builds upon established techniques (i.e., linking term count, concept specificity, and triple counts) and integrates information theoretic metrics that prioritize interesting and rare discoveries over common and spurious ones.

Our investigations have led to the development of AKG-LBD; an end-to-end semantic-based LBD framework that utilizes KG-based representation techniques to facilitate the discovery of biomedical knowledge from the literature.

## 6.2 AKG-LBD Compared to Established LBD Frameworks:

AKG-LBD contributes novel methodological advances to LBD in the biomedical domain. To our knowledge, this is the first semantics-based LBD framework that introduces a scalable KGC based approach focused on augmenting incomplete literature-based KGs with missing semantic relations between biomedical concepts. Prior LBD frameworks have incorporated network-based link prediction methods, such as Jaccard similarity, Adamic-Adar index, and common neighbours, to predict future links between non-co-occurring concepts (Kastrin et al., 2016). However, these methods do not address the fundamental problem of missing knowledge instances in LBD, as these methods can only quantify the likelihood of indirect associations rather than predicting new semantic relations. For example, the formalized LION-LBD system incorporates Jaccard similarity measures to quantify the strength of associations between source and target biomedical concepts. Recently, Zhang et al. employed KGC methods in semantic-based LBD to facilitate the discovery of novel treatments for COVID-19 via entity prediction – i.e., given a *subject* and a *predicate*, the goal is to predict a plausible *object* entity (Zhang et al., 2021). These approaches do not aim to address the fundamental problem of incompleteness in literature-based KGs. Instead, the objective is to enhance knowledge discovery by predicting connections between non-interacting source (A) and target (C) concepts. Additionally, this approach to knowledge discovery overlooks the existence of intermediate (B) concepts, which could provide further biomedical insights into the source-target interactions. Hence, compared to existing frameworks, AKG-LBD is novel as we utilize KGC to address a fundamental challenge in semantic-based LBD (i.e., incomplete knowledge extraction) by predicting missing relations between pre-existing nodes (i.e., concepts) with high precision. Further, our approach scales up to predict a wide range of semantic relations (i.e., predicates) as long as these relations are represented within the KG. Lastly, we introduced a relation prediction approach informed by co-occurrence found in the literature (i.e., MeSH descriptors) to focus on predicting relations between implicitly associated biomedical concepts.

In addition to these novel advances, AKG-LBD offers several enhancements to current semantic-based LBD frameworks:

- Improving the extraction of literature-based knowledge by combining the outputs of well-established biomedical knowledge extraction tools (SemRep and Pubtator) that are commonly used in LBD research
- Integrating multiple biomedical terminologies to consolidate and condense the representation of concepts extracted from the literature
- Leveraging curated knowledge from biomedical Knowledge Bases (KBs) to supplement the semantic triples extracted from the literature
- Proposing new knowledge filtering and ranking techniques focused on quantifying the information conveyed in ABC discovery paths

These advances exemplified by AKG-LBD demonstrated their significant impact on uncovering meaningful biomedical knowledge from the literature. The results presented in Chapter 5 showcased that AKG-LBD outperformed multiple established LBD systems in replicating discoveries published in peer-reviewed publications. Moreover, AKG-LBD was significantly better than the baseline LBD approach (i.e., without KGC) in repurposing existing drugs for cancers. We posit that the performance of AKG-LBD can be attributed to the following factors:

The knowledge extraction component, which incorporated concept disambiguation, addressed a common problem in semantic-based LBD – i.e., ambiguous concepts in semantic triples. While evaluating the established semantic-based LBD system MELODI-PRESTO, we encountered many instances ambiguous semantic triples. This made the task of reviewing output discoveries very challenging, as the *subject* and *object* were represented by multiple concepts.

The knowledge completion component provided clear benefits to semantics-based LBD by augmenting the incomplete literature-based KG with missing knowledge. Compared to the baseline approach (i.e., without knowledge completion), our approach resulted in a significantly better performance in the cancer discovery (20% vs. 100% recall) and drug repurposing tasks (46% vs. 71% recall). Likewise, compared to established semantics-based LBD systems (MELODI-PRESTO), our framework achieved significantly better results (0% vs. 100% recall).

We compared the performance of AKG-LBD with established co-occurrence-based systems, which are known to achieve better recall than semantics-based systems. The

results indicated that AKG-LBD achieves comparable results to co-occurrence-based systems due to its knowledge completion component. We consider this as an achievement, as AKG-LBD was capable of achieving high recall while still being a semantics-based framework.

Finally, we conducted a comparison of our knowledge filtering and ranking approaches with commonly used metrics, such as Odds Ratio, Log-Likelihood Ratio, and co-occurrence frequency. While we acknowledge that our methods outperformed the baselines, we also propose that Information Content (IC)-based metrics present a viable alternative to the commonly used metrics in the field. Our results suggest that IC-based metrics can be effective in addressing the limitations of traditionally used metrics, which are typically biased towards commonly occurring knowledge instances.

## 6.3 Practical Implications of AKG-LBD for Real-World Knowledge Discovery:

There are several implications for applying the AKG-LBD framework to real-world knowledge discovery tasks. Firstly, the literature curation component requires users to carefully determine keywords or phrases to retrieve relevant titles/abstracts from literature databases. In our approach, we utilized the MEDLINE database to acquire biomedical knowledge which provides an API to query the database using a combination of keywords and MeSH descriptors assigned to individual articles. Our implementation of AKG-LBD was focused on specific discovery tasks – i.e., molecular oncology and cancer drug repurposing – hence, we used a combination of keywords and MeSH descriptors related to the hallmarks of cancers. However, in situations where the discovery task is non-specific, determining relevant keywords and MeSH descriptors could be difficult and time-consuming. Additionally, selecting common keywords or MeSH descriptors will result in retrieving numerous titles/abstracts, thereby increasing the time and space complexity of the downstream knowledge completion task. Hence, we posit that the framework should be utilized in prespecified discovery tasks, and that domain experts should be consulted to determine a list of relevant keywords or MeSH descriptors to optimize literature retrieval. Secondly, the integration of curated knowledge entails careful consideration of the knowledge domain represented in biomedical Knowledge Bases (KBs) and an

understanding of the terminologies used to represent the knowledge. In our implementation of AKG-LBD, we utilized CTD and GO as the primary KBs which (a) cover a wide range of biomedical knowledge; and (b) represent knowledge using common terminological resources, such as MeSH. However, in situations where the discovery task requires specialized curated knowledge to augment the literature-based KG, alternative KBs should be considered since CTD and GO may not provide the relevant curated knowledge. Hence, we recommend conferring with domain experts to determine the most suitable KBs and facilitate mapping of concepts to the terminologies used by the baseline literature-based KG.

Thirdly, the relation prediction task requires careful planning of the types of MeSH descriptors to extract from MEDLINE to generate the subset of MeSH-MeSH associations for relation prediction. To reduce the time and space complexity of relation prediction, we recommend users to utilize co-occurrence statistics to reduce the effect of noisy associations and eliminate concepts that co-occur in few articles. In our experiments, we set the minimum number of co-occurrences to 5 – i.e., MeSH-MeSH associations are considered if they co-occur in at least 5 articles. This resulted in a significant decrease in the number of extracted associations from MEDLINE. Alternative approaches to filter such associations may include association metrics (e.g., chi-square, odds ratio, log-likelihood ratio), information content, or semantic type filtering.

Lastly, we proposed a novel knowledge ranking measure based on information theory, which performs well in prioritizing valid and interesting ABC discovery paths. However, in line with findings from prior LBD research, we recommend that users should not entirely rely on a single metric to rank the LBD output, as it offers the flexibility to rank discoveries based on different characteristics or attributes of knowledge instances (Thilakaratne et al., 2019).

## 6.4 Limitations and Future Work:

We identify several limitations in our work which can be mostly attributed to the use of external knowledge resource to represent and augment the literature-based KG.

Firstly, AKG-LBD relies on multiple terminologies to represent biomedical concepts. It is important to note that these terminologies undergo periodic updates which may involve the

addition, removal, or consolidation of biomedical concepts. Hence, ensuring the currency of the terminologies employed in AKG-LBD presents a substantial challenge, necessitating the need for frequent assessments to identify any updates that have been incorporated into the source terminologies. Additionally, the KBs utilized in this framework undergo similar periodic updates that need to be incorporated in the *augmented KG*. Hence, future work will address these limitations by investigating potential strategies and mechanisms to streamline the integration of updates, allowing for more efficient and timely synchronization between AKG-LBD and the evolving terminologies and KBs.

Secondly, the KGC methods presented in this research are classified as closed-world approaches which tend to utilize pre-existing knowledge within the KG to predict the missing instances (Z. Chen et al., 2020). Closed-world KGC implies that knowledge within the KG is fixed and that previously unseen concepts or relations cannot be predicted. In this setting, KGC aims to predict the most plausible missing relation from pre-existing relations in the KG. As such, this limitation makes closed-world KGC reliant on existing KG semantics and topography. As future work, we plan to address this limitation by exploring dynamic KGC methods that operate under the open-world assumption to predict external entities (i.e., relations and nodes).

Thirdly, the relation prediction task involves assigning a score to each relation used for the incomplete (*subject, ?, object*) triples based on inherent scoring functions employed by the Knowledge Graph Embeddings (KGEs). Typically, the scores are ranked from highest to lowest, with higher scores indicating plausible predictions. However, analyzing these scores without ranking them does not provide any insights into the credibility of predictions (Tabacof & Costabello, 2020; Zhu et al., 2022). Accordingly, probability calibration methods can be employed to transform prediction scores into interpretable probabilities that provide valuable insights into the trustworthiness of the predictions. As future work, we plan to investigate methods such as Platt scaling or isotonic regression, combined with curated ground truth negative samples, to calibrate the relation prediction scores and transform them into interpretable probabilities (Tabacof & Costabello, 2020).

Lastly, the AKG-LBD framework is not implemented currently as a formalized system and, therefore, it is not available for public use. In the future, we plan to develop a web-based application that provides access to all components of the AKG-LBD framework. Our

aim is to enable users to control the inputs and outputs of each component to provide a knowledge discovery process tailored to their needs and preferences. Additionally, we aim to leverage interactive graph-based visualizations to facilitate the review and exploration of generated discovery paths.

# Bibliography


Ahlers, C. B., Hristovski, D., Kilicoglu, H., & Rindflesch, T. C. (2007). Using the Literature-Based Discovery Paradigm to Investigate Drug Mechanisms. *AMIA Annual Symposium Proceedings*, *2007*, 6–10.

Aronson, A. R., & Lang, F.-M. (2010). An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association : JAMIA*, *17*(3), 229–236. https://doi.org/10.1136/jamia.2009.002733

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

Baker, N. C., & Hemminger, B. M. (2010). Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. *Journal of Biomedical Informatics*, *43*(4), 510–519. https://doi.org/10.1016/j.jbi.2010.03.008

Baumann, N. (2016). How to use the medical subject headings (MeSH). *International Journal of Clinical Practice*, *70*(2), 171–174. https://doi.org/10.1111/ijcp.12767

Biswas, S., Mitra, P., & Rao, K. S. (2021). Relation Prediction of Co-Morbid Diseases Using Knowledge Graph Completion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(2), 708–717. https://doi.org/10.1109/TCBB.2019.2927310

Bonner, S., Barrett, I. P., Ye, C., Swiers, R., Engkvist, O., Hoyt, C. T., & Hamilton, W. L. (2022). Understanding the Performance of Knowledge Graph Embeddings in Drug Discovery. *ArXiv:2105.10488 [Cs, q-Bio]*. http://arxiv.org/abs/2105.10488

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating Embeddings for Modeling Multi-relational Data. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

Bougiatiotis, K., Aisopos, F., Nentidis, A., Krithara, A., & Paliouras, G. (2020). Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph. In M. Michalowski & R. Moskovitch (Eds.), *Artificial Intelligence in Medicine* (Vol. 12299, pp. 122–132). Springer International Publishing. https://doi.org/10.1007/978-3-030-59137-3_12

Bruza, P., Cole, R., Song, D., & Bari, Z. (2006). Towards Operational Abduction from a Cognitive Perspective. *Logic Journal of the IGPL*, *14*(2), 161–177. https://doi.org/10.1093/jigpal/jzk012

Cairelli, M. J., Fiszman, M., Zhang, H., & Rindflesch, T. C. (2015). Networks of neuroinjury semantic predications to identify biomarkers for mild traumatic brain injury. *Journal of Biomedical Semantics*, *6*(1), 25. https://doi.org/10.1186/s13326-015-0022-4

Cairelli, M. J., Miller, C. M., Fiszman, M., Workman, T. E., & Rindflesch, T. C. (2013). Semantic MEDLINE for Discovery Browsing: Using Semantic Predications and the Literature-Based Discovery Paradigm to Elucidate a Mechanism for the Obesity Paradox. *AMIA Annual Symposium Proceedings*, *2013*, 164–173.

Cameron, D., Bodenreider, O., Yalamanchili, H., Danh, T., Vallabhaneni, S., Thirunarayan, K., Sheth, A. P., & Rindflesch, T. C. (2013). A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *Journal of Biomedical Informatics*, *46*(2), 238–251. https://doi.org/10.1016/j.jbi.2012.09.004

Chang, D., Balažević, I., Allen, C., Chawla, D., Brandt, C., & Taylor, A. (2020). Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 167–176. https://doi.org/10.18653/v1/2020.bionlp-1.18

Chen, C., & Song, M. (2017). Measuring Scholarly Impact. In C. Chen & M. Song (Eds.), *Representing Scientific Knowledge: The Role of Uncertainty* (pp. 139–204). Springer International Publishing. https://doi.org/10.1007/978-3-319-62543-0_4

Chen, G., Jia, Y., Zhu, L., Li, P., Zhang, L., Tao, C., & Jim Zheng, W. (2019). Gene fingerprint model for literature based detection of the associations among complex diseases: A case study of COPD. *BMC Medical Informatics and Decision Making*, *19*(Suppl 1), 20. https://doi.org/10.1186/s12911-019-0738-7

Chen, L., & Friedman, C. (2004). Extracting phenotypic information from the literature via natural language processing. *Studies in Health Technology and Informatics*, *107*(Pt 2), 758–762.

Chen, Q., Lee, K., Yan, S., Kim, S., Wei, C.-H., & Lu, Z. (2020). BioConceptVec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLOS Computational Biology*, *16*(4), e1007617. https://doi.org/10.1371/journal.pcbi.1007617

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., & Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, *11*(1), 492. https://doi.org/10.1186/1471-2105-11-492

Cohen, T., Schvaneveldt, R., & Widdows, D. (2010). Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, *43*(2), 240–256. https://doi.org/10.1016/j.jbi.2009.09.003

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

Crichton, G., Guo, Y., Pyysalo, S., & Korhonen, A. (2018). Neural networks for link prediction in realistic biomedical graphs: A multi-dimensional evaluation of graph embedding-based approaches. *BMC Bioinformatics*, *19*(1), 176. https://doi.org/10.1186/s12859-018-2163-9

Crichton, G., Pyysalo, S., Chiu, B., & Korhonen, A. (2017). A neural network multi-task learning approach to biomedical named entity recognition. *BMC Bioinformatics*, *18*(1), 368. https://doi.org/10.1186/s12859-017-1776-8

Dai, Z., Li, Q., Yang, G., Wang, Y., Liu, Y., Zheng, Z., Tu, Y., Yang, S., & Yu, B. (2019). Using literature-based discovery to identify candidate genes for the interaction between myocardial infarction and depression. *BMC Medical Genetics*, *20*, 104. https://doi.org/10.1186/s12881-019-0841-8

Daowd, A., Abidi, S., & Abidi, S. S. R. (2022). A Knowledge Graph Completion Method Applied to Literature-Based Discovery for Predicting Missing Links Targeting Cancer Drug Repurposing. In M. Michalowski, S. S. R. Abidi, & S. Abidi (Eds.), *Artificial Intelligence in Medicine* (pp. 24–34). Springer International Publishing. https://doi.org/10.1007/978-3-031-09342-5_3

Daowd, A., Barrett, M., Abidi, S., & Abidi, S. S. R. (2021a). A Framework To Build A Causal Knowledge Graph for Chronic Diseases and Cancers By Discovering Semantic Associations from Biomedical Literature. *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, 13–22. https://doi.org/10.1109/ICHI52183.2021.00016

Daowd, A., Barrett, M., Abidi, S., & Abidi, S. S. R. (2021b). Building a Knowledge Graph Representing Causal Associations Between Risk Factors and Incidence of Breast Cancer. *Public Health and Informatics*, 724–728. https://doi.org/10.3233/SHTI210267

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C., & Mattingly, C. J. (2019). The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Research*, *47*(Database issue), D948–D954. https://doi.org/10.1093/nar/gky868

Day, C. M., Hickey, S. M., Song, Y., Plush, S. E., & Garg, S. (2020). Novel Tamoxifen Nanoformulations for Improving Breast Cancer Treatment: Old Wine in New Bottles. *Molecules (Basel, Switzerland)*, *25*(5), 1182. https://doi.org/10.3390/molecules25051182

Deftereos, S. N., Andronis, C., Friedla, E. J., Persidis, A., & Persidis, A. (2011). Drug repurposing and adverse event prediction using high-throughput literature analysis. *WIREs Systems Biology and Medicine*, *3*(3), 323–334. https://doi.org/10.1002/wsbm.147

Demner-Fushman, D., Rogers, W. J., & Aronson, A. R. (2017). MetaMap Lite: An evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association : JAMIA*, *24*(4), 841–844. https://doi.org/10.1093/jamia/ocw177

DeNicola, G. M., Karreth, F. A., Humpton, T. J., Gopinathan, A., Wei, C., Frese, K., Mangal, D., Yu, K. H., Yeo, C. J., Calhoun, E. S., Scrimieri, F., Winter, J. M., Hruban, R. H., Iacobuzio-Donahue, C., Kern, S. E., Blair, I. A., & Tuveson, D. A. (2011). Oncogene-induced Nrf2 transcription promotes ROS detoxification and tumorigenesis. *Nature*, *475*(7354), Article 7354. https://doi.org/10.1038/nature10189

DiGiacomo, R. A., Kremer, J. M., & Shah, D. M. (1989). Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine*, *86*(2), 158–164. https://doi.org/10.1016/0002-9343(89)90261-1

Drury, B., Oliveira, H. G., & Lopes, A. de A. (2022). A survey of the extraction and applications of causal relations. *Natural Language Engineering*, *28*(3), 361–400. https://doi.org/10.1017/S135132492100036X

Du, J., & Li, X. (2020). A Knowledge Graph of Combined Drug Therapies Using Semantic Predications From Biomedical Literature: Algorithm Development. *JMIR Medical Informatics*, *8*(4), e18323. https://doi.org/10.2196/18323

Elsworth, B., Dawe, K., Vincent, E. E., Langdon, R., Lynch, B. M., Martin, R. M., Relton, C., Higgins, J. P. T., & Gaunt, T. R. (2018). MELODI: Mining Enriched Literature Objects to Derive Intermediates. *International Journal of Epidemiology*, *47*(2), 369–379. https://doi.org/10.1093/ije/dyx251

Elsworth, B., & Gaunt, T. R. (2021). MELODI Presto: A fast and agile tool to explore semantic triples derived from biomedical literature. *Bioinformatics*, *37*(4), 583–585. https://doi.org/10.1093/bioinformatics/btaa726

Eronen, L., & Toivonen, H. (2012). Biomine: Predicting links between biological entities using network models of heterogeneous databases. *BMC Bioinformatics*, *13*(1), 119. https://doi.org/10.1186/1471-2105-13-119

Francis, N., Green, A., Guagliardo, P., Libkin, L., Lindaaker, T., Marsault, V., Plantikow, S., Rydberg, M., Selmer, P., & Taylor, A. (2018). Cypher: An Evolving Query Language for Property Graphs. *Proceedings of the 2018 International Conference on Management of Data*, 1433–1445. https://doi.org/10.1145/3183713.3190657

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics*, *14*(5), 360–366. https://doi.org/10.1038/nrg3456

Gabetta, M., Larizza, C., & Bellazzi, R. (2013). A Unified Medical Language System (UMLS) based system for Literature-Based Discovery in medicine. *Studies in Health Technology and Informatics*, *192*, 412–416.

Gaffen, S. L., & McGeachy, M. J. (2015). Integrating p38α MAPK immune signals in nonimmune cells. *Science Signaling*, *8*(366), fs5–fs5. https://doi.org/10.1126/scisignal.aaa8398

The Gene Ontology Consortium. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338. https://doi.org/10.1093/nar/gky1055

Goodwin, J. C., Cohen, T., & Rindflesch, T. (2012). Discovery by scent: Discovery browsing system based on the Information Foraging Theory. *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 232–239. https://doi.org/10.1109/BIBMW.2012.6470309

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

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

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

Gordon, M. D., & Lindsay, R. K. (1996). Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. *Journal of the American Society for Information Science*, *47*(2), 116–128. https://doi.org/10.1002/(SICI)1097-4571(199602)47:2<116::AID-ASI3>3.0.CO;2-1

Henry, S. (2019). Indirect Relatedness, Evaluation, and Visualization for Literature Based Discovery. *VCU Theses and Dissertations*. https://doi.org/10.25772/C1P9-WG56

Henry, S., Cuffy, C., & McInnes, B. T. (2018). Vector representations of multi-word terms for semantic relatedness. *Journal of Biomedical Informatics*, *77*, 111–119. https://doi.org/10.1016/j.jbi.2017.12.006

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

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

Henry, S., McQuilkin, A., & McInnes, B. T. (2019). Association measures for estimating semantic similarity and relatedness between biomedical concepts. *Artificial Intelligence in Medicine*, *93*, 1–10. https://doi.org/10.1016/j.artmed.2018.08.006

Heo, G. E., Xie, Q., Song, M., & Lee, J.-H. (2019). Combining entity co-occurrence with specialized word embeddings to measure entity relation in Alzheimer's disease. *BMC Medical Informatics and Decision Making*, *19*(S5), 240. https://doi.org/10.1186/s12911-019-0934-5

Hoare, M., Ito, Y., Kang, T.-W., Weekes, M. P., Matheson, N. J., Patten, D. A., Shetty, S., Parry, A. J., Menon, S., Salama, R., Antrobus, R., Tomimatsu, K., Howat, W., Lehner, P. J., Zender, L., & Narita, M. (2016). NOTCH1 mediates a switch between two distinct secretomes during senescence. *Nature Cell Biology*, *18*(9), Article 9. https://doi.org/10.1038/ncb3397

Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. (2006). Exploiting Semantic Relations for Literature-Based Discovery. *AMIA Annual Symposium Proceedings*, *2006*, 349–353.

Hristovski, D., Kastrin, A., Dinevski, D., Burgun, A., Žiberna, L., & Rindflesch, T. C. (2016). Using Literature-Based Discovery to Explain Adverse Drug Effects. *Journal of Medical Systems*, *40*(8), 185. https://doi.org/10.1007/s10916-016-0544-z

Hristovski, D., Kastrin, A., Dinevski, D., & Rindflesch, T. C. (2015). *Towards Implementing Semantic Literature-Based Discovery with a Graph Database*. 5.

Hristovski, D., Kastrin, A., Peterlin, B., & Rindflesch, T. C. (2010). Combining Semantic Relations and DNA Microarray Data for Novel Hypotheses Generation. In C. Blaschke & H. Shatkay (Eds.), *Linking Literature, Information, and Knowledge for Biology* (pp. 53–61). Springer. https://doi.org/10.1007/978-3-642-13131-8_7

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

Huang, S., He, L., Yang, B., & Zhang, M. (2012). A compound correlation model for disjoint literature-based knowledge discovery. *Aslib Proceedings*, *64*(4), 423–436. https://doi.org/10.1108/00012531211244770

Ittipanuvat, V., Fujita, K., Sakata, I., & Kajikawa, Y. (2014). Finding linkage between technology and social issue: A Literature Based Discovery approach. *Journal of Engineering and Technology Management*, *32*, 160–184. https://doi.org/10.1016/j.jengtecman.2013.05.006

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2022). A Survey on Knowledge Graphs: Representation, Acquisition and Applications. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843

Jiang, H.-J., Huang, Y.-A., & You, Z.-H. (2020). SAEROF: An ensemble approach for large-scale drug-disease association prediction by incorporating rotation forest and sparse autoencoder deep neural network. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-61616-9

Kastrin, A., Ferk, P., & Leskošek, B. (2018). Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. *PLOS ONE*, *13*(5), e0196865. https://doi.org/10.1371/journal.pone.0196865

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

Kastrin, A., Rindflesch, T. C., & Hristovski, D. (2014). Link Prediction on the Semantic MEDLINE Network. In S. Džeroski, P. Panov, D. Kocev, & L. Todorovski (Eds.), *Discovery Science* (pp. 135–143). Springer International Publishing. https://doi.org/10.1007/978-3-319-11812-3_12

Kastrin, A., Rindflesch, T., & Hristovski, D. (2016). Link Prediction on a Network of Co-occurring MeSH Terms: Towards Literature-based Discovery. *Methods of Information in Medicine*, *55*(04), 340–346. https://doi.org/10.3414/ME15-01-0108

Katz, M. J. (Ed.). (2009). Composing the Sections of a Research Paper. In *From Research to Manuscript: A Guide to Scientific Writing* (pp. 79–162). Springer Netherlands. https://doi.org/10.1007/978-1-4020-9467-5_7

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Rindflesch, T. C. (2016). Sortal anaphora resolution to enhance relation extraction from biomedical literature. *BMC Bioinformatics*, *17*(1), 163. https://doi.org/10.1186/s12859-016-1009-6

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Shin, D. (2020). Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinformatics*, *21*(1), 188. https://doi.org/10.1186/s12859-020-3517-7

Kim, Y. H., Choi, Y. W., Lee, J., Soh, E. Y., Kim, J.-H., & Park, T. J. (2017). Senescent tumor cells lead the collective invasion in thyroid cancer. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15208

Kim, Y. H., & Park, T. J. (2019). Cellular senescence in cancer. *BMB Reports*, *52*(1), 42–46.

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

Knijnenburg, T. A., Bismeijer, T., Wessels, L. F. A., & Shmulevich, I. (2015). A multilevel pan-cancer map links gene mutations to cancer hallmarks. *Chinese Journal of Cancer*, *34*(3), 48. https://doi.org/10.1186/s40880-015-0050-6

Lever, J., Gakkhar, S., Gottlieb, M., Rashnavadi, T., Lin, S., Siu, C., Smith, M., Jones, M. R., Krzywinski, M., & Jones, S. J. M. (2018). A collaborative filtering-based approach to biomedical knowledge discovery. *Bioinformatics*, *34*(4), 652–659. https://doi.org/10.1093/bioinformatics/btx613

Li, J., Ichikawa, T., Villacorta, L., Janicki, J. S., Brower, G. L., Yamamoto, M., & Cui, T. (2009). Nrf2 protects against maladaptive cardiac responses to hemodynamic stress. *Arteriosclerosis, Thrombosis, and Vascular Biology*, *29*(11), 1843–1850. https://doi.org/10.1161/ATVBAHA.109.189480

Lindsay, R. K., & Gordon, M. D. (1999). Literature-based discovery by lexical statistics. *Journal of the American Society for Information Science*, *50*(7), 574–587.

Luo, Y., Uzuner, Ö., & Szolovits, P. (2017). Bridging semantics and syntax with graph algorithms—State-of-the-art of extracting biomedical relations. *Briefings in Bioinformatics*, *18*(1), 160–178. https://doi.org/10.1093/bib/bbw001

Marsi, E., Pinar Øzturk, P., & V. Ardelan, M. (2017). Marine Variable Linker: Exploring Relations between Changing Variables in Marine Science Literature. *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 91–94. https://aclanthology.org/E17-3023

McInnes, B. T., & Pedersen, T. (2015). Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of Biomedical Informatics*, *54*, 329–336. https://doi.org/10.1016/j.jbi.2014.11.014

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. http://arxiv.org/abs/1301.3781

Milošević, N., & Thielemann, W. (2023). Comparison of biomedical relationship extraction methods and models for knowledge graph creation. *Journal of Web Semantics*, *75*, 100756. https://doi.org/10.1016/j.websem.2022.100756

Miwa, M., Thompson, P., & Ananiadou, S. (2012). Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, *28*(13), 1759–1765. https://doi.org/10.1093/bioinformatics/bts237

Mohamed, S. K., Nounu, A., & Nováček, V. (2021). Biological applications of knowledge graph embedding models. *Briefings in Bioinformatics*, *22*(2), 1679–1693. https://doi.org/10.1093/bib/bbaa012

Moreau, E., Hardiman, O., Heverin, M., & O'Sullivan, D. (2021). *Literature-Based Discovery beyond the ABC paradigm: A contrastive approach* [Preprint]. Bioinformatics. https://doi.org/10.1101/2021.09.22.461375

Naderi Yeganeh, P., Richardson, C., Saule, E., Loraine, A., & Taghi Mostafavi, M. (2020). Revisiting the use of graph centrality models in biological pathway analysis. *BioData Mining*, *13*(1), 5. https://doi.org/10.1186/s13040-020-00214-x

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J. A., Bult, C. J., Caudy, M., Drabkin, H. J., D'Eustachio, P., Evsikov, A. V., Huang, H., Nchoutmboube, J., Roberts, N. V., Smith, B., Zhang, J., & Wu, C. H. (2011). The Protein Ontology: A structured representation of protein forms and complexes. *Nucleic Acids Research*, *39*(suppl_1), D539–D545. https://doi.org/10.1093/nar/gkq907

National Library of Medicine. (2022). *The MEDLINE Indexing Process: Determining Subject Content* [Training Material and Manuals]. U.S. National Library of Medicine. https://www.nlm.nih.gov/bsd/disted/meshtutorial/principlesofmedlinesubjectindexing/theindexingprocess/index.html

National Library of Medicine. (2023). *Medical Subject Headings* [Web Page]. Medical Subject Headings; U.S. National Library of Medicine. https://www.nlm.nih.gov/mesh/meshhome.html

Nian, Y., Hu, X., Zhang, R., Feng, J., Du, J., Li, F., Chen, Y., & Tao, C. (2022). Mining On Alzheimer's Diseases Related Knowledge Graph to Identity Potential AD-related Semantic Triples for Drug Repurposing. *ArXiv:2202.08712 [Cs]*. http://arxiv.org/abs/2202.08712

Nicholson, D. N., & Greene, C. S. (2020). Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, *18*, 1414–1428. https://doi.org/10.1016/j.csbj.2020.05.017

Özgür, A., Xiang, Z., Radev, D. R., & He, Y. (2010). Literature-Based Discovery of IFN-γ and Vaccine-Mediated Gene Interaction Networks. *Journal of Biomedicine and Biotechnology*, *2010*, 426479. https://doi.org/10.1155/2010/426479

Park, B. S., Kang, D., Kim, K. K., Jeong, B., Lee, T. H., Park, J. W., Kimura, S., Yeh, J.-Y., Roh, G. S., Lee, C.-J., Yang, S., Yang, S., Kim, J. G., & Lee, B. J. (2022). Hypothalamic TTF-1 orchestrates the sensitivity of leptin. *Molecular Metabolism*, *66*, 101636. https://doi.org/10.1016/j.molmet.2022.101636

Pedersen, T., Banerjee, S., McInnes, B., Kohli, S., Joshi, M., & Liu, Y. (2011). The Ngram Statistics Package (Text::NSP): A Flexible Tool for Identifying Ngrams, Collocations, and Word Associations. *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*, 131–133. https://aclanthology.org/W11-0821

Petric, I., Ligeti, B., Gyorffy, B., & Pongor, S. (2014). Biomedical hypothesis generation by text mining and gene prioritization. *Protein and Peptide Letters*, *21*(8), 847–857. https://doi.org/10.2174/0929866513209990063

Phang, C. S. J., Vong, W.-T., Sebastian, Y., Raman, V., & Then, P. H. H. (2022). Understanding the Usability of a Literature-Based Discovery System Among Clinical Researchers in Sarawak, Malaysia. *International Journal of Technology and Human Interaction (IJTHI)*, *18*(1), 1–21. https://doi.org/10.4018/IJTHI.304092

Preiss, J., & Stevenson, M. (2016). The effect of word sense disambiguation accuracy on literature based discovery. *BMC Medical Informatics and Decision Making*, *16*(1), 57. https://doi.org/10.1186/s12911-016-0296-1

Preiss, J., Stevenson, M., & Gaizauskas, R. (2015). Exploring relation types for literature-based discovery. *Journal of the American Medical Informatics Association*, *22*(5), 987–992. https://doi.org/10.1093/jamia/ocv002

Pyysalo, S., Baker, S., Ali, I., Haselwimmer, S., Shah, T., Young, A., Guo, Y., Högberg, J., Stenius, U., Narita, M., & 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

Qian, Q., Hong, N., & An, X. (2012). Structuring the Chinese disjointed literature-based knowledge discovery system: The key technologies to success. *Journal of Information Science*, *38*(6), 532–539. https://doi.org/10.1177/0165551512461104

Raja, K., Steill, J., Ross, I., Tsoi, L. C., Kuusisto, F., Ni, Z., Livny, M., Thomson, J., & Stewart, R. (2020). *SKiM - A generalized literature-based discovery system for uncovering novel biomedical knowledge from PubMed* (p. 2020.10.16.343012). bioRxiv. https://doi.org/10.1101/2020.10.16.343012

Rasmy, L., Tiryaki, F., Zhou, Y., Xiang, Y., Tao, C., Xu, H., & Zhi, D. (2020). Representation of EHR data for predictive modeling: A comparison between UMLS and other terminologies. *Journal of the American Medical Informatics Association*, *27*(10), 1593–1599. https://doi.org/10.1093/jamia/ocaa180

Rastegar-Mojarad, M., Elayavilli, R. K., Li, D., & Liu, H. (2015). Assessing the Need of Discourse-Level Analysis in Identifying Evidence of Drug-Disease Relations in Scientific Literature. *Studies in Health Technology and Informatics*, *216*, 539–543.

Rastegar-Mojarad, M., Elayavilli, R. K., Li, D., Prasad, R., & Liu, H. (2015). A new method for prioritizing drug repositioning candidates extracted by literature-based discovery. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 669–674. https://doi.org/10.1109/BIBM.2015.7359766

Ravikumar, K. E., Rastegar-Mojarad, M., & Liu, H. (2017). BELMiner: Adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences. *Database: The Journal of Biological Databases and Curation*, *2017*(1), baw156. https://doi.org/10.1093/database/baw156

Rindflesch, T. C., Blake, C. L., Cairelli, M. J., Fiszman, M., Zeiss, C. J., & Kilicoglu, H. (2018). Investigating the role of interleukin-1 beta and glutamate in inflammatory bowel disease and epilepsy using discovery browsing. *Journal of Biomedical Semantics*, *9*, 25. https://doi.org/10.1186/s13326-018-0192-y

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), Article 5. https://doi.org/10.1038/s42256-019-0048-x

Schroeder, J., Xu, J., Chen, H., & Chau, M. (2007). Automated criminal link analysis based on domain knowledge. *Journal of the American Society for Information Science and Technology*, *58*(6), 842–855. https://doi.org/10.1002/asi.20552

Sebastian, Y., Siew, E.-G., & Orimaye, S. O. (2017). Emerging approaches in literature-based discovery: Techniques and performance review. *The Knowledge Engineering Review*, *32*, e12. https://doi.org/10.1017/S0269888917000042

Singh, A., Bodas, M., Wakabayashi, N., Bunz, F., & Biswal, S. (2010). Gain of Nrf2 function in non-small-cell lung cancer cells confers radioresistance. *Antioxidants & Redox Signaling*, *13*(11), 1627–1637. https://doi.org/10.1089/ars.2010.3219

Smalheiser, N. R. (2005). The Arrowsmith Project: 2005 Status Report. In A. Hoffmann, H. Motoda, & T. Scheffer (Eds.), *Discovery Science* (pp. 26–43). Springer. https://doi.org/10.1007/11563983_5

Smalheiser, N. R. (2012). Literature-based discovery: Beyond the ABCs. *Journal of the American Society for Information Science and Technology*, *63*(2), 218–224. https://doi.org/10.1002/asi.21599

Smalheiser, N. R., & Swanson, D. R. (1996). Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology*, *47*(3), 809–810. https://doi.org/10.1212/wnl.47.3.809

Song, M., Kim, W. C., Lee, D., Heo, G. E., & Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of Biomedical Informatics*, *57*, 320–332. https://doi.org/10.1016/j.jbi.2015.08.008

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

Sun, D., Gao, W., Hu, H., & Zhou, S. (2022). Why 90% of clinical drug development fails and how to improve it? *Acta Pharmaceutica Sinica B*, *12*(7), 3049–3062. https://doi.org/10.1016/j.apsb.2022.02.002

Swanson, D. R. (1986a). 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

Swanson, D. R. (1986b). Undiscovered Public Knowledge. *The Library Quarterly: Information, Community, Policy*, *56*(2), 103–118.

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

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

Sybrandt, J., & Safro, I. (2018). *Validation and Topic-driven Ranking for Biomedical Hypothesis Generation Systems* (p. 263897). bioRxiv. https://doi.org/10.1101/263897

Tabacof, P., & Costabello, L. (2020). *Probability Calibration for Knowledge Graph Embedding Models* (arXiv:1912.10000). arXiv. http://arxiv.org/abs/1912.10000

Thilakaratne, M., Falkner, K., & Atapattu, T. (2019). A systematic review on literature-based discovery workflow. *PeerJ Computer Science*, *5*, e235. https://doi.org/10.7717/peerj-cs.235

Thilakaratne, M., Falkner, K., & Atapattu, T. (2020). Garbage In, Garbage Out? An Empirical Look at Information Richness of LBD Input Types. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 369–372. https://doi.org/10.1145/3383583.3398608

Torvik, V. I., & Smalheiser, N. R. (2007). A quantitative model for linking two disparate sets of articles in MEDLINE. *Bioinformatics*, *23*(13), 1658–1665. https://doi.org/10.1093/bioinformatics/btm161

Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., & Bouchard, G. (2016). Complex Embeddings for Simple Link Prediction. *Proceedings of The 33rd International Conference on Machine Learning*, 2071–2080. https://proceedings.mlr.press/v48/trouillon16.html

Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, *571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8

Tullu, M. S. (2019). Writing the title and abstract for a research paper: Being concise, precise, and meticulous is the key. *Saudi Journal of Anaesthesia*, *13*(Suppl 1), S12–S17. https://doi.org/10.4103/sja.SJA_685_18

van der Eijk, C. C., van Mulligen, E. M., Kors, J. A., Mons, B., & van den Berg, J. (2004). Constructing an associative concept space for literature-based discovery. *Journal of the American Society for Information Science and Technology*, *55*(5), 436–444. https://doi.org/10.1002/asi.10392

van der Heijden, M., Zimberlin, C. D., Nicholson, A. M., Colak, S., Kemp, R., Meijer, S. L., Medema, J. P., Greten, F. R., Jansen, M., Winton, D. J., & Vermeulen, L. (2016). Bcl-2 is a critical mediator of intestinal transformation. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms10916

Vicente-Gomila, J. M. (2014). The contribution of syntactic–semantic approach to the search for complementary literatures for scientific or technical discovery. *Scientometrics*, *100*(3), 659–673. https://doi.org/10.1007/s11192-014-1299-2

Vlietstra, W. J., Zielman, R., van Dongen, R. M., Schultes, E. A., Wiesman, F., Vos, R., van Mulligen, E. M., & Kors, J. A. (2017). Automated extraction of potential migraine biomarkers using a semantic graph. *Journal of Biomedical Informatics*, *71*, 178–189. https://doi.org/10.1016/j.jbi.2017.05.018

Vos, R., Aarts, S., van Mulligen, E., Metsemakers, J., van Boxtel, M. P., Verhey, F., & van den Akker, M. (2014). Finding potentially new multimorbidity patterns of psychiatric and somatic diseases: Exploring the use of literature-based discovery in primary care research. *Journal of the American Medical Informatics Association : JAMIA*, *21*(1), 139–145. https://doi.org/10.1136/amiajnl-2012-001448

Wang, J., Loberg, R., & Taichman, R. S. (2006). The pivotal role of CXCL12 (SDF-1)/CXCR4 axis in bone metastasis. *Cancer Metastasis Reviews*, *25*(4), 573–587. https://doi.org/10.1007/s10555-006-9019-x

Weeber, M., Klein, H., Aronson, A. R., Mork, J. G., de Jong-van den Berg, L. T., & Vos, R. (2000). Text-based discovery in biomedicine: The architecture of the DAD-system. *Proceedings of the AMIA Symposium*, 903–907.

Weeber, M., Klein, H., de Jong-van den Berg, L. T. W., & 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

Wei, C.-H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: Automated concept annotation for biomedical full text articles. *Nucleic Acids Research*, *47*(W1), W587–W593. https://doi.org/10.1093/nar/gkz389

Wilkowski, B., Fiszman, M., Miller, C. M., Hristovski, D., Arabandi, S., Rosemblat, G., & Rindflesch, T. C. (2011). Graph-Based Methods for Discovery Browsing with Semantic Predications. *AMIA Annual Symposium Proceedings*, *2011*, 1514–1523.

Wren, J. D. (2004). Extending the mutual information measure to rank inferred literature relationships. *BMC Bioinformatics*, *5*, 145. https://doi.org/10.1186/1471-2105-5-145

Yang, B., Yih, W., He, X., Gao, J., & Deng, L. (2015). *Embedding Entities and Relations for Learning and Inference in Knowledge Bases* (arXiv:1412.6575). arXiv. https://doi.org/10.48550/arXiv.1412.6575

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

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

Yetisgen-Yildiz, M., & Pratt, W. (2009). A new evaluation methodology for literature-based discovery systems. *Journal of Biomedical Informatics*, *42*(4), 633–643. https://doi.org/10.1016/j.jbi.2008.12.001

Zhang, R., Cairelli, M. J., Fiszman, M., Kilicoglu, H., Rindflesch, T. C., Pakhomov, S. V., & Melton, G. B. (2014). Exploiting Literature-derived Knowledge and Semantics to Identify Potential Prostate Cancer Drugs. *Cancer Informatics*, *13*(Suppl 1), 103–111. https://doi.org/10.4137/CIN.S13889

Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., & Kilicoglu, H. (2021). Drug repurposing for COVID-19 via knowledge graph completion. *Journal of Biomedical Informatics*, *115*, 103696. https://doi.org/10.1016/j.jbi.2021.103696

Zhu, R., Wang, F., Bundy, A., Li, X., Nuamah, K., Xu, L., Mauceri, S., & Pan, J. Z. (2022). A Closer Look at Probability Calibration of Knowledge Graph Embedding. *Proceedings of the 11th International Joint Conference on Knowledge Graphs*, 104–109. https://doi.org/10.1145/3579051.3579072

**APPENDIX A: Example of unprocessed SemRep output**

SE|29871641||ab|5|text|Parthenolide attenuated bleomycin-induced pulmonary fibrosis via the NF-κB/Snail signaling pathway.

SE|29871641||ab|5|entity|C0005740|bleomycin|aapp|||bleomycin|||0|1000|24|32

SE|29871641||ab|5|entity|C0034069|Pulmonary Fibrosis|dsyn|||pulmonary fibrosis|||0|1000|42|60

SE|29871641||ab|5|relation|||C0005740|bleomycin|aapp,antb|aapp|||bleomycin||||1000|24|32|VERB|CAUSES||||||C0034069|Pulmonary Fibrosis|dsyn|dsyn|||pulmonary fibrosis||||1000|42|60