Applying Domain Ontologies and Knowledge Graphs To Augment Literature-Based Discovery: Discovering Gene-Disease Associations Between COVID-19, Diabetes Mellitus, And Chronic Kidney Disease

by

Michael Barrett

Submitted in partial fulfillment of the requirements
for the degree of Master of Health Informatics

at

Dalhousie University
Halifax, Nova Scotia
March 2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Due to the rapid accumulation of scientific evidence related to coronavirus disease 2019 (COVID-19), there is a need to synthesize evidence to help researchers and clinicians understand the pathophysiologic mechanisms that lead to worse outcomes in patients with underlying chronic conditions. COVID-19 patients with Diabetes Mellitus (DM) and Chronic Kidney Disease (CKD) are more likely to experience severe forms of the disease and the reasons why are poorly understood, though the molecular mechanisms driving these diseases may overlap. This thesis proposes an automated knowledge synthesis and discovery framework that analyses published literature to identify and represent underlying mechanistic associations that aggravate chronic conditions due to COVID-19. We take a literature-based discovery approach that integrates text mining, medical ontologies, and knowledge graphs to identify novel pathophysiologic relations between COVID-19 and chronic disease mechanisms by integrating evidence dispersed across multiple literature databases. Our framework applies knowledge graph augmentation methods based on external knowledge (i.e., ontologies) to address the issue of incomplete knowledge captured in relations mined from text (called semantic associations) to improve literature-based discovery of complex mechanistic associations. We applied our approach to discover gene-disease associations between COVID-19, DM, and CKD. We discovered several novel associations that could help identify mechanisms driving the long-term impact of COVID-19 in patients with underlying conditions. We argue that our approach can serve as a useful tool for hypothesis generation by allowing researchers to benefit from the collective knowledge found in both structured and unstructured biological databases.

LIST OF ABBREVIATIONS USED

ACE2 - Angiotensin-Converting Enzyme 2

COVID-19 – Coronavirus Disease 2019

COVID-REdI - COVID-19 Renal and Endocrine Interactions

CKD – Chronic Kidney Disease

DM – Diabetes Mellitus

ESKD – End-Stage Kidney Disease

GO – Gene Ontology

HIF1A – Hypoxia-Inducible Factor 1-alpha

HUGO – Human Genome Organization

KG – Knowledge Graph

LBD – Literature-Based Discovery

LTC – Linking Term Count

MeSH – Medical Subject Headings

MySQL – My Structured Query Language

NLP – Natural Language Processing

PMID – PubMed article Identifier

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus 2

SemMedDB – Semantic Medline Database

T1DM – Type 1 Diabetes Mellitus

T2DM – Type 2 Diabetes Mellitus

TLR4 – Toll-Like Receptor 4

UMLS – Unified Medical Language System

ACKNOWLEDGEMENTS

This project could not have been completed without the help of my family, colleagues, advisers, friends, and countless others. I would like to give my sincerest thanks to Dr. Raza Abidi for offering his mentorship, insights, and guidance throughout this project. I also appreciate the help of Ali Daowd, who answered my (many) questions and helped me in many ways throughout. Further thanks are given to Dr. William Van Woensel for offering his knowledge and expertise. I would also like to extend thanks to Dr. John Harnett for kindly offering his support. Finally, I owe thanks to my parents, Mary and Brendan, whose unconditional love and support helped me throughout this significant task.

CHAPTER 1.0 INTRODUCTION

This thesis explores the use of Literature-Based Discovery (LBD) to develop a framework for uncovering novel biomedical knowledge. LBD is founded on the notion that large bodies of scientific literature contain mutually isolated fragments of interrelated knowledge [1] that can be combined to form novel hypotheses [2]. LBD is an evolving research field that has been successfully used in biomedical research settings to generate actionable insights for unanswered questions [3]–[5]. In the wake of a rapid increase of publications about Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [6], the causative agent of Coronavirus disease 2019 (COVID-19), there is an urgent need to accelerate the discovery of novel findings with regards to disease mechanisms. The effects of COVID-19 in patients with chronic comorbidities such as Diabetes Mellitus (DM) and Chronic Kidney Disease (CKD) is one research area that stands to benefit from this work.

In 2020, 10% of Canadians were living with DM [7] and in 2017 roughly 10% of Canadians were living with CKD [8]. Treating these disorders costs the Canadian healthcare system over $4 billion annually [7], [9]. In addition to the debilitating nature of these conditions, patients with DM or CKD are more likely to die from infections such as COVID-19 [10], [11]. COVID-19 is primarily seen as a lung disease but several recent studies point to extrapulmonary manifestations such as acute multiorgan injuries [12]–[14], which affect many of the same end organs damaged by DM and CKD. Various hypotheses for these overlaps have been proposed including immune dysfunction, cardiovascular comorbidities, and lifestyle risk factors [15]–[17]; however, the underlying mechanisms are presently unclear. It is possible that a combination of immune and metabolic disruptions, and gene-

environment and virus-host interactions—referred to as 'mechanistic associations'—predispose DM or CKD patients to worse COVID-19 outcomes.

The time and effort required to analyze evidence that is distributed across a large number of publications is a major barrier to biomedical researchers in the present context [18]. Increased awareness of this issue has opened the door to new tools and techniques. LBD often includes the use of text mining tools to automatically detect associations between biomedical concepts in unstructured text (e.g., gene-disease associations [19]) and have been used to pinpoint useful information with regards to disease mechanisms [20] and drug mechanisms [21]. These tools are currently part of several initiatives to help the biomedical and applied computing research communities make sense of the COVID-19 literature [22]. Importantly, LBD and text mining are used to automatically find interesting associations that could produce hidden knowledge.

Due to the complexity of COVID-19, researchers may benefit from intuitive content to help them interpret its mechanisms. Lately, visual knowledge representations have been a popular method for decomposing complex disease mechanisms, which comprise symbolic networks of biomedical concepts and their interrelations [23]. One application of this idea is the 'Knowledge Graph' (KG) which depicts concepts and relations as nodes and edges [24]. The advantage of a KG is that it allows rapid identification of important patterns that are indiscernible in traditional databases. The development of KGs concerning COVID-19 is already underway [25], [26], though to our knowledge none have considered specific comorbid conditions such as DM or CKD.

1.1 Problem Statement

Biomedical researchers need a deeper understanding of the underlying mechanisms of COVID-19 to help them make sense of its complex interactions with DM and CKD. We have set out to address this with a LBD framework that approaches the problem from systems medicine [27] and knowledge discovery [28] perspectives. The systems medicine component is a principled way of viewing illness as a complex network of disease states with molecular, physiologic, and pathologic levels. The knowledge discovery component uses LBD techniques to extract mechanistic associations and KGs to represent those associations to find hidden patterns.

1.2 Research Objectives

This thesis will develop an LBD-based knowledge discovery framework to deal with the aforementioned problem through the following objectives:

1) Identify biomedically relevant evidence on COVID-19, DM and CKD published from 2020 onwards
2) Integrate structured knowledge sources based on biomedical literature, ontologies, and public databases to create a KG
3) Extract mechanistic associations and evaluate these associations with respect to their relevance, interestingness, and plausibility

1.3 Research Approach

We pursued these objectives using a LBD approach to 1) collect, screen and analyze scientific articles from web-based servers including PubMed, 2) obtain a computable form of the knowledge contained in those articles and represent it in an interactive graph database, 3) develop methods to filter and rank associations such that relevant and

interesting findings are prioritized, 4) extend the knowledge base with structured biomedical data in the form of ontologies and information in public databases, and 5) extract series of related concepts or 'discovery chains' that contain genetic, molecular, physiologic, and pathologic knowledge to expand on existing hypotheses [29]. This approach was chosen since it allows for integration of knowledge across disparate research communities, thus supporting both clinical medicine and life sciences. The current situation warrants the use of LBD techniques due to an overabundance of literature that has overwhelmed the scientific community. However, this methodology is not specific to COVID-19, DM or CKD and could be applied to other diseases. As a result of the proposed framework, this thesis will proceed along these lines:

1) Design an LBD framework to meet the problem specifications
2) Develop a literature search process to identify evidence-based knowledge that is pertinent to the research questions
3) Leverage an existing database of knowledge extracted from the literature with natural language processing (NLP) and supplement it with external knowledge
4) Integrate those knowledge sources and represent the data in a KG
5) Implement filtering, ranking, and graph-based methods to produce discovery chains

1.4 CONTRIBUTIONS

This thesis describes the use of LBD to create a framework for automatic knowledge discovery. We developed a novel methodology for discovering hidden knowledge to achieve this outcome. Knowledge extracted from the literature, medical ontologies, and public databases was combined to create a semantically enriched KG that could be queried

with data mining algorithms. This allowed us to identify discovery chains that combine heterogeneous information beyond the content of published articles.

Previous studies have explored the use of LBD to automatically discover novel associations with regards to unanswered research questions [3]–[5], but they did so based on the literature only. To our knowledge, this is the first system that combines associations derived from the literature with external knowledge. One study [30] leveraged public databases to supplement literature-based associations with biomedical knowledge; however, it focused on indirect associations from two concepts away, whereas our system integrates data from multiple public databases to extract chains of related concepts.

The methodology described in this work can be used for knowledge discovery in other disease settings and is therefore a new method for synthesizing biomedical knowledge on complex comorbid conditions.

1.5 THESIS LAYOUT

Chapter 2 of this thesis outlines the background information and concepts related to this work. Chapter 3 describes the progression of the research solution from its initial conception through to the final evaluation. Chapter 4 presents the results of the analysis and evaluation, while Chapter 5 discusses the subsequent implications and identifies the potential avenues for future work on this topic. Finally, Chapter 6 summarizes the work and concludes the thesis.

CHAPTER 2.0 BACKGROUND

In this chapter, we review the necessary topics to help the reader understand the context of this work. In section 2.1, we focus on the state of recent clinical and biomedical research on COVID-19, DM, and CKD. In section 2.2, we discuss recent progress of health informatics tools to help translate knowledge into clinical practice. In section 2.3, we explore the topic of systems medicine and its current relevance to complex disorders such as COVID-19 and DM or CKD. Section 2.4 introduces LBD and discusses its applicability to the field of biomedicine. Section 2.5 gives an overview of KGs and their potential to help researchers understand complex disease mechanisms. Finally, in section 2.6 we summarize the chapter and provide a problem statement.

2.1 COVID-19, DM, AND CKD – REVIEW OF RECENT EVIDENCE

Diabetes Mellitus (DM) and Chronic Kidney Disease (CKD) are complex disorders that are intertwined in many ways. Type 2 Diabetes Mellitus (T2DM) and CKD are associated with cardiovascular, body-weight-related, age-related, genetic and environmental factors [31], [32]. Further, DM is the leading cause of kidney disease in the United States and more than a third of all individuals with DM develop CKD [31]. At a systemic level, DM and CKD are characterized by profound metabolic and immune dysregulation, whereby proinflammatory mechanisms trigger insulin resistance and vice versa [33], [34]. Metabolic functions such as glucose and lipid homeostasis are often disrupted in DM, which may lead to maintenance of systemic low-grade inflammation and, ultimately, an impaired immune response in the context of end-stage kidney disease (ESKD) [35]. Importantly, the underlying mechanisms of both DM and CKD are poorly understood, and may involve interactions between multiple genes and pathways [36], [37].

Like DM and CKD, COVID-19 is a systemic disease with many complications [38]. In this thesis, we focus on renal and metabolic comorbidities, though others have been reported elsewhere. SARS-CoV-2 targets angiotensin-converting enzyme 2 (ACE2) that is expressed in many tissues and, importantly, pancreatic β-cells [39]. It is postulated that direct involvement of SARS-CoV-2 in these cells could lead to worsening glycemic control and development of diabetic complications or new-onset diabetes [15]. Further, the virus's tendency to infect the kidneys is of concern since it often occurs in patients with pre-existing chronic conditions and may aggravate DM or CKD as a result [40]. Finally, multiple organs are impacted in both DM and CKD [33], [34], which may predispose some patients given that a history multiple organ dysfunction is a predictor of worse COVID-19 outcomes [14].

A concerning feature of COVID-19 is that it causes intense and, in some cases, systemic immune activation [41]. While the underlying mechanisms are presently unclear, it is postulated that the resultant systemic inflammation (as part of a larger, multifactorial pathogenesis) may be facilitated DM or CKD [15], [42]. Given that DM and CKD are associated with chronic proinflammatory biochemical milieus, it is possible that those patients are more susceptible to the COVID-19-induced inflammatory response [15]. Moreover, in terms of COVID-19 outcomes, there is heterogeneity among DM patients that may be predetermined by clinical history of complications (e.g., hyperglycemia or ketoacidosis), though evidence is still emerging at this time [43], [44].

2.2 Health Informatics For COVID-19

An overwhelming number of scientific publications has added to the stress faced by researchers during the pandemic. A recent study found that an average of 1,682 articles on

COVID-19 were published per week in 2020 in PubMed alone [6]. Researchers in biomedical fields have reported that they simply "don't have time" to keep on top of the evidence let alone absorb its content [18]. It is possible that recent findings related to pathophysiological mechanisms are dispersed across a range of journals and research fields, and a given researcher might not be able to easily locate overlapping evidence related to disease mechanisms based on article titles or abstracts. Another notable issue is the unprecedented speed at which articles have been published, leading to a large number of preprint studies and a wide range of quality in the available evidence [45], [46]. As a result, the time and effort required for researchers to realize connections between their work and existing high-quality evidence presents a significant burden at this time.

To address these issues, a variety of tools have been proposed to help health professionals keep up to date with recent evidence. Online platforms such as the University of Toronto's *Rapid Evidence Access Link* [47] crowdsource questions and quickly provide summaries of credible evidence on COVID-19 to an audience of healthcare decision makers and members of the public. Other applications have been proposed that extract information from articles and public databases represent it in a more manageable format (e.g., graph visualizations) [26], [48], [49]. The *COVID-19 Knowledge Graph* [26] is an exemplary application that displays manually encoded cause-and-effect relations between molecular and pathologic entities in a visual interface. Further, the *COVID-19 Disease Map* uses a community-based approach to involve experts to develop high quality disease models [49]. An important feature of these applications is that they allow users to piece together high-level knowledge (e.g., potential mechanisms) from a vast body of evidence. However, the speed at which these applications synthesize evidence is an issue given that they rely on

manually intensive data integration, which may cause them to lag recently published evidence. Moreover, to our knowledge there has not been an application designed for patients with DM or CKD, despite the fact that these groups are known to be more susceptible to poor COVID-19 outcomes [10], [11], [14].

The deleterious effects of COVID-19 are highly published but remain enigmatic to researchers and clinicians who must simultaneously manage chronic conditions including DM or CKD. Further, mechanistic knowledge continues to be integrated in an inefficient manner such that underlying pathophysiological mechanisms may not be better understood until later in the pandemic. In the following sections, we review knowledge- and data-driven approaches to determine the best solution to help make sense of the complex interactions between COVID-19 and DM or CKD.

2.3 Systems Medicine

Human diseases rarely follow from single biological abnormalities; rather, they are the result of various cross-linked pathological processes that interact in a complex network [50]. Generally, disruptions to this network at a molecular level can affect larger biological systems comprising cells, tissues, organs and other levels of organization, resulting in disease [27]. Systems medicine is an emerging approach that seeks to explain biological mechanisms underlying disease-related phenomena through the use of computational biology, network science, and graph-based visualization [51], [52]. It aims to integrate multiple levels of biological data (e.g., molecular interactions and physiologic processes) together with clinical, societal, and environmental factors to enable the translation of biomedical data back into a clinical setting [27]. In this context, diseases are more precisely

classified than by traditional (i.e., reductionist) methods, and can be interrogated with respect to comorbidity, disease severity and progression [53].

Lately, systems medicine has led key development areas such as predicting disease progression, tying candidate biomarkers to stages of diabetic kidney disease like onset of eGFR decline [54]; individual responses to treatment, showing links between urinary proteomic profile and response to spironolactone in treatment-resistant hypertensive DM patients [55]; and, novel therapeutic target discovery, elucidating a gene regulatory network involved in DM pathobiology [56], integrating multi-omics data on kidney diseases to support therapy-related hypothesis testing [57], [58], development of expert-curated disease maps [23] and multiple-protein biomarker panels [59]. Most recently, systems medicine approaches were used to identify biological pathways that are targeted by environmental stressors (i.e., endocrine-disrupting chemicals) [60], and dysregulated metabolic mediators of immunity [61], both of which may be implicated in severe COVID-19 with comorbid DM or CKD.

It is worthwhile to refocus attention on molecular factors underpinning COVID-19 in the context of DM or CKD. Due to the novelty of COVID-19, recent analyses of potential mechanisms that promote severe COVID-19 and exacerbation of DM or CKD have been limited to a few similar compounds and pathophysiological concepts [62]–[69]. Further, in terms of identifying the determinants of worse COVID-19 outcomes in DM or CKD patients, the predominant approach is focused on late-appearing or generic processes (e.g., inflammation, thrombosis, immunity) and isolated organs or cells (e.g., kidneys, pancreatic β-cells). Loscalzo et al [50] point out that this paradigm neglects underlying pathobiological mechanisms that are, by contrast, not specific to disease states, organ

systems, or phenotypes. Viewed in this way, progression to worse forms of disease is explained by the preponderance of interconnected processes at multiple biological levels. As such, there may be important mechanistic associations involved in COVID-19 and DM or CKD that defy explanation in terms of isolated processes and phenotypes or immediate viral impact (e.g., the effect on ACE2-expressing tissues without investigating related downstream effects).

The notion that diseases are driven by interconnected pathologic processes is supported by recent studies of the molecular behaviour of disease states. Using large-scale datasets and network-based methods, it has been shown in several relevant contexts that disease-associated genes and their products (e.g., proteins, RNAs) form distinct clusters, which interact separately and with each other [70]–[74]. In general, these interactions are used to infer pathophysiologic mechanisms (e.g., pathways). Interestingly, there appear to be similar patterns of clusters in similar diseases [70]. Further, these clusters are associated with known biological functions including those described in biomedical ontologies [75]. For example, a study of yeast used a gene interaction network to infer a set of concepts and relations similar to the Gene Ontology (GO) [76]. In this sense, gene interaction networks may embed hierarchical information consistent with manually curated knowledge of gene functions. Taken together, these findings imply that diseases such as COVID-19 and DM or CKD could be the result of highly structured and interrelated molecular mechanisms.

While systems medicine approaches often use genomic or multi-omic data, in this work we extract information from the literature and public databases. In this sense, the interrelations between genes, proteins, metabolites, compounds, and their locations in organs, tissues, and cells are analogously represented as a computable form of the knowledge found in

published studies or public databases [77]. Our method relies on semantically represented associations between biomedical concepts as opposed to studying statistical correlations (e.g., gene coexpression) or similarity metrics (e.g., phenotype similarity), as is often the case in systems medicine [51]. The details of this approach are described in the following sections.

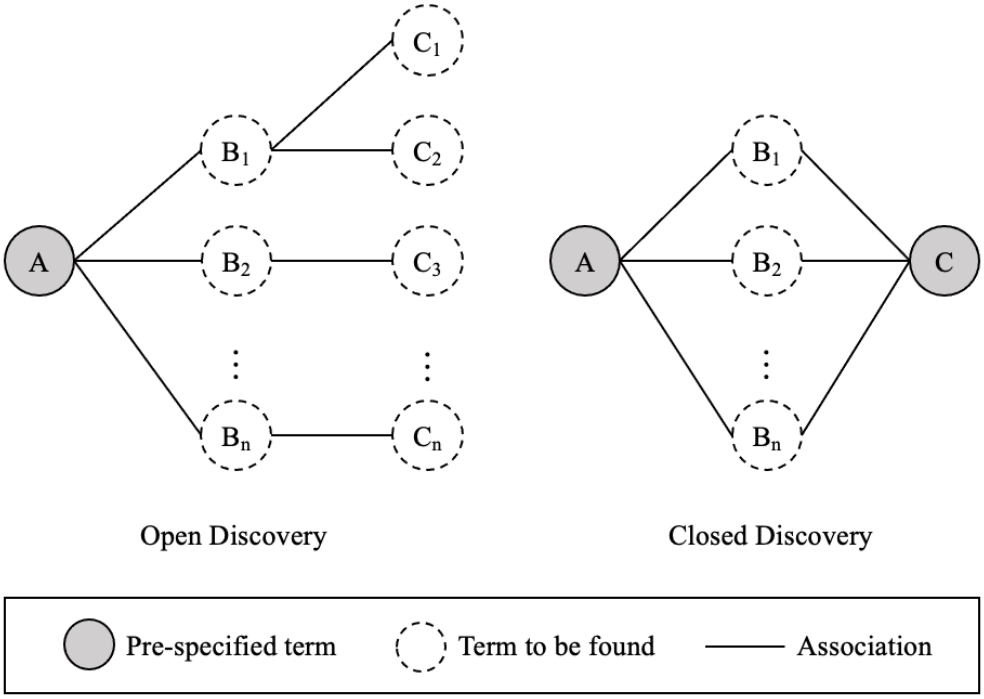
2.4 Literature-Based Discovery

Literature-based discovery (LBD) is an automated effort to expedite knowledge synthesis for the purpose of uncovering hidden knowledge [78]. It often uses text mining to identify relevant explicit knowledge (e.g., gene-disease associations), but the additional goal of LBD is to help researchers generate novel, implicit connections between topics of interest [2]. In this sense, 'discovery' is the result of hypothesis formulation using connectable knowledge fragments that are distributed in the literature [28]. These fragments may exist as (i) hidden refutations or modifications of a hypothesis, (ii) undrawn conclusions from two or more premises, (iii) cumulative evidence of weak tests, (iv) solutions to analogous problems, or (v) hidden correlations between factors [79]. LBD experiments have successfully facilitated the formulation of several hypotheses [80], [81], and prompted follow-up from clinical trials [82], [83]. LBD often takes form as a pipeline of methods, herein referred to as a 'system', that exploits associations between terms in text.

Most LBD systems build on a fundamental premise known as the *ABC discovery model*, which states that explicit knowledge found in text is used to generate direct associations between terms ("A implies B", "B implies C") that are connected using implicit knowledge ("therefore A implies C") to discover indirect associations [84]. A key assumption of this model, as noted by its creator Don Swanson, is that the two terms *A* and *C* should not be

discussed together in text, else there would be some attempt already to discern the mechanisms between them [81]. For example, if one study found that a disease (*A*) is caused by a certain gene (*B*), and a different study (with no connection to the *A* literature) found that *B* interacts with another gene (*C*), then *C* may also be implicated in *A*'s etiology. As such, interesting/novel indirect associations are those that few or no researchers know about since they link disjoint, non-interacting literatures.

Figure 2.1: Open and closed discovery based on the ABC model [85]



The discovery process, shown in Figure 2.1., proceeds through one of two modes; in open discovery, a source term (*A*) is used to generate a set of intermediate terms (B_1, B_2, … B_n) that are in turn used to produce a set of target terms (C_1, C_2, … C_n), creating indirect associations ($A - C$) as a result. In contrast, closed discovery begins with an indirect association and the aim is to generate a set of intermediate terms that interact with both A

and C (i.e., direct associations $A - B$ and $B - C$). Thus, while the first approach emphasizes finding both *B* and *C* terms, the second focuses on finding *B* terms only.

Whereas the purpose of open discovery is to support novel hypothesis generation, closed discovery is useful for elaborating on existing hypotheses [2]. Given two conditions that are indirectly associated by an unknown mechanism, a researcher could employ closed discovery to find an intermediate term (B_1) that shares mechanistic associations with both the source and target terms [4]. If, however, a disease (*A*) lacks explicit connections to biomarkers or genes and is instead directly associated with many comorbidities (B_1, B_2, … B_n), an open discovery approach would help to establish novel, indirect disease-chemical associations [30], [86]. Finally, given a unique disease or phenomenon for which there are few established associations, open discovery could be used to browse through high-level associations without a specific target term in mind and use domain knowledge to select associations to explore in greater detail (i.e., *discovery browsing*) [3]. The advantage of this approach is that it allows for spontaneous discoveries without relying on prior assumptions and is therefore highly generalizable.

2.4.1 ABC Model Limitations

The main limitation of the ABC model is that it creates exponential growth of associations [87], leaving the user to consider an immense number of candidate discoveries. Unfortunately, this issue is twofold in open discovery since the user must evaluate both *B*- and *C*-level associations. Several systems have proposed enhancements to the model's output and a pertinent example is the use of biological contexts as criteria for connecting direct associations [88]. A condition might be established, for instance, to return sets of associations that occur in similar organs, tissues, or cell types, thus reducing the number of

spurious connections [89]. A different method to limit the model's output is to incorporate term rarity. Petric et al [90] propose a system that, unlike previous methods, focuses on finding rare source terms that are then used to identify target terms of interest. Using closed discovery in the last step to find intermediate terms, the authors found a novel indirect association between autism and calcineurin that could contribute to better understanding of the condition.

Given that the ABC model was conceived for general-purpose reasoning, it does not stipulate the kinds of associations between terms. While this is seen as both a strength and a weakness since it is sufficiently vague to be applied to a variety of biomedical questions [1], it is nevertheless a source of ambiguity that hinders making sense of complex (i.e., involving multiple direct or indirect) associations. For instance, a commonly used association type is term co-occurrence, which merely reflects the presence of two terms in a document and not the relationship between them. Modern LBD approaches often rely on semantic associations to solve this issue [84]. Typically, these associations are determined with the help of expert-curated knowledge bases and by applying natural language processing (NLP) tools to free text from the literature [91]. Using semantic associations, the relationship between two terms (e.g., a gene and a disease) is made explicit in the form of a natural expression (i.e., a gene *causes* a disease). As such, current systems often make use of NLP methods to extract relationships between terms since these associations can be better interpreted or explained [84].

Semantics-based methods allow researchers to specify not only the kinds of associations but also logical patterns of associations leading to a coherent indirect association (i.e., *discovery pattern*) [92]. For example, consider the following 'may disrupt' pattern [93]:

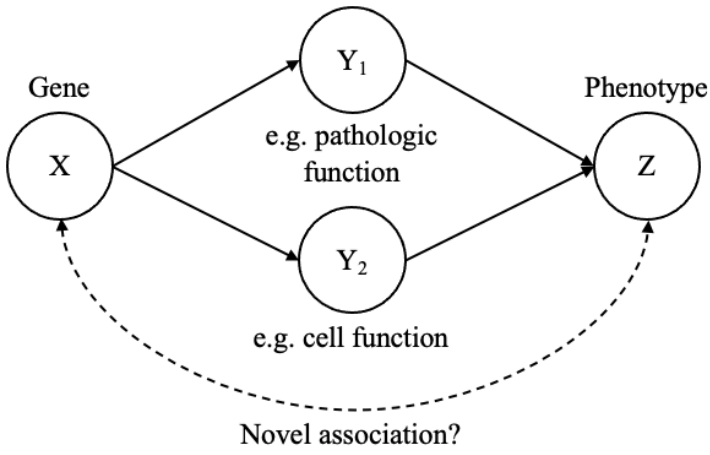
Substance X – inhibits – *Substance Y*

Substance Y – causes – *Pathology Z*

Substance X – may disrupt – *Pathology Z*

This pattern describes the potential therapeutic effect of *X* (e.g., a novel drug compound) whereby its action inhibits the pathologic effect of *Y* (e.g., a disease gene). While discovery patterns have traditionally been used to link drugs to target diseases [91], they were also recently adapted to explore gene-phenotype associations, which could help to explain underlying disease mechanisms [94]. An illustration of the gene-phenotype discovery pattern is shown in Figure 2.2.

Figure 2.2: Illustration of the gene-phenotype discovery pattern [94]

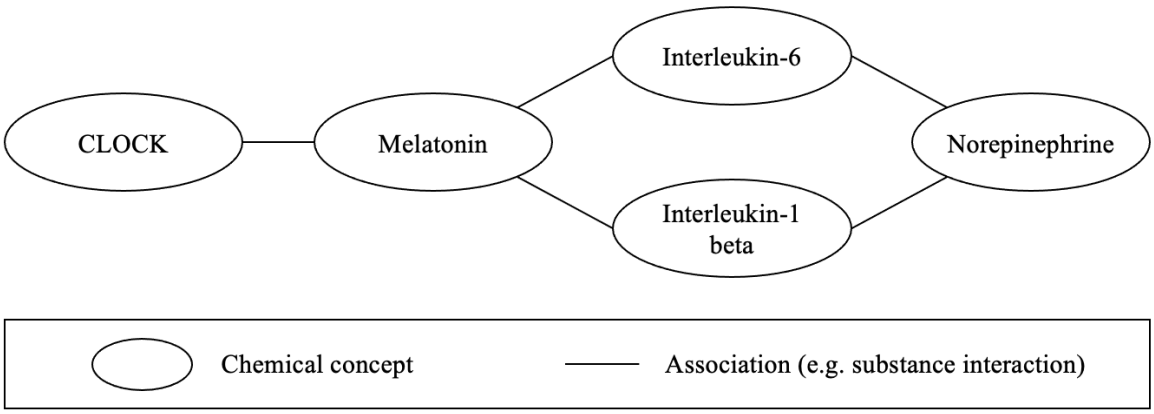


Pattern-based methods facilitate the discovery process, in these cases finding potential new drug treatments or disease mechanisms. In essence, discovery patterns use semantic associations to represent biomedical relationships between terms, allowing researchers to mine plausible indirect associations. At the same time, however, these patterns have only been used to find associations based on the ABC paradigm thus far, meaning that interesting and more complex associations will go undetected [95].

2.4.2 Discovery Chains

Finally, the ABC model does not address mechanisms involving more than three terms, though they are likely to exist in biomedicine. Wilkowski et al [96] proposed a solution to extend the model whereby source and target terms are linked by a series of intermediate terms ($A - B_1 - B_2 - \ldots - B_n - C$), referred to as a *discovery chain*. They selected terms by analyzing these chains in a graph to produce a subgraph, which is shown in Figure 2.3. As a result, they found a previously unnoticed set of mechanistic associations underlying depressive disorder.

Figure 2.3: Discovery chain subgraph [96]



In a different study, discovery chains were used to study metabolite-related disease pathways involving sequential biochemical interactions [29]. There, the authors were able to add two new associations to an existing hypothesis, which could have implications for underlying biological pathways or mechanisms. Importantly, discovery chains overcome the simplistic nature of the ABC model where only a single intermediate term is permitted [85]. Viewed another way, chains of associations can be used to find indirect links between distant literatures where novelty (and potential impact) of discoveries increases as a function of the chain's length [97].

2.4.3 Review Of Literature-Based Discovery Methods And Trends

LBD systems often proceed with a similar workflow where the first step is to select sections of articles (i.e., titles, abstracts, or full texts) for processing [28]. Titles and abstracts are the most commonly used sources of information, and there may not necessarily be an advantage to adding extra information (i.e., abstracts or full texts) [98], though it has been shown that document length has a greater effect on results than the number of articles included [99], [100]. Then, the researcher must decide how to extract and represent the knowledge.

2.4.3.1 Overview Of LBD Systems

We define five types of LBD systems that are characterized by the way they represent terms and associations found in text [101]:

1) Co-occurrence-based systems represent terms as single words or map them to ontology concepts. A direct association between two terms is defined as co-occurrence in the same document. Indirect associations are found manually or by using statistical models to combine different co-occurrence pairs.
2) Semantics-based systems represent terms as ontology concepts. Direct associations are extracted from text with the use of natural language processing (NLP). Unlike co-occurrence associations, semantic associations provide the meaning of the relationship between two concepts. Indirect associations are found manually using expert background knowledge or semi-automatically using discovery patterns.
3) Distributional systems represent terms as vectors using co-occurrence information. Direct associations are defined as co-occurrence or semantic associations and indirect associations are found automatically using vector operations and machine learning.

4) Graph-based systems represent terms as ontology concepts and depict them as nodes in a graph. Direct associations are defined as co-occurrence or semantic associations and denoted as edges between two nodes. Indirect associations are found semi-automatically as in the semantics-based approach or automatically by applying path-finding algorithms that search for series of related concepts (i.e., discovery chains).
5) Hybrid systems represent terms as ontology concepts and direct associations are defined as co-occurrence or semantic associations. To find indirect associations, these systems use some combination of previous methods (i.e., statistical, semantics-based, distributional, graph-based).

To help the reader understand different methods, we show an analysis of LBD approaches based on system types in Table 2.1.

Table 2.1: Analysis of LBD approaches based on system types

Approach	**Concept representation**	**Relation representation**	**Discovery method**	**Comments**
Co-occurrence-based	Words/ontology concepts	Mentioned in same document	Statistical analysis	High rate of false positives
Semantics-based	Ontology concepts	Semantic associations	Discovery patterns	Not highly scalable
Distributional	Words/ontology concepts	Co-occurrence or semantics	Machine learning	Difficult to interpret
Graph-based	Ontology concepts	Co-occurrence or semantics	Path-finding algorithms	Lack of agreed upon metrics
Hybrid	Ontology concepts	Co-occurrence or semantics	Combination of previous methods	Suitable for complex associations

2.4.3.2 Current Trends In LBD

Viewed over time, this ordering roughly resembles the development of LBD towards contemporary approaches [101]. The progression of LBD systems from co-occurrence-based (earliest) to hybrid approaches (latest) reflects an increasing trend towards rich concept representation [85]. While many early systems extracted terms directly from text, modern approaches automatically map terms to concepts from knowledge sources like the Gene Ontology (GO), Medical Subject Headings (MeSH), and the Unified Medical Language System (UMLS), a domain-independent collection of medical ontologies [29], [102]–[104]. The UMLS also contains relations between ontologies, which have been used to integrate disparate biomedical knowledge sources [30] and to mine semantic patterns for predicting new associations (e.g., drug treatments) [91]. Some newer systems also incorporate information from biomedical databases like UniProt for protein interactions, Gene Expression Omnibus (GEO) for disease-specific gene expression, Kyoto Encyclopedia of Genes and Genomes (KEGG) for metabolites and pathways, and Comparative Toxicogenomics Database (CTD) and DrugBank for chemicals and drugs [30], [105]–[107]. A notable example of the database-based approach is the Biomine system [105], which combined information from GO, KEGG, and several other sources to predict protein interactions and gene-disease associations. Importantly, public databases can be used to augment literature-based associations with external knowledge and to prioritize interesting concepts from databases that are currently overflowing with candidate genes, proteins, or other substances [108].

The use of domain knowledge to direct the LBD system is another important trend seen in recent approaches. This can be seen in the discovery pattern technique [93] that relies on known mechanisms between terms, and the discovery chain technique [29], [96] that relies

on user or expert input. Further, LBD systems that combine a variety of methods (e.g., knowledge-based, semantics-based) are better equipped to deal with complex associations and may be more capable of finding hidden knowledge compared with any single approach [85]. For instance, Cameron et al [95] combined semantic associations with semantically integrated expert knowledge to bridge knowledge gaps in a graph, allowing them to recover supplementary associations from an existing drug discovery problem. Importantly, this study showed that standalone techniques such as semantics, unless supplied with external knowledge, may fail to capture complex mechanisms.

2.4.3.3 Dealing With Spurious Associations

All LBD systems are faced with the issue of uncontrolled growth of candidate terms and each handles it in different ways. This step of the LBD workflow is referred to as *filtering/ranking* and involves applying metrics or heuristics to narrow the list of associations to those that are most relevant by removing spurious, general, uninteresting, or noisy terms/associations [28]. Whereas co-occurrence-based systems produce the largest and noisiest result sets, semantics-based systems achieve the highest precision at the expense of a large number of associations [84]. One of the earliest filtering methods in LBD is *semantic category filtering*, which assigns each term to a semantic type or group based on the Unified Medical Language System (UMLS) [28], [30]. For example, the term 'acute kidney injury' belongs to the semantic type *Injury or Poisoning*, and 'diabetes mellitus' belongs to *Disease or Syndrome*, both of which are part of the semantic group *Disorders* [109]. Filtering proceeds by selecting desired semantic types or groups to restrict the intermediate and target concepts returned during the discovery process [28]. Semantic category filtering is challenging to do correctly and relies on heuristics to avoid leaving out

interesting terms or including uninteresting terms whereby the filter is too narrow or broad, respectively [28].

Another filtering method based on the UMLS is *relation/predicate type filtering* that mostly considers direct associations assigned by the NLP tool SemRep [28]. These associations, referred to as *semantic predications*, are subject-predicate-object triplets based on the UMLS Semantic Network [110] where terms are mapped to ontology concepts (with corresponding semantic types) and assertions between subject and object concepts are represented by a high-level predicate type. For example, given the sentence 'The capacity for autophagy in both podocytes and renal tubular cells is markedly impaired in type 2 diabetes', SemRep provides:

1) Tubular Cells (subject) – LOCATION_OF (predicate) → Autophagy (object)
2) Podocytes (subject) – LOCATION_OF (predicate) → Autophagy (object)

These predications are referring to the fact that Autophagy (of type *Cell Function*) occurs in both Tubular Cells and Podocytes (both of type *Cell*). In general, predicates may refer to disease etiology (e.g., CAUSES, PREDISPOSES, ASSOCIATED_WITH), treatment (e.g., PREVENTS, TREATS), comorbidity (e.g., COEXISTS_WITH), molecular interactions (e.g., STIMULATES, INHIBITS), pharmacogenomic/physiologic processes (e.g., AFFECTS, DISRUPTS), and anatomic or static relations (e.g., LOCATION_OF, PART_OF, ISA). Relation/predicate type filtering techniques include removing certain predicates (e.g., negated predicates [111]), considering the direction of predications [29], and restricting the semantic type of the subject or object (e.g., using discovery patterns) [106]. Finally, *synonym mapping* involves grouping exactly or nearly related terms based

on resources like the UMLS to reduce the results [112]. In this technique, equivalent associations are merged, thus reducing the number of redundant associations [113].

Ranking or thresholding methods are used to downweigh or remove uninteresting or noisy associations and prioritize interesting or important associations when ordering a system's results [28]. The earliest approaches to this task relied on *conventional lexical statistics measures* such as frequency of occurrence or cosine similarity [29] and *non-conventional statistical measures* like linking term count (LTC), which considers the number of intermediate (*B*) terms where the source and target terms are connected by a single layer of *B* terms [114]. In contrast, distributional systems represent associations as vectors based on the underlying semantics and rank them using algorithms such as *nearest neighbours* [115]. Interestingly, this method eliminates the need to evaluate *B* terms, making it highly scalable; however, it has been criticized for producing results that are difficult to interpret [85].

Another notable technique is the use of *graph-based measures* such as degree centrality [77], [96] and PageRank [116] to score associations. This approach involves analyzing node- and edge-level attributes to prioritize highly connected terms [96] or to establish a minimum threshold of connectedness [77]. Alternatively, graph-based measures can be used to establish a maximum threshold of connectedness, whereby nodes that exceed a given score (i.e., highly-connected or overused terms like 'cell') are excluded [107]. While graph-based measures are a popular method [77], [96], [117], there is no consensus on whether a single measure or a combination of measures is more effective. Further, given that conventional lexical statistics and graph-based measures focus on frequency-based information from the literature to rank associations, it can be argued that they prioritize

well-known facts while disregarding latent, low-frequency phenomena that may hold more potential for discovery [118].

2.4.3.4 Output Representation And Evaluation

The final steps of the LBD workflow are *output representation* and *evaluation*. The output of an LBD system is often presented as a ranked list of associations [30], [119], which is problematic as it makes it difficult for researchers to understand the overall findings and to see how associations are linked with the source and target terms [28]. As such, producing intelligible visualizations of the system's output is an important technique that can reduce the burden of interpreting the results. *Graph-based visualizations* are one method that has been used to clearly illustrate LBD findings [28] and an advanced visualization technique was proposed by Cameron et al [88] whereby the results were decomposed into thematic, context-specific excerpts (i.e., subgraphs) and presented to the user. Subgraphs are an elegant way of representing complex associations between biomedical concepts [85]. Other methods include *using existing systems* such as Semantic Medline, which represents biomedical knowledge from Medline articles as a graph of semantic predications with links to images of the relevant text [4].

Once the output is in an acceptable format, the ultimate step in LBD is to evaluate the system's results. One such approach is *evidence-based evaluation*, which involves declaring the accuracy of each association based on reliable sources such as existing discoveries, literature, or public databases [28]. The most common technique is replication of existing medical discoveries (e.g., Swanson's Raynaud-Fish oil or Migraine-Magnesium findings) using only the literature published before those discoveries were made [88], [95], [113]. Discovery replication is limited in the sense that (i) it only proves the system is capable of rediscovering past *A-C* pairs and does not assess the actual associations in the

output, (ii) it is based on known associations and prone to user bias as a result, and (iii) it lacks generalizability and risks overfitting the LBD system [28], [85], [120].

A more objective method was therefore proposed by Yetisgen-Yildiz et al [120], which involves using future co-occurrence between terms as a standard for evaluation. In this sense, a system's performance is measured by its ability to predict future co-occurrence between terms that have never been mentioned together in the same document, using a time cutoff to define a point after which a co-occurrence reference set is constructed and compared with the pre-cutoff predictions [120]. The performance of the LBD method can then be evaluated using information retrieval metrics such as precision (i.e., fraction of associations in the reference set), recall (i.e., fraction of reference associations returned), F-score (harmonic mean of precision and recall), and Area Under Curve (AUC) [28].

The advantages of this approach are that it provides a quantitative measure by which to compare LBD systems and that it can be repeated with any set of previously known or unknown associations [28]. However, using co-occurrence as a placeholder for scientific discovery has been criticized for glazing over established findings elsewhere (e.g., in other literature or patent disclosure databases) [121]. An alternative evidence-based evaluation method is to compare the output against human-curated public databases, judging the LBD system by how many associations (e.g., drug-disease interactions) it can predict in a given database [111]. Unfortunately, public databases are never completely accurate and are subject to erroneous data entries for a variety of reasons [122]–[124]. Further, public databases such as DrugBank and the UMLS are limited in the sense that they do not comprise all possible associations between concepts [107]. Indeed, the creation of a gold

standard dataset of all knowledge and potential discoveries now and in the future is likely an impossible task, thus hindering precise, quantitative comparison of LBD systems [84].

Another evaluation approach is *expert-based evaluation*, which uses expert background knowledge to validate the system's results [29], [119]. This typically involves receiving input from one [119] or two [29] domain experts as to whether the suggested associations are meaningful or relevant. Aside from claiming the relevance of a system's findings, which is nevertheless subjective [28], it is also important to *evaluate the interestingness of the results*. Though there are few established methods to this end in LBD, in theory, an interestingness measure is applied to help identify novel associations [87]. Cameron et al [88] used a statistical measure (i.e., association rarity) to determine how often associations were mentioned in Medline articles, where those that were never mentioned were considered to be rare, and thus interesting to a given user. The limitation of this approach, as they go on to demonstrate, is that some rare associations are trivial and therefore uninteresting.

2.4.4 Review Of Literature-Based Discovery Systems In Biomedicine

One of the most common applications of LBD in biomedicine is drug discovery [125]. In one study [126], the authors extracted a set of genes and drugs related to prostate cancer from the literature to discover novel treatment options. Consequently, they identified 3 known prostate cancer drugs and 18 potential drugs that had not previously been used to treat the disease. More recently, Bakal et al [91] applied LBD and machine learning techniques to uncover causative and treatment associations between biomedical concepts. Using graph-based and distributional methods, they found five new *treatment associations* between medications and clinical phenotypes and three new *causative associations*

including: (1) Human Metapneumovirus *causes* Systemic Lupus Erythematosus, (2) Maternal Fetal Infection Transmission *causes* Autoimmune Diseases, and (3) Human Herpes Virus 6 *causes* IgG Gammopathy.

Several important papers have reported non-drug-related discoveries as well. An exemplar study by Srinivasan et al [102] began with two disjoint literature sets about (i) serum and (ii) salivary proteomes of DM patients. They developed a systematic search strategy based on critical components of DM (e.g., insulin resistance) with strict inclusion criteria to maximize recall and precision in PubMed. Using closed discovery, they automatically extracted a set of salivary biomarkers that could improve the prediction of DM. A different study [119] looked at two non-interacting literatures about (i) Alzheimer's disease (AD) and (ii) gut microbiota. They viewed closed discovery as a cross-domain issue whereby intermediate (*B*) terms are more likely to be found in highly unique articles that link two literature sets and used this approach in combination with expert-assisted filtering to narrow the search for interesting *B* terms. As a result, they were able to discover a novel biochemical association between AD and gut dysbiosis.

Rindflesch et al [103] began with an existing indirect association between inflammatory bowel disease and epilepsy. Given that the two diseases were known to be linked by a common biomarker, they used closed discovery to find other intermediate terms to expand the hypothesis. By looking for terms that also had an interesting relationship with the existing biomarker based on evidence, they were able to discern potential mechanistic associations. Finally, Sedler et al [104] used LBD and machine learning to discover sex-specific risk factors for the impact of smoking on bodily function. They combined biochemical, physiologic and pathologic data and were able to predict plausible

associations between smoking, cognition, and cardiovascular health. What both preceding studies have in common is their use of semantic predications to uncover interesting patterns relating to disease mechanisms, which could be explored in further depth using researchers' domain knowledge.

2.4.4.1 Related Work

Advanced LBD methods can find complex molecular associations involving a series of related concepts to discover unnoticed links between disease mechanisms. In previous work, Cameron et al [127] supplemented predications found in text with structured domain knowledge (i.e., ontology relations), allowing them to extract complex associations as chains of related biomedical concepts. Their approach considered a type of association not found in text by forming *intermediate relations* from two concepts away (i.e., A – B – C where B is an ontology term and A and C are predication concepts). In a more recent study, Cameron et al [88] extended their approach by integrating semantic associations with PubMed article MeSH terms to automatically identify meaningful associations in the form of context-based subgraphs. In this sense, they used ontologies to indirectly control the growth of associations. A limitation was the level of prior knowledge required to direct the system, which they felt could be overcome with additional background (external) knowledge. Further, Bakal et al [91] used ontology relations to predict semantic associations to uncover novel causative associations from predications mined from the literature, identifying indirectly related concepts based on similar contexts. Finally, Vlietstra et al [30] proposed an automatic biomarker discovery method that linked genes with disease concepts through annotations from public databases and relations found in text, which could improve the pathophysiologic understanding of migraine. It should be noted that previous works do not utilize the underlying connections between ontologies

and predications by integrating multiple ontology relations from background (external) knowledge to augment semantic associations to infer novel biomedical relations.

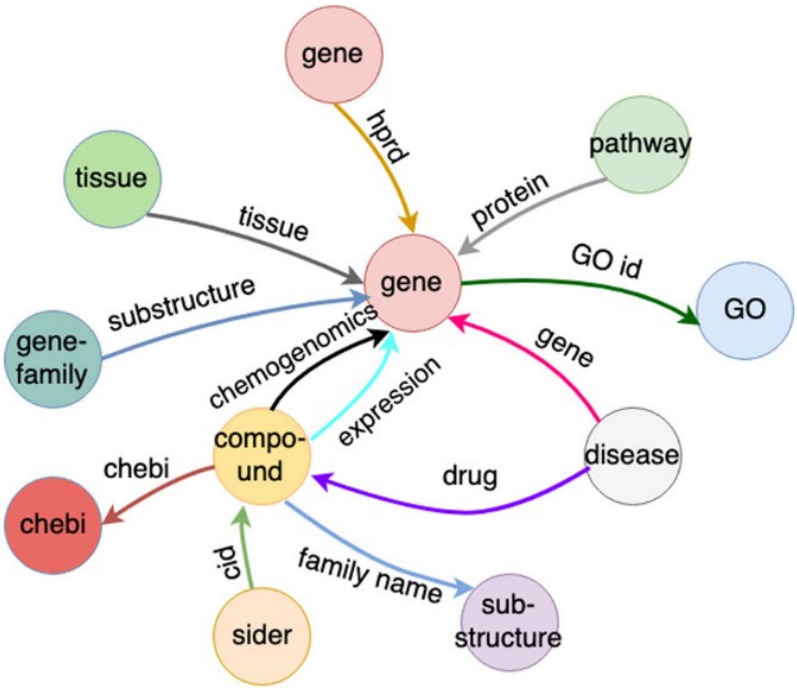
2.4.4.2 Requirements And Challenges

Taken together, these studies demonstrate the strength of LBD with respect to biomedical discovery and highlight several requirements and challenges. First, the creation of a robust evidence base via knowledge identification and screening; second, the fusion of disparate knowledge sources in a systematic way; third, the balance between prior knowledge and openness to new information; fourth, the use of a modified ABC model to extract highly relevant terms; fifth, the need to guide selection of candidate discoveries with expert advice and domain knowledge; sixth, the need to analyze novel hypotheses in the context of existing evidence.

2.5 KNOWLEDGE GRAPHS

A Knowledge Graph (KG) is a repository of semantically interrelated concepts that has a wide range of uses in biomedicine. An example of a biomedical knowledge graph structure is shown in Figure 2.4.

Figure 2.4: Biomedical knowledge graph example [128]



KGs integrate one or more knowledge sources and represent biomedical concepts and relations as nodes and edges [24]. This modeling strategy is used to represent subject-predicate-object triples according to basic standards such as resource description framework (RDF), thus allowing seamless transfer of information between KGs [129]; or to represent triples in a labeled property graph database like Neo4j [94]. For example, a semantic predication can be represented using RDF notation (i.e., UMLS concept (subject) – relation (predicate) → UMLS concept (object)) or using labeled nodes and edges (i.e., UMLS concept (node) – relation (edge) → UMLS concept (node)). The relative strength of KGs as contrasted with relational databases is that they include relations between concepts in their data structure, thus allowing users to easily find complex patterns or chains of concepts.

As an elegant solution to big biomedical data, KGs have been exploited for informatics and data mining tasks such as hypothesis formation [130] and data modeling [131]. For example, in [94] the authors combine literature and wet lab (i.e., high-throughput) datasets in a graph to predict clinically relevant genotype-phenotype associations. More recently, KGs have been used to represent large-scale heterogeneous datasets including genes, proteins, metabolites, drugs, diseases and phenotypes to predict complex biochemical interactions [128], [132]. Importantly, KGs complement systems medicine and LBD by representing associations from multiple evidence-based sources.

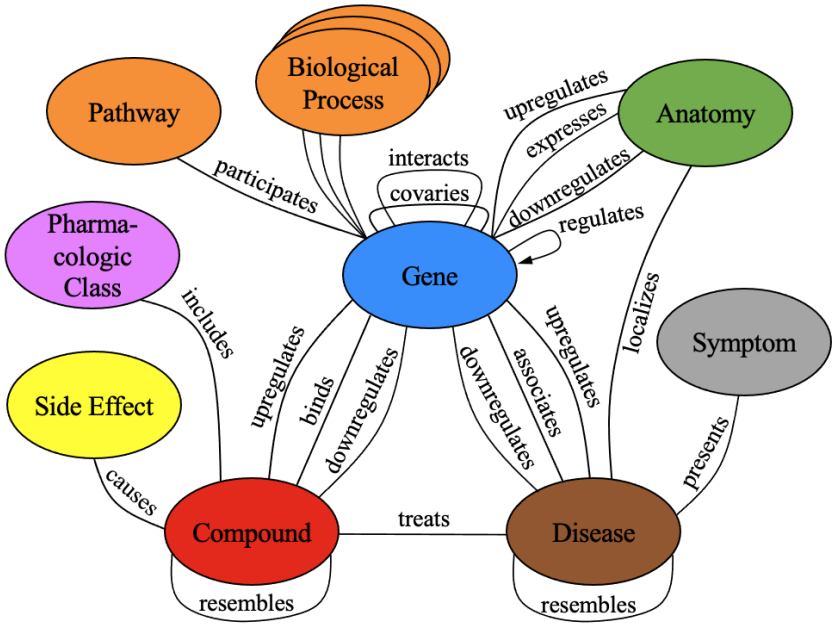
Unlike previous graph-based methods including those discussed in section 2.1, KGs do not define associations between concepts (i.e., genes, proteins, diseases, etc.) using only statistical- or similarity-based metrics. Moreover, KGs contain complex associations that cannot be captured by other representations such as hierarchies. Many KGs incorporate clinically and biologically meaningful relationships that are extracted from the literature, public databases, and ontologies [24]. These may refer to etiology, substance interactions, phenotypic manifestations, pathophysiological effects, and others [104], [130]. The advantages of this approach are that (i) it reduces the number of spurious connections between concepts and (ii) it preserves the context of each association [112]. As such, KGs are a rich representation of knowledge based on clinical and biological observations or facts.

2.5.1 Review Of Knowledge Graphs In Biomedicine

KGs have been used in biomedical research to perform drug repurposing [133], elucidate gene-disease associations [134], [135], and discover biomarkers [77], [136]. Himmelstein et al [133] integrated semantically represented knowledge from multiple public databases

to predict the probability of a set of drugs to treat epilepsy, which provided treatment insights. Their KG ('Hetionet'), shown in Figure 2.5, was particularly notable in that it brought together molecular, anatomical, physiological, and pathological data. Chen et al [134] integrated disease-phenotype relations from medical ontologies with genetic functional information from a public database to discover new therapeutic targets for Parkinson's disease. They compared their approach with similarity-based methods and showed that theirs performed significantly better at predicting gene associations across nine broad disease classes, especially with regards to complex diseases that share phenotypes with many other conditions. In a more recent study [135], the authors developed a KG from various evidence-based sources to predict gene-disease associations by computing semantic similarity between phenotypes and gene-phenotype associations. They found that this method was effective for identifying relevant genes that are directly or indirectly associated with disease, which could improve the prediction of disease mechanisms for rare or poorly understood conditions.

Figure 2.5: Schema of the Hetionet knowledge graph [133]



Cairelli et al [77] represented semantic associations from the literature in a KG to discover biomarkers for mild traumatic brain injury. Using graph-based metrics, they found a set of 17 potential biomarkers, comprising known and unknown associations. More recently, Vlietsra et al [30] combined semantic associations from the literature with manually curated associations from several public databases to extract potential migraine biomarkers. They demonstrated the use of the UMLS MRREL dataset to create a structured semantic graph, allowing them to retrieve 73% of the compounds from a reference set. Lastly, van Bilsen et al [136] created an integrative KG with evidence from the literature and multiple public databases to predict biomarkers of early life immune modulation. Using gene-disease and gene-function associations, they were able to infer connections between genes and immune health in six different phases of early life.

The aforementioned studies show that KGs, as an emerging concept in biomedicine, have a number of strengths and practical use cases. KGs are especially powerful when faced

with an immense amount of evidence but are also effective in making novel discoveries with limited information at hand. Integrating multiple knowledge sources causes biochemical data to be contextualized in a larger disease network, which is used to identify meaningful patterns that would otherwise be obscured by using a single dataset. These patterns connect heterogeneous concepts such as genes, drugs, phenotypes, diseases, biological processes, and anatomical structures in a way that is intuitive, unambiguous, and makes them amenable to data mining tasks. As a result, KGs can be used to harness the complexity of modern biomedical data to elucidate pathobiological mechanisms underlying various diseases.

2.6 Summary And Problem Statement

The interactions between COVID-19, DM and CKD are a major public health concern. Each condition is associated with multiple interconnected comorbidities and physiological disturbances, which may associate through similar severe manifestations (e.g., renal damage, hyperglycemia) [42], [137]. Yet, the underlying causes of severe COVID-19 and the consequences for DM and CKD patients are still unclear. Traditional medicine aims to decompose these illnesses into discrete phenotypes when in reality the underlying causes of disease are intricately connected. Systems medicine aims to exploit this complexity by viewing the entire network of molecular perturbations that gives rise to cross-linked pathological processes at multiple levels of biological organization. Applying this thinking to the immense body of literature on COVID-19, we aim to uncover hidden or neglected knowledge that could improve our understanding of the disease and how it affects or is affected by DM and CKD.

Presently, there is an overwhelming amount of literature from which to find associations about COVID-19 and DM or CKD. LBD is a state-of-the-art text mining approach that has been used to help researchers develop novel hypotheses. Recently, advanced LBD systems have shown promising results with respect to disease-related mechanistic discovery, making LBD well suited to the current context. As previous studies have shown, LBD can be applied with certain modifications to obtain highly relevant associations, thus providing novel insights to pathophysiological mechanisms. Finally, representation of biomedical knowledge in KGs is an emerging research area that complements the systems medicine and LBD paradigms. As rich sources of molecular, physiological, pathological, and anatomical information, KGs are an attractive tool for biomedical research. These findings provide the necessary foundation to propose a solution for discovering mechanistic associations related to COVID-19 and DM or CKD.

CHAPTER 3.0 METHODOLOGY AND METHODS

Given that biomedical researchers have reported feeling overwhelmed by the amount of literature on COVID-19, there is a need for tools that expedite knowledge discovery. Further, the approach to discovery must be based on credible evidence and should ideally draw from multiple knowledge sources due to the current paucity of mechanistic associations in the literature.

As discussed in the previous chapter, LBD can be used to uncover novel associations between knowledge fragments present in published literature. To meet our research objectives, this thesis develops a LBD system for discovering mechanistic associations between COVID-19 and DM or CKD through the following research activities:

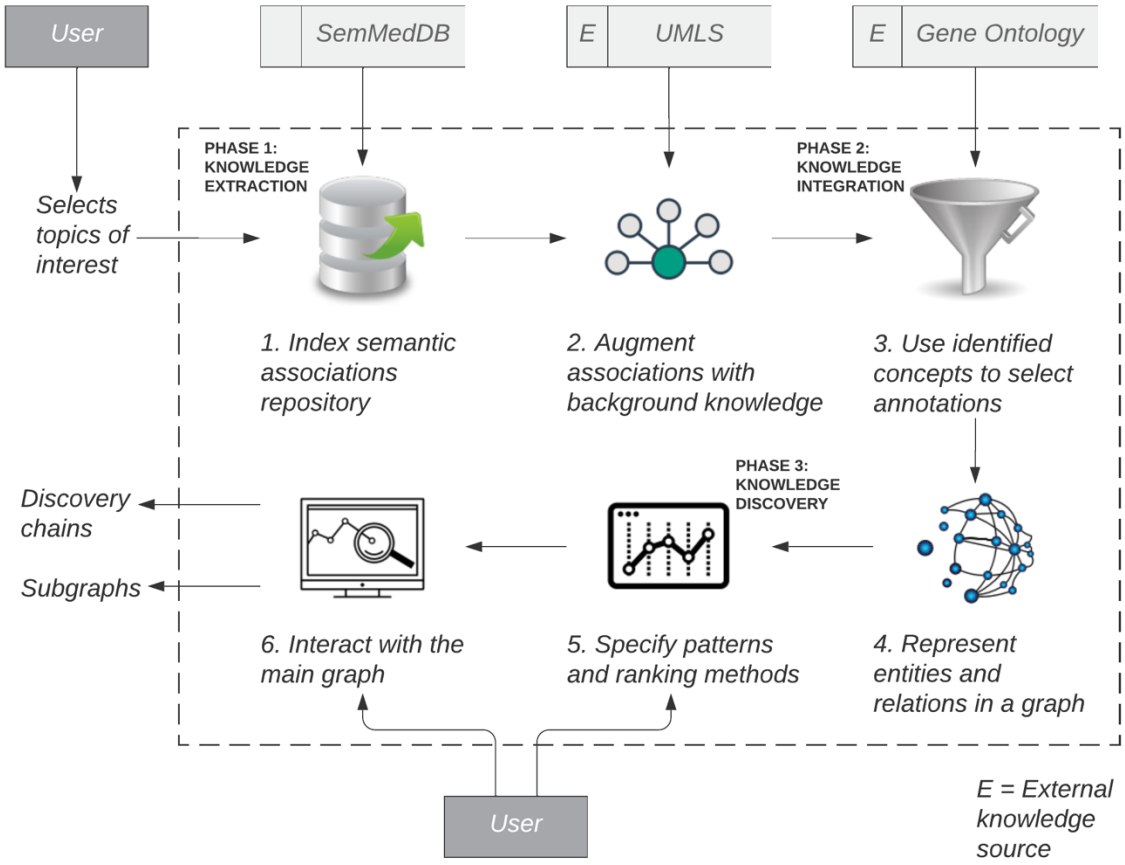
1) Synthesize biomedically relevant evidence on COVID-19, DM and CKD published from 2020 onwards
2) Integrate structured knowledge sources based on the literature and public databases to create a KG
3) Extract mechanistic associations and evaluate these associations with respect to their relevance, interestingness, and plausibility

3.1 METHODOLOGY

Our approach aims to address the issue of discovering limited knowledge by traditional LBD systems. In our approach, we extended existing LBD strategies to improve the discovery of (mechanistic) associations from the literature. Our strategy is to augment semantic associations (i.e., predications) retrieved from the literature using existing text mining sources with structured background knowledge from ontologies (i.e., implicit relations that are *not found in text*) [138]–[140] to discover new knowledge. Our approach

aims to (i) integrate knowledge sources to identify associations that are dispersed across multiple databases, (ii) extract implicit relations that remain hidden in background knowledge, and (iii) augment predications to discover plausible gene-disease associations. We implemented our strategy as an LBD framework [85] shown in Figure 3.1.

Figure 3.1: Schematic of the proposed LBD framework showing the 3 phases with their constituent activities



Our strategy combines knowledge from multiple evidence-based sources to identify plausible indirect associations between yet unconnected concepts by producing sequences of explicit and implicit relations (i.e., discovery chains). In this regard, our method aims to discover meaningful semantic associations between seemingly unrelated concepts where connections are dispersed across multiple knowledge sources.

We argue that including external knowledge will increase the number of mechanistic associations being found, and potentially the number of discoveries being made. Whereas previous LBD strategies [77], [96], [103] rely on expert knowledge to single out interesting discoveries, we build on related works [30], [88], [95] that demonstrated the potential of using structured background knowledge to supplement associations found in text. Unlike related works, however, we provide the means to draw out hidden discoveries as complex mechanistic associations between the literature and public databases.

The activities involved in our approach proceed in three phases as follows:

Phase 1 Knowledge Extraction

Activity 1 – Predication Extraction: Use article identifiers or keywords to index an existing repository of semantic associations, such as the Semantic Medline Database (SemMedDB), to extract relations found in text (i.e., predications) [141].

Activity 2 – Predication Extension: Extend the semantic associations using ontology relations from an external knowledge source such as the Unified Medical Language System (UMLS) [140] to extract common knowledge relations involving alternative (i.e., related) concepts, including Gene Ontology (GO) terms.

Phase 2 Knowledge Integration

Activity 3 – Annotation Selection And Pruning Annotations: Use the extracted concepts to index annotated entities from external knowledge, such as GO annotations [142], to select annotations. Repeat the previous activities with the identified concepts until no new associations are found to generate complex associations. Prune the resulting associations using ontology-based methods (e.g.,

[127], [143], [144]) to mitigate irrelevant information. A detailed diagram of the annotation selection and pruning activities is shown in Figure 3.2.

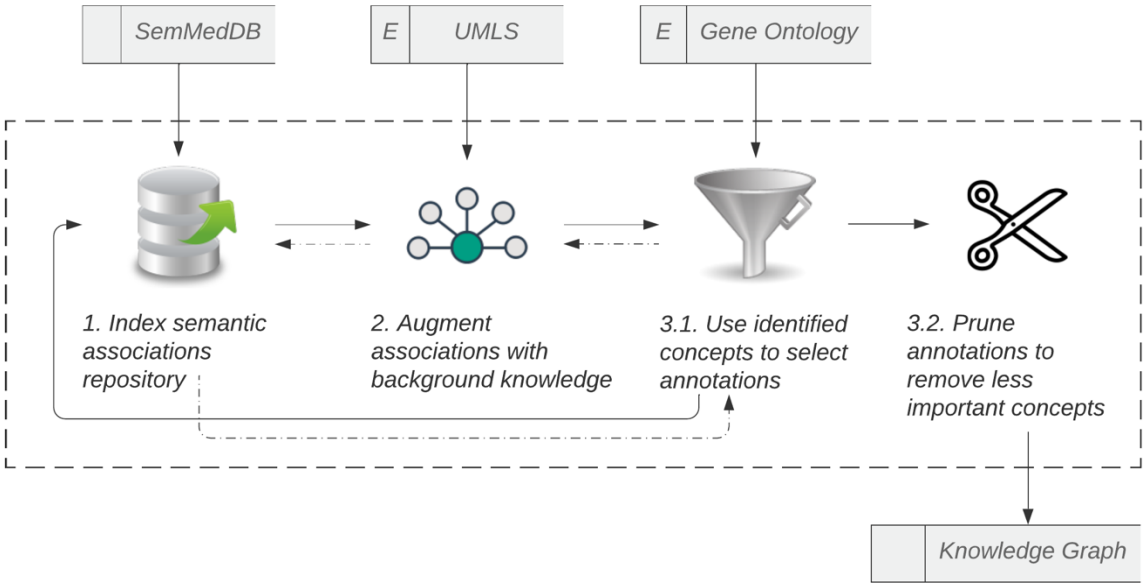
Phase 3 Knowledge Discovery

Activity 4 – Knowledge Representation: Represent the entities and relations using a property graph representation, such as a Knowledge Graph (KG), to represent the relationships between entities as complex associations.

Activity 5 – Pattern Mining: Apply pattern mining and ranking methods to capture sequences of related concepts (i.e., discovery patterns [92]) to identify novel indirect associations.

Activity 6 – Subgraph Generation: Generate subgraphs by interacting with the main graph to identify associations not found by pattern mining.

Figure 3.2: Detailed diagram of the annotation selection and pruning activities



These methods were applied to COVID-19, DM, and CKD but they could also be used to extract mechanistic associations for other diseases. Further, our proposed solution can be

continuously updated with biomedical knowledge as it is highly scalable and not constrained to a particular topic or period. The rest of this chapter will expand on these activities and the methods used to accomplish them. In each section, we refer to a general user of our framework (i.e., 'the user').

3.2 Phase 1: Knowledge Extraction

In this section, we analyze previous strategies that leveraged relations found in text to establish the rationale for (i) using existing methods of extracting semantic associations and (ii) developing a novel approach that extracts associations from both text-based and external knowledge sources.

3.2.1 Predication Extraction

Our approach uses predications mined from text to make sense of complex associations involving multiple concepts that are poorly explained by other types of associations (e.g., term co-occurrence). To extract predications from published articles, the user retrieves PubMed article identifiers (PMIDs) or defines keywords (UMLS concepts) and uses them to index the Semantic Medline Database (SemMedDB), a repository of semantic associations extracted from PubMed titles and abstracts [141]. This is a popular technique used in semantics-based LBD systems [30], [96], [145] as it provides access to a wide range of meaningful associations that are kept up to date with recent research. SemMedDB is made available as a MySQL database, which was exploited in this work.

When indexing SemMedDB with PMIDs, the user is exposed to associations within a pre-defined boundary (i.e., relations from abstracts of the selected articles). By comparison, previous studies use keywords [96], [103] to extract interesting associations between concepts and rely on expert knowledge to control the growth of information. The latter

method is equivalent to indexing PubMed titles and abstracts, using SemRep [146] to drive the identification of relevant studies, where each keyword is the subject or object of a given relation. In both cases, it is possible to access a wide range of biomedical knowledge by obtaining important predications, which could produce interesting hidden connections.

3.2.2 Predication Extension

In this activity, the user extracts ontology relations from the UMLS [140] to extend the reach of predications to fill knowledge gaps in relations mined from text, referred to in Figure 3.1 as 'background knowledge'. The logic behind our approach is that the associations needed to connect distantly related concepts may not be expressed in text (or known by the user), and that these connections if made known will allow important information to flow between previously unrelated contexts. As such, our approach aims to alleviate the issue of limited knowledge associated with the text mining tools being used. With the use of ontology relations, we extend predications with multiple alternative concepts to identify implicit associations from more than two concepts away. Given that predications identified by SemRep [146] have a wide range of granularities that do not account for different levels of user expertise, we believe that certain concepts (e.g., genes and pathophysiologic processes) remain unconnected despite there being logical connections between them that are considered as background (general) knowledge. As such, we expect that ontology relations will provide the user with useful domain knowledge and thusly allow them to explore hidden associations by considering biomedical relationships in finer detail.

The extension process augments predications that are simplistic or uninformative with regards to important biomedical relationships (e.g., Diabetes – AFFECTS → Immune

Response; Kidney Diseases – ASSOCIATED_WITH → Signal Transduction Pathways) by automatically providing related terms to represent simplistic concepts at multiple levels of detail. For example, the UMLS concept 'Immune Response' (*Organ or Tissue Function*) will be extended to include terms such as 'Complement Activation' (*Molecular Function*) [147].

We use hierarchical relations (i.e., *child/narrower*, *parent/broader*) and associative relations (i.e., *sibling/other*) relations to enumerate predications to provide alternative concepts. We use a set of examples below that are specific to COVID-19 and DM or CKD to elucidate the predication extension process for the reader. Other relations may need to be identified for different disorders, which can be found using the UMLS MRREL dataset.

Using the UMLS MRREL dataset, the user extracts 'child' and 'narrower' relations to extend high-level (i.e., broader) predication concepts to include more specific GO terms. An example of a predication being extended by this process is:

1) COVID-19 -CAUSES→ Inflammatory Response (*Pathologic Function*)
2) COVID-19 -CAUSES→ Inflammatory Response ←part_of- leukocyte activation involved in inflammatory response (*Organ or Tissue Function*)

'Inflammatory Response' is the concept being extended in the above example. The last term in 2 (which includes the augmented concept) is a biological process, which is one of three possible domains of GO terms (others being molecular functions and cellular components) [148].

There is an option to include other kinds of relations such as 'positively_regulates' or 'negatively_regulates'. An example of such an extension will be:

3) Metabolism (*Organism Function*) -ASSOCIATED_WITH→ COVID-19
4) positive regulation of metabolic process (*Organism Function*) -positively_regulates→ Metabolism -ASSOCIATED_WITH→ COVID-19

In 4) the first term is referring to any process that increases the rate or extent of biochemical reactions in an organism [149], which is not very informative compared to the original predication. Therefore, depending on the granularity of extracted predication concepts, it may be beneficial to extend high-level concepts using hierarchical relations to achieve a greater level of specificity.

The extension process continues by using previously extracted terms to include multiple layers of alternative concepts. Our aim is to improve the completeness of relations between genes and pathophysiologic concepts by using iterative extension steps. An example of an implicit relation with multiple alternative concepts is as follows:

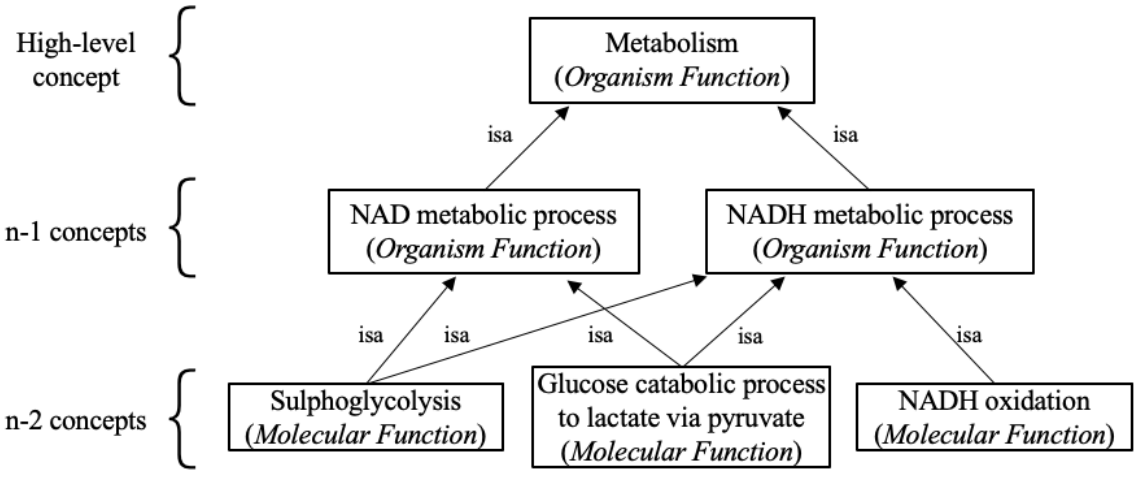
SIRT1 gene -involved_in→ cellular response to hypoxia -isa→ response to hypoxia -isa→ biological adaptation to stress -CAUSES→ Phosphorylation -COEXISTS_WITH→ Diabetes

Where relations in lowercase are referring to relations extracted from external knowledge (i.e., ontologies, annotations) and relations in all caps were extracted from medical literature by SemRep [146]. Here, the implicit (i.e., indirect) association is that SIRT1 is involved in the pathophysiology of Diabetes through a hypoxia-associated pathway. The

inclusion of alternative concepts as intermediate terms allows the user to explore a more complex, and perhaps more obscure, gene-disease association than would be found by considering explicit relations alone.

To avoid generating redundant relations, we do not include child/narrower terms of previously extracted parent/broader terms (or vice versa) or sibling/other terms of previously extracted sibling/other terms. As such, all combinations of ontology relations from up to three concepts away are included, apart from the following sequences: 1) C -*isa*→ B -*inverse_isa*→ A; C -*inverse_isa*→ B -*isa*→ A; and C -*sibling/other*→ B -*sibling/other*→ A, where A is a predication concept and B and C are alternative concepts. We refer to primary extensions as *n-1 concepts* and extensions of previously extracted terms as *n-2 concepts*. A diagram of the predication extension process is shown in Figure 3.3.

Figure 3.3: Diagram of the predication extension process



Thus, predication extension generates implicit associations between relations mined from text, which aims to provide interesting or novel biomedical relationships.

3.3 PHASE 2: KNOWLEDGE INTEGRATION

With the extended predications, the next phase of our approach exploits annotations from biomedical databases to uncover hidden (mechanistic) associations. In this section, we discuss a strategy to integrate entities and relations that are dispersed across multiple public databases using structured domain knowledge.

3.3.1 Annotation Selection And Pruning Annotations

In this activity, the user indexes ontology annotations from a public database using the concepts provided by predication extraction and extension. Ideally, the ontology annotations will come from a large source of reliable experimental relations that supplement the relations identified in previous activities. While we focus on annotations provided by GO [138], [139] (e.g., gene-function relations), other annotations would need to be used for different ontologies, which are made available through public resources (e.g., [150]–[152]). Indexing annotations from public databases with concepts obtained from predication extraction (e.g., genes) or predication extension (e.g., GO terms) allows the user to automatically generate complex associations. We argue that it is useful to select annotations through ontology relations as it could direct the user to associations that are complementary to those extracted from the literature.

With the extracted annotations, the user can repeat the extension process with concepts from previous activities to increase the chances of finding hidden associations. Our goal is to combine distant knowledge fragments (i.e., ontology relations, annotations) to generate meaningful indirect gene-disease associations. This method requires semantic filtering based on groups of concepts (e.g., UMLS semantic types), types of ontology relations, and concepts found in the literature to maintain closeness to the topic(s) of interest. We

iteratively select relations found in text or ontologies to extract additional entities of interest by anchoring a given entity (e.g., a gene) with other concepts in the literature (e.g., gene functions) to control the growth of associations. For this activity, we differentiate between two annotation selection methods, namely *gene selection* and *GO term selection*.

Gene selection (shown as solid arrows in Figure 3.2) takes as input a set of alternative concepts (i.e., GO terms) and outputs a list of genes from GO. Those genes are then used in a cycle of predication extraction followed by predication extension. The genes are filtered by including those directly associated with a target disorder in SemMedDB. The targets for **extension** are pathophysiologic predication concepts associated with said genes in SemMedDB, which are extended through the UMLS, and the resulting alternative concepts are fed back into GO to get a new list of genes. The selection process continues until no new extensions can be generated or no new genes are found.

The gene selection process is summarized by the following pseudocode:

#Inputs

- Target disorders: COVID-19, DM or CKD
- Target genes: genes linked to target disorders in SemMedDB
- Semantic associations: set of pathophysiologic concepts linked to target disorders in SemMedDB
- Alternative concepts: set of ontology extensions in UMLS
- Annotations: set of genes linked to alternative concepts in GO
- Molecular relations: set of pathophysiologic concepts linked to target genes in SemMedDB
- The current selection process generates a significant amount of noise

With

Semantic associations, molecular relations

Do

Get alternative concepts

Get annotations

Get molecular relations

Until

No new genes are found

GO term selection (shown as dashed arrows in Figure 3.2) takes as input a set of genes directly associated with the target disorders in SemMedDB and outputs a list of GO terms from GO. Those terms are used in a cycle of GO term extension followed by predication extraction i.e., the reverse of gene selection). The targets for **extension** are GO terms associated with each gene in GO. The extended GO terms are filtered by including those that are directly associated with a target disorder in SemMedDB. The resulting concepts are linked to a new set of genes in SemMedDB that are fed back into GO to get a new list of GO terms. The selection process continues until no new extensions can be generated or no new genes are found.

The GO term selection process is summarized by the following pseudocode:

#Inputs

- Target disorders: COVID-19, DM or CKD
- Target genes: genes linked to target disorders in SemMedDB
- Annotations: set of GO terms linked to target genes in GO
- Alternative concepts: set of ontology extensions in UMLS that are also linked to target disorders in SemMedDB
- Molecular relations: set of genes linked to alternative concepts in SemMedDB
- It is questionable whether multiple rounds of GO term selection are necessary

With

Target genes, molecular relations

Do

Get annotations

Get alternative concepts

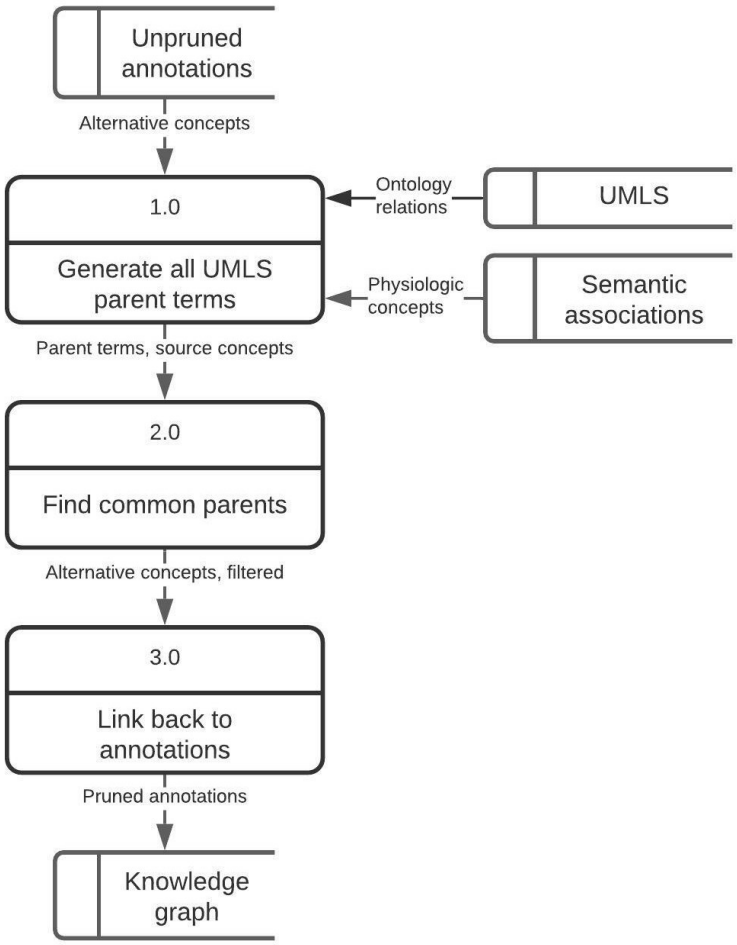
Get molecular relations

Until

No new genes are found

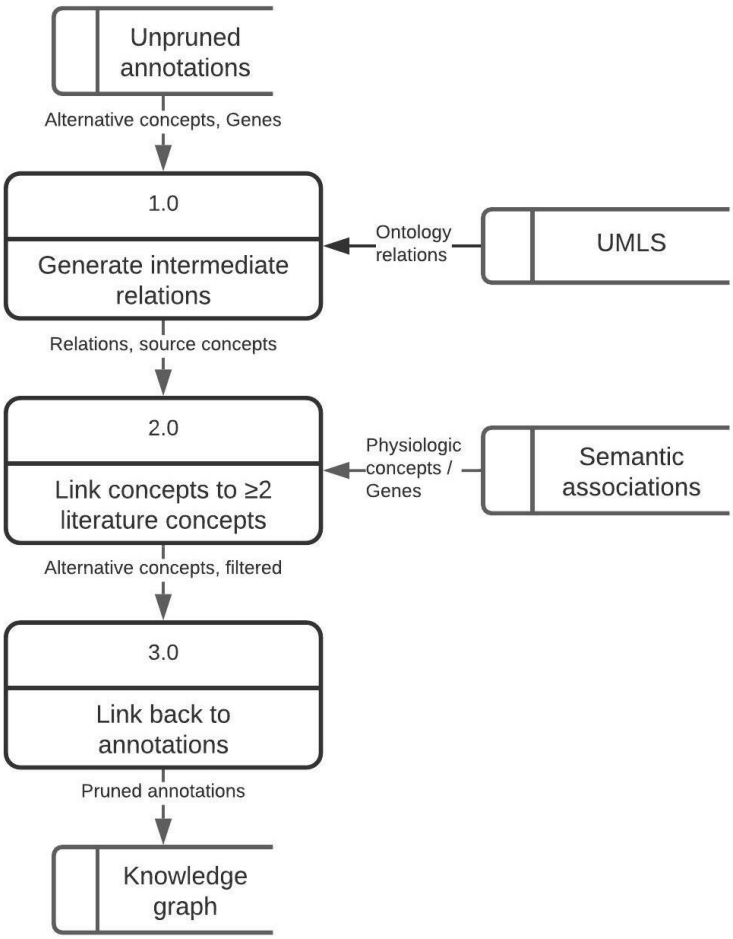
Given that we impose few restrictions with regards to allowable alternative concepts, we anticipate that our approach will require pruning to remove less important and useful concepts. To this end, we adapt the following methods: Common Parents [143] (referred to as 'CP') where an alternative concept shares a non-generic UMLS parent term with at least one predication concept. A diagram of the CP pruning method is shown in Figure 3.4.

Figure 3.4: Diagram of the CP pruning method



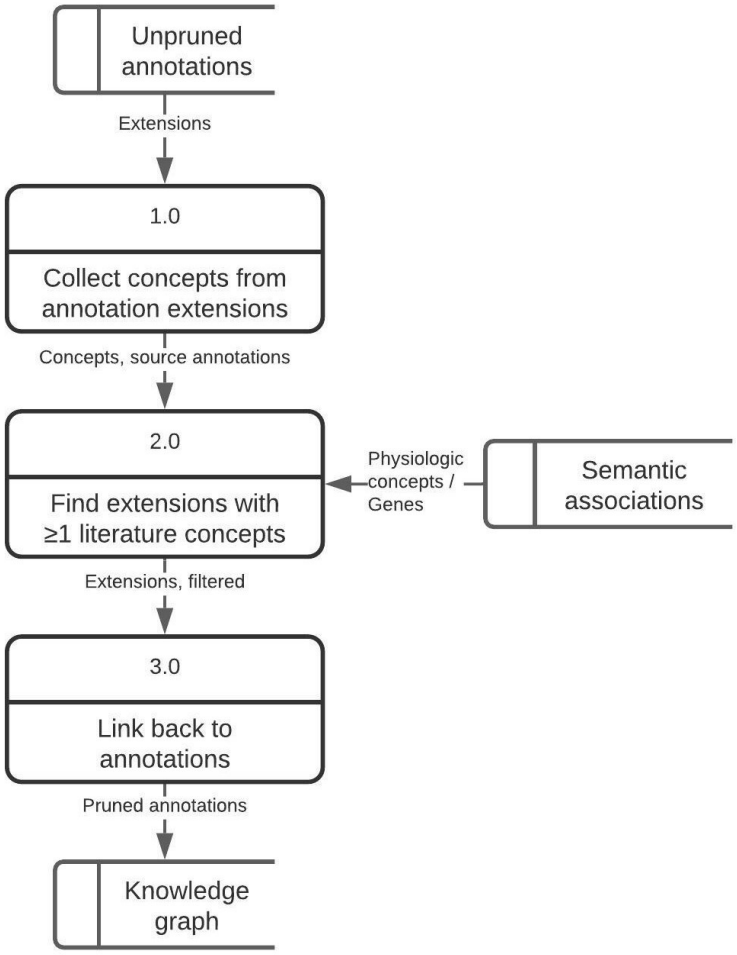
Intermediate Relations [127] (‘Intermediate’), where a given alternative concept occurs in at least two unique relations with predications concepts (i.e., concepts mined from the literature). A diagram of the intermediate pruning method is shown in Figure 3.5.

Figure 3.5: Diagram of the Intermediate pruning method



GO annotation extensions [144] (‘Link’), which are annotation subclasses that specify biologic context (e.g., the cellular location of a molecular function), where a given alternative concept must link to at least one gene or pathophysiologic concept from the literature through the GO extensions field. A diagram of the Link pruning method is shown in Figure 3.6.

Figure 3.6: Diagram of the Link pruning method



While the Intermediate and CP pruning methods are based on the intrinsic structure of ontologies, the Link method takes a completely different approach by recognizing that GO annotations refer to gene functions that occur under certain biological conditions [144]. As a result of performing predication extraction, predication extension, and annotation selection, associations mined from text will be augmented with external knowledge to produce a set of highly interrelated concepts.

3.4 PHASE 3: KNOWLEDGE DISCOVERY

The final phase of our approach aims to identify plausible associations by eliciting interesting or novel patterns from a graph. In the following subsections, we describe methods to (i) represent entities and relations for pattern analysis; (ii) extract sequences of related concepts (i.e., discovery patterns); and (iii) create subgraphs to expose additional associations not captured by patterns.

3.4.1 Knowledge Representation

To capture complex semantic associations, the user represents the identified entities and relations in a property graph, such as a KG. In this way, concepts and relations are represented as nodes and edges, respectively, using a standardized, interoperable format. KGs have been used for knowledge curation [77], semantic pattern analysis [91], [104], and discovery chain analysis [29], [96]. While several LBD systems discussed in chapter 2 use KGs to predict novel relations, they do not attempt to analyze complex patterns involving multiple concepts and relation types. Thus, utilizing a comprehensive representation such as a KG should facilitate pattern analysis by allowing complex indirect associations to be easily identified and explained.

The first part of this activity involves semantic integration with structured knowledge sources (e.g., generating triples based on UMLS-defined concepts and relations [30]) to represent the annotations, ontology relations, and predications from medical literature using a single representation format. Secondly, each relation is enriched with a reference to the source article or database entry (referred to as *provenance*) so that it can be validated in subsequent activities. Finally, the entities and relations are stored in a graph database, such as Neo4j.

3.4.2 Pattern Mining—Discovery Patterns

This activity involves specifying plausible indirect associations between distant knowledge fragments to identify novel relations. Subsequently, the results are ranked to prioritize interesting associations, which are regarded as discoveries or hypotheses. We use a set of examples below that are specific to molecular disease mechanisms to explain the pattern mining process, while acknowledging that different patterns may be applicable for other topics (e.g., [91], [126], [153]).

To extract mechanistic associations from the KG, we developed a search strategy using Neo4j's query language Cypher whereby we can specify both direct (concept A → B) or indirect (A → … → B) associations as multi-node patterns—called *discovery patterns*—that comprise substance interactions, physiologic disturbances, and disease-disease relations (e.g., comorbidity). Discovery patterns create restrictions on the semantic categories of given concepts and the relations between them based on the user's input [145]. To demonstrate our approach, we build on the gene-phenotype pattern proposed by Hristovski et al [94] as described in chapter 2. First, the user specifies direct associations by searching for a given pair of concepts with specific relations between them (e.g., a gene and its physiologic function). The semantic types and the direction of relations are specified as follows:

1) Gene A (*Gene or Genome*) – Relation → Function B (*Organ or Tissue Function*)
2) Function B – Relation → Disease C (*Disease or Syndrome*)

Direct associations are then combined by defining logical sequences of relations to find genes indirectly associated with a given disease:

3) Gene A – Relation → Function B – Relation → Disease C

Finally, the user can specify the types of relations to identify meaningful associations by making the pattern more specific:

4) Gene A – AUGMENTS → Function B – CAUSES → Disease C

Each indirect gene-disease association is found in fewer than 10 articles to focus on novel patterns. Additional search criteria are set to ensure that the source (i.e.., provenance) of each relation is different, else the discovery would be trivial to a given reader [154].

Discovery patterns based on the ABC model (i.e., involving only three concepts) may fail to capture complex associations [95]. To identify gene-disease associations from more than two concepts away, an intuitive method is to incorporate additional relations to create sequences of related concepts (i.e., discovery chains). Using the above example, a new pattern can be specified, such as:

5) Gene A – AUGMENTS → Function B_1 – is_a → Function B_2 – CAUSES → Disease C

Where the pattern now finds a gene-disease association involving two biologic functions, representing an underlying disease pathway. Thus, pattern generation involves specifying logical sequences of relations to hypothesize plausible associations between distantly connected concepts.

3.4.3 Discovery Pattern Ranking

To facilitate the process of finding interesting and important discovery patterns, ranking methods are applied to direct the user's attention to promising results. We focus on

frequency- and graph-based metrics due to their widespread use in biomedical LBD systems [77], [96], [117], [145]. We developed a ranking mechanism that compares (i) indirect association and (ii) graph-theoretic measures, namely Linking Term Count (LTC) [114] and PageRank [155], respectively. LTC considers the number of intermediate concepts when A and C are 2 concepts away to assess whether they are strongly correlated. LTC focuses external knowledge through relevant intermediate concepts in the literature or public databases. PageRank is a measure of each node's connectivity in the network, which tells if a given concept and its directly related concepts are highly important. We used the average scores of each pattern to rank the associations as a criterion for discovery of mechanistic associations, which was calculated as follows:

1) Count intermediate concepts (x) between A and C (i.e., A – x – C)
2) Calculate sum of PageRank of all nodes in the pattern and divide by the number of nodes
3) Compute average score using the two metrics

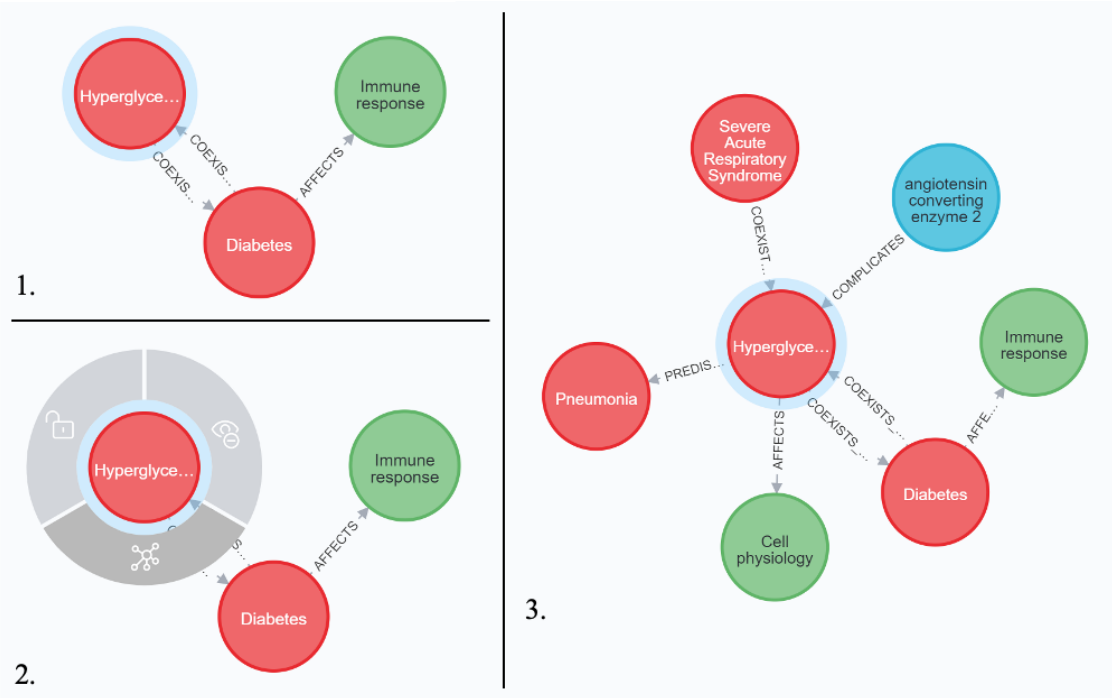
Given that ranking techniques are an essential part of the user's workflow, providing them with meaningful associations [28], the success of our LBD framework will be judged on the effectiveness of the ranking mechanism. To determine whether our method is successful at prioritizing important associations, we use the *precision at k* metric [120] to calculate the number of accurate discoveries up to rank k, divided by k. With this metric, we will evaluate the different methods proposed here to determine whether one is more advanced.

In lieu of a gold standard upon which to validate the discoveries, the user analyzes each discovery pattern by exploring them in a literature database, such as PubMed. We take a similar approach as recent LBD methods [29], [30], [156] that search for hidden associations to infer novel relations, relaxing the requirement of expert interpretation as the user self-validates by searching for supporting evidence. In this regard, we refer to evidence found in the abstract or full text of articles published after a pre-specified cutoff date [120].

3.4.4 Subgraph Generation

The final activity in our framework focuses on producing meaningful illustrations of interrelated concepts, which aims to uncover complex associations not found by pattern mining. In this activity, we focus on genes and pathophysiologic concepts that share at least two relations with concepts from each identified discovery pattern to simplify the results. Starting from a selected concept, the user interacts with the main graph, using iterative searching (i.e., 'discovery browsing') to focus the system's output on interesting concepts [96]. This process is driven by the user's prior background knowledge, where the interestingness of each association is determined by whether it is uncommon or unfamiliar in the given context [3]. A diagram of the discovery browsing process in Neo4j is shown in Figure 3.7. The activities shown in the figure are as follows: 1) select concept to be explored further; 2) expand concept; 3) view all related concepts.

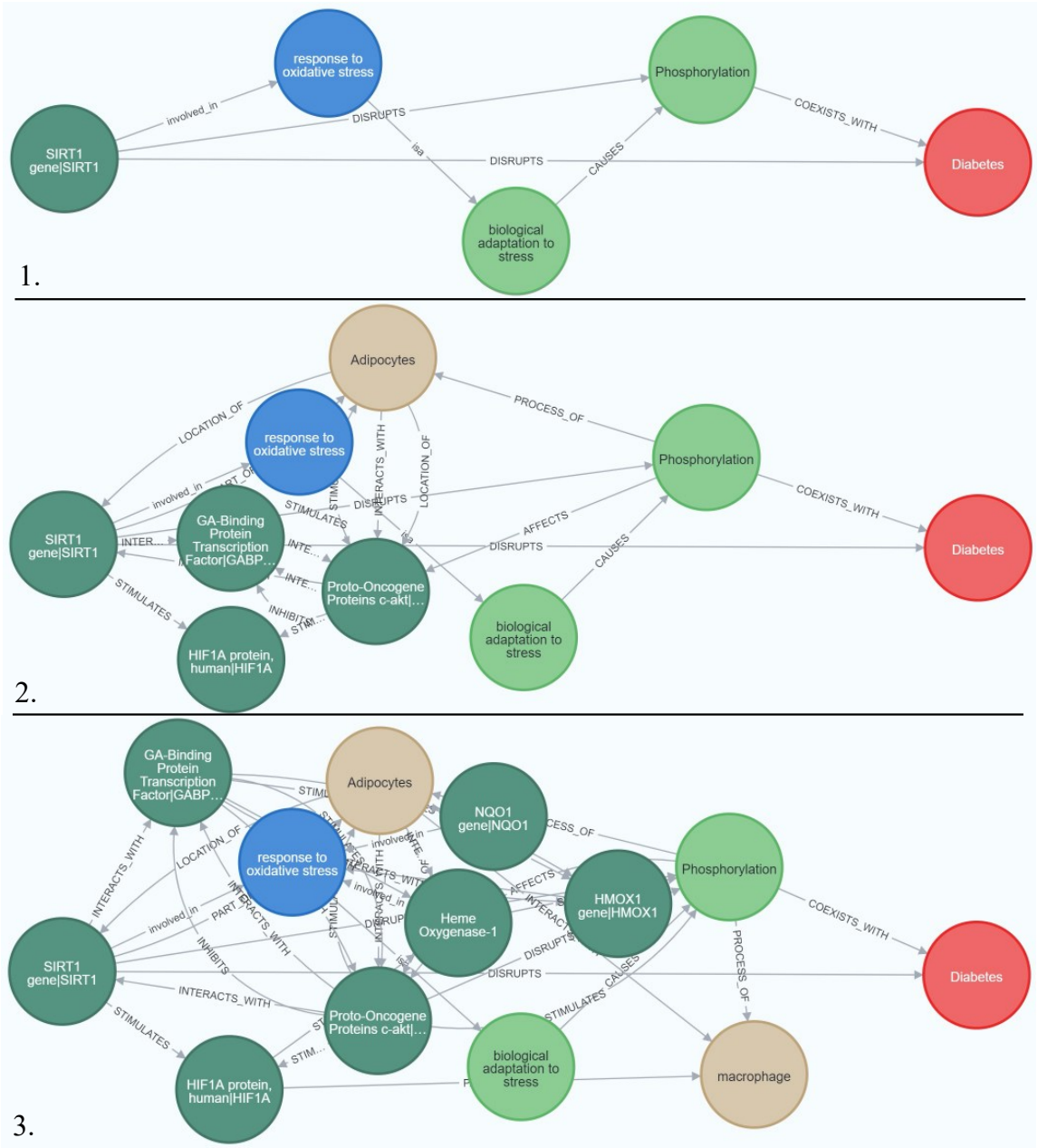
Figure 3.7: Diagram of the discovery browsing process in Neo4j



Previous subgraph generation methods have used different criteria to analyze complex associations to expose interesting links between concepts. Wilkowski et al [96] iteratively selected well-connected concepts in the literature, allowing them to identify important links between distantly connected concepts. Similarly, Vlietstra et al [30] created a high-level subgraph comprising a central disease concept and its neighbouring disease-related concepts. Finally, Cameron et al [88] considered a context-driven approach, whereby subgraphs are created using associations that meet minimum relatedness criteria based on their original publication context. Our approach differs from these previous methods in that we incorporate discovery browsing [96] to uncover novel associations between concepts, forgoing the use of preconceived inclusion criteria. We create detailed subgraphs of relations between genes and pathophysiologic concepts that, when viewed together, help with identifying disease mechanisms. Figure 3.8 shows an example of expanding a discovery pattern that is specific to DM to elucidate our approach for the reader. Relations

between concepts are shown as arrows and the original discovery pattern is shown in the first subgraph. The activities shown in the figure represent a different concept in the pattern being expanded from left to right. The discovery browsing process continues until each concept in the original discovery pattern has been expanded.

Figure 3.8: Expanding a discovery pattern in Neo4j



In Figure 3.8, genes are shown as dark green nodes, the alternative concept is blue, pathophysiologic concepts are light green, anatomic concepts are brown, and the disorder is red. Here, the hypothesis is that SIRT1 is involved in the pathophysiology of DM through a hyperglycemia-induced oxidative stress pathway [157]. Our approach identifies several additional concepts that may be affected by said pathway. For instance, the additional genes build on the original pattern as they are important mediators of downstream pathways of SIRT1 that may be perturbed in a disease state. As such, discovery browsing allows the user to explore the identified hypotheses further by expanding important concepts to uncover connections between disease pathways.

In summary, our approach aims to identify novel associations between distantly related concepts by integrating disparate knowledge sources in a KG to improve literature-based discovery of complex (mechanistic) associations. We augment relations mined from text (i.e., predications) using external knowledge (i.e., ontologies, annotations) to address identified knowledge gaps between genes and pathophysiologic concepts. We then apply pattern mining (i.e., discovery patterns and discovery browsing) as well as graph- and frequency-based ranking methods to uncover interesting and important associations, referred to as hypotheses. To improve the performance of our method, we propose the use of a pruning technique that filters out less important concepts based on their relevance to concepts mined from the literature. Finally, we validate our findings in the context of medical literature to determine whether a certain variation of our method performs best in terms of emphasizing accurate hypotheses.

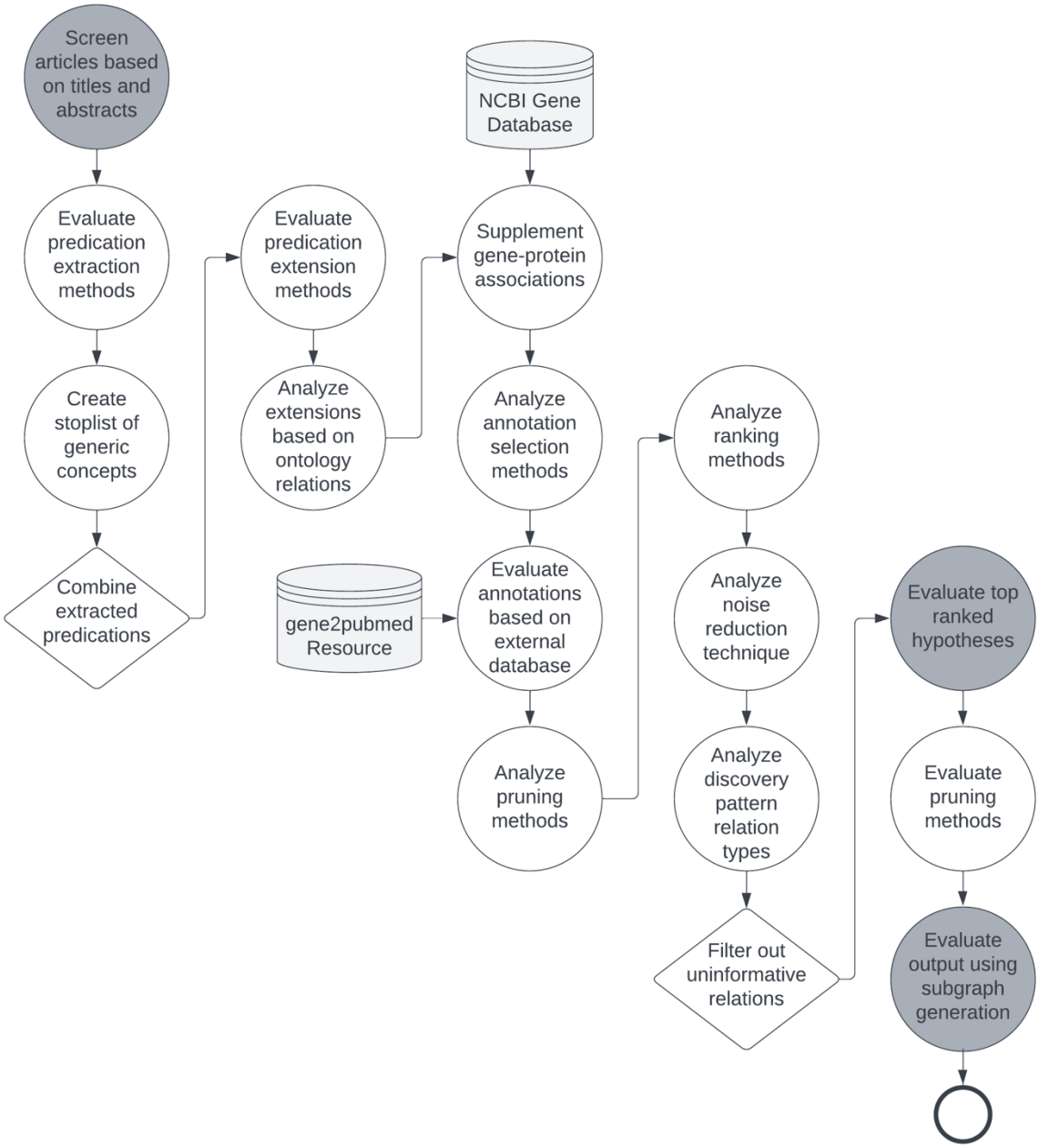
CHAPTER 4.0 EXPERIMENTS AND RESULTS

In this chapter, we will apply our LBD methodology to identify semantic associations from the literature and external knowledge sources to investigate two research questions; 1) how might the underlying disease states in DM or CKD predispose patients to worse COVID-19 outcomes, and 2) how might COVID-19, especially in severe cases, exacerbate DM or CKD? The resulting work has been named the COVID-19 Renal and Endocrine Interactions (COVID-REdI) system. The aim of this system is to provide novel or interesting gene-disease associations to discover plausible disease mechanisms from evidence-based sources that can help understand the pathobiology underlying disease impacts.

In line with our methodology described in the previous chapter, our work is organized in three phases that were designed to achieve strategic activities corresponding to the sections below. Each section describes a set of experiments and presents the results, thereafter, making comparisons with previous methods where possible. In the Knowledge Extraction section, we describe our approach to synthesize relevant evidence on COVID-19, DM, and CKD with regards to extracting relations found in text and supplementing those relations with external knowledge. In the Knowledge Integration section, we analyze our approach by comparing it with similar LBD methods. In the Knowledge Discovery section, we present the mechanistic associations identified by our approach and describe the optimal approach for the present work. Finally, we validate our approach in the Case Studies section, where we show how our LBD framework addresses identified COVID-19 research problems.

To clarify our work, we show a flowchart of the experiments done to derive the optimal approach in Figure 4.1. In the figure, dark grey nodes represent user actions, white nodes are system processes, light grey cylinders are data stores, and arrows are workflow activities. Optional processes are shown as diamonds, which may need to be modified for different research questions as they were implemented based on experimental findings.

Figure 4.1: Flowchart of the experiments done to derive the optimal approach



4.1 PHASE 1: KNOWLEDGE EXTRACTION

In the first part of this section, we identify articles published in PubMed that are relevant to disease mechanisms for patients with COVID-19 and DM or CKD. Secondly, we use the resulting articles to extract relations found in text (i.e., predications), comparing our approach with previous methods. Finally, we extend the extracted predications using structured background knowledge (i.e., ontologies) to generate informative mechanistic associations.

4.1.1 Literature Selection In Pubmed

We developed a systematic literature search process [158] to identify biomedical evidence on COVID-19 and DM or CKD published in PubMed (2020 – onwards). Similar to previous LBD studies [77], [102], the search process involved using a set of pre-defined MeSH index terms (i.e., descriptors). To specify article primary content, we used the MeSH descriptors *COVID-19*, *SARS-CoV-2*, *Diabetes Mellitus* and *Renal Insufficiency*, each of which comprised several narrower concepts (e.g., Type 1 and Type 2 DM for *Diabetes Mellitus*). Given that most of the available information points to Type 2 DM (T2DM) as a risk factor for poor COVID-19 outcomes [15], [66], we also included MeSH descriptors for critical components of T2DM, namely *Insulin Resistance*, *Glucose Intolerance*, and *Insulin-Secreting Cells* [102].

4.1.2 Selecting Topics Of Interest

Our initial searches contained many articles that were not relevant to disease mechanisms. To reduce the noisiness of results in PubMed, we used MeSH subheadings (i.e., qualifiers) to create topic-specific queries that include a wide range of studies (basic sciences, clinical studies) since mechanistic associations may be distributed across a variety of publication contexts. In previous work, Srinivasan et al [102] used MeSH descriptors to define topics

of interest (e.g., primary topic being DM, secondary being chemicals associated with DM). We considered a more generalized approach whereby secondary topics are replaced by MeSH qualifiers. These sets of qualifiers were combined with the primary set (i.e., COVID-19 AND [DM or CKD] AND [qualifier]), resulting in 10 topic-specific queries. The scope of each topic, retrieved from [159], is shown in Table 4.1. Examples of PubMed queries are included in Appendix A.

Table 4.1: Scope of query terms for biomedical topics

MeSH qualifier	**Scope**
Complications	Co-existing diseases, complications, or sequelae
Virology	Virologic studies of organs, animals, and diseases
Etiology	Causative agents of disease including viruses and environmental factors
Metabolism	Biochemical changes in organs, cells, and disease states
Physiopathology	Disordered function in disease states in organs and tissues
Biomarkers	Quantifiable biological parameters that serve as physiological indicators of disease risk
Pathology	Organ, tissue, or cell structure in disease states
Immunology	Immunologic studies of tissues, organs, and viruses, including immunologic aspects of disease
Genetics	Genetic basis of normal and pathologic states
Pathogenicity	Studies of the ability of viruses to cause disease

The topic-specific queries initially found less articles than expected, which we addressed by including MeSH term synonyms as free-text keywords in the title and abstract fields [158]. We also endeavoured to include articles published separately in the DM or CKD and COVID-19 literature, with the objective of finding supplementary evidence that was relevant to the intermediate literature (i.e., COVID-19 and DM or CKD), since recent

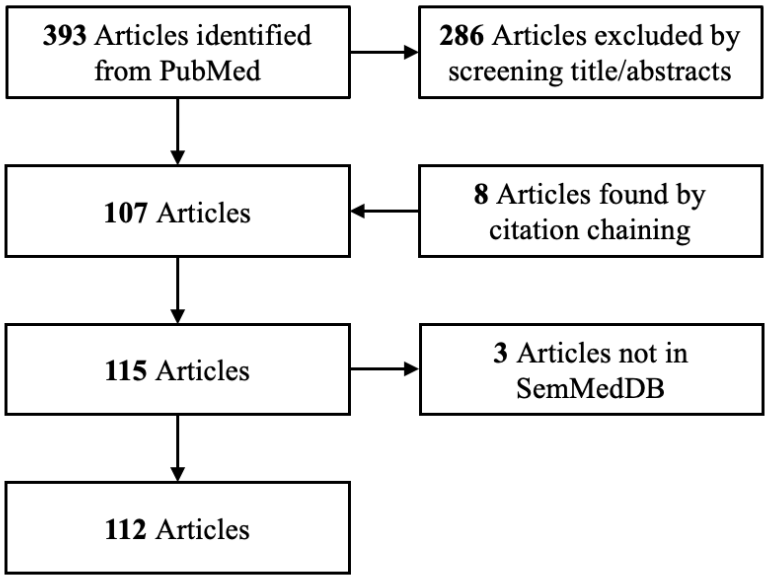
literature on DM or CKD could harbour interesting connections with COVID-19 that have gone unnoticed by readers of COVID-19 articles (or vice versa). We used the same search terms as above to retrieve articles for each target disorder. Ultimately, we identified three literature sets related to (i) DM or CKD (referred to as the *source* literature), (ii) COVID-19 and DM or CKD (*intermediate*), and (iii) COVID-19 (*target*). The number of articles found in PubMed for each literature set is shown in Table 4.2.

Table 4.2: Number of articles found in PubMed for each literature set

Literature set	**Articles**
Source (DM or CKD)	12,061
Intermediate (COVID-19 and DM or CKD)	393
Target (COVID-19)	11,834

The source and target literature sets were not screened due to the prohibitive size of the results. The titles and abstracts of articles in the intermediate set were screened to ensure that they were specific to DM or CKD, which was necessary since both disorders are associated with multiple comorbidities. Of the intermediate articles, 286 were removed as they were not specific to DM or CKD; made no reference to disease-specific parameters; focused on disease management or therapy; focused on unrelated conditions; or were too broadly or narrowly focused. An additional eight articles specific to COVID-19 and DM were found through citation chaining and included. Three articles were not found in the latest version of SemMedDB (version 43_R). This resulted in 112 intermediate articles to be included in the study. Article content ranged from mechanistic reviews to case reports of DM and CKD patients with COVID-19. A diagram of the inclusion process is shown in Figure 4.2.

Figure 4.2: Diagram of the article inclusion/exclusion process



4.1.3 Extracting Predications To Find Mechanistic Associations

Our approach uses relations mined from text (i.e., predications) to help make sense of complex associations between genes, chemicals, diseases, and pathophysiologic concepts. Using lists of PMIDs as input to SemMedDB [141] (version 43_R) in MySQL, we obtained predications for the three aforementioned literature sets. We then compared our method with two other methods to extract predications, namely (i) using the MeSH descriptors described above as *keywords* (UMLS concepts) to index SemMedDB [96]; and (ii) using PMIDs found in PubMed by browsing with MeSH descriptors in *all fields* [30]. To analyze each method, we calculated the number of articles, predications, and pathophysiologic (i.e., physiologic, pathologic, and anatomic) concepts and genes directly associated with COVID-19, DM, or CKD found in SemMedDB. The analysis of predication retrieval methods is shown in Table 4.3. A full list of semantic types used to group the concepts is

included in Appendix B. We excluded certain semantic types that were not relevant to biomedicine (e.g., *Healthcare Organization*, *Regulation or Law*, *Social Behaviour*).

Table 4.3: Analysis of predication retrieval methods

Metric	Keywords	PMIDs (all fields)	PMIDs (topic-specific)
Articles	56,988	20,472	15,147
Predications	198,400	67,558	55,009
Pathophysiologic concepts	880	413	427
Genes	631	211	296

Given that our initial method excluded several important relations, we integrated the results of the three methods, referred to hereafter as the extracted predications. We included 75,041 unique articles and 125,829 unique predications in our analysis, including several pathophysiologic concepts and genes that were relevant to the target disorders. To improve efficiency and reduce the search space, we used the filtering strategies described in [113] to create a stoplist of 458 uninformative concepts (e.g., 'Biological Processes') based on prior knowledge that is included in Appendix C. Finally, we removed relations pertaining to non-endogenous chemicals, including concepts assigned with the semantic types *Pharmaceutical Substance* and *Organic Chemical* [30] and the relation types 'TREATS', 'PREVENTS', and 'ADMINISTERED_TO'.

Using keywords as input generated many useful relations but only found articles that explicitly mentioned the disorders of interest. On the other hand, indexing with PMIDs identified 7,816 articles not found by keywords, which could comprise relations that are relevant to the target disorders. The *topic-specific* approach identified less articles and relations compared with the other approaches, which was expected as we applied multiple

filtering techniques to focus on relevant studies. Inspired by the results of the *keywords* approach [96], we supplemented our keywords by manually selecting additional (narrower) UMLS disease concepts that are included in Appendix D.

4.1.4 Extending Pathophysiologic Predication Concepts

In this activity, we augmented relations mined from text to cover concepts related to gene function. During our initial analysis of the extracted predications, we observed that interesting fine-grained physiologic concepts (e.g., biochemical pathways) corresponded with certain types of relations (gene-physiology) more often than others (physiology-disease). Our goal was to improve the completeness of relations with regards to molecular pathophysiologic concepts to address the noted sparsity of informative mechanistic associations.

To maintain closeness to the disorders of interest, we targeted pathophysiologic (i.e., pathologic, physiologic, and anatomic) concepts that were associated with COVID-19 **and** DM or CKD in SemMedDB as input ('joint associations'; N = 166). Further, we targeted pathophysiologic concepts that were directly associated with COVID-19, DM or CKD ('direct associations'; N = 787) to determine whether our approach works better with shared or disease-specific associations. We then compared our approach with a previous method to extend predications [127], extracting *intermediate* GO terms that link two predication concepts associated with the target disorders (i.e., A – B – C where B is a GO term and A and C are predication concepts that are jointly associated with the target disorders, directly associated with a target disorder, or jointly and directly associated, respectively).

To understand each method, we calculated: (i) the number of unique GO terms generated, referred to as 'Count'; (ii) the percentage of GO terms generated that were present in the extracted predications to assess relevance ('Overlap'); and (iii) the percentage of pathophysiologic predication concepts directly related to COVID-19, DM or CKD in SemMedDB (N = 787) that had common ancestors [143] (i.e., shared UMLS parent terms) with an extracted GO term to assess connectedness ('Similarity'). To reduce the possibility of meaningless associations, we excluded high-level ancestor terms that corresponded with the first two levels of the UMLS hierarchy. The analysis of predication extension methods is shown in Table 4.4.

Table 4.4: Analysis of predication extension methods

Method	Input	Count	Overlap	Similarity
Intermediate relations	Joint associations	26	23.1%	37.5%
	Direct associations	**183**	**21.9%**	**41.9%**
	Joint and direct associations	**106**	**23.6%**	**39.5%**
Child/narrower relations (*n-1*)	Joint associations	456	9.9%	41.7%
	Direct associations	1,497	7.4%	42.3%
Sibling/other relations (*n-1*)	Joint associations	104	3.9%	33.9%
	Direct associations	441	1.8%	35.2%
Parent/broader relations (*n-1*)	Joint associations	41	22.0%	34.5%
	Direct associations	226	19.0%	41.7%

Intermediate relations [127] was the best performing method with regards to the chosen metrics. It showed that pathophysiologic literature concepts were associated through various types of UMLS relations producing relatively small numbers of alternative GO terms. Notwithstanding, the connectedness between literature concepts and GO terms was

preserved as the target concepts were reachable through both narrower and broader relations. For said method, there was little difference between using direct associations or joint and direct associations as input (bolded in Table 4.4). Our approach allowed us to identify additional associations between pathophysiologic concepts but led to discrepancies between the number of existing terms and their closeness with predication concepts. Nevertheless, we found interesting links between pathophysiologic concepts, noting that direct associations generated a greater number disease-specific terms when compared with joint associations. We used direct associations as input hereafter since there was little difference between the two inputs and since disease-specific gene functions could help to explain non-disease-specific overlaps as was recently shown for DM [160].

Despite our approach being disjoint with relations mined from text, we found associations with certain target concepts that allowed us to extract highly specific gene functions. We observed similar results for the second layer of extensions (*n-2*), which are included in Appendix E. An analysis of extension layers and relation types is shown in Table 4.5.

Table 4.5: Analysis of extension layers and relation types

Extension layer	Relation type	Count	Overlap	Similarity
n-1	Taxonomic	1,387	6.0%	43.7%
	Non-taxonomic	726	1.3%	40.6%
n-2	Taxonomic	7,073	2.4%	57.4%
	Non-taxonomic	1,943	0.4%	40.3%

The extension process generated more relevant alternative concepts through parent/child relations (i.e., ‘taxonomic’) when compared with non-taxonomic ones. We observed that taxonomic relations accounted for more overlap and similarity with predication concepts

while non-taxonomic relations generated concepts that were not clearly associated with the target concepts. Notwithstanding, parent relations had significantly less overlap than child relations ($p<0.05$; two proportion z test), likely because irrelevant concepts were found by including all possible ancestors. Finally, the performance of our method diminished in the second extension layer, with a significant amount of noise produced thereafter.

4.2 Phase 2: Knowledge Integration

With the extracted relations from the literature and ontologies, the next phase of our approach aims to identify genetic contributions to COVID-19 and DM or CKD by synthesizing associations dispersed across isolated biomedical databases. Firstly, we elaborate on how gene-physiology associations from GO (i.e., GO annotations) were used to propagate the knowledge integration process by targeting important relations mined from text. Secondly, we show how our approach uncovers hidden mechanistic associations, building on a recently published method.

4.2.1 Selecting Entities Of Interest To Generate Complex Associations

To obtain mechanistic associations that complement relations mined from text, we iteratively selected gene-function associations from the Gene Ontology (GO) [139] using alternative concepts (i.e., GO terms) from the UMLS [140] and predications extracted from medical literature (i.e., SemMedDB [141]). We used the methods described in Chapter 3 to select entities that were directly associated with COVID-19, DM or CKD in SemMedDB to maintain closeness to the disorders of interest. Further, we targeted genes and pathophysiologic concepts that were indirectly associated with the target disorders, given that each disorder affects multiple cells, organs, and biological processes [34], [161], [162]. While the necessary knowledge fragments for our approach could be found in SemMedDB

and GO, there was a lack of up-to-date genetic knowledge for seamless integration of the two resources, which required gene-protein associations. To address that issue, we mapped genes to their corresponding protein products using a human-curated set of gene-protein associations provided by the National Center for Biotechnology Information (NCBI) [163].

We analyzed our methods by calculating (i) the number of GO annotations generated ('Count'), (ii) the percentage of annotations with a GO term that was present in the extracted predications to assess relevance ('Overlap'); and (iii) the percentage of annotations that were associated with a gene in GO and related to a pathophysiologic concept in UMLS, both of which were present in the extracted predications to assess coherence ('Continuity'). The analysis of cycles of annotation selection is shown in Table 4.6. In the table, we refer to significance values of a two-proportion z test of successive cycles whereby: * = $p<0.1$, ** = $p<0.05$, *** = $p<0.01$.

Table 4.6: Analysis of cycles of annotation selection

Method	Cycle No.	Count	Overlap	Continuity
Gene selection	1	41,662	10.8%	11.7%
	2	19,512	5.8%	9.6%
	3	455	2.0%	5.9%
GO term selection	1	6,979	13.5%	25.5%
	2	1,056	15.0%	30.2%***

The continuity between annotations and predication concepts was favourable after a second cycle of annotation selection, indicating that our approach was most effective until that point. Selecting GO annotations in this way allowed us to identify gene function relations at a high level of granularity, though the benefit of additional cycles was minimal due to the limited amount of new information generated. When the annotations were segmented

by relation types, parent relations performed best in GO term selection and child relations performed best in gene selection, likely due to differences in the specificity of terms between SemMedDB and GO with the latter providing more granular terms. Finally, gene selection continued for three cycles until no new extensions were generated and GO term selection continued for two cycles until no new genes were found.

4.2.2 Using Annotations To Uncover Hidden Mechanistic Associations

We compared our method with one proposed by Vlietstra et al [30] which was interpreted as selecting GO annotations based on *explicit relations* mined from text. To replicate the proposed method, we used entities (i.e., genes and pathophysiologic concepts) that were directly associated with a target disorder in SemMedDB as input to obtain annotations from GO.

To analyze each method, we calculated (i) the number of unique annotations generated ('Count'); (ii) the percentage of annotations with a GO term that was related to at least one gene and one pathophysiologic concept from the extracted predications to assess coherence ('Continuity'); and (iii) the percentage of genes associated with COVID-19, DM or CKD in PubMed [164] from 2020-present ('Recall'; N = 4,044) that were selected from GO. The latter calculation provided an independent evaluation of the ability of each method to generate relevant gene-disease associations which was performed by extracting gene-article associations using the *gene2pubmed* dataset at NCBI [163] and mapping articles to their corresponding disease MeSH terms via PubMed XML files. The analysis of annotation selection methods is shown in Table 4.7.

Table 4.7: Analysis of annotation selection methods

Method	Input	Count	Continuity	Recall
Explicit relations	GO terms	3,169	11.4%	23.4%
	Genes	6,979	25.5%	5.4%
Implicit relations	Alternative GO terms	61,629	11.0%	51.9%***
	Genes	8,035	26.1%	6.0%

Our method identified more potential disease genes from GO as recall improved significantly when using alternative GO terms to obtain annotations when compared with other methods ($p<0.0001$). Of those alternative terms, child concepts accounted for the greatest recall, consistent with our previous findings, by identifying 'narrower-to-broader' implicit relations. Finally, both methods generated GO annotations that were interrelated with relations mined from text, capturing hidden mechanistic associations at a higher level of granularity when compared with relations found in the literature.

4.2.3 Pruning Extensions To Include Relevant Associations

We developed techniques to mitigate irrelevant predication extensions as our initial results were inconsistent with phenomena described in the literature. Our goal was to generate cohesive indirect gene-disease associations by including annotations that were relevant to entities of interest found in SemMedDB. We considered three different pruning criteria for a given annotation: (1) has a common ancestor [143] (i.e., shared UMLS parent term) with at least one GO term or pathophysiologic concept that is directly associated with COVID-19, DM or CKD in SemMedDB (**c**ommon **p**arents or 'CP'); (2) has intermediate relations [127] (i.e., ontology relations or GO annotations) with a pair of GO terms or a GO term and a gene that were present in SemMedDB ('Intermediate'); and (3) links to a SemMedDB predication concept via the GO annotation extensions field [144] ('Link'). An analysis of

predication extension pruning is shown in Table 4.8. We calculated the GO annotation count, continuity, and recall as described in Section 4.2.2. For brevity, the condensed results are shown here, and the full results are included in Appendix F.

Table 4.8: Analysis of predication extension pruning

Method	Pruning	Count	Continuity	Recall
Gene selection	None	61,629	11.0%	51.9%
	CP	21,941	12.9%***	34.0%
	Intermediate	13,586	33.8%***	28.5%
	Link	1,224	14.1%***	3.9%
GO term selection	None	8,035	26.1%	6.0%
	CP	4,432	34.1%***	5.7%
	Intermediate	5,957	34.7%***	5.8%
	Link	206	17.5%	1.3%

The Intermediate method performed best in terms of continuity between annotations and concepts mined from the literature, though it omitted more potential disease genes than CP. There was a trade-off between continuity and recall, indicating that some potential disease genes were excluded by each method as they were too distant from the target concepts. The Link method excluded the most annotations overall due to a lack of molecular relations that linked genes and gene products from the selected literature. Finally, as in our previous experiments, child relations performed best in gene selection while parent relations performed best in GO term selection regardless of the pruning methods.

4.3 Phase 3: Knowledge Discovery—The COVID-REDI KG

Using the combined associations from medical literature and ontologies, the final phase of our approach aims to investigate testable indirect associations (i.e., patterns) to identify molecular mechanisms underlying COVID-19 and DM or CKD to explain disease impacts.

Firstly, we use graph-based methods to capture complex mechanistic associations, analyzing the results of our pattern ranking approach. Secondly, we validate the top ranked discovery patterns found by each method described in the previous section and evaluate the ranking techniques. Thirdly, we discuss our findings in the context of recently published medical literature. Finally, we show visualizations of the KG to expand on patterns by using discovery browsing.

4.3.1 Using A Knowledge Graph To Analyze Complex Associations

With the ranked multi-node patterns (i.e., discovery patterns) from the KG, we noticed that the proposed ranking methods often gave conflicting results. Table 4.9 shows an analysis to understand how each method prioritized the top 500 ranked patterns using as input our implicit relations method without pruning. We classified patterns into three different types based on the number of nodes in each pattern, where A = genes, B_n = pathophysiologic concepts, and C = target disorders, and calculated the number of unique genes, pathophysiologic concepts, and disorders.

Table 4.9: Analysis of multi-node discovery patterns

Pattern Type	Method	# of Genes	# of Physiologic	# of Disorders
A → B → C	PageRank	164	68	1
	LTC	12	88	17
	Average	21	94	11
A→ B_1 → B_2 → C	PageRank	104	50	1
	LTC	4	117	2
	Average	9	111	1
A → B_1 → B_2 → B_3 → C	PageRank	74	52	1
	LTC	1	148	1
	Average	2	123	1

Whereas PageRank tended to favour a smaller number of highly important pathophysiologic concepts from SemMedDB, LTC rankings were more diverse, though they focused on highly cited genes or disease concepts (e.g., 'COVID-19', 'Diabetes'), generating some redundant gene-disease associations as pattern length increased. Averaging the two metrics mitigated some of this redundancy and narrowed down the list of genes, disfavouring genes that were not mentioned in the literature. The two ranking methods initially gave contrasting scores, which was expected since PageRank and LTC are calculated in different ways. To address that issue, we used min-max normalization [165] to rescale the two metrics to be in the range of 0 to 1. Finally, to reduce noise we only considered patterns where the two initial scores deviated by less than a factor of 10. A summary of pattern ranking noise reduction is shown in Table 4.10.

Table 4.10: Summary of pattern ranking noise reduction

Pattern Type	# of Patterns	# of Omitted	Noise
A → B → C	20,826	17,231	82.7%
A→ B_1 → B_2 → C	161,252	157,196	97.5%
A → B_1 → B_2 → B_3 → C	545,377	538,853	98.8%

The noise reduction technique caused several patterns with high PageRank scores to be omitted by removing genes that did not have any linking terms (i.e., A – *x* – B) with the target disorders. LTC favoured genes that had several indirect associations with COVID-19, DM, or CKD, causing well-known associations to be highly ranked due to the abundance of connections between recent literature concepts.

4.3.2 Analyzing Relation Types To Capture Important Patterns

Upon closer inspection of the top 500 ranked patterns using the 'Average' method, we noted that certain relations tended to occur more than others, often drowning out more

interesting patterns. To demonstrate this, we provide a summary of the top 500 discovery patterns' relation types in Table 4.11. In the table, relation types in italics refer to relations that were generated by SemRep [146] while lowercase relations were found using GO [139]. We calculated the cumulative count of each relation type based on its position in each pattern where r_1 = relations between A and B_1, r_2 = relations between B_1 and B_2, and so on.

Table 4.11: Analysis of discovery pattern relation types

Relation Type	**Count (r_1)**	**Count (r_2)**	**Count (r_3)**	**Count (r_4)**
Affects	780	965	623	330
Coexists with	0	343	133	0
Causes	141	99	49	0
Augments	334	0	0	0
Disrupts	209	0	0	0
Stimulates	23	0	0	0
involved in	349	0	0	0
acts upstream of	10	0	0	0
Associated with	0	90	169	156
Predisposes	0	2	20	14
isa	0	190	63	0
part of	0	8	12	0
regulates	0	45	38	0

Certain relation types were generalized ('AFFECTS') while others were significant to disease mechanisms ('CAUSES', 'AUGMENTS') and high-level pathophysiologic relationships ('COEXISTS_WITH', 'ASSOCIATED_WITH'). Whereas causal relations tended to dissipate with increased pattern length, weaker, associative relations became more common. To capture meaningful indirect associations, we filtered the patterns by

specifying active relations [156] (e.g., 'CAUSES', 'AUGMENTS', 'STIMULATES') between gene and pathophysiologic concepts at r_1 to focus on disease-specific mechanisms. We then filtered out uninformative relations between pathophysiologic concepts at r_2 and r_3 ('AFFECTS', 'inverse_isa'), focusing on more granular relationships. Finally, we removed 'ASSOCIATED_WITH' relations at all positions as they were often weak or redundant.

To further understand our ranking methods, we compared the ranking distributions of different types of implicit relations to analyze their relative importance. We calculated the percentage of patterns in the top 10th percentile of rankings of two groups of patterns, namely those comprising only taxonomic implicit relations ('isa') and those comprising other types of relations (e.g., 'part_of', 'has_sibling', 'regulates'). We focused on patterns containing four or more nodes since they involved implicit relations. The analysis of pattern ranking distributions is shown in Table 4.12. In the table, we refer to the percentage of top-ranked patterns as 'Top' with the pattern group shown adjacently in brackets.

Table 4.12: Analysis of pattern ranking distributions

Pattern Type	Method	Top (taxonomic)	Top (other)
A → B_1 → B_2 → C	PageRank	0.8%	4.4%*
	LTC	5.9%	4.7%
	Average	0%	5%**
A → B_1 → B_2 → B_3 → C	PageRank	5%	5.6%
	LTC	7.9%	7.2%
	Average	6.0%	7.6%
A → B_1 → B_2 → B_3 → B_4 → C	PageRank	7.1%	12.5%
	LTC	10.4%*	9.5%
	Average	5.2%	0%

While PageRank showed preference for other (i.e., non-taxonomic) relations, LTC rankings were in favour of child (i.e., taxonomic) relations, though this distinction was not always statistically significant. Further, the importance of both taxonomic and non-taxonomic relations increased with pattern length, with taxonomic patterns becoming relatively more important. Taxonomic relations performed the best out of all relation types as patterns remained coherent as the number of nodes increased. Our analysis thusly indicates that patterns comprising implicit relations may become more meaningful as the granularity of alternative concepts increases, and that LTC may be better at emphasizing such patterns.

4.3.3 Evaluating The Accuracy Of Top Ranked Discovery Patterns

We provide mechanistic associations (i.e., indirect gene-disease associations) that require confirmatory evidence as we expedite discoveries as bases for future biomedical research. To avoid confusion, we refer to novel indirect associations as *hypotheses*. Table 4.13 presents the top 10 ranked discovery patterns found by our LBD methodology without pruning. To simplify our analysis, we focused on patterns with less than five nodes as there were several redundant hypotheses across the different pattern types.

Table 4.13: Top 10 ranked discovery patterns

Node 1	Relation 1	Node 2	Relation 2	Node 3	Relation 3	Node 4
ACE2 protein	*Disrupts*	Immunoglobulin binding	*Coexists with*	COVID-19	-	-
Spike protein, SARS-CoV-2	*Augments*	Angiotensin converting enzyme activity	*Coexists with*	COVID-19	-	-
TNF protein	involved in	Extrinsic apoptotic signaling pathway	has sibling	Extrinsic apoptosis	*Coexists with*	COVID-19
NF-kappa B	*Augments*	Excretory function	*Causes*	Complement activation	*Coexists with*	COVID-19
NF-kappa B	*Stimulates*	Signal transduction	*Coexists with*	COVID-19	-	-
NFE2L2 gene	*Stimulates*	Antioxidant activity	*Coexists with*	COVID-19	-	-
N protein, SARS-Cov-2	*Augments*	Angiotensin converting enzyme activity	*Coexists with*	COVID-19	-	-
Leptin	involved in	Regulation of steroid biosynthetic process	regulates	Steroid biosynthesis	*Coexists with*	Diabetes
Interleukin-6	*Disrupts*	Phosphorylation	*Coexists with*	Diabetes	-	-
Adiponectin	*Stimulates*	Phosphorylation	*Coexists with*	T2DM	-	-

There were some inaccuracies among the discovery patterns with regards to relations between adjacent concepts. Further, some hypotheses were uncertain as mechanistic associations did not coincide with the disease-related literature. To avoid misinterpreting each relation, we indexed SemMedDB for the original sentence that was read by SemRep [146]. Subsequently, we investigated each hypothesis by reviewing abstracts and full texts of relevant articles using keyword searching in PubMed [103]. An analysis of discovery pattern hypotheses is shown in Table 4.14. Similar to a previous work [156], we created three groups of patterns where Type 1 = valid relations and valid hypothesis, Type 2 = invalid relations but valid hypothesis, and Type 3 = invalid relations or invalid hypothesis. For brevity, we show the top 10 hypotheses returned by our method here and a list of the top 20 hypotheses is included in Appendix G.

Table 4.14: Analysis of discovery pattern hypotheses

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	3 nodes	ACE2 protein	COVID-19		x	
2	3 nodes	Spike protein, SARS-CoV-2	COVID-19		x	
3	4 nodes	TNF protein	COVID-19	x		
4	4 nodes	NF-kappa B	COVID-19			x
5	3 nodes	NF-kappa B	COVID-19		x	
6	3 nodes	NFE2L2 gene	COVID-19		x	
7	3 nodes	N protein, SARS-CoV-2	COVID-19	x		
8	4 nodes	Leptin	Diabetes	x		
9	3 nodes	Interleukin-6	Diabetes		x	
10	3 nodes	Adiponectin	T2DM		x	

Among the top 10 hypotheses, there were several type 2 patterns where the relations between concepts were uncertain due to errors made by SemRep. On the other hand, we captured three type 1 patterns that were entirely accurate, including two patterns comprising implicit relations, though the resulting hypotheses were somewhat obvious. Although longer patterns (5 or 6 nodes) were a source of valid relations and more obscure hypotheses, they were given low average scores due to our calculation methods. To address this issue, we considered using a weighted average whereby PageRank and LTC scores were assigned weight values of 0.25 and 1.75, respectively, to focus on longer (i.e., more experimental) patterns. We then compared our method with the explicit relations method [30], using the precision at k metric [120] (P@K) to determine whether implicit relations provide more accurate hypotheses to show if our method is more advanced. Type 1 patterns were considered to be accurate while all other patterns were inaccurate. An evaluation of the top ranked hypotheses is shown in Table 4.15, comparing the results of the two ranking mechanisms. In the table, we refer to the precision of the original and modified ranking mechanisms as 'Average' and 'Weighted', respectively. A P@K of 1 indicates ideal performance while a score of 0 indicates poor performance. Lists of the top 20 hypotheses for each ranking method are included in Appendix H.

Table 4.15: Evaluation of top ranked hypotheses

Method	Metric	Average	Weighted
Explicit relations	P@5	0	0
	P@10	0	0
	P@20	0	0
Implicit relations	P@5	0.2	0.4
	P@10	0.3	0.4
	P@20	0.3	0.25

Implicit relations accounted for more accurate hypotheses than patterns comprising two or more explicit (SemRep) relations. Explicit relations generated several type 2 patterns that were inaccurate due to invalid relations between concepts, resulting in poor performance, although the gene-disease associations were supported in relevant literature. The weighted average technique favoured longer patterns, though it was biased toward highly cited disease concepts (e.g., 'Diabetes'). There was no notable difference between the two ranking techniques as neither allowed us to focus on more obscure hypotheses. However, the top 10 weighted ranked hypotheses were more accurate than the average rankings, which may be useful in cases where researchers investigate a handful of hypotheses [156]. Moreover, implicit patterns were often hard to judge due to circumstantial evidence (e.g., a pathway was only mentioned, rather than discussed, in related literature). Despite all the relations in a pattern being valid, the hypothesis was doubtful as there was no theoretical overlap between relations in the same pattern.

4.3.4 Evaluating Pruning Methods To Improve Our Approach

To further understand our method to determine whether the augmentation process can be refined, we assessed the accuracy of top ranked patterns after applying the pruning methods. We excluded 3-node patterns from our analysis to focus on more obscure patterns without altering the ranking mechanism. Further, we excluded patterns containing more than two explicit (SemRep) relations to mitigate the number of type 2 patterns. An evaluation of top ranked hypotheses with pruning is shown in Table 4.16. In the table, we refer to the precision of top ranked hypotheses as 'Result', where scores of 1 and 0 indicate ideal and poor performance, respectively. Lists of the top 20 hypotheses for each pruning method are included in Appendix I.

Table 4.16: Evaluation of top ranked hypotheses with pruning

Pruning	Metric	Result
None	P@5	0.6
	P@10	0.4
	P@20	0.5
Common Parents (CP)	P@5	0.6
	P@10	0.7
	P@20	0.6
Intermediate	P@5	0.8
	P@10	0.6
	P@20	0.55
Link	P@5	0.2
	P@10	0.5
	P@20	0.5

There was a slight increase in precision after pruning and CP was the best method in terms of overall performance. In general, the top 10 hypotheses were well known, with more interesting associations occurring in the 10-20 range. CP and Intermediate offered a mix of interesting and well-known results, and Link provided the most obscure hypotheses, though the latter method produced mostly narrow (i.e., disease-specific) patterns. In most cases, the relations between concepts were entirely accurate, though some patterns were uncertain as non-taxonomic relations (e.g., negatively_regulates) were inconsistent with mechanisms described in the literature. We noted that pattern length decreased in parallel with the pruning methods, whereby longer patterns were most frequent with no pruning and least frequent with Link pruning. Our results indicate that Intermediate and Link pruning are more applicable for implicit relations with coarser granularity while CP is suited to a wider variety of relations.

It was somewhat difficult to narrow down interesting hypotheses as there were multiple patterns comprising the same gene-disease pair but with different intermediate (pathophysiologic) concepts. To simplify our analysis, we omitted duplicate gene-disease associations for patterns of a given length, validating only the highest ranked association of its kind. We observed that implicit concepts (e.g., underlying pathways) were often upstream of disease pathways discussed in the literature, meaning that they may be implicated in the pathophysiology of COVID-19, DM, or CKD. We found few articles discussing links between said pathways, suggesting that the discovered patterns are poorly understood in pathophysiological conditions.

4.3.5 Analyzing Medical Literature To Validate Mechanistic Associations

We used PubMed and Google to find supporting evidence for the discovery patterns. Discoveries were considered valid (i.e., interesting) if they were substantiated in at least one article after the cutoff date, which in our case is August 18th, 2021 since this was the latest release of SemMedDB. To ensure that each hypothesis was justifiable, we only considered studies where the underlying pathway was clearly relevant and not simply mentioned in passing. Patterns of all lengths and relation types (except those excluded as described in the previous section) were used to inform the literature analysis as we sought to identify pathways that made sense in light of multiple studies, with a specific focus on mechanisms that are relevant to COVID-19 **and** DM or CKD. We focused on implicit patterns with CP pruning, our best performing method, for the remainder of our experiments. With the selected discoveries, we explored potential disease pathways as follows:

The TLR4-NF-kappaB-NLRP3 pathway – We found several patterns linking innate immunity to the target disorders. For instance, we explored a pattern linking the NLR family Pyrin domain-containing 3 (NLRP3) protein to T2DM through innate immune pathways [166]. Cytokine overproduction in severe COVID-19 may be caused by innate immune receptors such as Toll-Like Receptors (TLRs), which coordinate the activation of NLRP3 through the transcription factor NF-kappaB, amplifying the immune response [167], [168]. Interestingly, the TLR4 protein, part of the host's innate immunity normally associated with bacterial infections, has been discussed as playing a role in the progression of COVID-19 [169] and DM [170], causing an inflammation cascade that induces tissue damage and insulin resistance in the pancreas and kidneys. We found one published article [171] after the cutoff date that discusses TLR4 activity in COVID-19, indicating that the underlying mechanisms are still poorly understood and that this is a potentially interesting association.

The SIRT1-HIF1A pathway – Abnormal inflammation and oxidative stress may cause patients with underlying chronic conditions to experience worse COVID-19 outcomes [172]. We explored a pattern linking Sirtuin 1 (SIRT1) to T2DM through regulation of glucose metabolism [173]. SARS-CoV-2 infection may cause rewiring of immune cells toward glycolysis through the action of Hypoxia-Inducible Factor-1-Alpha (HIF1A), allowing the virus to replicate more quickly [174], [175]. Further, the HIF1A-glycolysis pathway may be modulated by the SIRT1 transcription factor in immune cells and other cells and SIRT1 is downregulated under conditions of oxidative stress [176]. There was one article published after the cutoff date [177] and one article published before the cutoff date [178] investigating the link between SIRT1 and COVID-19 progression in T2DM

patients. The SIRT1-HIF1A hypothesis is considered interesting as we did not find any articles discussing the role of HIF1A and SIRT1 in COVID-19 patients with underlying chronic conditions.

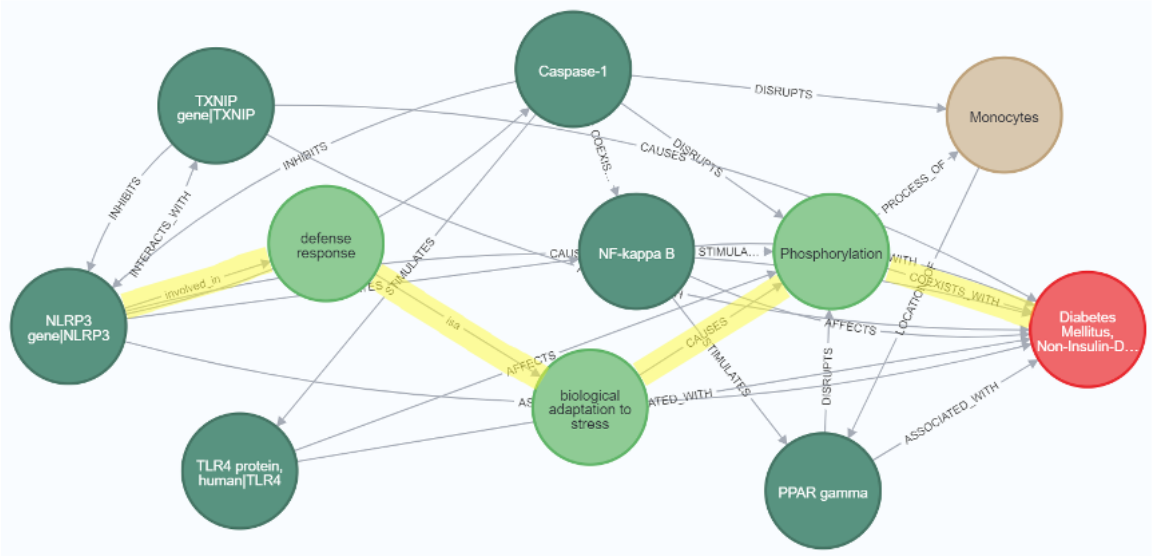
The Cx43-ADAMTS13 pathway – COVID-19 patients and DM patients have a high risk of developing lethal blood clots known as microthrombi [179], [180]. We explored a pattern linking the Gap Junction Alpha-1 gene (GJA1, encodes the Cx43 protein) to Diabetes through endothelial stress [157]. In an experimental model of diabetes, mice that lacked ADAMTS13 protein (that may be implicated in DM) had altered distribution of Cx43 in association with an increased propensity for sudden cardiac arrhythmia [181]. Interestingly, a study published before the cutoff date found that COVID-19 severity correlated with decreased levels of ADAMTS13, confirming a prothrombotic status [179]. While the authors did not anticipate that ADAMTS13 activity was altered, only its expression, the findings presented in [181] suggest that there may be a role of ADAMTS13 in diabetes beyond antithrombotic activity. Moreover, we found one article published after the cutoff date showing that SARS-CoV-2 proteins are capable of degrading Cx43, and that diabetic endothelial cells are susceptible to these effects [182]. However, there were no articles listed in PubMed or Google discussing the role of Cx43 and ADAMTS13 in COVID-19 and DM, indicating the interestingness of this hypothesis.

4.3.6 Generating Subgraphs To Explore Associations Beyond Patterns

To explore complex associations that were not captured by patterns, we created subgraphs of interrelated genes and pathophysiologic concepts using a combination of structured (i.e., query-based) and unstructured (i.e., open-ended) browsing, referred to as discovery browsing. We explored top ranked hypotheses by using the graph interface to expand each

concept in a given pattern to find associations that were interesting (i.e., relevant and previously unknown) in that context. To ensure background information was relevant to each hypothesis, we supplemented the KG by using interesting concepts as input for queries in PubMed to determine whether any research exists for a given association [103] (e.g., we searched for 'TLR4' AND 'Type 2 Diabetes Mellitus' AND 'Monocytes'). Figure 4.3 shows a visualization of a subgraph exploring the association between NLRP3 and T2DM.

Figure 4.3: Visualization of a subgraph for the NLRP3-T2DM hypothesis

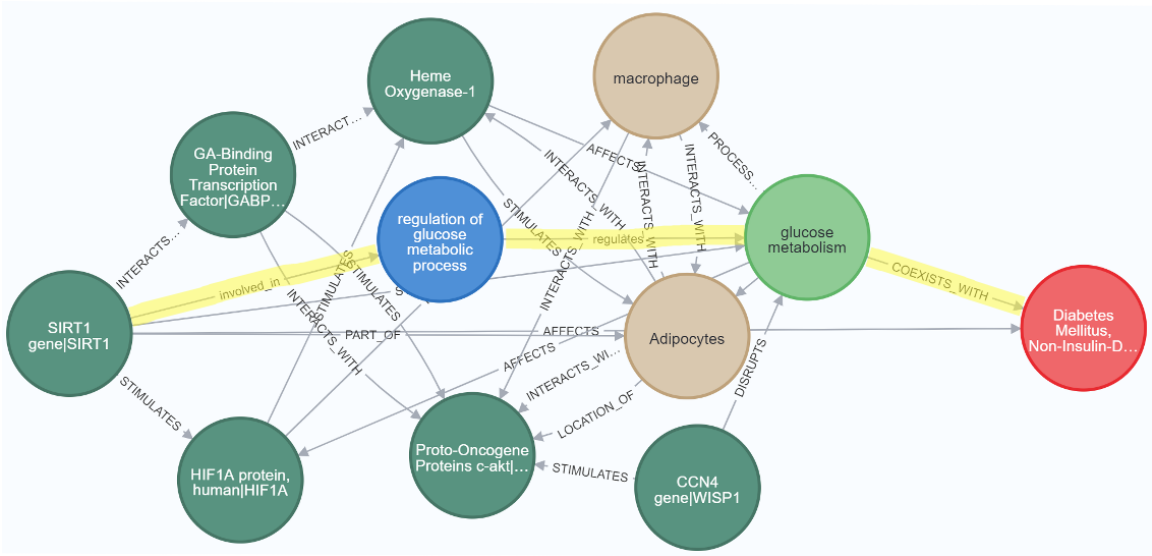


In Figure 4.3, genes are shown as dark green nodes, pathophysiologic concepts are light green, anatomic concepts are light brown, and disorders are red. The highlighted arrows are relations that show the original discovery pattern linking NLRP3 (left) to T2DM (right) through innate immune pathways, represented as implicit and explicit relations. Note that some of the relations adjacent to the original pattern are inaccurate as they were captured by SemRep without manual review. While it was difficult to pinpoint other pathophysiologic concepts (i.e., pathways) as the related literature was too expansive, we

found three additional genes that may be relevant to the TLR4-NF-kappaB-NLRP3 hypothesis, namely TXNIP, Caspase-1, and PPAR gamma. The TXNIP gene is involved in activating inflammatory responses (including NLRP3 signaling) in circulating immune cells and its expression may be upregulated in T2DM [183]. Activated NLRP3 stimulates NF-kappaB and Caspase-1, the latter of which mediates inflammation and cell death in infected cells and may contribute to insulin resistance in adipose tissue [184]. Further, SARS-CoV-2 blocks the action of Caspase-1 in monocytes, which may cause an accumulation of inflammatory cytokines that are subsequently released, causing severe illness in patients [185]. Finally, PPAR gamma regulates inflammatory genes such as NF-kappaB in macrophages and may be implicated in the progression of T2DM and COVID-19 [186], [187]. Our results thusly indicate that discovery browsing was effective by identifying previously unknown associations that build on an identified hypothesis.

To further explore our results, we generated a subgraph to expand the association between SIRT1 and T2DM through regulation of glucose metabolism. We used the following search strategy in PubMed: ‘SIRT1’ AND ‘Type 2 Diabetes Mellitus’ AND ‘Oxidative Stress’ AND ‘X’, where X was a gene found by expanding a concept in the original discovery pattern. A visualization of a subgraph for the SIRT1-T2DM hypothesis is shown in Figure 4.4.

Figure 4.4: Visualization of a subgraph for the SIRT1-T2DM hypothesis

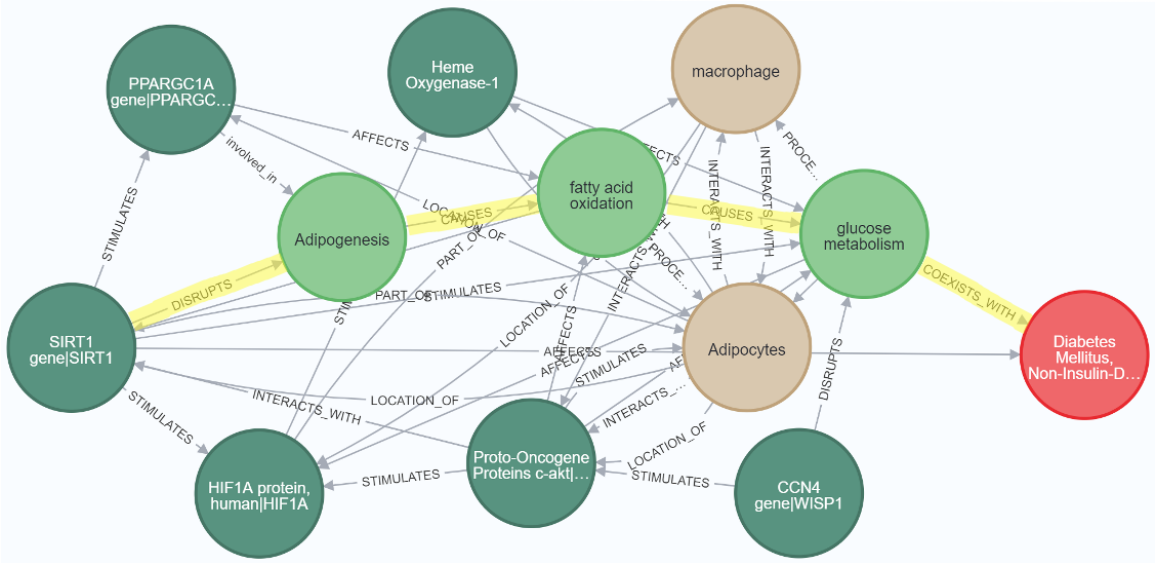


We found four additional genes that appear to be interrelated through SIRT1-mediated glucose metabolism pathways in T2DM, namely GABPA/Nrf2, Heme Oxygenase-1 (HO-1), AKT, and CCN4/WISP1. There was less research in PubMed concerning these genes as compared with those shown in the previous subgraph, with each query returning only one relevant article. We noted that the discovery browsing process generated several genes that were not relevant to the hypothesis, which was especially noticeable after we expanded implicit concepts (e.g., regulation of glucose metabolic process, shown in blue), indicating that there was minimal benefit of exploring ontology relations beyond the context of the identified pattern. To simplify our analysis, we omitted genes that had fewer than two valid relations with genes and gene functions in the original pattern. For brevity, we do not discuss the individual genes here, though it is worth mentioning that they are involved in diverse pathways as mediators of oxidative stress and glucose homeostasis in immune cells and adipocytes [188]–[190], potentially contributing to insulin resistance in a disease state.

4.3.7 Using Subgraphs To Evaluate LBD System Output

To assess the efficacy of our method, we compared it to the explicit relations method [30] by repeating subgraph generation to expand a similar hypothesis linking SIRT1 to T2DM through glucose metabolism. We chose this hypothesis as it was the only one where a similar pattern existed in the top ranked hypotheses from the explicit relations KG. To ensure that our methods were consistent, we applied an identical search strategy in PubMed as for the previous subgraph. Figure 4.5 shows a visualization of a subgraph of explicit relations for the SIRT1-T2DM hypothesis.

Figure 4.5: Visualization of a subgraph of explicit relations

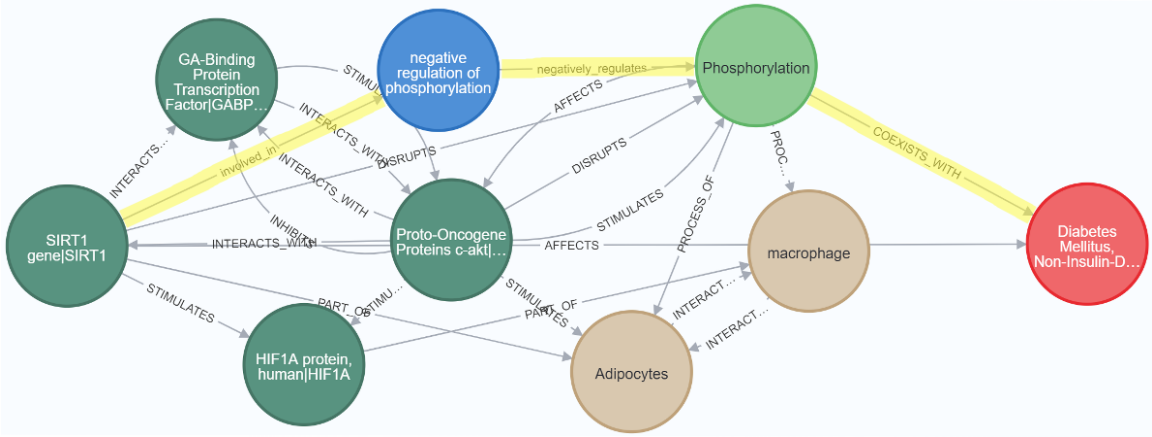


The original discovery pattern, highlighted in Figure 4.5, contains inaccurate relations that were generated by SemRep, and is therefore misleading. For instance, fatty acid oxidation does not cause glucose metabolism; rather, it is an alternate pathway that is preferred by cells under normal conditions [191]. Following this pattern from SIRT1 to T2DM, we encountered associations that were difficult to interpret as there was no shared context from one relation to the next. Note, however, that the subgraphs generated by explicit and

implicit relations are similar aside from few differences in the genes and gene functions present. We found that GAPBA/Nrf2 was replaced the PPARGC1A gene, the latter of which may have renoprotective effects in T2DM and CKD [192], though both genes are important as they are involved in SIRT-1 mediated pathways. Our results indicate that the implicit relations method is better at identifying coherent gene-disease associations than a previous method, though the number of discoveries made is similar.

To further assess our method to understand whether pruning affects the coherence of mechanistic associations, we repeated the subgraph generation process to expand the SIRT1-T2DM hypothesis without pruning. We focused on the highest ranked pattern as there were multiple patterns comprising the same gene-disease pair with different mechanistic associations. A visualization of a subgraph of implicit relations without pruning is shown in Figure 4.6.

Figure 4.6: Visualization of a subgraph of implicit relations without pruning



The SIRT1-T2DM discovery pattern without pruning involved a different set of mechanistic associations as the previous subgraphs, but the hypothesis was similar as it pertained to insulin resistance of monocytes in T2DM [166]. Here, the hypothesis is that

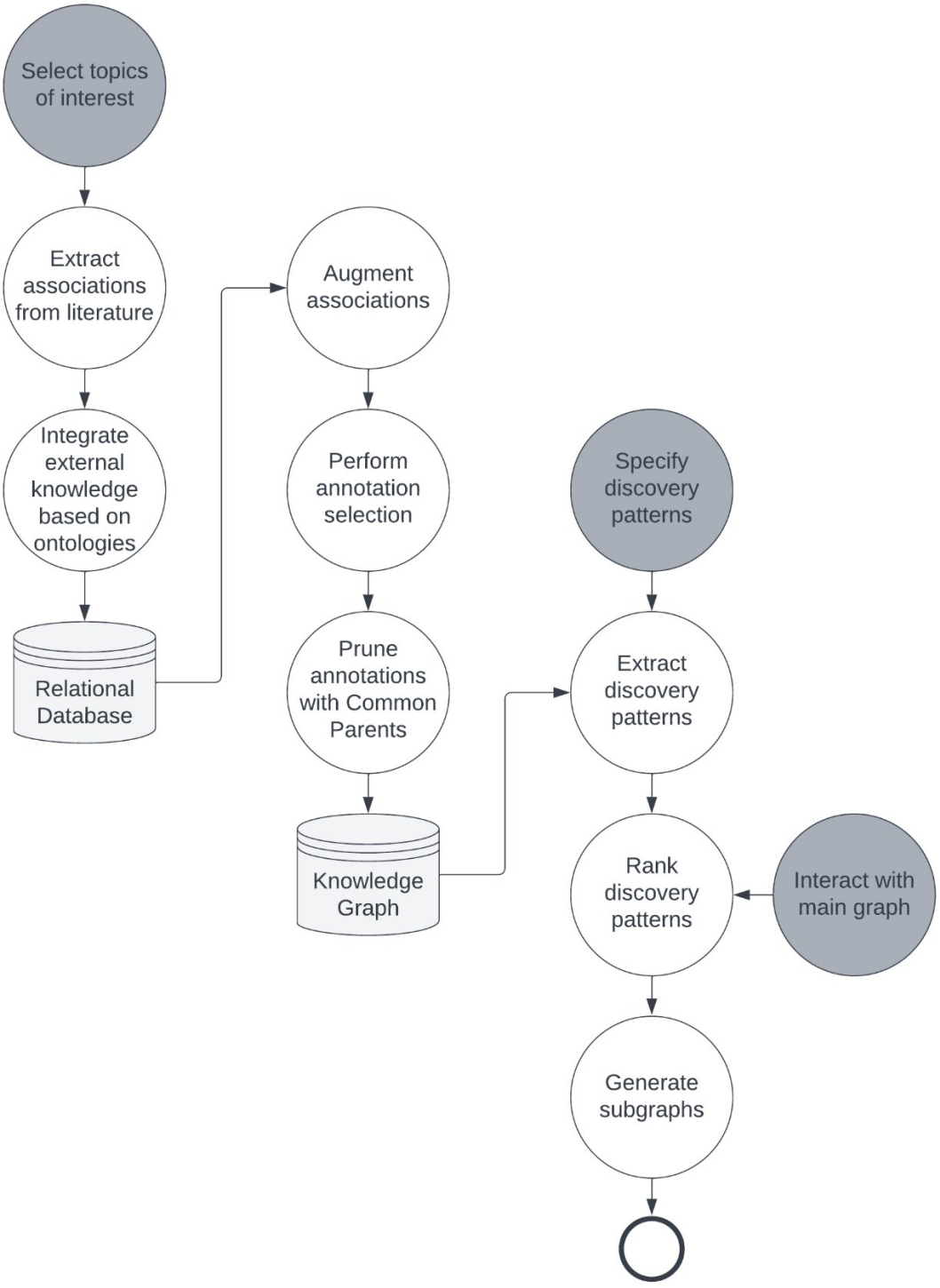
SIRT1 negatively regulates an insulin signaling pathway that is altered in T2DM. Firstly, the ontology relation shown in the center of pattern is inaccurate, as SIRT1 probably has a positive effect on insulin signaling [193], instead of a negative one. This is interesting because the same pattern did not exist in the KG with CP pruning. Secondly, we found fewer relevant genes through discovery browsing compared to CP. We observed that the HO-1 and WISP1 genes were missing even though all other genes in the subgraph were identical to the previous ones. Our results thusly indicate that pruning is beneficial to our method as it emphasizes more coherent patterns that are conducive to discovery browsing.

In summary, our approach generated more accurate hypotheses than a previous LBD method [30], referred to as explicit relations, by integrating relations from multiple public databases to augment mechanistic associations from medical literature. We demonstrated that external knowledge (i.e., ontologies, annotations) was a useful source of background information as it allowed the user to explore interesting associations that formed coherent links as indirect gene-disease associations. In this regard, ontology relations provided hidden intermediate concepts that created links between previously unrelated concepts by improving the granularity of mechanistic associations (e.g., pathophysiologic processes), though these relations generated a significant amount of noise. Our approach performed best with taxonomic relations, where alternative (i.e., related) concepts were connected by narrower (i.e., ‘isa’) relationships, and performance was improved by filtering concepts with Common Parents (‘CP’ pruning) to focus on patterns with semantically similar concepts. Moreover, combining graph- and frequency-based ranking methods emphasized interesting and important patterns, but it also caused most experimental relations (i.e., ontology annotations) to be omitted as noise. Finally, the use of discovery pattern mining

techniques led to several interesting hypotheses that were validated and further explored by interacting with the KG, using evidence from PubMed articles to ensure that the hypotheses stayed relevant.

To help the reader understand the final approach based on our evaluation results, we show a flowchart of the optimal approach for the featured work in Figure 4.7. In the figure, dark grey nodes represent user actions, white nodes are system processes, light grey cylinders are data stores, and arrows are workflow activities.

Figure 4.7: Flowchart of the optimal approach for the featured work



4.4 CASE STUDIES

In this section, we evaluate our approach by applying our LBD framework to address two unanswered COVID-19 research questions, with the aim of uncovering plausible associations that are not evident by querying a medical literature database (i.e., PubMed). In each subsection, we introduce a research question (i.e., the 'input') and show how it is processed by briefly describing the activities of our framework and their outcomes (i.e., the 'outputs'). Finally, with the associations identified by our approach, we discuss whether each result addresses the corresponding research question.

4.4.1 SARS-Cov-2 Virulence In COVID-19 Patients With DM

Several studies have questioned whether the virulence of SARS-CoV-2 is increased due to disease states associated with DM (e.g., glycemic instability) [66], [194], [195]. It was postulated that SARS-CoV-2 interferes with HIF1A-mediated glucose control in host cells, allowing it to replicate more quickly [174], yet it is unclear how this leads to the pathophysiology of severe COVID-19. We hypothesize that there is an indirect association (i.e., discovery pattern or subgraph) involving HIF1A and COVID-19 that may explain underlying mechanisms in COVID-19 patients with DM.

Predication Extraction & Predication Extension – Firstly, we searched for relations in studies of patients with COVID-19 in SemMedDB [141] linking HIF1A to pathophysiologic concepts that were in turn associated with COVID-19 (i.e., mechanistic associations). Secondly, we augmented mechanistic associations by searching for alternative (i.e., related) concepts in external knowledge (i.e., UMLS ontologies [140], GO annotations [139]). Thirdly, we pruned alternative concepts using CP pruning. To compare the results of these three activities to see if our approach improves the output of mechanistic

associations, we calculated the number of unique paths (i.e., sets of relations) linking HIF1A to COVID-19 through pathophysiologic concepts before and after augmentation. An analysis of mechanistic associations between HIF1A and COVID-19 is shown in Table 4.17. In the table, we refer to the number of non-augmented, augmented, and pruned paths as 'Before', 'After', and 'Pruned', respectively.

Table 4.17: Analysis of mechanistic associations between HIF1A and COVID-19

Path length	Before	After	Pruned
3 nodes	1	2	2
4 nodes	1	7	5
5 nodes	3	20	16
6 nodes	7	89	48

Augmenting mechanistic associations created more paths between HIF1A and COVID-19, though alternative concepts were often not relevant to pathways described in the literature (i.e., glucose metabolism). As such, the augmentation process increased the quantity, but not necessarily the quality, of mechanistic associations. Pruning associations improved the outcome as it reduced the number of redundant paths, and we use CP pruning for the remainder of our experiments.

Discovery Patterns – We represented the relations in a KG along with other relations from studies of DM and CKD. To ensure that mechanistic associations were meaningful, we used pattern mining to identify interesting associations (referred to as hypotheses). This activity reduced the number of associations as we chose to focus on more granular relationships. For instance, we ruled out paths such as:

HIF1A -involved_in→ HIF1A signaling pathway -isa→ Signal Transduction -AFFECTS→ COVID-19

We then applied ranking metrics (i.e., PageRank and LTC) to prioritize the results of pattern mining to focus on important hypotheses. Pattern mining caused most paths between HIF1A and COVID-19 to be omitted as they were too vague, leaving one pattern with 3 nodes and one with 5 nodes. The remaining patterns were ranked favourably (i.e., placed in the top 10^{th} percentile of rankings) by PageRank and LTC, but combining PageRank and LTC (i.e., Average) excluded them as noise. We decided to recover these patterns for further investigation as we suspect the noise calculations were biased toward highly cited concepts (e.g., 'Diabetes'). After validating each relation in SemMedDB and PubMed, we selected the following pattern:

HIF1A protein -STIMULATES→ Glycolysis -CAUSES→ lactate biosynthesis -isa→ lactate metabolic process -COEXISTS_WITH→ COVID-19

Where explicit relations (i.e., relations captured by SemRep [146]) are shown in all caps and the relation in lowercase was extracted from UMLS ontologies. All the relations in the pattern are valid and the hypothesis is that HIF1A contributes to severe COVID-19 through lactate production [196]. We found one article in PubMed that discusses our hypothesis [197], indicating that the association may be interesting to researchers.

Subgraphs – We interacted with the KG by using a combination of structured and unstructured browsing to further explore the HIF1A-COVID-19 hypothesis. We used the following search strategy in PubMed: 'HIF1A' AND 'lactate' AND 'COVID-19' AND 'X' where X was a gene found by expanding concepts in the KG. After reviewing evidence

from PubMed, we were unable to find any additional associations, indicating that underlying mechanisms of HIF1A in COVID-19 are poorly understood.

4.4.2 Immune Response To SARS-Cov-2 In COVID-19 Patients With DM

COVID-19 patients with DM or CKD may have an abnormal immune response to SARS-CoV-2 for a variety of reasons (e.g., immune dysregulation) [64], [171], [194]. It has been postulated that innate immune signaling by TLR4 may contribute to the progression of COVID-19 [167] and DM [170], but the exact mechanisms are still unclear. We hypothesize that there is a discovery pattern or subgraph involving TLR4 and COVID-19 that may explain underlying mechanisms in COVID-19 patients with DM or CKD.

Predication Extraction & Predication Extension – We repeated the first three activities described in the previous section to identify paths (i.e., sets of relations) between TLR4 and COVID-19. To see if our approach improves the output of mechanistic associations, we calculated the number of unique paths linking TLR4 to COVID-19 through pathophysiologic concepts before and after augmentation. An analysis of mechanistic associations between TLR4 and COVID-19 is shown in Table 4.18. In the table, we refer to the number of non-augmented, augmented, and pruned paths as 'Before', 'After', and 'Pruned', respectively.

Table 4.18: Analysis of mechanistic associations between TLR4 and COVID-19

Path length	Before	After	Pruned
3 nodes	0	0	0
4 nodes	0	2	0
5 nodes	1	24	12
6 nodes	2	62	31

We observed similar results from the augmentation and pruning activities as described in the previous case study. There was a wide variety of pathophysiologic processes linking TLR to COVID-19 due to the augmentation process, which may be explained by the fact that TLR4 regulates several important genes.

Discovery Patterns – After representing the relations in a KG along with relations from studies of DM and CKD patients, we applied pattern mining and ranking activities to extract interesting and important associations (i.e., hypotheses). Following the removal of unimportant patterns, PageRank and LTC rankings were in favour of two remaining TLR4-COVID-19 patterns, though the Average technique ruled them out as noise. After closer inspection, we determined that the two patterns were equivalent, and we selected the shorter pattern for further investigation, which is as follows:

> TLR4 protein -involved_in→ activation of innate immune response -isa→ positive regulation of innate immune response -positively_regulates→ Immunity, Innate -PREDISPOSES → COVID-19

All the relations in the pattern are valid and the hypothesis is that TLR4 is involved in the innate immune response to COVID-19 [198]. Our hypothesis is uncertain as we found one study in support of [199], and one against [200], TLR4 activation by SARS-CoV-2.

Subgraphs – Given that TLR4 may determine the activity of important downstream genes, we explored the TLR4-COVID-19 hypothesis further by interacting with the KG. We were unable to identify additional associations that could explain the role of TLR4 in COVID-19.

In summary, our approach was applied to two unanswered COVID-19 research questions that could explain disease impacts in patients with DM or CKD. While we identified interesting patterns that were validated by recent studies, we did not find any associations that would be considered novel. Given that the association between HIF1A and increased lactate production could be considered as general knowledge [197], the first hypothesis does not address the question of how HIF1A-mediated pathways lead to the pathophysiology of severe COVID-19. Further, since the TLR4-COVID-19 hypothesis did not lead to any new information, our approach was unable to address the question of how TLR4 signaling affects the progression of COVID-19. Finally, our approach generated several associations that were not relevant to pathways described in the literature and there was limited evidence to help identify interesting associations due to the novelty of COVID-19.

CHAPTER 5.0 DISCUSSION AND FUTURE WORK

The COVID-REdI system was designed to uncover hidden associations between the literature and external knowledge sources in a way that facilitates researchers' understanding of pathophysiologic mechanisms underlying COVID-19 and DM or CKD. The addition of external knowledge (ontologies, annotations) to semantic associations extracted from medical literature allowed high-level pathophysiologic mechanisms to be explored in depth by using intuitive patterns and visual representations to identify interesting and previously unknown overlaps between immune and metabolic pathways in these disorders.

COVID-REdI is an implementation of a novel knowledge synthesis and discovery approach. It uses a combination of literature-based discovery, medical ontologies, and knowledge graphs to represent pathophysiologic mechanisms as semantic associations where, through simple techniques and interactions, the user can identify interesting hypotheses, infer relations between previously unconnected concepts, and make sense of complex associations. The graph database and semantic associations provide the user with the ability to find complex relationships between biomedical concepts, which could be useful for understanding disease mechanisms.

5.1 THESIS CONTRIBUTIONS

The integration of semantic associations from an existing resource (i.e., SemMedDB) with external knowledge allowed us to identify three novel gene-disease associations that were supported by medical literature, awaiting validation by expert review. By comparison, a recent LBD method brought together three biochemical relations to represent a novel biological pathway which was validated in clinical tests [29]. To our knowledge, ours is

the first attempt to address the issue of incomplete knowledge in relations mined from text by combining different ranking metrics to link distant literature sources and thus increase the likelihood of discovering hidden or unknown associations. The benefit of our approach was demonstrated as it allowed us to generate mechanistic associations that were hidden from previous LBD methods [30], [127] by uncovering contextually relevant implicit (i.e., ontology) relations between pathophysiologic concepts, though the number of discoveries being made was unchanged. Moreover, the ranking and evaluation methods used in our analysis may have neglected the importance of rare or obscure hypotheses. We discuss these issues and other issues further in the remainder of this chapter.

While previous LBD studies only consider indirect links as two concepts away, such as the work to discover biomarkers for migraine [30], we generated patterns (i.e., testable indirect associations) that synthesize important biological relationships to uncover interesting associations between distantly related concepts. Further, our approach builds on the intermediate relations method proposed by Cameron et al [127] by extending the coverage of alternative concepts to form intelligible associations as chains of related concepts. Our method differs from previous works by Cameron et al [88], [95] in that it aims to merge background (external) knowledge with relations mined from text instead of indirectly integrating the two based on pre-existing structures (i.e., expert knowledge or PubMed article MeSH terms). The augmentation process is not perfect, however, as it generates a considerable amount of noise. Our method can be improved by using a basic semantic similarity method [143], referred to as Common Parents (CP), to reduce the number of meaningless associations. Given that several studies have adapted semantic similarity

methods for the UMLS [156], [201], [202], we believe this is an excellent area for experimentation. We discuss this topic further in the Future Work section.

5.2 Usefulness Of External Knowledge

Related works [30], [127] have encountered an issue regarding the usefulness of external knowledge, due to a large amount of redundant information in knowledge repositories such as the UMLS. Cameron et al [127] noted that ontology concepts are rarely used in common language, indicating that some alternative concepts are too obscure to form meaningful associations with concepts in the literature. Further, Vlietstra et al [30] questioned whether alternative concepts are always necessary as they create long lists of results to comb through, some of which may be equivalent with each other, increasing the workload for non-expert users. While our observations appear to echo these concerns, we also noted that patterns comprising multiple alternative concepts yielded more accurate hypotheses, suggesting that these concepts may be beneficial as they allow the user to explore associations that would otherwise be found by considering multiple studies, saving time during the discovery phase. Our approach closely resembles discovery browsing, where the user is equipped with knowledge to navigate and uncover insights in selected area of interest [96]. To facilitate discovery browsing, it may be necessary to focus on a specific group of relations (e.g., pathophysiology) to reduce the number of associations that the user must consider.

5.3 Interestingness Of Ranked Associations

There were some issues with the ranking mechanism as top ranked patterns tended to focus on highly cited concepts. While these patterns offered some insight into disease mechanisms, it can be argued that meaningful discoveries are not restricted to well-known

concepts. Indeed, combining different metrics allowed us to narrow down important patterns, but our technique may have diminished the importance of rare or obscure hypotheses [118]. Cameron et al [127] noted that ranking techniques are a key requirement as alternative concepts are seldom of critical importance in the context of medical literature. It may be useful to consider ranking techniques that emphasize rare associations [88], [90] to balance the importance of alternative concepts with those mined from the literature. Further, some metrics described in [97] (e.g., transitivity) may be useful in terms of assessing whether our approach emphasizes a variety of associations between previously disconnected concepts, mitigating highly cited (or obvious) associations. Moreover, given that PageRank and LTC provide contrasting scores, there may be better combinations of metrics that are applicable to complex mechanistic associations.

Recent LBD studies [156], [201] have generated chains (i.e., patterns) of related biochemical concepts (e.g., drug side effects) by using semantic similarity to extract meaningful associations from large bodies of literature. Employing semantic similarity techniques may be preferable to frequency- and graph-based metrics that neglect the underlying semantics of associations. Other LBD studies [105], [203] have incorporated confidence scores from external databases like STRING into their ranking techniques, which may be useful for tasks such as link prediction. It is unclear what metrics are best suited to external knowledge since we were unable to find any studies published in this area. Therefore, more work is needed to determine what kinds of measures should be used to identify meaningful extensions of relations found in text.

5.4 Appropriateness Of The COVID-REDI Evaluation

Using PubMed to evaluate LBD methods is probably not ideal. There was a limited amount of evidence to support each hypothesis described in Chapter 4, and it is possible that we missed important studies that could have impacted the assessment of a given method. The lack of a gold standard for comparing the performance of LBD systems is a well-known issue [28]. Nevertheless, it would be beneficial to find a more efficient technique to evaluate our LBD framework. Cameron et al [88] developed an interestingness metric based on PubMed articles, referred to as association rarity, that measures whether LBD results are interesting to a given reader. While we considered using this metric in our preliminary experiments, it became susceptible to bias as we could have selected rare associations to form more interesting hypotheses. We found that implicit relations captured more indirect gene-disease associations compared to a previous method by utilizing an external database of relevant associations (i.e., the gene2pubmed resource [163]), but the importance of these associations was unclear as many were subsequently excluded as noise. Moreover, it is uncertain whether the use of pattern types [156] to judge the validity of hypotheses (i.e., Type 1 = valid relations and valid hypothesis) is a fair assessment technique when the relations were generated by different methods, such as SemRep [146] or the Gene Ontology (GO) [139], since each method captures biological relations with different levels of validity. We discuss this issue further in the Future Work section.

5.5 Scalability Of COVID-REDI

Discovery patterns require foreknowledge of potentially interesting relations, which may limit the complexity of discoveries being made [95]. Although patterns may allow the user to narrow down interesting gene-disease associations by limiting the number of intermediate terms, it is difficult to predict important patterns in a given domain, especially

when underlying structures (e.g., disease mechanisms) are poorly understood. We explored patterns by interacting with the KG to further expand interesting gene-disease associations, but this process created too many associations to consider. Previous LBD methods [95], [96] have viewed this issue as predicting intermediate concepts (B), where the source (A) and target (C) concepts are already known and can be used as input in path-finding algorithms. However, these methods neglect the possibility of open-based discovery, where only A or C is known. Recent LBD studies employ machine learning (ML) methods, such as the work to understand the impacts of smoking in males and females [104], using an unsupervised learning algorithm with multiple ranking metrics to investigate several targets of interest in an open-ended manner. Further, Henry et al [204] employed hierarchical clustering to group target concepts, allowing the user to explore branches of the hierarchy they find most interesting. Given that the augmentation process generates too many associations to be practical for manual investigation, it may be worthwhile to consider the use of ML techniques to aid the process of identifying interesting hypotheses.

Interacting with the KG, we found interesting and complex associations that were more than two concepts away from the target concept, but exploring alternative concepts often led to associations that were irrelevant to the original hypothesis. We observed that ontology relations were inconsistent with phenomena described in the literature, potentially stemming from a lack of contextual cues. Consequently, our analysis was limited to shorter patterns (i.e., 5 nodes or less) to control the growth of associations, which is not ideal since longer patterns (i.e., 6 nodes) could harbor more interesting and complex associations. Previous LBD studies have employed storytelling algorithms to explore complex patterns, such as the work to understand cytokine networks in disease states [205], using a context

overlap filter to ensure that the story remains coherent as it moves from one study to the next. Moreover, it may be possible to use different forms of external knowledge, such as the work to capture clusters of relations that exceeded a threshold of relatedness based on article content (i.e., MeSH index terms) [88], to further improve our method. Unfortunately, external knowledge (i.e., ontologies, annotations) is limited by its completeness, causing the user to infer relations based on well-known associations. Related works [91], [206] attempt to solve this issue by utilizing intrinsic patterns found in ontologies in combination with expert domain knowledge to predict new relations, which could be useful in cases where external knowledge is incomplete. More work is needed to determine if previous methods can be adapted for implicit relations, such that alternative concepts are better aligned with up-to-date research and expert knowledge.

5.6 Implications Of COVID-REDI For COVID-19 Research

Although biological markers of severe disease are reported extensively in the COVID-19 literature, few studies attempt to explain their roles in disease mechanisms. Our approach identified hypothetical genetic contributions to disease mechanisms of COVID-19 and DM or CKD, an ongoing research area that is burdened with sparse knowledge [42], [195]. Given the inherent complexity of disorders like COVID-19, DM, and CKD, there is an urgent need for studies that investigate molecular disease pathways involving multiple genes and gene products. We generated indirect gene-disease associations between distant literature sources that may be relevant to mechanisms driving the long-term impact of COVID-19 on patients with underlying chronic conditions. Through our research, we provide hypotheses that are discussed in line with our research questions below. While we

were able to find supporting evidence for these hypotheses in PubMed, we await expert validation to determine whether they are meaningful.

How might the underlying disease states in DM or CKD predispose patients to worse COVID-19 outcomes? – COVID-19 triggers an exaggerated proinflammatory cytokine response in patients with DM [207]. In particular, diabetic patients seem to have higher levels of interleukin-6 (IL6) as well as increased serum C-reactive protein (CRP) and D-dimer compared to non-diabetics [208]. Interestingly, severe COVID-19 patients exhibit increased TLR4 expression, which may lead to hyperactivation of upstream innate immune pathways that trigger inflammatory cytokine and oxidative stress mechanisms [171], [199]. While evidence connecting TLR4 to diabetes-associated severe COVID-19 is elusive, the ability of SARS-CoV-2 to disrupt vascular physiology, which is already disrupted in diabetic patients, may enable crosstalk with reactive oxygen species (ROS) production, leading to vascular complications and progression to severe illness [171]. Further, TLR4 appears to induce NLRP3 and its downstream proteins (e.g., Caspase-1) that may be involved in driving the progression of COVID-19 and T2DM through inflammation and oxidative stress [184], [209].

How might COVID-19, especially in severe cases, exacerbate DM or CKD? – Given that individuals with DM experience chronic inflammation and oxidative stress, it is possible that they will have a dysregulated cellular response to COVID-19 [178]. TLR4 activity may be particularly important to the long-term impact of COVID-19 on patients with DM as it may promote insulin resistance and tissue damage in the pancreas and kidneys, driving the progression of chronic illness [170], [171]. Further, the ability of SARS-CoV-2 to alter metabolism in circulating immune cells including pathways such as the SIRT1-HIF1A

glycolysis axis [174] may be of concern for patients with DM as it precedes insulin resistance, though further investigation is needed in severe COVID-19 to support this hypothesis.

5.7 Significance Of COVID-REDI For Biomedical Research

We envisage two ways in which our approach could be useful to biomedical researchers. Firstly, by bringing together disparate knowledge fragments to find associations between components of similar diseases, which may be of interest to clinician researchers and to researchers studying novel or poorly understood diseases. Our method aims to fill the gaps between genetic and pathophysiologic processes, which has the potential to explain complex gene-disease associations [50]. In this regard, the main challenge facing our work is to distinguish meaningful associations from those that are misleading, both of which could occur in the context of sparse knowledge. Previous LBD studies [29], [119] have dealt with this issue by seeking expert input to help identify promising hypotheses that are pertinent to current research interests. Given that our LBD framework is designed to identify meaningful targets to support biomedical research, we believe that including knowledgeable experts from different fields in the process of selecting these targets is an excellent area for future work.

Secondly, we provide a means to narrow down potential disease genes from a large list of candidates, which could be beneficial to genomics researchers. There is a related body of literature on this issue, where the aim is to identify and rank a list of genes given a disease query [210]. Here, previous studies consider concept co-occurrence as direct and indirect associations between genes and diseases, using statistical metrics (e.g., cosine similarity) to rank associations [210], [211]. Our approach has the advantage of identifying links

between gene and pathophysiologic concepts that are enriched with contextual information (i.e., the nature of those relationships) from SemMedDB [141] and external knowledge (i.e., ontologies, annotations). In this context, the main challenge facing our work is to ensure that it emphasizes accurate associations. Recent studies have shown that ontologies such as GO can be augmented with phenotype ontologies and text mining of published evidence to predict gene-disease associations [105], [212]. In this sense, augmenting semantic associations with multiple ontologies could improve the output of our approach by providing additional support for novel associations.

5.8 LIMITATIONS

There are several limitations to this work. Certain studies of COVID-19 patients with DM or CKD involved small sample sizes, which could limit the applicability of our findings to a wider population. Further, the evaluation of COVID-REdI was based on studies from a single literature database (i.e., PubMed), meaning that the potential for false discoveries is higher than it would have been if multiple databases had been used.

Extending semantic associations to cover related concepts relies on multiple domain knowledge sources (ontologies, annotations) that may not be available in other research contexts. As such, the augmentation process may only be applicable to research domains where structured knowledge resources are well developed, such as biomedicine.

Creating patterns is a time-consuming process based on relations that are known ahead of time, which limits the generalizability of our methods. While patterns have been described for a variety of biomedical research questions [91], [93], [153], it is difficult to predict new patterns based on prior knowledge. This means that researchers studying a novel disease

could miss important associations if the underlying mechanisms cannot be easily inferred from the given associations.

Finally, interacting with a KG to construct interesting visualizations is manually intensive and may not be feasible with large datasets. The amount of time required to focus background information is prohibitive and may preclude forming interesting hypotheses in a timely manner.

5.9 Future Work

First, COVID-REdI will be revised based on the results of the pruning methods. Additional techniques to filter alternative concepts to reduce noise should be considered. This could involve applying semantic similarity techniques to select alternative concepts that exceed a pre-specified similarity threshold. It would be beneficial to compare different methods, such as Common Parents (CP) [143] and ontology-based similarity [202], to determine whether one is more effective at producing coherent associations.

The ranking methods will be refined to emphasize interesting and important hypotheses by investigating different combinations of metrics. LTC is a proven metric that could be compared with other metrics found in [213], and there is an option to use PageRank in combination with similar metrics such as the HITS algorithm [214]. It would also be worthwhile to experiment with semantic similarity metrics, such as those described in [202], association rarity metrics [88], graph-based metrics [97], and/or confidence scores from external databases [105] by embedding scores as weighted edges in the KG.

A better evaluation method is needed to compare the performance of explicit and implicit relations. This could involve replicating historic discoveries, as is often done in LBD, or

replicating recent discoveries (e.g., gene-disease associations) in PubMed [210]. Generating a reference set of concepts from the literature (e.g., disease biomarkers) [30] and using it to measure precision of the output is another option. Alternatively, it is possible to use association rarity metrics [88] to assess the interestingness of the results, provided that associations were not gathered by a human, which could introduce bias.

There should be a way to create subgraphs that requires less manual input. This could involve using a combination of graph traversal, graph metrics, and interestingness measures to model the behaviour of discovery browsing [215]. Ideally, the subgraph generation process would be fully automated, requiring the user to simply input target concepts. In this regard, mitigating computational requirements should be a key priority. For instance, to mitigate uninformative associations it was necessary to rule out certain associations (e.g., HIF1A -involved_in→ HIF1A signaling pathway -isa→ Signal Transduction -AFFECTS→ COVID-19) that would otherwise lead to too many associations to consider, which may be useful with regards to setting a 'ceiling' for the granularity of alternative concepts. Depending on the research context, it may be possible to employ ML techniques, which can support complex tasks such as open-ended discovery by assisting the researcher in identifying interesting hypotheses [204].

The current literature selection process could be improved. It would be beneficial to include expert input to align with current research interests. Recent knowledge synthesis platforms use crowdsourcing to identify research questions, such as the COVID-19 Rapid Evidence Access Link (REAL) [47], which could be a useful approach for future implementations. Moreover, manually defining PubMed search terms (i.e., MeSH terms) is time-consuming and could be assisted by techniques such as topic modeling [216].

Finally, a more comprehensive knowledge integration approach is needed to provide useful information that is not captured in relations mined from text. The integration task should include multiple public databases of relevant information (e.g., gene and protein interactions). Including phenotype and other functional information (e.g., gene pathways through the Kyoto Encyclopedia of Genes and Genomes (KEGG)) could further improve the breadth of the knowledge base.

CHAPTER 6.0 CONCLUSION

This thesis describes the design, implementation, and preliminary evaluation of the COVID-REdI system for studying molecular interactions between COVID-19 and DM or CKD. The results from this work show that the system holds potential to address some of the current gaps faced by other systems in LBD. The methodology used to develop the COVID-REdI system is novel in several ways. While other semantics-based LBD systems [77], [96], [103] rely on information found in text to establish novel associations, to our knowledge this is the first time that external knowledge sources beyond published literature have been incorporated and used to find complex indirect associations. One other system integrated external knowledge sources [30] but it only considered indirect associations as 2 concepts away. The COVID-REdI system is novel in that it automatically integrates structured knowledge sources to generate series of related concepts (i.e., discovery chains) from up to 6 concepts away. The methods used in this work are generalizable to other conditions, and thus we present a novel method for knowledge synthesis and discovery in those conditions.

The potential positive impacts of an updated COVID-REdI system include discovery of indirect gene-disease associations in the literature [90], [96], novel mechanistic hypotheses [3], [29], [103], disease gene prediction [105], and new knowledge to help researchers understand their domain of interest [29], [119]. Additionally, the COVID-REdI system was able to identify interesting pathophysiological pathways that may be perturbed in COVID-19 patients with DM or CKD. This could lead to improved speed of knowledge translation of recent findings to researchers who are overcome by the current overabundance of

COVID-19 literature [6], and to help understand the underlying mechanisms of severe COVID-19 [65], [178], [195], [217], [218].

In addition to these potential academic benefits, the findings of this thesis are relevant from a health informatics perspective. The extension of predications was largely successful by integrating previously unrelated knowledge sources to form interesting hypotheses as coherent indirect associations. Further, these associations were successfully explored using illustrative subgraphs, though the subgraph generation process still requires some work. The use of frequency- and graph-based ranking metrics [114], [155] yielded several uninformative associations, and as such a different way of ranking the system's output is needed to emphasize rare or obscure associations.

The use of external knowledge was successful at identifying complex associations that were dispersed across multiple studies, as was shown by previous works [30], [127]. It was not useful for open discovery, however, as it created too many associations for the user to consider. More work is needed to improve the COVID-REdI system to support this mode of discovery.

The success of our LBD system was demonstrated as we identified several interesting hypotheses that are relevant to COVID-19 patients with DM or CKD, which may be valuable to biomedical researchers studying these conditions. With further experiments and refinements, and input from experts of varied backgrounds, our system will be updated and applied to new research opportunities.

REFERENCES

[1] D. R. Swanson, “Literature-Based Discovery? The Very Idea,” Springer, Berlin, Heidelberg, 2008, pp. 3–11.

[2] M. Weeber, J. A. Kors, and B. Mons, “Online tools to support literature-based discovery in the life sciences,” *Brief. Bioinform.*, vol. 6, no. 3, pp. 277–286, Sep. 2005.

[3] M. J. Cairelli, C. M. Miller, M. Fiszman, T. E. Workman, and T. C. Rindflesch, “Semantic MEDLINE for discovery browsing: using semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox.,” *AMIA Annu. Symp. Proc.*, vol. 2013, pp. 164–173, 2013.

[4] C. M. Miller *et al.*, “A Closed Literature-Based Discovery Technique Finds a Mechanistic Link Between Hypogonadism and Diminished Sleep Quality in Aging Men,” *Sleep*, vol. 35, no. 2, pp. 279–285, Feb. 2012.

[5] J. Hur *et al.*, “Literature-based discovery of diabetes- and ROS-related targets,” *BMC Med. Genomics*, vol. 3, no. 1, pp. 1–11, Oct. 2010.

[6] M. Kang, S. S. Gurbani, and J. A. Kempker, “The Published Scientific Literature on COVID-19: An Analysis of PubMed Abstracts,” *J. Med. Syst.*, vol. 45, no. 1, 2021.

[7] Diabetes Canada, “Diabetes in Canada: Backgrounder,” 2020.

[8] B. Bikbov *et al.*, “Global, regional, and national burden of chronic kidney disease, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” *Lancet*, vol. 395, no. 10225, pp. 709–733, 2020.

[9] B. Manns, S. Q. McKenzie, F. Au, P. M. Gignac, and L. I. Geller, “The financial impact of advanced kidney disease on Canada Pension Plan and private disability insurance costs,” *Can. J. Kidney Heal. Dis.*, vol. 4, pp. 1–11, 2017.

[10] A. K. Singh *et al.*, “Prevalence of co-morbidities and their association with mortality in patients with <scp>COVID</scp> -19: A systematic review and meta-analysis,” *Diabetes, Obes. Metab.*, vol. 22, no. 10, pp. 1915–1924, Oct. 2020.

[11] I. Huang, M. A. Lim, and R. Pranata, “Diabetes mellitus is associated with increased mortality and severity of disease in COVID-19 pneumonia - A systematic review, meta- analysis, and meta-regression,” *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, no. 4, pp. 395–403, Jul. 2020.

[12] F. Zhou *et al.*, “Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study,” *Lancet*, vol. 395, no. 10229, pp. 1054–1062, Mar. 2020.

[13] N. Potere *et al.*, “Acute complications and mortality in hospitalized patients with coronavirus disease 2019: a systematic review and meta-analysis,” *Crit. Care*, vol. 24, no. 1, p. 389, Dec. 2020.

[14] T. Wu *et al.*, “Multi-organ Dysfunction in Patients with COVID-19: A Systematic Review and Meta-analysis,” *Aging Dis.*, vol. 11, no. 4, p. 874, Jul. 2020.

[15] M. Apicella, M. C. Campopiano, M. Mantuano, L. Mazoni, A. Coppelli, and S. Del Prato, “COVID-19 in people with diabetes: understanding the reasons for worse outcomes,” *The Lancet Diabetes and Endocrinology*, vol. 8, no. 9. Lancet Publishing Group, pp. 782–792, 01-Sep-2020.

[16] M. Asgharpour, E. Zare, M. Mubarak, and A. Alirezaei, “COVID-19 and Kidney Disease: Update on Epidemiology, Clinical Manifestations, Pathophysiology and Management,” *J. Coll. Physicians Surg. Pak.*, vol. 30, no. 6, pp. 19–25, Jun. 2020.

[17] S. Knapp, “Diabetes and Infection: Is There a Link? - A Mini-Review,” *Gerontology*, vol. 59, no. 2, pp. 99–104, Feb. 2013.

[18] J. Brainard, “Scientists are drowning in COVID-19 papers. Can new tools keep them afloat?,” *Science (80-.).*, May 2020.

[19] À. Bravo, J. Piñero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, “Extraction of relations between genes and diseases from text and large-scale data analysis: Implications for translational research,” *BMC Bioinformatics*, vol. 16, no. 1, Feb. 2015.

[20] D. Xi, J. Zhao, W. Lai, and Z. Guo, “Systematic analysis of the molecular mechanism underlying atherosclerosis using a text mining approach,” *Hum. Genomics*, vol. 10, no. 1, p. 14, Jun. 2016.

[21] M. Pham, S. Wilson, H. Govindarajan, C. H. Lin, and O. Lichtarge, “Discovery of disease- And drug-specific pathways through community structures of a literature network,” *Bioinformatics*, vol. 36, no. 6, pp. 1881–1888, Mar. 2020.

[22] L. L. Wang and K. Lo, “Text mining approaches for dealing with the rapidly expanding literature on COVID-19,” *Brief. Bioinform.*, vol. 2020, no. 0, pp. 1–19, Dec. 2020.

[23] A. Mazein *et al.*, “Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms,” *npj Syst. Biol. Appl.*, vol. 4, no. 1, p. 21, Dec. 2018.

[24] D. N. Nicholson and C. S. Greene, “Constructing knowledge graphs and their biomedical applications,” *Computational and Structural Biotechnology Journal*, vol. 18. Elsevier B.V., pp. 1414–1428, 01-Jan-2020.

[25] "#GraphCast: Graphs4Good Knowledge Graph to Fight COVID-19." [Online]. Available: https://neo4j.com/blog/graphcast-graphs4good-covid-19-knowledge-graph/. [Accessed: 06-Jun-2020].

[26] D. Domingo-Fernandez *et al.*, "COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology," *bioRxiv*, p. 2020.04.14.040667, Apr. 2020.

[27] D. Vandamme, W. Fitzmaurice, B. Kholodenko, and W. Kolch, "Systems medicine: Helping us understand the complexity of disease," *QJM*, vol. 106, no. 10, pp. 891–895, 2013.

[28] M. Thilakaratne, K. Falkner, and T. Atapattu, "A systematic review on literature-based discovery workflow," *PeerJ Comput. Sci.*, vol. 5, p. e235, Nov. 2019.

[29] S. H. Baek, D. Lee, M. Kim, J. H. Lee, and M. Song, "Enriching plausible new hypothesis generation in PubMed," *PLoS One*, vol. 12, no. 7, p. e0180539, Jul. 2017.

[30] W. J. Vlietstra *et al.*, "Automated extraction of potential migraine biomarkers using a semantic graph.," *J. Biomed. Inform.*, vol. 71, pp. 178–189, Jul. 2017.

[31] "Risk Factors for Type 2 Diabetes | NIDDK." [Online]. Available: https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes. [Accessed: 08-Jan-2021].

[32] R. Kazancioğlu, "Risk factors for chronic kidney disease: An update," in *Kidney International Supplements*, 2013, vol. 3, no. 4, pp. 368–371.

[33] S. E. Kahn, M. E. Cooper, and S. Del Prato, "Pathophysiology and treatment of type 2 diabetes: Perspectives on the past, present, and future," *The Lancet*, vol. 383, no. 9922. Lancet Publishing Group, pp. 1068–1083, 2014.

[34] C. Zoccali *et al.*, "The systemic nature of CKD," *Nature Reviews Nephrology*, vol. 13, no. 6. Nature Publishing Group, pp. 344–358, 01-Jun-2017.

[35] F. Bonacina, A. Baragetti, A. L. Catapano, and G. D. Norata, "The Interconnection Between Immuno-Metabolism, Diabetes, and CKD," *Current Diabetes Reports*, vol. 19, no. 5. Current Medicine Group LLC 1, pp. 1–8, 01-May-2019.

[36] J. Loscalzo, "Network medicine and type 2 diabetes mellitus: insights into disease mechanism and guide to precision medicine," *Endocrine*, vol. 66, no. 3. Springer, pp. 456–459, 01-Dec-2019.

[37] Y. Guo *et al.*, "Identification of key pathways and genes in different types of chronic kidney disease based on WGCNA," *Mol. Med. Rep.*, vol. 20, no. 3, pp. 2245–2257, 2019.

[38] A. Gupta *et al.*, “Extrapulmonary manifestations of COVID-19,” *Nat. Med.*, vol. 26, no. 7, pp. 1017–1032, 2020.

[39] M. Iwai and M. Horiuchi, “Devil and angel in the renin-angiotensin system: ACE-angiotensin II-AT1 receptor axis vs. ACE2-angiotensin-(1-7)-Mas receptor axis,” *Hypertens. Res.*, vol. 32, no. 7, pp. 533–536, 2009.

[40] V. G. Puelles *et al.*, “Multiorgan and Renal Tropism of SARS-CoV-2,” *N. Engl. J. Med.*, vol. 383, no. 6, pp. 590–592, Aug. 2020.

[41] D. Cyranoski, “Profile of a killer: the complex biology powering the coronavirus pandemic,” *Nature*, vol. 581, no. 7806. pp. 22–26, 2020.

[42] S. K. Kunutsor and J. A. Laukkanen, “Renal complications in COVID-19: a systematic review and meta-analysis,” *Ann. Med.*, vol. 52, no. 7, pp. 1–9, 2020.

[43] S. J. McGurnaghan *et al.*, “Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland,” *Lancet Diabetes Endocrinol.*, vol. 8587, no. 20, pp. 1–12, 2020.

[44] B. Cariou *et al.*, “Phenotypic characteristics and prognosis of inpatients with COVID-19 and diabetes: the CORONADO study,” *Diabetologia*, vol. 63, no. 8, pp. 1500–1515, 2020.

[45] S. P. J. M. Horbach, “Pandemic publishing: Medical journals drastically speed up their publication process for Covid-19,” *bioRxiv*, 2020.

[46] M. S. Majumder and K. D. Mandl, “Early in the epidemic: impact of preprints on global discourse about COVID-19 transmissibility,” *The Lancet Global Health*, vol. 8, no. 5. Elsevier Ltd, pp. e627–e630, 01-May-2020.

[47] “COVID-19 Rapid Evidence Access Link | Questions • Evidence • Answers.” [Online]. Available: https://www.dlsph.utoronto.ca/covid19real/. [Accessed: 28-Sep-2020].

[48] “Coviz19 | Visualization dashboard for coronavirus pandemic.” [Online]. Available: https://www.coviz19.com/. [Accessed: 01-Oct-2020].

[49] M. Ostaszewski *et al.*, “COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms,” *Scientific Data*, vol. 7, no. 1. Nature Research, pp. 1–4, 01-Dec-2020.

[50] J. Loscalzo and A.-L. Barabasi, “Systems biology and the future of medicine,” *Wiley Interdiscip. Rev. Syst. Biol. Med.*, vol. 3, no. 6, pp. 619–627, Nov. 2011.

[51] L. Y. H. Lee and J. Loscalzo, “Network Medicine in Pathobiology,” *American Journal of Pathology*, vol. 189, no. 7. Elsevier Inc., pp. 1311–1326, 01-Jul-2019.

[52] F. Kramer, S. Just, and T. Zeller, “New perspectives: Systems medicine in cardiovascular disease,” *BMC Syst. Biol.*, vol. 12, no. 1, pp. 1–13, 2018.

[53] J. Bousquet *et al.*, “Systems medicine and integrated care to combat chronic noncommunicable diseases,” *Genome Medicine*, vol. 3, no. 7. BioMed Central, pp. 1–12, 06-Jul-2011.

[54] S. Mulder, H. Hamidi, M. Kretzler, and W. Ju, “An integrative systems biology approach for precision medicine in diabetic kidney disease,” *Diabetes, Obes. Metab.*, vol. 20, pp. 6–13, Oct. 2018.

[55] M. Lindhardt *et al.*, “Predicting albuminuria response to spironolactone treatment with urinary proteomics in patients with type 2 diabetes and hypertension,” *Nephrol. Dial. Transplant.*, vol. 33, no. 2, pp. 296–303, Feb. 2018.

[56] A. Sharma *et al.*, “Controllability in an islet specific regulatory network identifies the transcriptional factor NFATC4, which regulates Type 2 Diabetes associated genes,” *npj Syst. Biol. Appl.*, vol. 4, no. 1, pp. 1–11, Dec. 2018.

[57] T. Papadopoulos *et al.*, “Omics databases on kidney disease: Where they can be found and how to benefit from them,” *Clinical Kidney Journal*, vol. 9, no. 3. Oxford University Press, pp. 343–352, 01-Jun-2016.

[58] K. Cisek, M. Krochmal, J. Klein, and H. Mischak, “The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease,” *Nephrology, dialysis, transplantation : official publication of the European Dialysis and Transplant Association - European Renal Association*, vol. 31, no. 12. Oxford Academic, pp. 2003–2011, 01-Dec-2016.

[59] M. J. Pena, H. Mischak, and H. J. L. Heerspink, “Proteomics for prediction of disease progression and response to therapy in diabetic kidney disease,” *Diabetologia*, vol. 59, no. 9. Springer Verlag, pp. 1819–1831, 01-Sep-2016.

[60] Q. Wu, X. Coumoul, P. Grandjean, R. Barouki, and K. Audouze, “Endocrine disrupting chemicals and COVID-19 relationships: A computational systems biology approach,” *Environ. Int.*, no. xxxx, p. 106232, 2020.

[61] B. Schwarz *et al.*, “Cutting Edge: Severe SARS-CoV-2 Infection in Humans Is Defined by a Shift in the Serum Lipidome, Resulting in Dysregulation of Eicosanoid Immune Mediators,” *J. Immunol.*, vol. 206, no. 2, pp. 329–334, 2021.

[62] A. K. Singh, R. Gupta, A. Ghosh, and A. Misra, “Diabetes in COVID-19: Prevalence, pathophysiology, prognosis and practical considerations,” *Diabetes Metab Syndr.*, vol. 14, pp. 303–310, 2020.

[63] R. Albulescu *et al.*, “COVID-19 and diabetes mellitus: Unraveling the hypotheses that worsen the prognosis (Review),” *Exp. Ther. Med.*, vol. 20, no. 6, pp. 1–1, 2020.

[64] S. Lim, J. H. Bae, H. S. Kwon, and M. A. Nauck, “COVID-19 and diabetes mellitus: from pathophysiology to clinical management,” *Nature Reviews Endocrinology*, vol. 17, no. 1. Nature Research, 01-Jan-2020.

[65] H. Yaribeygi, T. Sathyapalan, T. Jamialahmadi, and A. Sahebkar, “The Impact of Diabetes Mellitus in COVID-19: A Mechanistic Review of Molecular Interactions,” *J. Diabetes Res.*, vol. 2020, pp. 1–9, 2020.

[66] G. Lisco, A. De Tullio, V. A. Giagulli, E. Guastamacchia, G. De Pergola, and V. Triggiani, “Hypothesized mechanisms explaining poor prognosis in type 2 diabetes patients with COVID-19: a review,” *Endocrine*, vol. 70, no. 3, pp. 441–453, Dec. 2020.

[67] K. Renu, P. L. Prasanna, and A. Valsala Gopalakrishnan, “Coronaviruses pathogenesis, comorbidities and multi-organ damage – A review,” *Life Sci.*, vol. 255, no. January, p. 117839, Aug. 2020.

[68] M. A. Martinez-Rojas, O. Vega-Vega, and X. N. A. Bobadilla, “Is the kidney a target of SARS-CoV-2?,” *Am. J. Physiol. - Ren. Physiol.*, vol. 318, no. 6, pp. F1454–F1462, 2020.

[69] S. S. Farouk, E. Fiaccadori, P. Cravedi, and K. N. Campbell, “COVID-19 and the kidney: what we think we know so far and what we don’t,” *J. Nephrol.*, vol. 33, no. 6, pp. 1213–1218, 2020.

[70] J. Yang, S. J. Wu, W. T. Dai, Y. X. Li, and Y. Y. Li, “The human disease network in terms of dysfunctional regulatory mechanisms,” *Biol. Direct*, vol. 10, no. 1, Oct. 2015.

[71] S. Wahl *et al.*, “Multi-omic signature of body weight change: Results from a population-based cohort study,” *BMC Med.*, vol. 13, no. 1, pp. 1–17, Dec. 2015.

[72] K. T. Do *et al.*, “Phenotype-driven identification of modules in a hierarchical map of multifluid metabolic correlations,” *npj Syst. Biol. Appl.*, vol. 3, no. 1, p. 28, Dec. 2017.

[73] L. Yu, S. Yao, L. Gao, and Y. Zha, “Conserved disease modules extracted from multilayer heterogeneous disease and gene networks for understanding disease mechanisms and predicting disease treatments,” *Front. Genet.*, vol. 10, no. JAN, p. 745, Jan. 2019.

[74] M. Ji *et al.*, “Long noncoding RNA-mRNA expression profiles and validation in pancreatic neuroendocrine neoplasms,” *Clin. Endocrinol. (Oxf).*, vol. 92, no. 4, pp. 312–322, Apr. 2020.

[75] S. Meng *et al.*, “Functional clusters analysis and research based on differential coexpression networks,” *Biotechnol. Biotechnol. Equip.*, vol. 32, no. 1, pp. 171–182, 2018.

[76] J. Dutkowski *et al.*, “A gene ontology inferred from molecular networks,” *Nat Biotechnol*, vol. 31, no. 1, pp. 1–20, 2013.

[77] M. J. Cairelli, M. Fiszman, H. Zhang, and T. C. Rindflesch, “Networks of neuroinjury semantic predications to identify biomarkers for mild traumatic brain injury,” *J. Biomed. Semantics*, vol. 6, no. 1, May 2015.

[78] A. Ahmed, “Literature-Based Discovery: Critical Analysis and Future Directions,” 2016.

[79] D. ROY, “THE CREATION OF NEW KNOWLEDGE BY INFORMATION RETRIEVAL AND CLASSIFICATION,” *J. Doc.*, vol. 45, no. 4, pp. 273–301, Jan. 1989.

[80] D. R. Swanson, “Fish oil, Raynaud’s syndrome, and undiscovered public knowledge.,” *Perspect. Biol. Med.*, vol. 30, no. 1, pp. 7–18, 1986.

[81] D. R. Swanson, “Migraine and magnesium: eleven neglected connections.,” *Perspectives in biology and medicine*, vol. 31, no. 4. pp. 526–557, 1988.

[82] R. A. Digiacomo, J. M. Kremer, and D. M. Shah, “Fish-oil dietary supplementation in patients with Raynaud’s phenomenon: A double-blind, controlled, prospective study,” *Am. J. Med.*, vol. 86, no. 2 C, pp. 158–164, Jan. 1989.

[83] A. Peikert, C. Wilimzig, and R. Köhne-Volland, “Prophylaxis of migraine with oral magnesium: results from a prospective, multi-center, placebo-controlled and double-blind randomized study.,” *Cephalalgia*, vol. 16, no. 4, pp. 257–63, Jun. 1996.

[84] S. Henry and B. T. McInnes, “Literature Based Discovery: Models, methods, and trends,” *J. Biomed. Inform.*, vol. 74, pp. 20–32, 2017.

[85] Y. Sebastian, E. G. Siew, and S. O. Orimaye, “Emerging approaches in literature-based discovery: techniques and performance review,” *Knowl. Eng. Rev.*, vol. 32, pp. 1–35, May 2017.

[86] M. Gabetta, C. Larizza, and R. Bellazzi, “A unified medical language system (UMLS) based system for literature-based discovery in medicine,” *Stud. Health Technol. Inform.*, vol. 192, no. 1–2, pp. 412–416, 2013.

[87] N. R. Smalheiser, “Literature-based discovery: Beyond the ABCs,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, pp. 218–224, Feb. 2012.

[88] D. Cameron, R. Kavuluru, T. C. Rindflesch, A. P. Sheth, K. Thirunarayan, and O. Bodenreider, “Context-driven automatic subgraph creation for literature-based discovery,” *J. Biomed. Inform.*, vol. 54, pp. 141–157, Apr. 2015.

[89] Y. H. Kim and M. Song, “A context-based ABC model for literature-based discovery,” *PLoS One*, vol. 14, no. 4, pp. 1–25, 2019.

[90] I. Petrič, T. Urbančič, B. Cestnik, and M. Macedoni-Lukšič, “Literature mining method RaJoLink for uncovering relations between biomedical concepts,” *J. Biomed. Inform.*, vol. 42, no. 2, pp. 219–227, Apr. 2009.

[91] G. Bakal, P. Talari, E. V Kakani, and R. Kavuluru, “Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations,” *J. Biomed. Inform.*, vol. 82, no. 3, pp. 189–199, Jun. 2018.

[92] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin, “Exploiting semantic relations for literature-based discovery.,” *AMIA Annu. Symp. Proc.*, vol. 2006, pp. 349–353, 2006.

[93] C. B. Ahlers, D. Hristovski, H. Kilicoglu, and T. C. Rindflesch, “Using the literature-based discovery paradigm to investigate drug mechanisms.,” *AMIA Annu. Symp. Proc.*, pp. 6–10, 2007.

[94] D. Hristovski and A. Kastrin, “Towards using a Graph Database and Literature-based Discovery for Interpretation of Next Generation Sequencing Results,” no. c, pp. 68–70, 2018.

[95] D. Cameron *et al.*, “A graph-based recovery and decomposition of Swanson’s hypothesis using semantic predications,” *J. Biomed. Inform.*, vol. 46, no. 2, pp. 238–251, 2013.

[96] B. Wilkowski *et al.*, “Graph-based methods for discovery browsing with semantic predications.,” *AMIA Annu. Symp. Proc.*, vol. 2011, pp. 1514–1523, 2011.

[97] V. Novacek, “Formalising Hypothesis Virtues in Knowledge Graphs: A General Theoretical Framework and its Validation in Literature-Based Discovery Experiments,” Mar. 2015.

[98] Y. Sebastian, E.-G. Siew, and S. O. Orimaye, “Learning the heterogeneous bibliographic information network for literature-based discovery,” *Knowledge-Based Syst.*, vol. 115, pp. 66–79, Jan. 2017.

[99] J. Sybrandt, A. Carrabba, A. Herzog, and I. Safro, “Are Abstracts Enough for Hypothesis Generation?,” in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019, pp. 1504–1513.

[100] D. Westergaard, H. H. Stærfeldt, C. Tønsberg, L. J. Jensen, and S. Brunak, “A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts,” *PLoS Comput. Biol.*, vol. 14, no. 2, Feb. 2018.

[101] V. Gopalakrishnan, K. Jha, W. Jin, and A. Zhang, “A survey on literature based discovery approaches in biomedical domain,” *J. Biomed. Inform.*, vol. 93, p. 103141, May 2019.

[102] M. Srinivasan, C. Blackburn, M. Mohamed, A. V. Sivagami, and J. Blum, “Literature-based discovery of salivary biomarkers for type 2 diabetes mellitus,” *Biomark. Insights*, vol. 10, pp. 39–45, Jan. 2015.

[103] T. C. Rindflesch, C. L. Blake, M. J. Cairelli, M. Fiszman, C. J. Zeiss, and H. Kilicoglu, “Investigating the role of interleukin-1 beta and glutamate in inflammatory bowel disease and epilepsy using discovery browsing,” *J. Biomed. Semantics*, vol. 9, no. 1, pp. 1–14, Dec. 2018.

[104] A. R. Sedler and C. S. Mitchell, “SemNet: Using local features to navigate the biomedical concept graph,” *Front. Bioeng. Biotechnol.*, vol. 7, no. JUL, p. 156, Jul. 2019.

[105] L. Eronen and H. Toivonen, “Biomine: predicting links between biological entities using network models of heterogeneous databases,” *BMC Bioinformatics*, vol. 13, no. 1, pp. 1–22, 2012.

[106] D. Hristovski, A. Kastrin, B. Peterlin, and T. C. Rindflesch, “Combining semantic relations and DNA microarray data for novel hypotheses generation,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6004 LNBI, pp. 53–61.

[107] D. Weissenborn, M. Schroeder, and G. Tsatsaronis, “Discovering relations between indirectly connected biomedical concepts,” *J. Biomed. Semantics*, vol. 6, no. 1, p. 28, Jul. 2015.

[108] K. Hassani-Pak and C. Rawlings, “Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes,” *J. Integr. Bioinform.*, vol. 14, no. 1, pp. 1–9, 2017.

[109] US National Library of Medicine, “The UMLS Semantic Groups.” [Online]. Available: http://wayback.archive-it.org/org-350/20130703100754/http://semanticnetwork.nlm.nih.gov/SemGroups/. [Accessed: 10-Jul-2020].

[110] “Semantic Network,” 2009.

[111] M. Rastegar-Mojarad, R. K. Elayavilli, L. Wang, R. Prasad, and H. Liu, “Prioritizing adverse drug reaction and drug repositioning candidates generated by literature-based discovery,” *ACM-BCB 2016 - 7th ACM Conf. Bioinformatics, Comput. Biol. Heal. Informatics*, pp. 289–296, 2016.

[112] J. Preiss, M. Stevenson, and R. Gaizauskas, “Exploring relation types for literature-based discovery,” *J. Am. Med. Informatics Assoc.*, vol. 22, no. 5, pp. 987–992, 2015.

[113] J. Preiss and M. Stevenson, “Quantifying and filtering knowledge generated by literature based discovery,” *BMC Bioinformatics*, vol. 18, no. 7, pp. 59–67, May 2017.

[114] D. R. Swanson and N. R. Smalheiser, “An interactive system for finding complementary literatures: A stimulus to scientific discovery,” *Artif. Intell.*, vol. 91, no. 2, pp. 183–203, 1997.

[115] T. Cohen, R. Schvaneveldt, and D. Widdows, “Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections,” *J. Biomed. Inform.*, vol. 43, no. 2, pp. 240–256, 2010.

[116] I. Petric, B. Ligeti, and B. G. and S. Pongor, “Biomedical Hypothesis Generation by Text Mining and Gene Prioritization,” *Protein & Peptide Letters*, vol. 21, no. 8. pp. 847–857, 2014.

[117] J. Hur, A. Özgür, and Y. He, “Ontology-based literature mining of E. coli vaccine-associated gene interaction networks,” *J. Biomed. Semantics*, vol. 8, no. 1, pp. 1–10, 2017.

[118] R. N. Kostoff *et al.*, “Literature-related discovery,” *Annual Review of Information Science and Technology*, vol. 43, no. 1. Information Today, pp. 1–71, 01-Jan-2009.

[119] D. Gubiani, E. Fabbretti, B. Cestnik, N. Lavrač, and T. Urbančič, “Outlier based literature exploration for cross-domain linking of Alzheimer’s disease and gut microbiota,” *Expert Syst. Appl.*, vol. 85, pp. 386–396, 2017.

[120] M. Yetisgen-Yildiz and W. Pratt, “A new evaluation methodology for literature-based discovery systems,” *J. Biomed. Inform.*, vol. 42, no. 4, pp. 633–643, Aug. 2009.

[121] R. N. Kostoff, “Validating discovery in literature-based discovery,” *J. Biomed. Inform.*, vol. 40, no. 4, pp. 448–450, Aug. 2007.

[122] R. Percudani, D. Carnevali, and V. Puggioni, “Ureidoglycolate hydrolase, amidohydrolase, lyase: How errors in biological databases are incorporated in scientific papers and vice versa,” *Database*, vol. 2013, pp. 1–9, 2013.

[123] N. Škunca, A. Altenhoff, and C. Dessimoz, “Quality of Computationally Inferred Gene Ontology Annotations,” *PLoS Comput. Biol.*, vol. 8, no. 5, p. e1002533, May 2012.

[124] S. Poux, M. Magrane, C. N. Arighi, A. Bridge, C. O'Donovan, and K. Laiho, "Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data," *Database*, vol. 2014, no. 0, pp. bau016–bau016, Mar. 2014.

[125] A. Kastrin and D. Hristovski, "Scientometric analysis and knowledge mapping of literature-based discovery (1986–2020)," Nov. 2020.

[126] R. Zhang *et al.*, "Exploiting Literature-derived Knowledge and Semantics to Identify Potential Prostate Cancer Drugs," *Cancer Inform.*, vol. 13s1, p. CIN.S13889, 2014.

[127] D. Cameron, R. Kavuluru, O. Bodenreider, P. N. Mendes, A. P. Sheth, and K. Thirunarayan, "Semantic predications for complex information needs in biomedical literature," in *Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011*, 2011, vol. 2011, pp. 512–519.

[128] Z. Gao *et al.*, "edge2vec: Representation learning using edge semantics for biomedical knowledge discovery," *arXiv*, pp. 1–15, 2018.

[129] "What is an RDF Triplestore? | Ontotext Fundamentals." [Online]. Available: https://www.ontotext.com/knowledgehub/fundamentals/what-is-rdf-triplestore/. [Accessed: 11-Feb-2021].

[130] T. C. Rindflesch *et al.*, "Informatics support for basic research in biomedicine," *ILAR J.*, vol. 58, no. 1, pp. 80–89, Jul. 2017.

[131] A. Lysenko, I. A. RoznovĂţ, M. Saqi, A. Mazein, C. J. Rawlings, and C. Auffray, "Representing and querying disease networks using graph databases," *BioData Mining*, vol. 9, no. 1. BioMed Central Ltd., pp. 1–19, 25-Jul-2016.

[132] N. Swainston *et al.*, "biochem4j: Integrated and extensible biochemical knowledge through graph databases," *PLoS One*, vol. 12, no. 7, Jul. 2017.

[133] D. S. Himmelstein *et al.*, "Systematic integration of biomedical knowledge prioritizes drugs for repurposing," *Elife*, vol. 6, pp. 1–35, 2017.

[134] Y. Chen and R. Xu, "Context-sensitive network-based disease genetics prediction and its implications in drug discovery," *Bioinformatics*, vol. 33, no. 7, p. btw737, Jan. 2017.

[135] M. Alshahrani and R. Hoehndorf, "Semantic Disease Gene Embeddings (SmuDGE): Phenotype-based disease gene prioritization without phenotypes," *Bioinformatics*, vol. 34, no. 17, pp. i901–i907, 2018.

[136] J. H. M. van Bilsen *et al.*, "Seeking Windows of Opportunity to Shape Lifelong Immune Health: A Network-Based Strategy to Predict and Prioritize Markers of Early Life Immune Modulation," *Front. Immunol.*, vol. 11, p. 644, Apr. 2020.

[137] A. Rajpal, L. Rahimi, and F. Ismail-Beigi, “Factors leading to high morbidity and mortality of COVID-19 in patients with type 2 diabetes,” *J. Diabetes*, vol. 12, no. 12, pp. 895–908, 2020.

[138] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, pp. 25–29, May 2000.

[139] S. Carbon *et al.*, “The Gene Ontology resource: Enriching a GOld mine,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, 2021.

[140] O. Bodenreider, “The Unified Medical Language System (UMLS): Integrating biomedical terminology,” *Nucleic Acids Res.*, vol. 32, no. DATABASE ISS., p. D267, Jan. 2004.

[141] H. Kilicoglu, D. Shin, M. Fiszman, G. Rosemblat, and T. C. Rindflesch, “SemMedDB: A PubMed-scale repository of biomedical semantic predications,” *Bioinformatics*, vol. 28, no. 23, pp. 3158–3160, 2012.

[142] “Introduction to GO annotations.” [Online]. Available: http://geneontology.org/docs/go-annotations/. [Accessed: 02-Jan-2021].

[143] K. Anyanwu and A. Sheth, “ρ-Queries: Enabling querying for semantic associations on the semantic web,” *Proc. 12th Int. Conf. World Wide Web, WWW 2003*, pp. 690–699, 2003.

[144] R. P. Huntley and R. C. Lovering, “The Gene Ontology Handbook,” in *Methods in Molecular Biology*, vol. 1446, 2017, pp. 233–243.

[145] D. Hristovski, A. Kastrin, D. Dinevski, and T. C. Rindflesch, “Towards implementing semantic literature-based discovery with a graph database,” in *Proceedings of the The Seventh International Conference on Advances in Databases, Knowledge, and Data Applications*, 2015, pp. 180--184.

[146] H. Kilicoglu, G. Rosemblat, M. Fiszman, and D. Shin, “Broad-coverage biomedical relation extraction with SemRep,” *BMC Bioinformatics*, vol. 21, no. 1, 2020.

[147] “UMLS Metathesaurus Browser.” [Online]. Available: https://uts.nlm.nih.gov/uts/umls/concept/C0009528. [Accessed: 16-Feb-2021].

[148] “Gene Ontology overview.” [Online]. Available: http://geneontology.org/docs/ontology-documentation/. [Accessed: 13-Feb-2021].

[149] “UMLS Metathesaurus Browser.” [Online]. Available: https://uts.nlm.nih.gov/uts/umls/concept/C1159404. [Accessed: 13-Feb-2021].

[150] D. S. Wishart *et al.*, “DrugBank 5.0: A major update to the DrugBank database for 2018,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018.

[151] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, “OMIM.org: Leveraging knowledge across phenotype-gene relationships,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D1038–D1043, Jan. 2019.

[152] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: New perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D353–D361, Jan. 2017.

[153] R. Zhang *et al.*, “Using semantic predications to uncover drug-drug interactions in clinical data,” *J. Biomed. Inform.*, vol. 49, pp. 134–147, Jun. 2014.

[154] A. Daowd, “Personal Communication.” 2020.

[155] L. Page and S. Brin, “The anatomy of a large-scale hypertextual Web search engine,” *Comput. Networks*, vol. 30, no. 1–7, pp. 107–117, 1998.

[156] M. Song, S. H. Baek, G. E. Heo, and J. H. Lee, “Inferring drug-protein-side effect relationships from biomedical text,” *Genes (Basel).*, vol. 10, no. 2, Feb. 2019.

[157] B. Hegyi *et al.*, “CaMKII Serine 280 O-GlcNAcylation Links Diabetic Hyperglycemia to Proarrhythmia,” *Circ. Res.*, vol. 129, no. 1, pp. 98–113, 2021.

[158] W. M. Bramer, G. B. de Jonge, M. L. Rethlefsen, F. Mast, and J. Kleijnen, “A systematic approach to searching: An efficient and complete method to develop literature searches,” *J. Med. Libr. Assoc.*, vol. 106, no. 4, pp. 531–541, Oct. 2018.

[159] “MeSH Qualifiers with Scope Notes.” [Online]. Available: https://www.nlm.nih.gov/mesh/qualifiers_scopenotes.html. [Accessed: 08-Feb-2021].

[160] S. Mathur and D. Dinakarpandian, “Finding disease similarity based on implicit semantic similarity,” *J. Biomed. Inform.*, vol. 45, no. 2, pp. 363–371, 2012.

[161] M. Akbari and V. Hassan-Zadeh, “IL-6 signalling pathways and the development of type 2 diabetes,” *Inflammopharmacology*, vol. 26, no. 3. Birkhauser Verlag AG, pp. 685–698, 01-Jun-2018.

[162] M. K. Bohn, A. Hall, L. Sepiashvili, B. Jung, S. Steele, and K. Adeli, “Pathophysiology of COVID-19: Mechanisms underlying disease severity and progression,” *Physiology*, vol. 35, no. 5. American Physiological Society, pp. 288–301, 01-Sep-2020.

[163] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, “Entrez gene: Gene-centered information at NCBI,” *Nucleic Acids Res.*, vol. 39, no. SUPPL. 1, p. D52, Jan. 2011.

[164] G. Chen *et al.*, “Gene fingerprint model for literature based detection of the associations among complex diseases: A case study of COPD,” *BMC Med. Inform. Decis. Mak.*, vol. 19, no. S1, p. 20, Jan. 2019.

[165] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012.

[166] S. Nakamura, K. Mori, H. Okuma, T. Sekine, A. Miyazaki, and K. Tsuchiya, “Age-associated decline of monocyte insulin sensitivity in diabetic and healthy individuals,” *Diabetes Vasc. Dis. Res.*, vol. 18, no. 1, 2021.

[167] R. Root-Bernstein, “Innate Receptor Activation Patterns Involving TLR and NLR Synergisms in COVID-19, ALI/ARDS and Sepsis Cytokine Storms: A Review and Model Making Novel Predictions and Therapeutic Suggestions,” *Int. J. Mol. Sci.*, vol. 22, no. 4, pp. 1–47, Feb. 2021.

[168] A. Hariharan, A. R. Hakeem, S. Radhakrishnan, M. S. Reddy, and M. Rela, “The Role and Therapeutic Potential of NF-kappa-B Pathway in Severe COVID-19 Patients,” *Inflammopharmacology*, vol. 1. Springer Science and Business Media Deutschland GmbH, p. 1, 01-Feb-2020.

[169] D. Birra *et al.*, “COVID 19: a clue from innate immunity,” *Immunol. Res.*, vol. 68, no. 3, p. 1, Jun. 2020.

[170] J. Wada and H. Makino, “Innate immunity in diabetes and diabetic nephropathy,” *Nature Reviews Nephrology*, vol. 12, no. 1. Nature Publishing Group, pp. 13–26, 01-Jan-2016.

[171] A. A. de Oliveira and K. P. Nunes, “Crosstalk of TLR4, vascular NADPH oxidase, and COVID-19 in diabetes: What are the potential implications?,” *Vascul. Pharmacol.*, vol. 139, p. 106879, Aug. 2021.

[172] C. Lammi and A. Arnoldi, “Food-derived antioxidants and COVID-19,” *J. Food Biochem.*, vol. 45, no. 1, pp. 1–6, 2021.

[173] C. D. T. de Freitas *et al.*, “Characterization of Three Osmotin-Like Proteins from Plumeria rubra and Prospection for Adiponectin Peptidomimetics,” *Protein Pept. Lett.*, vol. 27, no. 7, pp. 593–603, Jan. 2020.

[174] A. C. Codo *et al.*, “Elevated Glucose Levels Favor SARS-CoV-2 Infection and Monocyte Response through a HIF-1α/Glycolysis-Dependent Axis,” *Cell Metab.*, vol. 32, no. 3, pp. 437-446.e5, Sep. 2020.

[175] H. Medini, A. Zirman, and D. Mishmar, “Immune system cells from COVID-19 patients display compromised mitochondrial-nuclear expression co-regulation and rewiring toward glycolysis,” *iScience*, vol. 24, no. 12, Dec. 2021.

[176] J. H. Lim, Y. M. Lee, Y. S. Chun, J. Chen, J. E. Kim, and J. W. Park, “Sirtuin 1 Modulates Cellular Responses to Hypoxia by Deacetylating Hypoxia-Inducible Factor 1α,” *Mol. Cell*, vol. 38, no. 6, pp. 864–878, Jun. 2010.

[177] M. Mahmudpour, K. Vahdat, M. Keshavarz, and I. Nabipour, “The COVID-19-diabetes mellitus molecular tetrahedron,” *Mol. Biol. Rep.*, vol. 1, p. 1, Jan. 2022.

[178] R. Miller, A. R. Wentzel, and G. A. Richards, “COVID-19: NAD+ deficiency may predispose the aged, obese and type2 diabetics to mortality through its effect on SIRT1 activity,” *Med. Hypotheses*, vol. 144, Nov. 2020.

[179] I. Mancini *et al.*, “The ADAMTS13-von Willebrand factor axis in COVID-19 patients,” *J. Thromb. Haemost.*, vol. 19, no. 2, pp. 513–521, Feb. 2021.

[180] P. J. Grant, “Diabetes mellitus as a prothrombotic condition,” in *Journal of Internal Medicine*, 2007, vol. 262, no. 2, pp. 157–172.

[181] P. Cassis *et al.*, “ADAMTS13 Deficiency Shortens the Life Span of Mice With Experimental Diabetes,” *Diabetes*, vol. 67, no. 10, pp. 2069–2083, Oct. 2018.

[182] S. Raghavan, D. B. Kenchappa, and M. D. Leo, “SARS-CoV-2 Spike Protein Induces Degradation of Junctional Proteins That Maintain Endothelial Barrier Integrity,” *Front. Cardiovasc. Med.*, vol. 8, Jun. 2021.

[183] A. Szpigel *et al.*, “Lipid environment induces ER stress, TXNIP expression and inflammation in immune cells of individuals with type 2 diabetes,” *Diabetologia*, vol. 61, no. 2, pp. 399–412, 2018.

[184] R. H. Pirzada, N. Javaid, and S. Choi, “The Roles of the NLRP3 Inflammasome in Neurodegenerative and Metabolic Diseases and in Relevant Advanced Therapeutic Interventions,” *Genes (Basel).*, vol. 11, no. 2, p. 131, Jan. 2020.

[185] J. Ma *et al.*, “SARS-CoV-2 nucleocapsid suppresses host pyroptosis by blocking Gasdermin D cleavage,” *EMBO J.*, vol. 40, no. 18, Sep. 2021.

[186] S. Cataldi, V. Costa, A. Ciccodicola, and M. Aprile, “PPARγ and Diabetes: Beyond the Genome and Towards Personalized Medicine,” *Curr. Diab. Rep.*, vol. 21, no. 6, 2021.

[187] C. Desterke, A. G. Turhan, A. Bennaceur-Griscelli, and F. Griscelli, “PPARγ Cistrome Repression during Activation of Lung Monocyte-Macrophages in Severe COVID-19,” *iScience*, vol. 23, no. 10, p. 101611, 2020.

[188] K. Maiese, “Prospects and Perspectives for WISP1 (CCN4) in Diabetes Mellitus,” *Curr. Neurovasc. Res.*, vol. 17, no. 3, pp. 327–331, Mar. 2020.

[189] W. Cai, M. Ramdas, L. Zhu, X. Chen, G. E. Striker, and H. Vlassara, “Oral advanced glycation endproducts (AGEs) promote insulin resistance and diabetes by depleting the antioxidant defenses AGE receptor-1 and sirtuin 1,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 109, no. 39, pp. 15888–15893, Sep. 2012.

[190] F. Wang, Y. Shang, R. Zhang, X. Gao, and Q. Zeng, “A SIRT1 agonist reduces cognitive decline in type 2 diabetic rats through antioxidative and anti-inflammatory mechanisms,” *Mol. Med. Rep.*, vol. 19, no. 2, pp. 1040–1048, Feb. 2019.

[191] L. S. Gewin, “Sugar or Fat? Renal Tubular Metabolism Reviewed in Health and Disease,” *Nutrients*, vol. 13, no. 5, May 2021.

[192] Y. Y. Cai *et al.*, “Renoprotective effects of brown adipose tissue activation in diabetic mice,” *J. Diabetes*, vol. 11, no. 12, p. 958, Dec. 2019.

[193] M. Kitada, Y. Ogura, I. Monno, and D. Koya, “Sirtuins and type 2 diabetes: Role in inflammation, oxidative stress, and mitochondrial function,” *Frontiers in Endocrinology*, vol. 10, no. MAR. Frontiers Media S.A., p. 187, 27-Mar-2019.

[194] R. Gupta, A. Hussain, and A. Misra, “Diabetes and COVID-19: evidence, current status and unanswered research questions,” *Eur. J. Clin. Nutr.*, vol. 74, no. 6, pp. 864–870, Jun. 2020.

[195] A. Hussain, B. Bhowmik, and N. C. do Vale Moreira, “COVID-19 and diabetes: Knowledge in progress,” *Diabetes Research and Clinical Practice*, vol. 162. Elsevier Ireland Ltd, 01-Apr-2020.

[196] U. W. Iepsen *et al.*, “The role of lactate in sepsis and COVID-19: Perspective from contracting skeletal muscle metabolism,” *Exp. Physiol.*, 2021.

[197] O. J. O. F. McElvaney *et al.*, “Characterization of the inflammatory response to severe COVID-19 Illness,” *Am. J. Respir. Crit. Care Med.*, vol. 202, no. 6, pp. 812–821, Sep. 2020.

[198] E. Farshi, B. Kasmapur, and A. Arad, “Investigation of immune cells on elimination of pulmonary-Infected COVID-19 and important role of innate immunity, phagocytes,” *Rev. Med. Virol.*, vol. 31, no. 2, Mar. 2021.

[199] M. C. S. Menezes *et al.*, “Lower peripheral blood Toll-like receptor 3 expression is associated with an unfavorable outcome in severe COVID-19 patients,” vol. 11, no. 1, p. 15223, Dec. 2021.

[200] L. E. H. van der Donk *et al.*, “SARS-CoV-2 infection activates dendritic cells via cytosolic receptors rather than extracellular TLRs,” *Eur. J. Immunol.*, Feb. 2022.

[201] M. Song, G. E. Heo, and Y. Ding, “SemPathFinder: Semantic path analysis for discovering publicly unknown knowledge,” *J. Informetr.*, vol. 9, no. 4, pp. 686–703, 2015.

[202] V. N. Garla and C. Brandt, “Semantic similarity in the biomedical domain: An evaluation across knowledge sources,” *BMC Bioinformatics*, vol. 13, no. 1, p. 261, Oct. 2012.

[203] Y. Ding *et al.*, “Entitymetrics: Measuring the Impact of Entities,” *PLoS One*, vol. 8, no. 8, p. e71416, Aug. 2013.

[204] S. Henry, A. Panahi, D. S. Wijesinghe, and B. T. McInnes, “A Literature Based Discovery Visualization System with Hierarchical Clustering and Linking Set Associations.,” *AMIA Jt. Summits Transl. Sci. proceedings. AMIA Jt. Summits Transl. Sci.*, vol. 2019, pp. 582–591, 2019.

[205] M. S. Hossain, J. Gresock, Y. Edmonds, R. Helm, M. Potts, and N. Ramakrishnan, “Connecting the Dots between PubMed Abstracts,” *PLoS One*, vol. 7, no. 1, p. e29509, Jan. 2012.

[206] H. Mohammadhassanzadeh, W. Van Woensel, S. R. Abidi, and S. S. R. Abidi, “Semantics-based plausible reasoning to extend the knowledge coverage of medical knowledge bases for improved clinical decision support,” *BioData Min.*, vol. 10, no. 1, p. 7, Dec. 2017.

[207] W. Guo *et al.*, “Diabetes is a risk factor for the progression and prognosis of COVID-19,” *Diabetes. Metab. Res. Rev.*, vol. 36, no. 7, pp. 1–9, 2020.

[208] H. Debi, Z. T. Itu, M. T. Amin, F. Hussain, and M. S. Hossain, “Association of serum C-reactive protein (CRP) and D-dimer concentration on the severity of COVID-19 cases with or without diabetes: a systematic review and meta-analysis,” *Expert Rev. Endocrinol. Metab.*, 2021.

[209] S. L. Lage *et al.*, “Persistent Oxidative Stress and Inflammasome Activation in CD14highCD16− Monocytes From COVID-19 Patients,” *Front. Immunol.*, vol. 12, p. 1, Jan. 2021.

[210] S. ElShal, L. C. Tranchevent, A. Sifrim, A. Ardeshirdavani, J. Davis, and Y. Moreau, “Beegle: From literature mining to disease-gene discovery,” *Nucleic Acids Res.*, vol. 44, no. 2, p. e18, Jan. 2016.

[211] K. M. Hettne *et al.*, “The Implicitome: A Resource for Rationalizing Gene-Disease Associations,” *PLoS One*, vol. 11, no. 2, p. e0149621, Feb. 2016.

[212] P. Li, Y. Nie, and J. Yu, “Fusing literature and full network data improves disease similarity computation,” *BMC Bioinformatics*, vol. 17, no. 1, pp. 1–13, Aug. 2016.

[213] S. Henry and B. T. McInnes, “Indirect association and ranking hypotheses for literature based discovery,” *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–19, Aug. 2019.

[214] R. Kaalia and I. Ghosh, “Semantics based approach for analyzing disease-target associations,” *J. Biomed. Inform.*, vol. 62, pp. 125–135, 2016.

[215] J. C. Goodwin, T. Cohen, and T. Rindflesch, “Discovery by scent: Discovery browsing system based on the Information Foraging Theory,” *Proc. - 2012 IEEE Int. Conf. Bioinforma. Biomed. Work. BIBMW 2012*, pp. 232–239, 2012.

[216] A. Älgå, O. Eriksson, and M. Nordberg, “Analysis of scientific publications during the early phase of the COVID-19 pandemic: Topic modeling study,” *J. Med. Internet Res.*, vol. 22, no. 11, 2020.

[217] N. P. Somasundaram *et al.*, “The Impact of SARS-Cov-2 Virus Infection on the Endocrine System,” *J. Endocr. Soc.*, vol. 4, no. 8, Aug. 2020.

[218] V. Lambadiari, F. Kousathana, A. Raptis, K. Katogiannis, A. Kokkinos, and I. Ikonomidis, “Pre-Existing Cytokine and NLRP3 Inflammasome Activation and Increased Vascular Permeability in Diabetes: A Possible Fatal Link With Worst COVID-19 Infection Outcomes?,” *Front. Immunol.*, vol. 11, p. 3063, Nov. 2020.

APPENDIX A: EXAMPLES OF PUBMED QUERIES

COVID-19 and Virology (subheading) – (LitCMechanism[Filter] OR LitCGeneral[Filter]) AND ((COVID-19[Supplementary Concept]) OR (COVID-19[majr]) OR (SARS-CoV-2[majr]) OR (2019 novel coronavirus disease[tiab]) OR (COVID19[tiab]) OR (COVID-19 virus disease[tiab]) OR (coronavirus disease 2019[tiab]) OR (coronavirus disease-19[tiab]) OR (2019-nCoV disease[tiab]) OR (2019 novel coronavirus infection*[tiab]) OR (2019-nCoV[tiab]) OR (novel coronavirus[tiab]) OR (SARS-CoV-2[tiab])) AND ((insulin-secreting cells[majr]) OR (pancreatic b cell*[tiab]) OR (pancreatic beta cell*[tiab]) OR (insulin resistance[majr]) OR (insulin sensitivity[tiab]) OR (glucose intolerance[majr]) OR (impaired glucose tolerance[tiab]) OR (diabetes mellitus[majr:noexp]) OR (diabetes mellitus, type 1[majr]) OR (diabetes mellitus, type 2[majr]) OR (Autoimmune Diabetes[tiab]) OR (IDDM[tiab]) OR (Type 1 Diabet*[tiab]) OR (adult-onset diabetes mellitus[tiab]) OR (MODY[tiab]) OR (NIDDM[tiab]) OR (Type 2 Diabet*[tiab]) OR (diabetes mellitus[tiab]) OR (T2DM[tiab]) OR (diabetic ketoacidosis[tiab]) OR (diabetic ketosis[tiab]) OR (diabetic acidosis[tiab]) OR (diabetic vascular complication*[tiab]) OR (diabetic vascular disease*[tiab]) OR (diabetic microangiopath*[tiab]) OR (Diabetic Glomerulosclerosis[tiab]) OR (Diabetic Kidney Disease[tiab]) OR (Diabetic Nephropath*[tiab]) OR (renal insufficiency[majr]) OR (Kidney Failure[tiab]) OR (chronic kidney insufficienc*[tiab]) OR (chronic renal insufficienc*[tiab]) OR (Renal Failure[tiab]) OR (end-stage kidney disease[tiab]) OR (end-stage renal disease[tiab]) OR (CKD[tiab]) OR (chronic kidney disease*[tiab]) OR (chronic renal disease*[tiab])) AND ((Virology[sh]) OR (virus assembly[tiab]) OR (viral assembly) OR (virus entry[tiab]) OR (viral internalization[tiab]) OR (viral entry[tiab]) OR (viral membrane fusion[tiab]) OR (viral replication[tiab]) OR (virus release[tiab]) OR (virus budding[tiab]) OR (virus egress[tiab]) OR (viral tropism[tiab]) OR (host cell tropism[tiab]) OR (host tissue tropism[tiab]) OR (host microbial interactions[mh]) OR (host virus interaction*[tiab]) OR (viral-host interaction*[tiab]) OR (virus host interaction*[tiab])) NOT (pregnan*) NOT (child*) NOT (infant*) NOT (lung diseases, obstructive[mh]) NOT (fatty liver[mh]) NOT (disease management[mh]) NOT (sex factors[mh]) NOT (race factors[mh]) NOT (age factors[mh]) NOT (social determinants of health[mh]) NOT (risk assessment[mh]) NOT (telemedicine[mh]) NOT (drug therapy[sh]) NOT (cancer*) NOT (Diagnosis[sh]) NOT (Pharmacology[sh]) NOT (Standards[sh]) NOT (preprint[pt])

Diabetes Mellitus – ((insulin-secreting cells[majr]) OR (pancreatic b cell*[tiab]) OR (pancreatic beta cell*[tiab]) OR (insulin resistance[majr]) OR (insulin sensitivity[tiab]) OR (glucose intolerance[majr]) OR (impaired glucose tolerance[tiab]) OR (diabetes mellitus[majr:noexp]) OR (diabetes mellitus, type 1[majr]) OR (Autoimmune Diabetes[tiab]) OR (IDDM[tiab]) OR (Type 1 Diabet*[tiab]) OR (adult-onset diabetes mellitus[tiab]) OR (MODY[tiab]) OR (NIDDM[tiab]) OR (diabetes mellitus, type 2[majr]) OR (Type 2 Diabet*[tiab]) OR (diabetes mellitus[tiab]) OR (T2DM[tiab])) AND ((Metabolism[sh]) OR (biochemical pathway*[tiab]) OR (biodegradation[tiab]) OR (biotransformation[tiab]) OR (catabolism[tiab]) OR (degradation[tiab]) OR (incorporation[tiab]) OR (mobilization[tiab]) OR (secretion[tiab]) OR (turnover[tiab]) OR (anabolism[tiab]) OR (bioformation[tiab]) OR (enzyme activity[tiab]) OR (enzyme*[tiab]) OR (urinary aspects[tiab]) OR (urinary levels[tiab]) OR (lipids[mh]) OR

(carbohydrates[mh]) OR (amino acids[mh]) OR (Immunology[sh]) OR (non-specific immun*[tiab]) OR (innate immun*[tiab]) OR (native immun*[tiab]) OR (natural immun*[tiab]) OR (adaptive immun*[tiab]) OR (humoral immun*[tiab]) OR (humoural immun*[tiab]) OR (cellular immun*[tiab]) OR (cell-mediated immun*[tiab]) OR (primary immun*[tiab]) OR (antiviral immun*[tiab]) OR (anti-viral immun*[tiab]) OR (peripheral immun*[tiab]) OR (phagocytosis[mh]) OR (complement activation[mh]) OR (neutrophil activation[mh]) OR (chemotaxis, leukocyte[mh]) OR (Genetics[sh]) OR (heredit*[tiab]) OR (epigenetic*[tiab]) OR (polymorphism*[tiab]) OR (gene-environment interaction*[tiab]) OR (environment-gene interaction*[tiab]) OR (genotype-environment interaction*[tiab]) OR (genotype-phenotype[tiab]) OR (Pathology[sh]) OR (biopsy[tiab]) OR (biopsies[tiab]) OR (cytopatholog*[tiab]) OR (histopatholog*[tiab]) OR (immunopatholog*[tiab]) OR (autopsy[tiab]) OR (autopsies[tiab]) OR (ultrastructur*[tiab]) OR (subcellular structure*[tiab]) OR (Physiopathology[sh]) OR (dysfunction[tiab]) OR (pathophysiolog*[tiab]) OR (molecular mechanism*[tiab]) OR (biomarkers[mh]) OR (biomarker[tiab]) OR (biochemical marker*[tiab]) OR (biologic marker*[tiab]) OR (clinical marker*[tiab]) OR (immune marker*[tiab]) OR (immunologic marker*[tiab]) OR (laboratory marker*[tiab]) OR (surrogate marker*[tiab]) OR (surrogate end point*[tiab]) OR (surrogate endpoint*[tiab]) OR (albumins[mh]) OR (blood[mh]) OR (blood[sh]) OR (inflammation mediators[mh]) OR (adipokines[mh]) OR (glycoproteins[mh]) OR (blood coagulation factors[mh]) OR (iron-binding proteins[mh]) OR (free fatty acid*[tiab]) OR (amyloid[mh]) OR (micrornas[mh]) OR (etiology[sh:noexp]) OR (causality[tiab]) OR (causes[tiab]) OR (pathogenesis[tiab]) OR (oxidative stress[mh]) OR (endothelium[mh])) NOT ((COVID-19[Supplementary Concept]) OR (COVID-19[mh]) OR (SARS-CoV-2[mh]) OR (2019 novel coronavirus disease[tiab]) OR (COVID19[tiab]) OR (COVID-19 virus disease[tiab]) OR (coronavirus disease 2019[tiab]) OR (coronavirus disease-19[tiab]) OR (2019-nCoV disease[tiab]) OR (2019 novel coronavirus infection*[tiab]) OR (2019-nCoV[tiab]) OR (novel coronavirus[tiab]) OR (SARS-CoV-2[tiab])) NOT (pregnan*) NOT (child*) NOT (infant*) NOT (lung diseases, obstructive[mh]) NOT (fatty liver[mh]) NOT (disease management[mh]) NOT (trends[sh]) NOT (statistics & numerical data[sh]) NOT (sex factors[mh]) NOT (race factors[mh]) NOT (age factors[mh]) NOT (social determinants of health[mh]) NOT (risk assessment[mh]) NOT (telemedicine[mh]) NOT (instrumentation[sh]) NOT (psychology[sh]) NOT (rehabilitation[sh]) NOT (economics[sh]) NOT (animals[mh:noexp]) NOT (methods[sh]) NOT (drug therapy[sh]) NOT (cancer*) NOT (Diagnosis[sh]) NOT (Pharmacology[sh]) NOT (Standards[sh]) NOT (prevention & control[sh]) NOT (preprint[pt])

Chronic Kidney Disease – ((renal insufficiency[majr]) OR (Kidney Failure[tiab]) OR (chronic kidney insufficienc*[tiab]) OR (chronic renal insufficienc*[tiab]) OR (Renal Failure[tiab]) OR (end-stage kidney disease[tiab]) OR (end-stage renal disease[tiab]) OR (CKD[tiab]) OR (chronic kidney disease*[tiab]) OR (chronic renal disease*[tiab])) AND ((Metabolism[sh]) OR (biochemical pathway*[tiab]) OR (biodegradation[tiab]) OR (biotransformation[tiab]) OR (catabolism[tiab]) OR (degradation[tiab]) OR (incorporation[tiab]) OR (mobilization[tiab]) OR (secretion[tiab]) OR (turnover[tiab]) OR (anabolism[tiab]) OR (bioformation[tiab]) OR (enzyme activity[tiab]) OR (enzyme*[tiab]) OR (urinary aspects[tiab]) OR (urinary levels[tiab]) OR

(Immunology[sh]) OR (non-specific immun*[tiab]) OR (innate immun*[tiab]) OR (native immun*[tiab]) OR (natural immun*[tiab]) OR (adaptive immun*[tiab]) OR (humoral immun*[tiab]) OR (humoural immun*[tiab]) OR (cellular immun*[tiab]) OR (cell-mediated immun*[tiab]) OR (primary immun*[tiab]) OR (antiviral immun*[tiab]) OR (anti-viral immun*[tiab]) OR (peripheral immun*[tiab]) OR (Genetics[sh]) OR (heredit*[tiab]) OR (epigenetic*[tiab]) OR (polymorphism*[tiab]) OR (gene-environment interaction*[tiab]) OR (environment-gene interaction*[tiab]) OR (genotype-environment interaction*[tiab]) OR (genotype-phenotype[tiab]) OR (Pathology[sh]) OR (biopsy[tiab]) OR (biopsies[tiab]) OR (cytopatholog*[tiab]) OR (histopatholog*[tiab]) OR (immunopatholog*[tiab]) OR (autopsy[tiab]) OR (autopsies[tiab]) OR (ultrastructur*[tiab]) OR (subcellular structure*[tiab]) OR (Physiopathology[sh]) OR (dysfunction[tiab]) OR (pathophysiolog*[tiab]) OR (molecular mechanism*[tiab]) OR (biomarkers[mh]) OR (biomarker[tiab]) OR (biochemical marker*[tiab]) OR (biologic marker*[tiab]) OR (clinical marker*[tiab]) OR (immune marker*[tiab]) OR (immunologic marker*[tiab]) OR (laboratory marker*[tiab]) OR (surrogate marker*[tiab]) OR (surrogate end point*[tiab]) OR (surrogate endpoint*[tiab]) OR (etiology[sh:noexp]) OR (causality[tiab]) OR (causes[tiab]) OR (pathogenesis[tiab])) NOT ((COVID-19[Supplementary Concept]) OR (COVID-19[mh]) OR (SARS-CoV-2[mh]) OR (2019 novel coronavirus disease[tiab]) OR (COVID19[tiab]) OR (COVID-19 virus disease[tiab]) OR (coronavirus disease 2019[tiab]) OR (coronavirus disease-19[tiab]) OR (2019-nCoV disease[tiab]) OR (2019 novel coronavirus infection*[tiab]) OR (2019-nCoV[tiab]) OR (novel coronavirus[tiab]) OR (SARS-CoV-2[tiab])) NOT (pregnan*) NOT (child*) NOT (infant*) NOT (lung diseases, obstructive[mh]) NOT (fatty liver[mh]) NOT (disease management[mh]) NOT (trends[sh]) NOT (statistics & numerical data[sh]) NOT (sex factors[mh]) NOT (race factors[mh]) NOT (age factors[mh]) NOT (social determinants of health[mh]) NOT (risk assessment[mh]) NOT (telemedicine[mh]) NOT (instrumentation[sh]) NOT (psychology[sh]) NOT (rehabilitation[sh]) NOT (animals[mh:noexp]) NOT (methods[sh]) NOT (drug therapy[sh]) NOT (cancer*) NOT (Diagnosis[sh]) NOT (Pharmacology[sh]) NOT (Standards[sh]) NOT (prevention & control[sh]) NOT (economics[sh]) NOT (preprint[pt])

APPENDIX B: SEMANTIC GROUPS AND TYPES

Semantic Group	Semantic Types	Example
Physiology	Physiologic Function; Molecular Function; Organism Function; Organ or Tissue Function; Cell Function; Genetic Function; Organism Attribute; Clinical Attribute	Glucose metabolism
Anatomy	Cell; Cell Component; Tissue; Body Substance	Adipocytes
Disorders	Disease or Syndrome; Pathologic Function; Cell or Molecular Dysfunction	COVID-19
Chemicals and Drugs	Amino Acid, Peptide, or Protein; Hormone; Immunologic Factor; Enzyme; Biologically Active Substance; Receptor; Nucleic Acid, Nucleoside, or Nucleotide; Element, Ion, or Isotope; Vitamin; Carbohydrate; Lipid; Hazardous or Poisonous Substance; Inorganic Chemical; Organic Chemical	ACE2 protein
Genes and Molecular Sequences	Gene or Genome	IL6 gene
Phenomena	Biologic Function; Natural Phenomenon or Process	Virus replication
Concepts and Ideas	Conceptual Entity; Functional Concept	JNK pathway

APPENDIX C: STOPLIST OF GENERIC CONCEPTS

Concept Identifier	Name
C0012634	Disease
C0042776	Virus
C1099354	RNA, Small Interfering
C0017262	Gene Expression
C1334043	Homologous Gene
C0752046	Single Nucleotide Polymorphism
C0040549	Toxin
C0751973	Proteome
C0026882	Mutation
C0011065	Cessation of life
C0162326	DNA Sequence
C0003086	Ankle
C0037313	Sleep
C0040648	TRANSCRIPTION FACTOR
C0024109	Lung
C1101610	MicroRNAs
C0012854	DNA
C1140618	Upper Extremity
C0015385	Limb structure
C0079189	cytokine
C0006104	Brain
C0013081	Down-Regulation
C0035696	RNA, Messenger
C0682523	Human Cell Line
C0040300	Body tissue
C0015392	Eye
C0035298	Retina
C0042789	Vision
C0007600	Cell Line
C0042798	Visual impairment
C0175996	Protoplasm
C0678951	gene polymorphism
C0021359	Infertility
C0440744	Human tissue
C0450442	Agent
C0314657	Genetic Predisposition to Disease
C0178784	Organ
C0597357	receptor
C0013470	Eating
C0042210	Vaccines
C0018563	Hand
C0231170	Disability NOS

Concept Identifier	Name
C0041904	Up-Regulation (Physiology)
C0005456	Binding Sites
C0014653	Equilibrium
C1522240	Process
C0443640	Specific antibody
C0085639	Falls
C0041755	Adverse drug effect
C0018284	Growth Factor
C0006147	Breast Feeding
C0699680	Metric
C0241863	Diabetic
C0678544	wave
C1512488	Homologous Protein
C0033413	Promoter Regions (Genetics)
C0029235	Organism
C0003241	Antibodies
C0015733	Feces
C0023317	Lens, Crystalline
C0817096	Chest
C0038250	Stem cells
C0000726	Abdomen
C0221198	Lesion
C0205400	Thickened
C0032529	Polymorphism, Genetic
C0043251	Wounds and Injuries
C0814002	Neural Development
C0086287	Females
C0027428	Structure of mucous membrane of nose
C0024032	Low Birth Weights
C0600688	Toxic effect
C0039597	Testis
C0038432	Streptozocin
C0009450	Communicable Diseases
C0037817	Speech
C0231303	Distress
C0022742	Knee
C0025552	Metals
C0027442	Nasopharynx
C1550101	Supernatant
C0035203	Respiration
C0015745	Feeding behaviors
C0015930	Fetal Distress
C0025320	Menopause
C0016658	Fracture

Concept Identifier	Name
C0028778	Obstruction
C0599779	Animal Model
C0599732	cell injury
C0162327	RNA Sequence
C0023216	Lower Extremity
C0700276	Anatomic structures
C0010957	Tissue damage
C0025255	Tissue membrane
C0038435	Stress
C0036866	Sex Characteristics
C0243076	antagonists
C0085080	Chinese Hamster Ovary Cell
C0041582	Ulcer
C0796494	Lobe
C0015450	Face
C0006159	Breeding
C0243077	inhibitors
C0032961	Pregnancy
C0042149	Uterus
C1171362	Protein Expression
C1446377	Mental health problem
C0005615	Birth
C0456909	Blind Vision
C0040480	Musculoskeletal torsion, function
C0015895	Fertility
C0683321	poor health
C0017260	Gene Deletion
C0003316	Epitopes
C0022864	Labor (Childbirth)
C0042449	Veins
C0597360	receptor expression
C0028429	Nose
C0013203	Drug resistance
C0035245	Respiratory physiology
C0229089	Right eye
C0010357	Cross Reactions
C1328819	small molecule
C0851346	Radiation
C0226896	Oral cavity
C0443158	Brain activity
C0242786	High-Risk Pregnancy
C0008946	Climate
C0206419	Genus: Coronavirus
C0597177	particle

Concept Identifier	Name
C0016701	Freezing
C0042333	Variation (Genetics)
C0599220	Protein Subunits
C0016504	Pes
C0042542	Vero Cells
C0000934	Acclimatization
C0035150	Reproduction
C0442692	Reproductive process
C0009637	Conception
C0032787	Postoperative Complications
C0005898	Body Regions
C1566558	Natural Products
C0041004	Triglycerides
C0282554	chemokine
C0234451	Sleep, Slow-Wave
C0086860	Promoter (Genetics)
C0870935	Napping
C0030660	Pathologic Processes
C0151526	Premature Birth
C1332838	Candidate Disease Gene
C0543419	Sequela of disorder
C0205949	Sexual Orientation
C0005889	Body Fluids
C0008269	Chloroquine
C0086582	Males
C0000786	Spontaneous abortion
C1515670	mRNA Expression
C0033640	PROTEIN KINASE
C0025329	Menstrual cycle
C0013790	Electricity
C0025274	Menarche
C0489786	Height
C0278092	Sexual function
C0037361	Smell Perception
C0042939	Voice
C0032914	Pre-Eclampsia
C0032931	Precipitation
C0220898	Predisposition
C0010031	Cornea
C0743925	Fetal Growth
C0677874	In complete remission
C0442749	05-Jun
C0341950	Severe pre-eclampsia unspecified
C0233481	Worried

Concept Identifier	Name
C0678933	genetic locus
C0012652	Disease Outbreaks
C0017504	Gestational Age
C0586688	Tissue specimen from liver
C0085732	Ability
C1318963	Readiness
C0020167	Humidity
C0428692	Ambient temperature
C0028877	Odontogenesis
C0002151	Alloxan
C0009253	Coitus
C0004886	BCG Vaccine
C0023974	Loneliness
C0162358	Ecosystem
C0229962	Body part
C0032962	Pregnancy Complications
C0595939	Stillbirth
C0242640	Multi-Drug Resistance
C0206076	Reproductive History
C0442759	06-Mar
C0392534	Ruptured ectopic pregnancy
C0037420	Social Interaction
C1522002	RNA Recognition Motif
C0024888	Mastication
C0038442	Stress, Mechanical
C0231224	Crisis
C0233324	Term Birth
C0015944	Fetal Membranes, Premature Rupture
C0079399	Gender
C0277787	Social stigmata
C0806140	Flow
C0679225	multiple pathologies
C1166607	cellular_component
C0015927	Fetal Death
C0032994	Pregnancy, Tubal
C0235280	Ototoxicity
C0221082	Etiology, operative procedure, as cause of
C0020336	Hydroxychloroquine
C0678568	cooling
C0878751	Late pregnancy
C0042034	Urination
C0442752	12-Jun
C0278054	Male reproductive function
C0000832	Abruptio Placentae

Concept Identifier	Name
C0683140	Drug Metabolism
C0032987	Pregnancy, Ectopic
C1148523	Childbirth
C0000936	Visual Accommodation
C0681779	atmospheric condition
C0086685	Natural Selection
C0678881	testicular function
C0014259	Corneal Endothelium
C1268086	Body structure
C0030847	Penile Erection
C0001168	Complete obstruction
C0025323	Menorrhagia
C0021294	Infant, Premature
C1515300	Testicular Tissue
C0031104	Periodontium
C0025594	Meteorological Factors
C0016248	Floods
C0010813	Cytokinesis
C0029164	Dental Hygiene
C0032984	Pregnancy, Abdominal
C0011135	Defecation
C0450448	Waveforms
C0450254	Pathogenic organism
C0149744	Oral lesion
C0456057	Fetal stress
C0021361	Female infertility
C0336996	Physical force
C0419437	High risk infant
C1148560	molecular_function
C0033421	Pronation
C0868933	Climatic factors
C0683954	research outcome
C0028884	Odors
C0035154	Reproductive Behavior
C0425152	Engaged to be married
C0574765	Grey hair
C0559477	Perinatal asphyxia
C0005612	Birth Weight
C0241889	Family history of
C0043085	Weather
C0807745	RESISTANCE.INDEX
C0014499	Epidemic
C0337014	Avalanche
C0178292	Complications of pregnancy, childbirth and the puerperium

Concept Identifier	Name
C0946401	SPHERE
C0871747	Fetal Exposure
C0456149	Intelligence quotient
C0012618	Disasters
C1615608	Pandemics
C1096243	Central line infection
C0027485	Natural Disasters
C0678659	biochemical mechanism
C0150312	Present
C0370003	Specimen
C0233426	Personal appearance
C0337000	Cyclone
C0005520	Biological Phenomena
C1444662	Discontinued
C0038941	Surgical Wound Infection
C0458827	Airway structure
C0037088	Signs and Symptoms
C0442768	1/60
C0450030	Fog
C0678723	Biologic Development
C0934502	anatomical layer
C0444868	All
C0012644	Animal Disease Models
C0035648	risk factors
C1608383	whole blood specimen
C0441655	Activities
C0598197	Contagion
C0349482	High birth weight infant
C1510610	Globalization
C0595998	Household composition
C0520930	Late menarche
C0425119	Child at risk
C0405136	[X]Multiple delivery, unspecified
C0232515	Spitting
C0233894	Femininity
C0041276	Ruptured tubal pregnancy
C0278095	Male sexual function
C1326169	microglial cell activation
C0033213	Problem
C0848898	Morbidity, newborn
C0332149	Possible
C0086312	Forests
C0009488	Comorbidity
C1545588	Protection

Concept Identifier	Name
C4321237	High Level
C1825598	IMPACT gene
C4281807	Vitronectin, human
C3178810	Transcriptome
C0439663	Infected
C3272283	American College of Cardiology/American Heart Association Lesion Complexity Score C
C2987634	Agonist
C3272281	American College of Cardiology/American Heart Association Lesion Complexity Score A
C2825142	Experimental Result
C4699158	Increased risk
C0750484	Confirmation
C2926735	Duration
C0205160	Negative
C3714738	Compliance
C4050466	Borg Category-Ratio 10 Perceived Exertion Score 5
C0600457	Gravidity
C3714514	Infection
C1457868	Worse
C0087130	Uncertainty
C0231170	Disability
C3816499	Pathogenic
C0518609	Consideration
C0683525	treatment options
C3536832	air
C0007600	Cultured Cell Line
C0016504	Foot
C3842672	Day 7
C1550100	Specimen Type - Serum
C4505065	Noncommunicable Diseases
C1883559	Wild Type
C0184511	Improved
C1829822	Mental health.status
C1821461	Close Relationship
C3840880	Traffic
C4738506	Operating
C0679215	health and disease
C4534363	At home
C0080048	Privacy
C3662030	Spontaneous cerebral hemorrhage
C0032529	Genetic Polymorphism
C1999216	Inhibitor
C4321351	Low Level

Concept Identifier	Name
C1764827	Isolate - microorganism
C4743777	Activator
C2584321	Reaching
C1882365	Phenomenon
C2985438	Novel Mutation
C1516998	Exogenous Factors
C1299586	Has difficulty doing (qualifier value)
C4330475	Immune Cell
C0439662	Immune
C3843156	Less often
C0041755	Adverse reaction to drug
C0033640	Protein Kinases
C0220898	Predisposition -- attribute
C0024819	Marital Status
C0683954	research results
C3494405	Maternal Death
C2718051	Climate Change
C4035627	2 times
C0456909	Blindness
C0042798	Low Vision
C0158915	Exceptionally large baby (disorder)
C0086860	Promoter
C3843309	1 time
C4055506	Accumulation
C4035626	3 times
C4722602	Underlying
C3687742	Oropharyngeal swab
C3845288	Strong positive
C1709157	Negative Surgical Margin
C3714634	Biological Processes
C0678568	Cool - action
C0029162	Oral health
C2986594	Mouse Model
C2717940	Hep G2 Cells
C4698664	Rocky
C1998720	Effective Communication
C0848898	Neonatal morbidity
C0033413	Promoter Regions, Genetic
C0018748	Health Services Accessibility
C2348693	Flux
C3841448	Much worse
C1948023	Stimulation (motivation)
C3533236	Mean score
C1704241	complex (molecular entity)

Concept Identifier	Name
C1610733	Urine - SpecimenType
C1749467	soluble
C3887486	Interstitial lung fibrosis
C4699604	<12 months
C0024888	Chewing
C0241889	Family history
C0032972	Pregnancy Outcome
C0814002	Neurogenesis
C0150637	assessment.initial
C1398625	Gestational Weight Gain
C0678544	wave - physical agent
C0341950	Severe pre-eclampsia
C2364172	Adherence To Medication Regime
C2939181	Motor vehicle accident
C0699680	Metric (substance)
C2370955	Smell function
C0948496	abortion late
C0032987	Ectopic Pregnancy
C3853758	Metabolic Profile
C3898092	Oral Complication
C0238617	High altitude (physical force)
C2362326	Sexual Health
C0221082	adverse effect due to surgery
C0349482	High birth weight
C3843647	> 2 years
C0678933	Genetic Loci
C2886794	Catheter related bloodstream infection
C0442768	20/1200
C1881717	Medical Device Mechanical Issue
C2921106	Recurrent pregnancy loss
C3687582	Produces milk for human food
C3843645	10-Jan
C1720845	Maternal Nutritional Physiological Phenomena
C2350828	Physiological Phenomena
C3272282	American College of Cardiology/American Heart Association Lesion Complexity Score B
C0558187	Lactation established (finding)
C0337000	Cyclonic Storms
C3662302	Deep incisional surgical site infection
C2985294	Fourth Stage of Labor
C5204818	Global Response
C0522534	Saturated
C3852980	Drug Activation
C3641827	Agree

Concept Identifier	Name
C3494255	Water Resources
C0946401	ocular sphere
C1551358	Incident
C4021819	Phenotypic abnormality
C0442752	Distance vision 6/12
C0405136	Multiple delivery
C4021061	Testicular fibrosis
C4698298	Genotype 3
C1998926	Tsunamis
C3898097	Ophthalmologic Complication
C0444089	Umbilical cord tissue sample
C0178237	diseases and injuries
C0178477	Animal Breeding
C2880858	Bilateral occlusion of central retinal arteries
C0574765	Gray hair
C4049706	Borg Category-Ratio 10 Perceived Exertion Score 3
C5236984	Responsive Disease
C3816499	Pathogenic Variant
C0221082	Surgical Complication
C0337014	Avalanches
C5392851	Water Scarcity
C5392245	Water Insecurity
C5380405	cellular anatomical entity

APPENDIX D: UMLS TARGET DISEASE CONCEPTS

Concept Identifier	Name	Source Vocabulary
C5397144	Acute respiratory distress syndrome due to disease caused by Severe acute respiratory syndrome coronavirus 2	SNOMEDCT_US
C3875082	Chronic kidney disease due to type 1 diabetes mellitus	SNOMEDCT_US
C3662038	Chronic kidney disease due to type 2 diabetes mellitus	SNOMEDCT_US
C2316401	Chronic kidney disease stage 1	SNOMEDCT_US
C2316786	Chronic kidney disease stage 2	SNOMEDCT_US
C2316787	Chronic kidney disease stage 3	SNOMEDCT_US
C3839533	Chronic kidney disease stage 3A	SNOMEDCT_US
C3839870	Chronic kidney disease stage 3B	SNOMEDCT_US
C2317473	Chronic kidney disease stage 4	SNOMEDCT_US
C2316810	Chronic kidney disease stage 5	MTH
C1561638	Chronic kidney disease, stage I	ICD9CM
C1561639	Chronic kidney disease, stage 2 (mild)	ICD10CM
C1561640	Chronic kidney disease, Stage III (moderate)	ICD10CM
C1561641	Chronic kidney disease, stage 4 (severe)	ICD10CM
C1561642	Chronic kidney disease, stage V	ICD9CM
C5439539	Chronic post-COVID-19 syndrome	SNOMEDCT_US
C5419164	COVID-19-Associated Acute Respiratory Distress Syndrome	NCI
C5419163	COVID-19-Associated Pneumonia	NCI
C0205734	Diabetes, Autoimmune	MSH
C3149273	Diabetes, nonautoimmune	OMIM
C0011880	Diabetic Ketoacidosis	MTH
C0011881	Diabetic Nephropathy	MTH
C0854110	Insulin-resistant diabetes mellitus	HPO
C1739108	Latent Autoimmune Diabetes in Adults	MSH
C5397146	Lower respiratory infection caused by SARS-CoV-2	SNOMEDCT_US
C0342276	Maturity onset diabetes mellitus in young	MTH
C5431835	Multisystem inflammatory syndrome associated with COVID-19	MTH
C5244027	Pneumonia caused by SARS-CoV-2	SNOMEDCT_US
C5439525	Post-acute COVID-19	SNOMEDCT_US
C5433293	Post-acute COVID-19 syndrome	MSH
C1720457	Renal disorder due to type 2 diabetes mellitus	SNOMEDCT_US
C5400365	SARS-CoV-2 viremia	SNOMEDCT_US

Concept Identifier	Name	Source Vocabulary
C3250571	Type 2 diabetes mellitus with diabetic chronic kidney disease	ICD10CM
C5391473	Envelope protein, SARS-CoV-2	MSH
C5391474	Membrane protein, SARS-CoV-2	MSH
C5391850	Ns7b protein, SARS-CoV-2	MSH
C5391854	NS8 protein, SARS-CoV-2	MSH
C5391563	NSP1 protein, SARS-CoV-2	MSH
C5392373	Nsp2 protein, SARS-CoV-2	MSH
C5433456	NSP3 protein, SARS-CoV-2	MSH
C5391562	NSP4 protein, SARS-CoV-2	MSH
C5433352	NSP5A protein, SARS-CoV-2	MSH
C5433353	NSP5B protein, SARS-CoV-2	MSH
C5391507	NSP6 protein, SARS-CoV-2	MSH
C5391489	NSP7 protein, SARS-CoV-2	MSH
C5391490	NSP8 protein, SARS-CoV-2	MSH
C5391508	NSP9 protein, SARS-CoV-2	MSH
C5391509	NSP10 protein, SARS-CoV-2	MSH
C5433521	NSP12 protein, SARS-CoV-2	MSH
C5435227	NSP14 protein, SARS-CoV-2	MSH
C5391445	NSP15 protein, SARS-CoV-2	MSH
C5433590	NSP16 protein, SARS-CoV-2	MSH
C5433528	Nucleocapsid phosphoprotein, SARS-CoV-2	MSH
C5392670	ORF1a polyprotein protein, SARS-CoV-2	MSH
C5391478	ORF3a protein, SARS-CoV-2	MSH
C5391479	ORF6 protein, SARS-CoV-2	MSH
C5391480	ORF7a protein, SARS-CoV-2	MSH
C5391481	ORF7b protein, SARS-CoV-2	MSH
C5391482	ORF8 protein, SARS-CoV-2	MSH
C5435294	PLpro protein, SARS-CoV-2	MSH
C5391442	Spike protein, SARS-CoV-2	MSH

APPENDIX E: RESULTS OF PREDICATION EXTENSION (N-2)

Method	Input (*n-1*)	Existing	Overlap	Similarity
Child/narrower relations	Joint associations	2,327	3.9%	42.3%
	Direct associations	7,098	2.4%	45.6%
Sibling/other relations	Joint associations	517	0.6%	34.6%
	Direct associations	1,348	1.0%	38.9%
Parent/broader relations	Joint associations	177	5.1%	35.5%
	Direct associations	682	5.4%	40.2%

APPENDIX F: ANALYSIS OF PREDICATION EXTENSION PRUNING

Method	Cycle No.	Pruning	Count	Continuity	Recall
Gene selection	1	None	41,662	11.7%	43.9%
		CP	18,561	13.6%	32.0%
		Intermediate	9,686	35.8%	22.3%
		Link	877	16.7%	3.1%
	2, 3	None	19,967	9.6%	35.6%
		CP	3,380	9.3%	9.2%
		Intermediate	3,900	28.9%	14.0%
		Link	347	7.8%	1.4%
GO term selection	1	None	6,979	25.5%	5.4%
		CP	3,841	33.2%	5.1%
		Intermediate	5,516	31.7%	5.2%
		Link	172	15.7%	1.2%
	2	None	1,056	30.2%	0.6%
		CP	591	39.8%	0.5%
		Intermediate	441	71.9%	0.5%
		Link	34	26.5%	0.2%

APPENDIX G: DISCOVERY PATTERN HYPOTHESES

Input: Implicit relations (no pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	3 nodes	ACE2 protein	COVID-19		x	
2	3 nodes	Spike protein, SARS-CoV-2	COVID-19		x	
3	4 nodes	TNF protein	COVID-19	x		
4	4 nodes	NF-kappa B	COVID-19			x
5	3 nodes	NF-kappa B	COVID-19		x	
6	3 nodes	NFE2L2 gene	COVID-19		x	
7	5 nodes	Estrogen receptor alpha	COVID-19			x
8	6 nodes	Leptin	COVID-19			x
9	3 nodes	N protein, SARS-CoV-2	COVID-19	x		
10	4 nodes	Leptin	Diabetes	x		
11	5 nodes	Interleukin-6	Diabetes		x	
12	6 nodes	ACE2 gene	T2DM		x	
13	6 nodes	SIRT1 gene	Diabetes	x		
14	5 nodes	Adiponectin	T2DM		x	
15	4 nodes	NF-kappa B	T2DM		x	
16	5 nodes	GJA1 gene	Diabetes	x		
17	4 nodes	SIRT1 gene	T2DM	x		
18	5 nodes	TNF protein	Diabetes			x
19	6 nodes	ACE2 protein	Diabetes		x	
20	6 nodes	ACE2 protein	Obesity		x	

Input: Explicit relations

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	3 nodes	ACE2 protein	COVID-19		x	
2	5 nodes	Interleukin-6	Diabetes		x	
3	5 nodes	Leptin	Diabetes		x	
4	5 nodes	Adiponectin	T2DM		x	
5	5 nodes	NF-kappa B	Diabetes		x	
6	5 nodes	TNF protein	Diabetes			x
7	5 nodes	SIRT1 gene	T2DM		x	
8	3 nodes	Spike protein, SARS-CoV-2	COVID-19		x	
9	6 nodes	ACE2 gene	Diabetes		x	
10	4 nodes	NF-kappa B	COVID-19			x
11	5 nodes	IGF-1 protein	T2DM		x	
12	5 nodes	Adiponectin	Obesity		x	
13	5 nodes	Leptin	Obesity		x	
14	3 nodes	NF-kappa B	COVID-19		x	

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
15	5 nodes	FBN1 gene	T2DM		x	
16	6 nodes	ACE2 protein	T2DM		x	
17	5 nodes	AKT protein	T2DM		x	
18	5 nodes	TGF beta	Diabetes		x	
19	5 nodes	PPAR gamma	Diabetes		x	
20	5 nodes	KL protein	T2DM		x	

APPENDIX H: DISCOVERY PATTERN HYPOTHESES (WEIGHTED)

Input: Implicit relations (no pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	3 nodes	ACE2 protein	COVID-19		x	
2	6 nodes	Leptin	Diabetes	x		
3	4 nodes	TNF protein	COVID-19	x		
4	3 nodes	Spike protein, SARS-CoV-2	COVID-19		x	
5	4 nodes	NF-kappa B	COVID-19			x
6	6 nodes	Leptin	Obesity	x		
7	5 nodes	Interleukin-6	Diabetes		x	
8	6 nodes	ACE2 gene	Diabetes		x	
9	6 nodes	ACE2 protein	Obesity		x	
10	5 nodes	GJA1 gene	Diabetes	x		
11	5 nodes	Adiponectin	T2DM		x	
12	3 nodes	NF-kappa B	COVID-19		x	
13	6 nodes	SIRT1 gene	Diabetes	x		
14	5 nodes	NF-kappa B	Diabetes		x	
15	5 nodes	Adiponectin	Obesity		x	
16	5 nodes	Estrogen receptor alpha	COVID-19			x
17	6 nodes	Leptin	COVID-19			x
18	5 nodes	TNF protein	Diabetes			x
19	3 nodes	NFE2L2 gene	COVID-19		x	
20	6 nodes	ACE2 protein	Diabetes		x	

Input: Explicit relations

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	3 nodes	ACE2 protein	COVID-19		x	
2	5 nodes	Interleukin-6	Diabetes		x	
3	5 nodes	Leptin	Diabetes		x	
4	5 nodes	Adiponectin	T2DM		x	
5	5 nodes	NF-kappa B	Diabetes		x	
6	5 nodes	Adiponectin	Obesity		x	
7	5 nodes	TNF protein	Diabetes			x
8	5 nodes	Leptin	Obesity		x	
9	5 nodes	SIRT1 gene	Diabetes		x	
10	6 nodes	ACE2 gene	Diabetes		x	
11	3 nodes	Spike protein, SARS-CoV-2	COVID-19		x	
12	4 nodes	NF-kappa B	COVID-19			x
13	5 nodes	IGF-1 protein	Diabetes		x	
14	6 nodes	ACE2 protein	Obesity		x	

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
15	5 nodes	KL protein	Diabetic Nephropathy		x	
16	5 nodes	TGF beta	Diabetes		x	
17	5 nodes	PPAR gamma	Obesity		x	
18	5 nodes	FBN1 gene	T2DM		x	
19	5 nodes	PPAR gamma	Diabetes		x	
20	5 nodes	Interleukin-6	Obesity		x	

APPENDIX I: DISCOVERY PATTERN HYPOTHESES (PRUNING)

Input: Implicit relations (no pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	4 nodes	TNF protein	COVID-19	x		
2	5 nodes	Estrogen receptor alpha	COVID-19			x
3	6 nodes	Leptin	COVID-19			x
4	4 nodes	Leptin	Diabetes	x		
5	5 nodes	GJA1 gene	Diabetes	x		
6	4 nodes	SIRT1 gene	T2DM		x	
7	6 nodes	SIRT1 gene	Diabetes	x		
8	5 nodes	TNF protein	Diabetes			x
9	5 nodes	Leptin	T2DM			x
10	5 nodes	Leptin	Obesity			x
11	6 nodes	PPAR gamma	Diabetes	x		
12	4 nodes	TNF protein	T2DM	x		
13	5 nodes	SIRT1 gene	Diabetes	x		
14	6 nodes	FGF21 protein	Diabetes		x	
15	6 nodes	Ghrelin	Diabetes	x		
16	4 nodes	APLN gene	Diabetes			x
17	6 nodes	TP53 gene	Diabetes			x
18	4 nodes	PPAR gamma	Diabetes		x	
19	6 nodes	HIF1A protein	Diabetes	x		
20	5 nodes	NLRP3 gene	T2DM	x		

Input: Implicit relations (Common Parents pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	4 nodes	TNF protein	COVID-19	x		
2	5 nodes	Leptin	Diabetes			x
3	6 nodes	SIRT1 gene	Diabetes	x		
4	5 nodes	GJA1 gene	Diabetes	x		
5	5 nodes	Leptin	T2DM			x
6	5 nodes	SIRT1 gene	Diabetes	x		
7	4 nodes	SIRT1 gene	T2DM	x		
8	5 nodes	Leptin	Obesity			x
9	5 nodes	SIRT1 gene	Diabetic Nephropathy	x		
10	4 nodes	TNF protein	T2DM	x		
11	5 nodes	GJA1 gene	T2DM			x
12	6 nodes	TP53 gene	Diabetes			x
13	6 nodes	HIF1A protein	Diabetes	x		
14	5 nodes	NLRP3 gene	T2DM	x		
15	6 nodes	mTOR gene	Diabetes			x

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
16	5 nodes	SMAD3 gene	Diabetic Nephropathy	x		
17	4 nodes	FOXO1 gene	T2DM	x		
18	5 nodes	LCN2 protein	T2DM			x
19	5 nodes	HIF1A protein	Diabetes			x
20	6 nodes	Heme Oxygenase-1	Diabetes	x		

Input: Implicit relations (Intermediate Relations pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	4 nodes	TNF protein	COVID-19	x		
2	4 nodes	Leptin	Diabetes	x		
3	5 nodes	TLR4 protein	COVID-19	x		
4	5 nodes	Leptin	Diabetes			x
5	5 nodes	GJA1 gene	Diabetes	x		
6	5 nodes	TNF protein	Diabetes			x
7	4 nodes	Leptin	T2DM	x		
8	5 nodes	Leptin	Obesity			x
9	5 nodes	Leptin	T2DM			x
10	6 nodes	SIRT1 gene	Diabetes	x		
11	4 nodes	SIRT1 gene	Diabetes	x		
12	4 nodes	SIRT1 gene	T2DM		x	
13	5 nodes	SIRT1 gene	Diabetes	x		
14	5 nodes	SIRT1 gene	T2DM	x		
15	5 nodes	GJA1 gene	T2DM			x
16	4 nodes	TNF protein	T2DM	x		
17	4 nodes	APLN gene	Diabetes			x
18	5 nodes	TNF protein	T2DM	x		
19	6 nodes	SIRT1 gene	Diabetic Nephropathy			x
20	6 nodes	TP53 gene	Diabetes			x

Input: Implicit relations (Annotation Extensions pruning)

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
1	5 nodes	Leptin	Diabetes			x
2	6 nodes	TNF protein	T2DM		x	
3	5 nodes	Leptin	Obesity			x
4	4 nodes	TGFB1 protein	Diabetic Nephropathy	x		
5	5 nodes	Leptin	DM			x
6	4 nodes	TNF protein	Diabetic Nephropathy	x		

Pattern No.	Length	Gene	Disorder	Type 1	Type 2	Type 3
7	4 nodes	GJA1 gene	Diabetic Nephropathy	x		
8	4 nodes	NOS3 protein	Diabetic Nephropathy			x
9	4 nodes	TP53 gene	Diabetic Nephropathy	x		
10	4 nodes	TLR4 protein	Diabetic Nephropathy	x		
11	4 nodes	mTOR gene	Diabetic Nephropathy	x		
12	5 nodes	STAT3 gene	Diabetic Nephropathy	x		
13	4 nodes	HIF1A protein	Diabetic Nephropathy			x
14	4 nodes	STAT3 protein	Diabetic Nephropathy	x		
15	4 nodes	CD36 protein	Diabetic Nephropathy			x
16	5 nodes	Leptin	Diabetic Nephropathy			x
17	5 nodes	GAS6 gene	Obesity		x	
18	4 nodes	PPARGC1A gene	Diabetic Nephropathy	x		
19	4 nodes	HGS protein	Diabetic Nephropathy			x
20	5 nodes	FGF19 gene	Obesity	x		