NUTRITIVE KNOWLEDGE BASED DISCOVERY: ENHANCING PRECISION NUTRITION HYPOTHESIS GENERATION.

by

Aaron Trautman

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2022

Approved by:

_____
Dr. Cory Brouwer

_____
Dr. Robert Reid

_____
Dr. Cynthia Gibas

_____
Dr. Way Sung

_____
Dr. Xiaoxia Newton

ABSTRACT

AARON TRAUTMAN. Nutritive Knowledge based discovery: enhancing precision nutrition hypothesis generation. . (Under the direction of DR. CORY BROUWER)

Diet-related diseases like obesity and type-2 diabetes are on the rise. Precision nutrition, a way to tailor dietary requirements for each individual, is heralded as a solution to these problems. However, nutritional research is held within sparse, siloed resources that rarely connect, which leads to significant barriers hindering the progress of precision nutrition. Three knowledgebases were produced as a result of this work. The ABCkb 1.0 overcomes these barriers by linking 11 separate resources in the path from plants to disease through molecular mechanisms. This resource is built in Neo4j and provides a web-based interface available for browsing (https://abckb.charlotte.edu). A second knowledgebase, ABCkb 2.0 connects microbiota information to diet and human health through the incorporation of text-mined associations from full text articles. The final knowledgebase produced links long-covid to dietary components through possible molecular mechanisms. These three knowledgebases promote progress in precision nutrition to tackle the rise in diet-related disease.

DEDICATION

To my wife. Thank you is not enough for all of the support you have given to me over the years.

ACKNOWLEDGEMENTS

A huge thank you to all who have supported this work and my educational journey. I fully appreciate my advisor whose guidance was invaluable to completing this work. I appreciate the librarians from UNC Charlotte for securing rights to text-mine full text articles and the patience they demonstrated. The Plant Pathways Elucidation Project provided graduate funding along with interns who supported the knowledgebase work. A huge thank you to Steven Blanchard for all the IT support provided and guidance writing better code. Of course this work would not be possible without the help of all my fellow classmates and co-laborers but specifically: Richard Linchangco, and Aneeta Uppal. Thank you to all of the P2EP interns who worked on this project: Logan Kavanaugh, Chris Williams, Aswani Unnikrishnan, Kiera Patanella, Natalie Kratts, Kevin Selles, and Daniel Firo.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HetERel Heterogeneous Edge-adjusted Relevance

IBD Inflammatory Bowel Disease

IBS Irritable Bowel Syndrome

KB KnowledgeBase

NCBI National Center for Biotechnology Information

NLP Natural Language Processing

T2D Type 2 Diabetes

WGS Whole Genome Sequencing

WMS Whole Metagenomic Sequencing

CHAPTER 1: INTRODUCTION

Diet-related, noncommunicable diseases are rising in developing countries worldwide. One of these diseases, Irritable Bowel Syndrome (IBS) a functional disorder defined primarily by its symptoms, affects around 11% of the world's population with prevalence in America reaching as high as 20%[1]. Other diet-related noncommunicable diseases including heart disease, type 2 diabetes (T2D), various respiratory diseases, and some cancers contribute to 71% of all deaths worldwide[2]. Several studies have revealed nutritive connections to these diseases which has driven many public-policy decisions in an attempt to combat this rise[3, 4, 5]. However, many of these studies have not established the molecular mechanisms contributing to these nutritive connections. Further complicating understanding the connections is the interaction of the human gut microbiome with dietary components which produce secondary and tertiary metabolites that impose specific symptoms on host biology[6, 7]. This lack of knowledge hinders the ability to define and characterize optimal human health.

1.1 Knowledge Based Discovery

Literature-based discovery was first described by Don Swanson after discovering the molecular connections between Raynaud syndrome and fish oil in 1986[8]. He did this by manually reading a corpus of abstracts on Raynaud syndrome, a disease that affects blood flow to the extremities under extreme temperatures and stress. He discovered the associated effects of Raynaud's on blood viscosity, platelet aggregation, and further read through corpii of abstracts on the associated topics. This lead to the discovery of Fish Oil as a candidate therapeutic for the symptoms of Raynaud syndrome[9]. This process, though tedious and time-consuming, is how scientific dis-

covery usually occurs. Scientific discoveries are communicated through unstructured literature, journal articles, and then rediscovered by readers. However, structured resources like databases and queriable ontologies facilitate a more efficient method of discovery.

Significant effort has been put into the development of resources to catalogue and expedite scientific discovery. The National Center for Biotechnology Information (NCBI) currently maintains a total of 39 databases [10]. The Nucleic Acids Research journal from Oxford Academic hosts a repository of over 1,600 databases with their categories and publishes an annual database report [11].

Often researchers develop databases surrounding an individual topic of study. Interfaces are subsequently developed to browse and query the contents of these databases. To explore the contents, users perform searches and then download the results. Cross-referencing these results with other resources, requires manual connection of these resources. Difficulties arise when researchers attempt to cross-reference separate resources that use unique identifiers specific to each resource. The same resources generated to aid scientific discovery create a significant barrier to produce testable hypotheses. The urgency of increasing and prevalent diet-related diseases combined with limited knowledge of molecular mechanisms in the pathway of plants to human health, along with the significant barrier of a multiplicity of resources create an interesting challenge to solve.

1.2 Nutritional Resources

Scientific resources are scattered throughout many platforms, and none fully capture the molecular mechanisms by which plants and plant compounds affect human health [12]. NutriChem was developed by text-mining MEDLINE abstracts using a naive bayesian classifier to identify pairs of plants, from NCBI Taxonomy database, and human disease phenotypes [13]. More recently, NutriChem 2.0 has focused primarily on drug-plant associations which can provide informative information on how

efficacy of specific drugs may be affected by consumption of particular plants [14]. One major drawback of the resulting database is the lack of chemical-gene associations, a key component to determine known molecular mechanisms and develop nutritive hypotheses. In addition, the database is no longer available from their web-browser, a key feature of many biological databases (Figure 1.1).

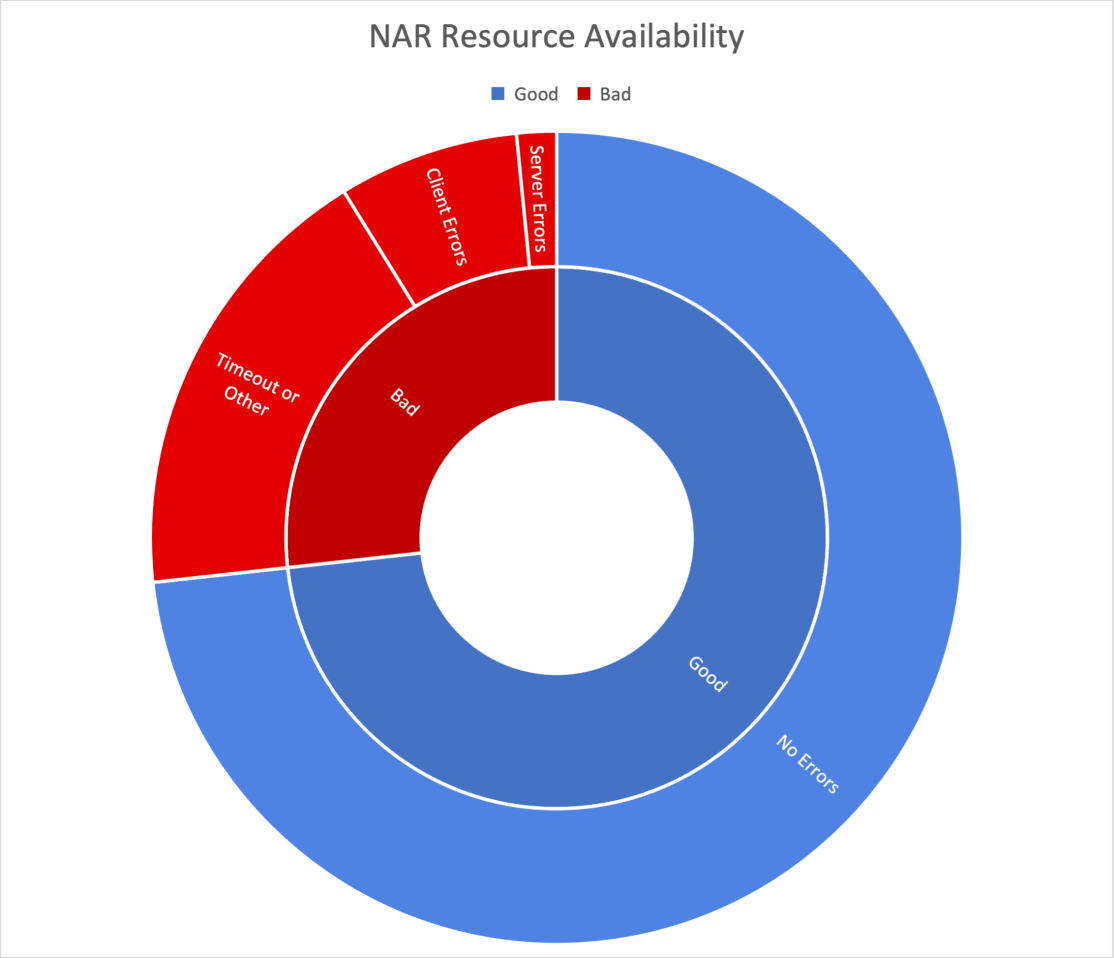

Figure 1.1: Going through all of the available biological resources and databases listed on the NAR, all of the links found on the page were examined for availability and categorized as good, with a url response code of 200 or bad, with response codes grouped by type.

One of the more comprehensive resources available is the Comparative Toxicogenomics Database (CTD) [15]. The CTD provides users with manually curated associations from chemicals, genes, pathways, and phenotypes. Apart from the time and

resources required to develop manually curated connections, the main disadvantage of utilizing this resource for nutritive research is the lack of plant-chemical associations.

Plant-chemical associations can be found in other databases, but are often highly specialized to a specific plant or chemical group. The PhenolExplorer database catalogues plant polyphenol components with quantitative amounts [16]. However, PhenolExplorer was last updated in 2016 and only contains phenolic compounds. Other resources exist to categorize flavor components of food without the medical components. As a general resource, FooDB provides many possible chemical components of plants with quantitative information and mappings to external resources [17]. This resource is frequently updated, however the quantitative chemical amounts are often less informatively provided as a "trace amount."

1.3 Previous Knowledgebase Research

Manually connecting data from separate, isolated resources delays research in the information gathering phase. Previous research from Richard Linchangco resulted in a Knowledgebase (KB) which could provide nutrition research a path to develop new evidence-based hypotheses aiding the discovery of molecular mechanisms between dietary plants on human phenotypes through a path of Plant -> Chemical -> Gene -> Pathway -> Phenotype [18]. The KB was created in Neo4j, a graph database management system and contained nodes with labels: Organisms, Plants, Chemicals, Genes, Pathways, and Phenotypes all extracted from a combination of resources. Connections were extracted from both public resources and associations derived from Natural Language Processing (NLP) of over 30 million abstracts from the MEDLINE index. This KB was not without limitations.

One major drawback of this KB is the technical skill required to utilize it. Neo4j uses Cypher Query Language to retrieve data within the graph database. Users are required to create complex queries to extract data, and connect associations. Additionally, the data contained within was quickly out of date with no way to extract,

transform, and load new, up-to-date data. Another issue with the KB design is the large edge density of the graph database which significantly hindered graph traversal and exploration. Edge density is simply the relative amount of edges, or connections, between nodes. Significant effort was spent reducing predicates into more informative edge labels which yielded some node pairs containing upwards of 20 edges. Graph traversal time is dependent on both the amount of nodes, and connections. Thus, queries from Plants to Phenotypes required significant computational resources to complete, and an unreasonable amount of time (Figure 1.2).

The KB was also never published, never fully implemented, and contained data that added little value to the original goal, providing molecular mechanisms from plants to human health. Finally, as microbiome research has progressed, an additional piece of nutritional research missing from the KB is the contribution from gut microbiota on diet and human health.

1.4 Microbiome Research

The gut microbiome is formally defined as the collection of bacterial genomes contained within the gut, while the term microbiota refers to the set of bacterial species [19]. The two terms are used interchangeably, but it is worth noting the difference. There is a rich history of microbiome research yielding knowledge of a functioning bacterial community existing within healthy organisms[20, 21]. *Escherichia coli* was found naturally occurring in the gut of infants in the 1800's[22]. However, it wasn't until the 1960's that the acceptance of a gut microbiome became more widespread[23]. Many microbiota have only recently been discovered due to their inability to grow in culture. This is in part, due to the technological advancements in genome sequencing over the past 20 years. These advancements have lead to cheaper sequencing costs, greater depth and coverage, better assemblers, and the incorporation of microbiome strategies into many biological analyses. Increases in the specificity of assays have also led to taxinomic reclassification of known species like *alistipes* from *bacteroides*

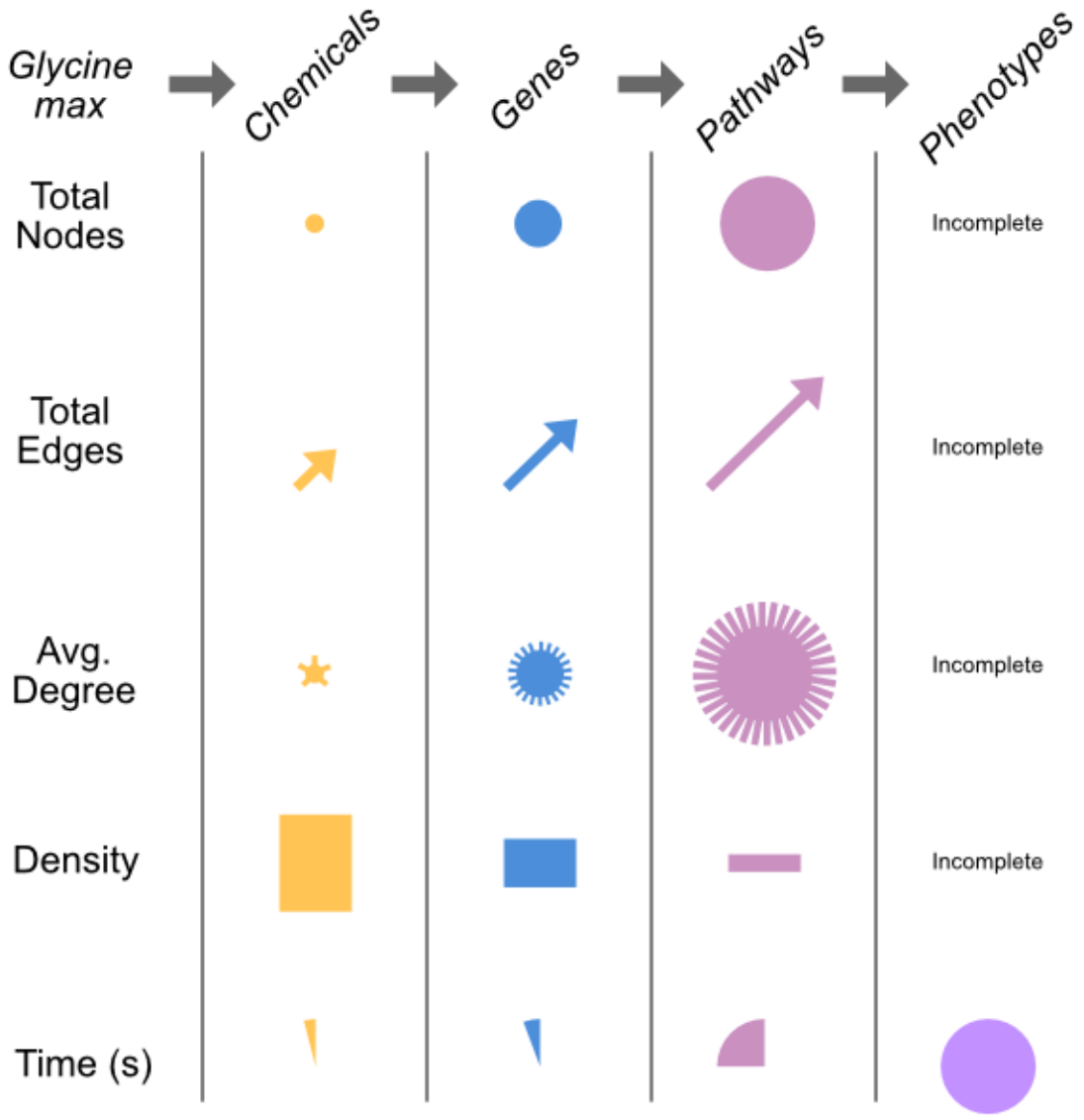

Figure 1.2: A query was run on the old ABCkb from Soybean (*Glycine max*) to Chemicals, Genes, Pathways, and Phenotypes on a 1 TB memory cluster at UNC Charlotte. The time it took to produce the results was logged and the query to phenotypes maxed out after a day, returning no results.

[24]. There are several microbiome pipelines available with strengths and weaknesses alike.

1.4.1 Microbiome Pipelines

Researchers have many options when performing a gut microbiome study, most of which borrow from traditional genomic analyses. 16S sequencing, the most preva-

lent approach, quantifies a region of the 16S gene from bacterial taxa found in stool samples [25]. Shotgun Whole Metagenomic Sequencing (WMS) is becoming more common and attempts to reconstruct whole metagenomes from stool samples [26]. A challenge with WMS is assembling multiple small genomes with many fragments of random sizes. Another pipeline is metatranscriptomics which attempts to characterize all of the expressed genes within a sample [27]. This allows detection of active community members, and may provide insights on how diet affects microbiota gene expression.

1.4.1.1 16S Sequencing

16S sequencing targets ribosomal DNA of bacteria in a sample, is cost effective, works with most sample types, and provides taxonomic classification into the family level [28]. This is useful for detecting broad shifts in microbiota composition. The 16S gene contains nine hypervariable regions that can be amplified through simple Polymerase Chain Reaction (PCR) methods, however the v4 region is often the target of these analyses [29]. Primers are generally universally applicable, as the 16S gene is highly conserved which further decreases the costs of microbiome analyses. In addition, Illumina sequencers are frequently chosen for sequencing amplified 16S regions. However, 16S sequencing does not separate bacterial strains well, which is necessary for a more detailed look at microbiota composition [30].

1.4.1.2 Whole Metagenome Sequencing

To quantify microbiome composition at a species-strain level, Whole Metagenome Sequencing (WMS), also called shotgun sequencing, is used to capture all of the genomic data within a sample [30]. This method utilizes any high-throughput sequencer with Illumina or Pac-Bio being the most common. Sequenced WMS reads are mapped to a reference database to remove host contamination and identify strains and species present within the sample [31]. Abundances are estimated from mapped reads that

align to bacterial reference genomes in a greater resolution than 16S sequencing [30]. Limitations of shotgun sequencing include the generation of random DNA sequences that are either too small to align with any confidence, or too low in quality [31]. Different versions of WMS attempt to capitalize on the resolution of this method while adding additional strengths. Phase Genomics adds a Hi-C method in conjunction with WMS to tagment DNA sequences which facilitates genome assembly after sequencing [32]. This tagmentation method aids in genome fragment reconstruction by linking sequences that are close in proximity. Though sequencing costs have decreased in recent years, WMS is still an expensive method for a large volume of samples [30].

1.4.2 Current State

Associations between specific gut microbiota and human phenotypic outcomes have been established from these different methods. Increases in bifidobacteria in the human gut are associated with reductions in colorectal cancer [33]. Clear links between *Helicobacter pylori* and stomach ulcers, which often leads to gastric cancer, have also been established [34]. Given the discovery of gut disbiosis, an increase in pathogenic bacteria, opportunistic bacteria, or decreases in beneficial bacteria, determining optimal gut microbial composition was a goal of the human microbiome project [35]. Controversially, no one specific microbiota composition has been identified to provide favorable phenotypic outcomes over another. Increased abundances of beneficial microbiota, called probiotics, have been found to provide benefits to the host which has led to an increasing interest in the discovery of interaction with dietary components [36]. Strong interacting components driving microbiota variance include diet, host gender, environment, geographical location, and ethnicity [37].

Some associations between phytochemicals and these beneficial gut microbiota have been established which has led to the development of prebiotics. These are specific nondigestible nutrients that are selectively fermented to facilitate the growth of probiotic bacteria and can be supplied by diet[38]. For example, members of the genus

Bifidobacterium commonly supplied in probiotic supplements are benefited by non-digestible oligosaccharides that can be found in oats[39]. Both Avena sativa and Bifidobacteria have been linked to a reduction in colon cancer[38, 40]. This raises interesting questions for nutrition researchers. Are the phytochemical associations produced from in vitro experiments fully representing the health benefits of plants? How does the microbiome contribute to human health and what are the associations from the foods we eat to the bacteria in our gut, and how does this all correspond to human health? Experimental designs that exclude microbiota assays will simply miss the full picture of health and the contributing factors. Therefore, providing researchers with a tool to represent the associative effects of gut microbiota is crucial to developing hypotheses and representing the overall contribution of diet on human health.

1.4.3 Microbiome Resources

Mentioned previously, the Human Microbiome Project elucidated baseline species present within the gut and other microbiomes [35]. There were two stages to the project and all of the data is available for researchers to use. There are some reference databases publicly available for mapping assembled metagenomic reads and taxonomic identification through a user-friendly portal [41]. There are currently two databases that link microbiota to human phenotypes, Disbiome and gutMDisorder [42, 43]. Disbiome, originally published in 2018, last updated in 2020 was produced from researchers at Ghent University in Belgium. Associations were generated by manually curating connections from microbiota to their effect (increase or decrease) on a disease. In comparison, gutMDisorder was originally made available online in 2019, with no indication of a date last updated, from researchers at the Harbin Medical University in China. These relationships between microbiota and food are also manually curated and specify the host organism. Generating connections through manual curation, as previously mentioned, is a difficult and time and resource con-

suming task. Clearly, a better solution is necessary.

1.5 Data

The life sciences fields have generated massive quantities of data for both storage and analysis [44]. Different types of data collections provide researchers with necessary components to compare data sets or generate hypotheses. Repositories that contain raw data provide researchers with data useful to test new analytic pipelines or compare to new data. In contrast, analyzed data repositories allow researchers to query what is currently known. There are three primary types of analyzed data repositories: databases, ontologies, and structured vocabularies.

1.5.1 Databases

Many databases exist, spanning various parts of life sciences domains [11]. Some databases contain raw data, others contain analyzed data. Some databases are made with Relational Database Management systems, others in Graph Database systems. Some are manually populated, others are populated automatically. Clearly, there is no "one size fits all" solution to database solutions. Though biological databases are not an exact science there are general recommended guidelines to follow [45]. Before developing a biological database, there are important questions to answer: "What questions are being answered," "How will the data be accessed," and "What is the data model."

1.5.1.1 Criteria for database developers

Developers must determine the research questions before developing a database to get a clear direction. The answer determines the data sources necessary to populate the database, and gives researchers a starting point to decide if the data exists already, or needs to be generated through experimental procedures. Scope creep will occur without a clear direction, which yields production delays or systems that are unusable [45]. Though other questions may arise from exploring the data, and the

database, it is crucial to keep to the original goal.

The question of how the database will be accessed is answered by understanding for whom the database is developed. If the database is intended for other developers with internet access, an instance running on a cloud web-server platform accessed remotely is likely useful. This method may employ the use of docker to containerize the database and add data protections where necessary. In contrast, if the database is developed for individuals with limited computer science experience, an interface with pre-programmed or "fill-in-the-blank" queries may be required. Other considerations include: limited internet access, which may require a database hosted on a workstation, restricted areas of the database, who will be contributing to the database, and the data model. Determining the best database provider will be aided by all of these considerations.

Additional criteria for developers stem from the data being used within the database and the copyright claims which originators may hold. Not all public resources are available for extraction and use in a separate resource without either proper attribution or appropriate licenses. Databases with patient data require special care and considerations which include data-masking, to remove personal identifiers, limiting access, and hard drive encryption.

1.5.1.2 Database Management Systems

Relational Database Management Systems (RDBMS) use Structured Query Language (SQL) as the management language which can be used to retrieve or alter data within the database. Many platforms have been developed to use version specific SQL with the most common being: SQL Server, Oracle, MySQL, PostgreSQL, and SQLite. Data within RDBMS are organized into tables with columns containing attributes and rows with values. Relational algebra can be used to connect tables and extract useful relationships. The most used function of SQL is the join which operates on tables with shared attributes. While RDBMS are useful for modeling

relational data, they are not well suited for modeling graphical networks. In cases where RDBMS are utilized to store graph data, linking tables are created to join multiple types of data and simulate a graphical network. As the network grows more complex, the number of joins required grows further which increases query time complexity. As a general rule, for each relationship, two joins are required. This can be mitigated by creating specific linking tables to bypass intermediate steps and reduce the number of joins. These tables must be recreated any time a table within the network is updated, which requires time and computational resources dependent on database complexity and size.

Graph Database Management Systems GDBMS use various forms of Query Languages to retrieve and alter data within the database. One of the earliest forms of graph databases was in Resource Description Framework (RDF) written in XML and navigated using SparQL. Commonly, RDF is used to describe web pages and knowledge management systems. Popular GDBMS solutions are: Neo4j, RedisGraph, TerminusDB, and AllegroGraph. Additionally, both Oracle database and SQL Server provide enterprise level products to model graph relationships. Factors that affect performance will be related to both the GDBMS solution used, and the data within. Graph traversal is an $O(N + E)$ computational problem where N is the number of nodes and E, the number of edges. Thus, traversing a dense graph, a graph in which the number of edges is significantly greater than the number of nodes, will require significant computational resources and time.

1.5.2 Ontologies

The second form of data used is ontologies, which borrow from the branch of philosophy and organize concepts into parent-child, hierarchical relationships. This is the most strict type of data collection, as concepts can only exist in one place in the tree. There are four primary features of ontologies: class and relationship identifiers, a domain vocabulary, metadata and descriptions, and finally formal definition and

axioms [46].

1.5.2.1 Class and Relationship Identifiers

Class and relationship identifiers allow for each node in the ontology to have a unique ID that allows reference by other data sources. This allows connections to be made across databases, which is useful when researchers desire to probe relationships over multiple sources.

1.5.2.2 Domain Vocabulary and Metadata

Ontologies are often restricted to a specific domain, or part of, which requires a specific vocabulary for each ontology in the form of labels. For instance, the Disease Ontology categorizes human diseases with phenotypic characteristics where the Human Phenotype Ontology provides phenotypic abnormalities not found within the Disease Ontology [47, 48]. A query for Colorectal cancer from the Disease Ontology returns a node representing colorectal cancer with a definition and cross-references to other databases and ontologies. In contrast, a query from the Human Phenotype Ontology returns inheritance patterns and specific neoplasms related to colorectal cancer with related genes. Similar labels or categories across ontologies are: name, synonyms, and definition. The metadata and descriptions of an ontology represent the form and style of data held within each class.

1.5.2.3 Formal Definition and Axioms

The fourth and final feature of an ontology, is the formal definition and axioms which allow it to be analyzed or represented using graph networks. Typical formats of ontologies include Web Ontology Language (OWL), OBO flat-file, and in some cases JavaScript Object Notation. Given the domain restrictions of ontologies, troubles arise when researchers refer to colon cancer and colorectal cancer interchangeably in literature, two distinct classes within the Disease Ontology. The difficulty mapping which of the two results is being discussed can be aided by using structured

vocabularies.

1.5.3 Structured Vocabularies

Medical Subject Headings (MeSH), a structured vocabulary maintained by the NCBI, aides researchers with literature searches to match as many terms as possible [49]. Where ontologies are more restrictive, structured vocabularies are more relaxed in the positioning of classes and relationships. The class Colonic Neoplasms, or colon cancer, exists in multiple tree locations within MeSH where Disease Ontology contains a singular tree location. Issues arise when graphically representing the data or mass assignment of labels based on a tree location. For example the vocabulary term "Metabolic Side Effects of Drugs and Substances" (D065606) is listed under two root terms, "Chemically-Induced Disorders" and "Chemical Actions and Uses" [49]. Both structured vocabularies and ontologies are useful for literature searches and Natural Language Processing queries.

1.6 Natural Language Processing

Perhaps, one of the most difficult tasks in text-mining is extracting associations from scientific literature. Natural language is flexible, which adds considerable difficulty to process. Simple concepts to extract from natural language are: subjects, predicates and objects. Often these are referred to as a triple. In a graph model, a subject and object are nodes, and the relationships are the edges that connect them. A challenge for NLP tools is identifying the directionality of the triple. Due to the nature of literature, position is not an adequate predictor of directionality. The phrase "Bacillus subtilis affects diabetes" confers a different relationship than "Bacillus subtilis is affected by diabetes." Other challenges include subject/object identification, multi-sentence parsing and negative relationship identification. Many NLP methods and tools exist that can aid researchers with extracting these triples, both open-source and enterprise. Two tasks any NLP researcher will utilize are Information Retrieval

and Information Extraction (Figure 1.3).

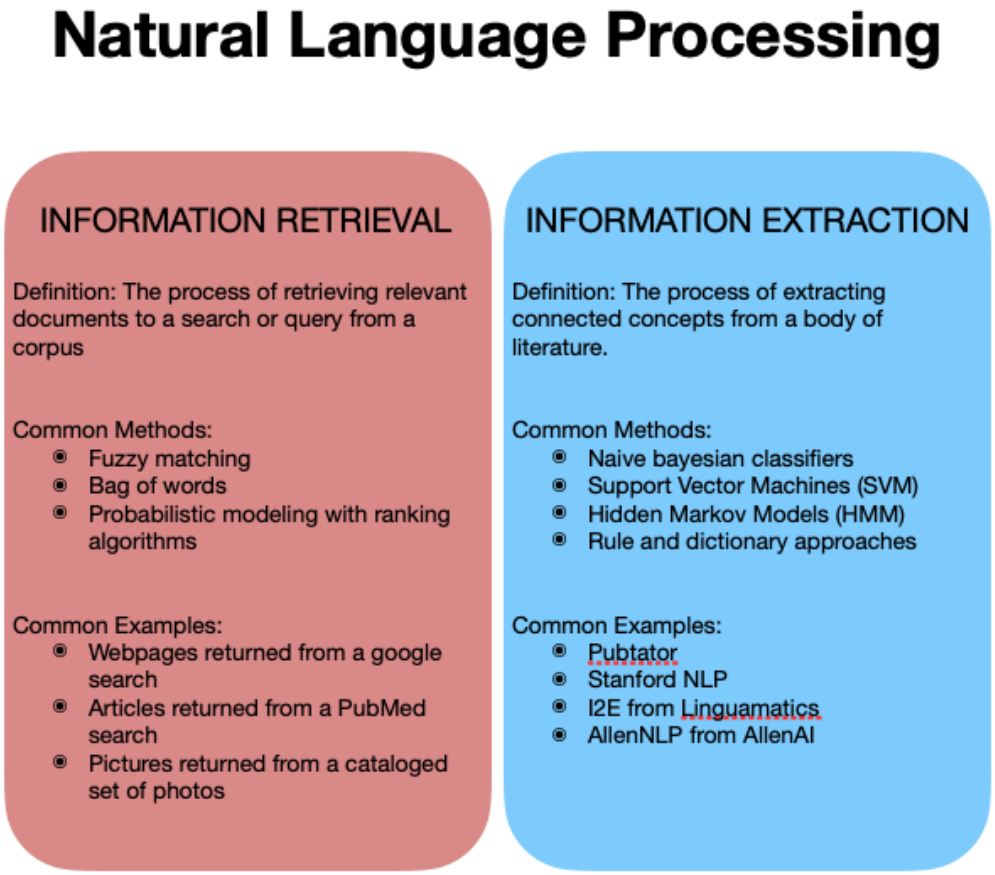

Figure 1.3: The two primary tasks of natural language processing are Information Retrieval and Extraction.

1.6.1 Information Retrieval

The task of Information Retrieval is summarized as the process of returning relevant documents based on an individual request. This task is often analogous to a google search, which returns web-pages relevant to the entered search query. Pubmed papers in the MEDLINE repository are tagged with MeSH terms that map to scientific entities and related terms. This can help narrow down paper subsets on specific topic. A query of "Bifidobacterium" quickly returns over 9,000 results where querying the MEDLINE index with the related MeSH term reduces output papers. MeSH also

gives researchers the ability to add subheadings with primary query terms to further limit resulting documents. Ranking metrics can be employed to order results by importance or relevance to the search query.

1.6.1.1 Ranking Algorithms

One popular algorithm utilized by Google and Twitter is called PageRank (PR), which attempts to highlight more relevant pages with a search query[50]. This works by weighting the rank of a page based on the pages that link to it using a random walk to simulate a web surfer visiting pages. The simplified algorithm is defined in equation 1.1 with the Page Rank PR for the selected page denoted u, PR for each page linking to u denoted v, the set of links connecting to the page u denoted $B(u)$, and the sum total links on page v denoted $L(v)$. There is also a damping factor added to the equation to approximate the probability of randomly selecting another page within the subset of pages. This algorithm works well for ranking pages, ordering nodes within a graph, and in biological applications for ranking protein networks[51, 52].

$$PR(u) = \sum_{v \in B(u)} \frac{PR(v)}{L(v)} \tag{1.1}$$

Another algorithm utilized for ranking web-pages is the Hyperlink-Induced Topic Search HITS algorithm which calculates an authority score and a hub score [53]. The authority score for a given link is calculated by summing the hub scores of pages that point to it. In contrast, each hub score is calculated by summing authority scores of pages that point to that hub. This algorithm is run at query time, which increases the time required for the database to return results.

1.6.2 Information Extraction

Literature searches often return thousands of documents, an overwhelming number for a single researcher to effectively probe before beginning research. Information extraction techniques can be employed to help researchers examine a large corpus

of documents and derive meaning or extract connected concepts. Popular methods utilize Machine Learning methods, rule or dictionary based approaches, or hybrid approaches.

1.6.2.1 Machine Learning Approach

ML or statistical methods mathematically predict the entities within and the relationships they share. The problem is treated as a classification problem. Three primary techniques used are: naive bayes, Support Vector Machines (SVM), and Hidden Markov Models.

Naive bayesian classifiers, often called the bag of words approach, are used to classify an input text with a class from a fixed set of classes. This technique is beneficial for classifying texts as belonging to a specific topic for example, distinguishing positive from negative. A pre-classified subset of training texts is required for the bayesian approach. Unfamiliar or unseen terms in the training set will be ranked with a zero, which require use of an additive algorithm to remove zero probabilities. This approach assumes word positions within the text are irrelevant to final scores, thus the informal bag-of-words name. Additionally, it is assumed that probabilities of terms within the text are independent. Both assumptions are not always true which leads to problems with sentences containing opposite classes.

Support Vector Machines are a supervised learning approach and are commonly used in NLP. They determine decision boundaries between vectors for text classification. This requires natural language text to be transformed into a vector. Standard practice is to select a maximum number of features in a training dataset, and vectorize Term Frequency, Inverse Document Frequency scores. This method builds a target vocabulary that other texts can be weighted against to determine if the threshold is high enough to cross the decision boundary. SVMs excel in classifying high-dimensional data but require labelled and balanced training datasets to perform effectively, which is a time-consuming task to manually develop when not available.

An unsupervised approach can aid when supervised approaches are not optimal.

Hidden Markov Models are a common unsupervised learning method that observes a predefined number of states, or observations, and computes sequence probabilities of hidden information. This method utilizes the viterbi algorithm to decode these hidden sequence probabilities. Assigning parts of speech tags given a sentence, or input text with observed words, is a common NLP use-case. This is beneficial when limited training data is available or with specialized vocabularies. However, HMMs cannot provide assertions within text and must be combined with another method such as a rule and dictionary approach.

1.6.2.2 Rule and Dictionary Approach

Dictionary approaches group terms into a singular value from curated synonym lists [54]. These dictionaries can be applied on a corpus to highlight key terms within a text and classify a text based on these terms or aid researchers view groupings of terms. Pubtator is a common example that highlights chemicals, genes, pathways and phenotypes in published abstracts from the MEDLINE index [55]. Rule based approaches depend on user-chosen rules to classify texts or extract assertions. Extracting all nouns with verbs or verb phrases sandwiched in-between from assigned POS tags in a corpus, is an example of this approach. This can be informative for extracting assertions, however POS taggers must be carefully chosen based on the corpus type and are not the best choice for scientific literature.

1.6.2.3 I2E

Few NLP tools are able to extract assertions from scientific literature. Open-source tools trained on general datasets like the brown corpus work well with twitter and other sources of general knowledge but are inadequate for use on scientific literature or any body of knowledge with a highly-specialized vocabulary. I2E from Linguamatics (https://www.linguamatics.com/products/i2e) has a long history of application

in the life-sciences domain and focuses on extracting assertions using a rule and dictionary approach. Assertions are extracted through queries from indexed datasets. Queries are a model of how each individual assertion is represented in the literature. These are useful on both labeled and unlabeled datasets. With labeled datasets, precision, recall, and F measures are calculated to indicate accuracy of each query. Unlabelled datasets require an iterative approach to examine the extracted assertions and refine queries. I2E provides model queries trained on MEDLINE abstracts that extract "class affects class" assertions. A "class" in I2E is a dictionary object that is used by a query to locate terms in the literature under that object. For example, all instances of *Vaccinium corymbosum* are found through the NCBI taxonomy node, which contains synonyms American blueberry, highbush blueberry, and the preferred term Vaccinium corymbosum. Users will select all terms under a specific node to identify all leaf terms and synonyms from a specific dictionary, if desired. Class matches are provided confidence values using a proprietary disambiguation algorithm to filter out ambiguous matches. Benefits of using I2E include, the Graphical User Interface that is used to develop queries, quickly returned results from indexed datasets, use of many ontologies without the need to convert from various file formats, and the ability to link across separate data sources in published literature. Drawbacks include the cost to purchase a license, and the learning curve to manage indexes and ontologies into I2E specific language. Often, I2E users opt to have Linguamatics handle index creation and ontology or class parsing into I2E specific formats.

1.7 Aims

Aim 1: Develop a knowledgebase on molecular mechanisms from plants to human health by connecting siloed, sparse data from public resources and text-mined assertions.

> Given the rise in noncommunicable diseases like heart disease and diabetes, an integrated resource is necessary. This resource should combine

structured and unstructured resources to connect plants to human health in a graph database. Another feature is an interface to aid users not skilled in computational biology, browse resource contents. All of this should be contained in an easy to use, easy to install, system that requires little user input. This aim will further nutrition research towards personalized solutions.

Aim 2: Incorporate microbiota effects on diet and human health into the knowledgebase to identify key factors driving IBS

The compound effects of microbiota on diet are a required addition to the integrated resource from Aim 1. These effects should include both the effect of microbiota on phytochemicals, and the effect of phytochemicals on microbiota. Given that microbiota data is not as prevalent in abstracts, associations will be extracted from full-text articles to maximize discovery of hidden relationships. Microbiota associations will also be extracted from structured resources where available. This will provide nutritional researchers a tool to examine dietary effects on microbiota and human health.

Aim 3: Apply the knowledgebase and the methods of text-mining literature to elucidate dietary effects on SARS-CoV-2 infections.

The SARS-CoV-2 pandemic quickly enveloped researchers, which yielded a multitude of full-text preprint publications contained in a Coronavirus specific literature set. These publications will be indexed in I2E and mined to extract the Coronavirus and associated human genes. These associations will be inserted into the knowledgebase and will be analyzed to functionally group affected human genes and pathways. Finally, plant

phytochemical connections linked to the molecular mechanisms of long-COVID from coronavirus infections will be analyzed in the graph network.

CHAPTER 2: THE ALIMENT TO BODILY CONDITION KNOWLEDGEBASE (ABCKB): A DATABASE CONNECTING PLANTS AND HUMAN HEALTH

Introduction

The growth of obesity worldwide correlates strongly with overconsumption of processed foods [56].This has contributed to an increase in chronic diet-related diseases like type 2 diabetes (T2D), heart disease, and some cancers [57]. Exercise and diets high in fruit, vegetables, whole grains, and nuts have been linked with healthier outcomes and reduce the risk of developing these diseases [58]. Unfortunately, the specific mechanisms driving these associations are poorly understood. The Plant Pathways Elucidation Project (P2EP) was a collaboration started to uncover the mechanisms between plant-pathway products and human health [59]. Three questions drove this collaboration: "What do plants make," "How do they make them," and "What is their effect on human health?" The ABCkb was developed to capture the information required to answer these questions and provide researchers with a tool to build informed, nutritive hypotheses with molecular mechanisms as the linking factor between dietary plants and human health.

These questions closely align to the recently released "2020-2030 Strategic Plan for NIH Nutrition Research.". This plan contains 4 strategic goals for further study to move closer to a precision nutrition approach including foundational research into "What do we eat and how does it affect us?" as well as understanding "How can we improve the use of food as medicine?" A cornerstone for answering these questions and the questions of the P2EP collaboration is an understanding of the mechanism of action of how our diet affects our health.

However, manually capturing this information is a difficult, time-consuming task

due to scattered bodies of scientific knowledge. Currently available resources contain partial information to answer these questions, but they do not address mechanism of action. For example, the Comparative Toxicogenomics Database (CTD) connects chemicals to human health through human genes by manually curating associations between chemicals, genes, pathways and phenotypes but excludes nutritional data [15]. Specialized nutritional databases like FooDB (https://foodb.ca) and Phenol-Explorer aid researchers in estimating quantity of phytochemical content, but lack human phenotypic information [16]. NutriChem was developed to bridge the gap between plant-based nutrition and human disease through the chemicals contained in those plants, but does not contain gene-chemical associations, a key part of the driving molecular mechanisms between diet and human health [13]. While a small proportion of assertions are in available databases, others are hidden in published research and can only be extracted through extensive reading or by natural language processing (NLP) the literature. Given the rise in diet-related diseases, and the pursuit of personalized nutrition, an integrated resource to develop nutritive hypotheses is necessary.

Main Text

We have developed the Aliment to Bodily Condition Knowledgebase (ABCkb) to address the gap of connecting plant compounds to human indications through their mechanism of action. The ABCkb integrates multiple resources for building informed hypotheses with molecular mechanisms as the linking factor between dietary plants and human health. To accomplish this, the ABCkb uses both structured and unstructured data sources (Fig. 2.1). The structured resources are publicly accessible, curated databases and the unstructured data is in the form of Medline abstracts. Since this data, composed of entities and relationships or nodes and edges, composes a graphical network, we extracted, transformed, and then loaded into a Neo4j graph database. To help users begin discovering these nutritive connections, the knowledge-

base is available on GitHub and a simplified online web interface.

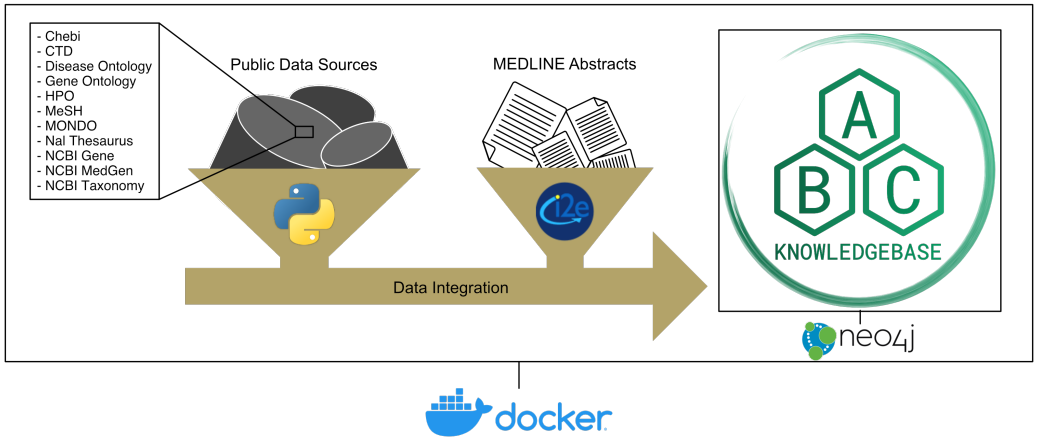

Figure 2.1: The architectural diagram of our Knowledgebase shows the various tools and resources utilized to generate the database.

Structured resource collection

Structured data from 11 resources (Fig. 2.2) produce five major node types (Plant, Chemical, Gene, Pathway, Phenotype) in a Neo4j graph database. Connections, or edges between these nodes are provided by both structured data, and unstructured MEDLINE Abstracts through NLP. The ABCkb utilizes three types of structured data sources: ontologies, structured vocabularies, and databases.

	Node label					Relationship	
Source	Plant	Chemical	Gene	Pathway	Phenotype	External	Internal
Chemical Entities of Biological Interest		✓					✓
Comparative Toxicogenomics Database						✓	
Disease Ontology					✓	✓	✓
Gene Ontology				✓		✓	✓
Human Phenotype Ontology					✓	✓	✓
Medical Subject Headings	✓	✓		✓	✓		✓
MONDO					✓	✓	✓
NAL Thesaurus	✓	✓			✓		✓
NCBI Gene			✓			✓	
NCBi MedGen						✓	
NCBI Taxonomy	✓						✓

Figure 2.2: Data from each source is transformed into one of the 5 labels and may provide external and internal references to nodes within the knowledgebase. The CTD provides manually curated references between labels with no original node labels

Ontologies and Structured Vocabularies

The ontologies and structured vocabularies create well-controlled edges between chemicals, pathways, and phenotypes. The Chemical Entities of Biological Interest provide chemical nodes and semantic connections (edges) between chemicals [60]. Genes are grouped into pathways from the Gene Ontology resource [61, 62]. Human phenotypes are represented from three sources. The Disease Ontology categorizes human diseases with phenotypic characteristics [47]. The Human Phenotype Ontology provides phenotypic abnormalities not found within the Disease Ontology which allows researchers to focus on specific phenotypic symptoms and the associated molecular mechanisms [48]. Finally, the MONDO Disease Ontology was used to collapse similar phenotype nodes from multiple sources using their source identifiers [63]. The Medical Subject Headings resource provided nodes and connections for all major labels with the exception of Genes [49]. Additional plant, chemical, and phenotype nodes were extracted from the National Agricultural Library Thesaurus [64]. Terms from different ontologies or vocabularies with the same identifiers are collapsed into the same node. All other nodes are left separate to retain their hierarchical relationships.

Databases

Several databases were utilized to increase molecular mechanisms from plant to human disease in the ABCkb. The Comparative Toxicogenomics Database added over 7.4 million manually curated edges between chemicals, genes, pathways, and phenotype nodes [15]. We utilized three public databases from The National Center for Biotechnology Information. All plants under the Embryophyta clade from the NCBI Taxonomy database produced plant nodes and phylogenetic relationships between plants [65, 66]. The Gene database provided gene names, types, and synonyms [67]. Finally, additional edges were added utilizing NCBI gene nodes and MONDO phe-

notypes were extracted from the NCBI MedGen database [68]. The compendium of structured data sources provide many of the node and edges connecting plants to disease. However, unstructured literature contains informative relationships not contained within these sources, leaving many gaps in our understanding.

Unstructured NLP collection

To uncover relationships in literature, elucidate molecular mechanisms, and answer the three questions of the P2EP, we mined the literature using Linguamatics' I2E NLP text mining platform (https://www.linguamatics.com/products/i2e). This platform utilizes ontologies and structured vocabularies to transform unstructured text into structured assertions (nodes and edges).

Natural Language Processing of MEDLINE Abstracts

The I2E platform employs a graphical user interface for NLP query development, where each query extracts a set of subjects, objects, and predicates, or relationships from user-specified ontologies and structured vocabularies. From published abstracts, and titles extracted from MEDLINE in May, 2019, NLP queries were developed with I2E for each of the 4 steps (plant to chemical, chemical to gene, gene to pathway, pathway to phenotype) with an additional query from genes to phenotypes. All I2E assertions generated are provided to users of the ABCkb as source files and are parsed when the graph database is built.

Statistics and application

Extracted public data sources generated over 957,000 nodes with over 11 million edge relationships. NLP results from I2E queries make up 1.26 million of the overall relationship count, of which 1.25 million relationships were novel, not from structured public data sources. Figure 2.3 gives a visual presentation of (a) the relative number of each node type and their source, (b) the edge relationships from each source and (c) the relative comparison of edge relationship types between each type of node.

This collection of nodes and edge relationships forming semantic triples, naturally forms a biological network of knowledge that is best stored in a graph database like Neo4j. Chaining these triples together in the ABCkb highlights connections between dietary plants and human phenotypes that would otherwise go unseen if left in their original sources, particularly unstructured literature sources. The intention of the knowledgebase is for information in the network to flow from plants to phenotypes/disease indications, however, assertions are maintained in both directions, which allows for query flexibility of relationships between any nodes. Start and end node types are not enforced which allows queries from any point, to any point. All associations are kept along with references to the original source allowing the user to evaluate potential inconsistencies using the original evidence. To explore the database and discover connections, users have two choices. One, use the online interface (available at https://abckb.charlotte.edu). Otherwise, download from GitHub and build the database on a local machine which can then be queried in the Neo4j interface, or on the command line. A prebuilt data folder with the neo4j database is also available [12].

The provided user-friendly interface aids users unfamiliar with Neo4j query language (Cypher) to browse the contents within and examine nutritive connections (Fig. 2.4). On the home page, users are provided a search box to enter in a search term. This scans the nodes in the knowledgebase and returns results ranked by similarity to search term. Users can select nodes and continue to build a query to any end point within the knowledgebase (plant, chemical, gene, pathway, or phenotype). Running the query scans the database for all paths to the selected end point and returns them to the user, which are available to download. Additionally, a Cypher query is available to users that can be used in the built in Neo4j interface or the terminal for further exploration.

Oat and Type 2 Diabetes

To demonstrate how the ABCkb connects dietary plants to separate human indications through molecular mechanisms, a graph was created in the ABCkb, through the Neo4j browser, depicting the diet-disease network between *Avena sativa*, T2D, and heart failure (Fig 2.5). Connections from the CTD indicate genes commonly associated with cholesterol and heart failure. However, text-mining indicates that consumption of oats affects cholesterol levels in the body, which is associated with the gene HSD11B1 that affects lipid metabolic processes with both positive and negative impacts on the incidence of T2D. These relationships are due to the presence of beta-glucan in oat grains. Consumption of beta-glucan-containing oat can help lower LDL cholesterol [69]. The cholesterol lowering effects of oat can also be attributed to the presence of certain lipids and proteins [70]. The proteins in oat with low lysine-arginine and methionine-glycine ratios contribute to lower total cholesterol and LDL cholesterol levels. Hypocholesterolemic properties of oat cannot simply be attributed to one factor, but a combination of many, including oleic acid, vitamin E, and plant sterols [70].

T2D patients frequently have abnormal levels of many different lipids, as well as abnormal qualities to these lipids, for example, T2D patients experience normal or slightly elevated LDL cholesterol with increased LDL oxidation and glycation [71]. Dyslipidemia in T2D patients is associated with cardiovascular disease [72, 73]. This creates an elevated risk for cardiovascular diseases including atherosclerosis, and dislipidemia may play a role in these risks [73]. In the graph, HSD11B1 is the human gene connecting this relationship. HSD11B1 expression is increased in adipose tissues of obese individuals [74]. Dysregulation of HSD11B1 is associated with an imbalance of glucocorticoid in adipose tissues, glucose imbalance, and visceral fat accumulation [75]. These factors contribute to metabolic syndrome, which puts patients at a higher risk for cardiac diseases [76]. Various SNPs in HSD11B1 have associations

with T2D, metabolic syndrome, and hypertension [77, 78, 79, 80].

Due to the established relationship between oat beta-glucans, cholesterol, and weight, the connection to T2D is logical [69, 75]. Decreased weight, specifically visceral fat in the abdomen, would result in reduced expression of HSD11B1, which would improve regulation of cortisol. Further examination of the oat - cholesterol - HSD11B1 relationship could be very informative to both patients and doctors in making more informed dietary choices and reducing the risk of developing T2D. This example demonstrates the ABCkb ability to connect seemingly separate conditions through the molecular mechanistic links within.

Discussion

The ABCkb integrates structured and unstructured resources in a network that connects plants to human disease through molecular mechanisms. This reduces the time required to manually connect these links through each individual resource. Additionally, knowledge discovery is aided by the development of a user-friendly interface. All of these components provide precision nutrition a path to better understand the mechanisms behind diet-related conditions. The ABCkb is available from the interface (https://abckb.charlotte.edu).

Limitations

- Microbiota contributions to diet and human disease. Bacteria within the gut are known to affect disease both through the production of metabolites and the conversion of plant phytochemicals. In addition, gut bacteria are affected by diet. Future implementations of the ABCkb will contain microbiota associations to enhance precision nutrition hypotheses.

- Mining abstracts versus full text. Abstracts contain valuable associations, however associations full text articles would provide a greater number of associations.

- Incorporating genomic data. Precision nutrition hypotheses and treatment plans will depend on patient genomic data, to provide optimal dietary solutions for each individual. Future versions of the ABCkb should incorporate human genomic data.

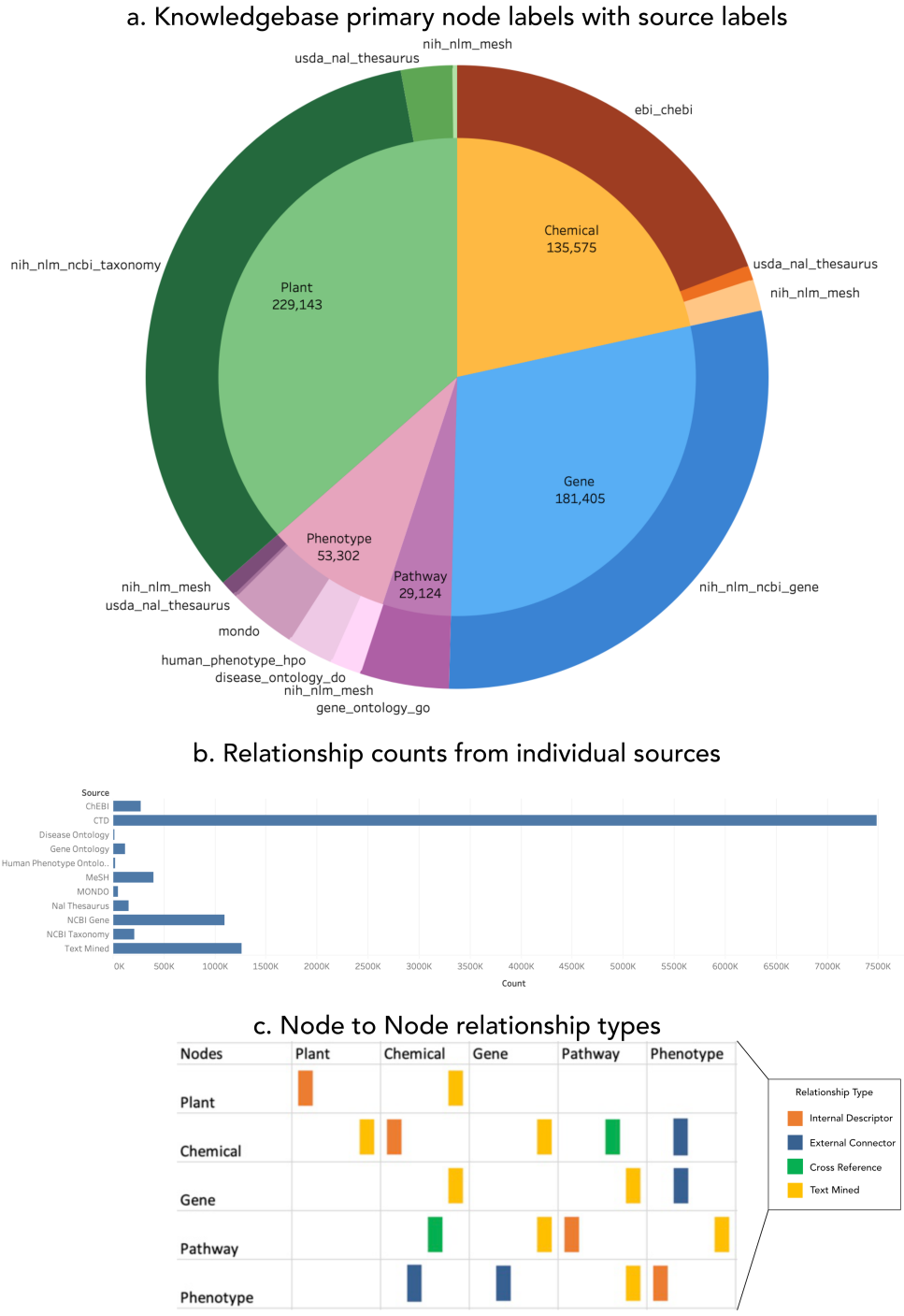

Figure 2.3: a. The pie chart shows primary labels indicated by color with named secondary (source) labels, shaded and sized by proportion of total nodes in the knowledgebase. b. The sum of relationship counts for each source is indicated by the bar chart. c. Relative relationship counts indicated from node-node in rows, columns in a bar chart in order by type (Internal Descriptor, External Connector, Cross Reference, and Text Mined)

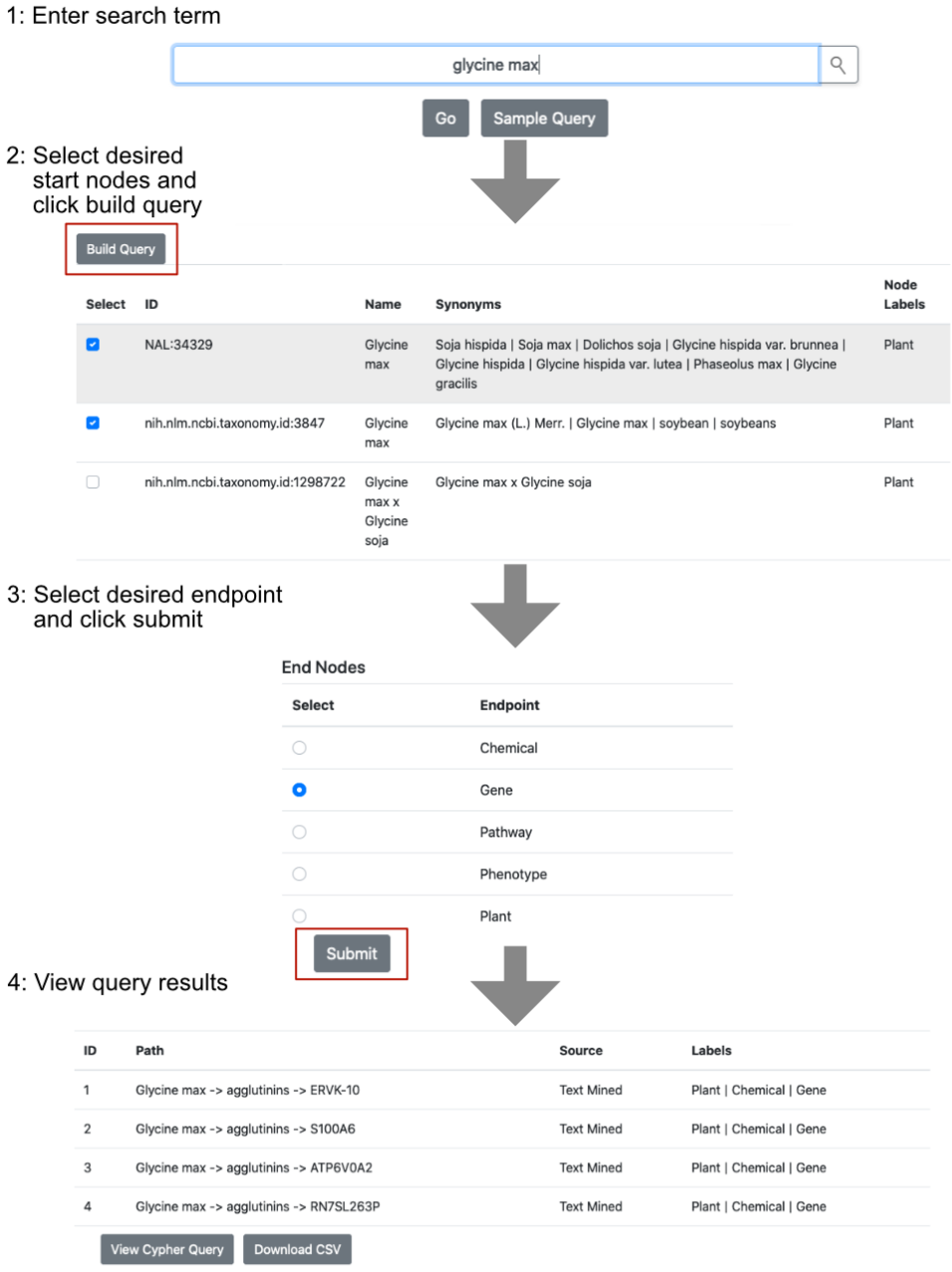

Figure 2.4: There are 4 primary steps to browsing using the provided interface. Once the query endpoint is selected and the user clicks submit, they have the option of downloading all results as a csv, or viewing the Cypher query.

Figure 2.5: This meta-path highlights the connectivity between oats, diabetes, and heart failure through the gene HSD11B1 from the ABCkb.

CHAPTER 3: ADDITION OF MICROBIOTA DATA TO ABCKB FROM FULL TEXT LITERATURE

Modeling diet to disease includes more than the phytochemicals from plants and the human genes they affect upon ingestion. A community of micro-organisms exists within the digestive system of humans that maintains a symbiotic relationship, though not always beneficial for the host. Two primary diseases associated with this community, Irritable Bowel Syndrome (IBS) and Inflammatory Bowel Disease (IBD), are both increasing in prevalence and diet related [81]. The term microbiota refers to the collection of bacterial species within the gut, where the microbiome refers to the collective genomic data present in the gut [19]. Specific microbiota have been implicated in various digestive disorders, and the impact of microbiota on diet and compounds descending the alimentary canal has been a topic of recent interest.

3.1 History of Microbiome Research

Historically, an abiotic existence was argued for by the majority of biologists as microbiota were thought to be solely drivers of diseases. This thought prevailed through the 1950's until researchers discovered bacteria present within healthy mice grown in sterile conditions [20]. This, along with improvements in sequencing technologies, allowed for the identification of bacteria present within the human gut and the discovery that these bacteria exist in communities [19]. The shift in health research to include microbiota resulted in the human microbiome project that set out to elucidate the single microbiota formulation for optimal human health [35]. Controversially, no one, single formulation is found to be the best but rather a conglomeration of abundances provides protective benefits to the host. Increased abundances of opportunistic bac-

teria result in digestive disorders due to the fermentation of sugars within the gut and promotion of inflammatory molecules [82]. The goal of microbiome research is to maximize the abundances of good bacteria, or probiotics. The levels of bacteria are largely attributed to dietary components and compounds, called prebiotics.

3.1.1 Probiotics

Probiotics have risen in popularity in the 21st century. The widely accepted definition of probiotics from the WHO is that they are "live strains of strictly selected microorganisms which, when administered in adequate amounts, confer a health benefit on the host" [83]. Often probiotic strains are specifically developed for use in commercial products like supplements, fermented beverages, or yoghurts.

3.1.2 Prebiotics and diet

Prebiotic compounds encourage the growth of probiotic bacteria and are generally fermented by probiotic bacteria. Common examples of prebiotics are inulin, resistant starches, and dietary fibers [84]. These prebiotic compounds can be found in supplements, or via natural sources. Complications arise in patients with IBS and IBD that are advised to follow a low FODMaP (Fermentable Oligo, Di, Mono and Polysaccharide) diet, which reduces digestive symptoms along with probiotic bacteria [85]. As prebiotics impact probiotic abundances, so also does diet as a whole. Many studies have found that microbiota communities are largely affected by diet, stronger than genetic drivers often seen segregating by geographic regions [86, 87, 88]. This is likely due to the availability of various foods in regions and cultural differences. However regional cultural or otherwise defined, diet is largely constructed of chemical components whether naturally derived or synthetically produced.

Therefore, we set out to provide a knowledgebase through which the effects of diet on human health could be examined in conjunction with the various interactions between specific bacterial taxa within the gut, and the phytochemicals contained in

plant compounds.

3.2 Structured Data Sources

Few resources are available on the interaction between microbiota and human health. Mentioned previously, the Human Microbiome Project provides raw data to researchers, which can be used to compare to raw data from experimental conditions, or reanalyzed to drive new insights[35]. However, no associations are provided in a human readable format. Recently developed, the Disbiome database provides a queriable resource to catalogue the interaction between microbiota and disease [42]. No molecular mechanisms are provided with Disbiome. There are no databases that provide microbiota to diet, or dietary compounds, currently available. The NCBI Taxonomy will provide bacteria nodes which will be used in NLP queries to generate links between microbiota, diet, and human health.

3.3 Unstructured Data Sources

The majority of microbiota interactions are not held within structured data sources, but rather in published, unstructured research. Therefore, a method is necessary to extract and place those assertions into the ABCkb. Previously, we used I2E to extract assertions for every step in the pathway from MEDLINE abstracts. Recent research has highlighted the need to develop methods that can extract assertions from full text, as additional assertions are not contained within the abstract [89, 90]. The addition of microbiota data to the knowledgebase is a key component to determine how diet affects human health and move towards a personalized solution to nutrition.

3.4 ABCkb 2.0 Methods

Using the ABCkb as a starting point, the microbiome information integrates with the existing knowledgebase (Figure 3.3). The visualized schema shows the bi-directionality of the microbiota on plant compounds, along with general assertions provided from the ABCkb 1.0 linking diet and disease. The logic is simple. Often,

links in literature are provided, much like the disbiome resource, where a bacteria is linked with a disease. These links are generated through association studies and molecular mechanisms are simply unknown. The ABCkb 2.0 will provide both the general links (from bacteria to disease) and possible molecular mechanisms for those diseases (from genes to disease). Researchers will be able to use this knowledgebase to generate testable hypotheses about the unknown from what is known. However, unlike the first iteration of the ABCkb, we will be extracting microbiome assertions from full text articles.

3.4.1 Identifying Relevant Full-Text Articles

To identify a suitable subset of full-text articles for text mining, a query was developed that locates instances in PUBMED abstracts of microbiome, with related synonyms from MeSH, NAL Thesaurus, and NCBI Taxonomy, and one of either *Homo sapiens*, *Ratticus norvegicus*, and *Mus musculus* with their respective synonyms. The sources that provided these matches with synonyms were the NAL Thesaurus, MeSH, and the NCBI Taxonomy database. These three hosts were chosen for their relevance to human health and to remove any non-related biological hosts. This query identified over 30,000 unique articles for full-text extraction from abstracts in MEDLINE in May, 2019.

3.4.2 Extracting Full-Text Articles

Full text articles were extracted first through obtaining permission from the publishers. This process took approximately 6 months to complete. Permissions to extract and mine through 23,000 of the original 30,000 was acquired with the assistance of the UNC Charlotte Library from current and updated license agreements. An automated paper scraper was developed to pull pdf, xml, and txt files from PUBMED article identifiers. This tool is available from GitHub as a standalone resource.

3.4.3 Text Mining Full-Text Articles

Extracting assertions from articles was performed with two primary functions, entity recognition and relationship extraction. Entity recognition is the function of recognizing specific entities (Plant, Chemical, Gene, Pathway, Phenotype, Bacteria, Diet), which yield start and end nodes in the knowledgebase. I2E employs a user-specified dictionary based approach to entity recognition. We utilized a set of ontologies and structured vocabularies for this process (Figure 3.1). Each of these sources were used in I2E queries to extract the 7 entity types (nodes) provided in the knowledgebase.

Entity Recognition Sources

Source	Plant	Chemical	Gene	Pathway	Phenotype	Bacteria	Diet
Chemical Entities of Biological Interest		✓					
Disease Ontology					✓		
Gene Ontology				✓			
Human Phenotype Ontology					✓		
Medical Subject Headings	✓	✓		✓	✓	✓	✓
NAL Thesaurus	✓	✓			✓	✓	✓
NCBI Gene			✓				
NCBI Taxonomy	✓					✓	

Figure 3.1: There are a total of 8 sources used for entity recognition in I2E. These 8 sources provide all of the nodes in the Knowledgebase as indicated in the figure.

Bacterium in literature are often abbreviated for brevity which is a hurdle for dictionary based text mining tools to overcome given the dictionary, in many cases, does not contain the abbreviation. This is a result of multiple genera sharing the same species name, in some cases an indication of function (*Halolactibacillus halophilus* and *Halobacillus halophilus*).

The abbreviated bacterium identification hurdle was overcome by using, a combined process of text mining with the rule/dictionary method in I2E to identify all mentioned bacterium in the article, and a separate regular expressions query to identify abbreviated matches. Each abbreviated match in a single article is compared to the set of bacterium mentioned for each article to map back to the original node

identifier and thus extract the specific bacterium for each relationship. The additional specific entities were well-defined within the structured vocabularies used for text-mining and did not require additional mappings.

Once the entities were categorized, relationship extraction was the next step. I2E by Linguamatics provided 21 sentence models for extracting assertions, however, a specified predicate ontology was used and iteratively refined to remove false positive results. Text-mining through full-text articles inherently adds noise through additional text brought in and methods were employed to combat this. Reference matches, though desirable in specific text-mining cases, were not immediately useful to the goal of developing this knowledgebase, and a query was developed to negate matches within the references section.

The microbiome queries were developed through an iterative process both for entity recognition and relationship extraction (Figure 3.2). In the summer of 2018, we manually categorized 100 assertions for validation of I2E queries. I2E queries were run against the MEDLINE index and the full-text index and judged by measures of precision and accuracy. Manually we refined the queries until reaching a F-score threshold of 85% for entity recognition and F2-score threshold of 85%. The F2 score was chosen for a higher recall rate over precision given our focus of hypothesis generation, to maximize possible connections between entities.

3.4.4 Building the ABCkb 2.0

After extracting Nodes and relationships from public sources, node and relationship reduction steps were necessary. Duplicate nodes were limited by ensuring node identifiers appeared only once, and collapsing nodes with matching names. To ensure unique nodes present, all node id's from all sources were compared to determine nodes were not defined multiple times. Nodes were further reduced by iterating through the types (Plant, Chemical, Gene, Pathway, Bacteria, Phenotype, Diet) and linking identifiers of a type with the same name into a new identifier. These new node ids are

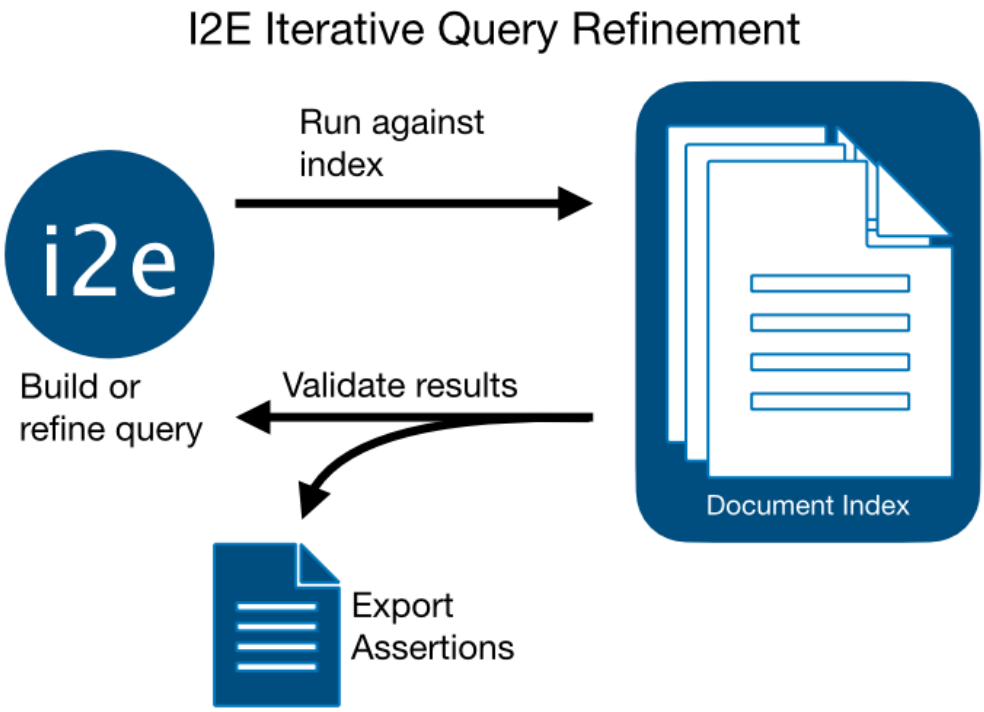

Figure 3.2: Iteratively developing I2E queries is a process of 1. Building a query, 2. Running the query against the document index, and 3. Validating the results returning either to refine the query or export the results when reaching a satisfactory threshold.

set up by combining "abckb-" with node type (e.g plant, chemical) and an auto-incrementing id. All original source identifiers are retained as a property of the new node along with name and synonyms converted to lowercase for convenience. Relationships originally matched to these original source ids were then remapped to the new knowledgebase ids. All of these extracted, collapsed relationships from full-text and the structured data sources were inserted into the knowledgebase to create the ABCkb 2.0 (Figure 3.3). This tool will aid researchers to elucidate the molecular mechanisms driving the health effects of the microbiome and diet.

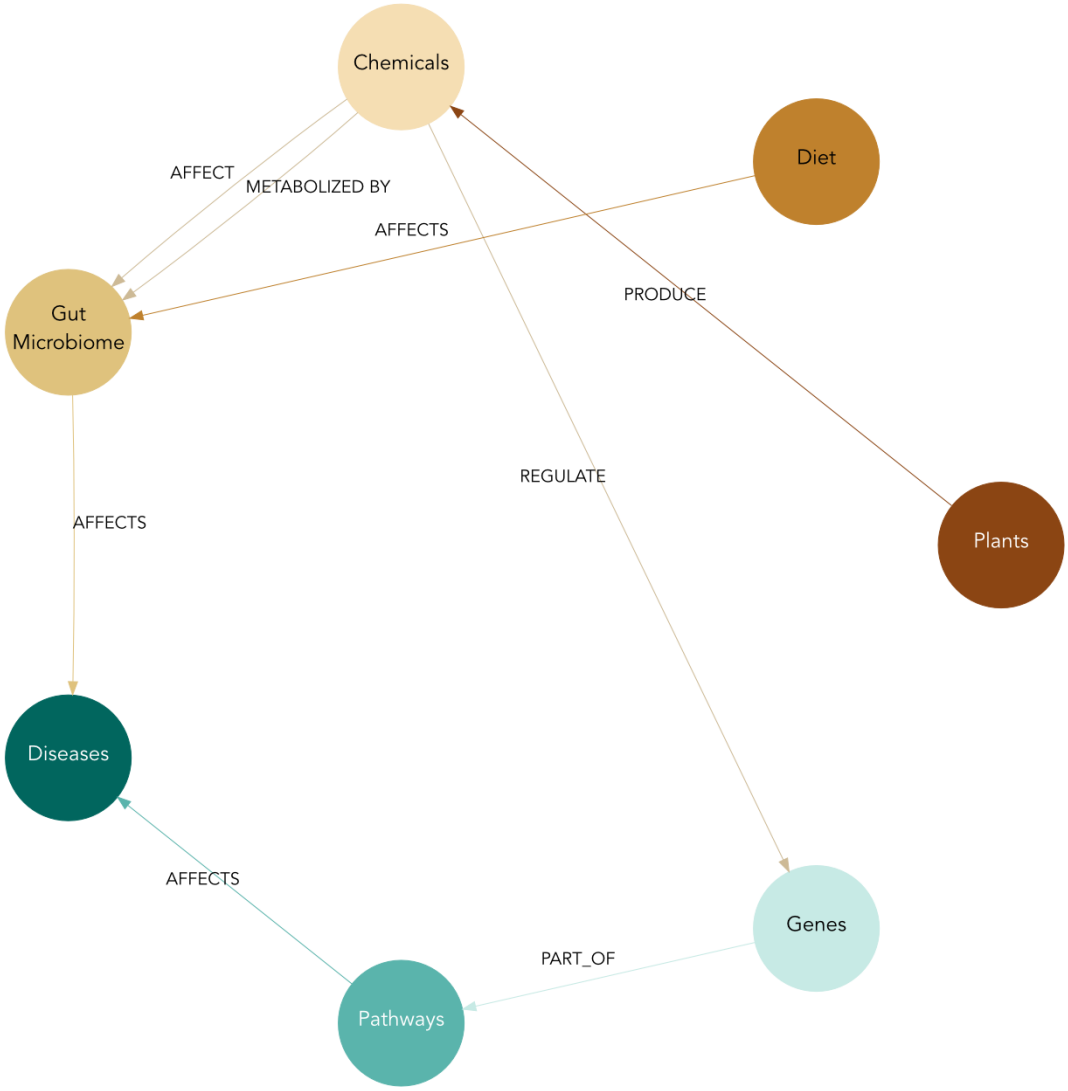

Figure 3.3: The schema for the ABCkb 2.0 showing connectivity between the additional microbiome portions and the ABCkb 1.0

3.5 Statistics and Browser

The ABCkb 2.0 contains 1,155,603 nodes and 7,338,259 relationships, a 60 percent decrease in edges from the first version (table 3.1). This is due to the reduction of similar nodes. Full text, text-mining yielded 625,975 hits, which produced 69,206 edges. Over 24,000 novel relationships were found solely in full-text articles and not found in abstracts. This indicates that mining through the full text increased the

amount of associations by a factor of 35 percent.

Table 3.1: The statistics showing the breakdown of node types between the two knowledgebase versions.

Node Type	ABCkb v1.0	ABCkb v2.0
Plant	237,350	222,087
Chemical	156,297	151,336
Gene	180,669	192,076
Pathway	28,654	28,759
Phenotype	54,769	43,428
Bacteria	0	517,888
Diet	0	29

The interface browser was updated to include the most recent version of the knowledgebase as a separate browsable data source. Open-discovery queries are the only available query type from the interface. The process begins with the user, entering in a search term in the browser search-bar. This search bar queries a full-text index in Neo4j on node names and node terms, or synonyms. This query produces tabulated results, ranked on search term relevance and limited to the top 10 results to increase efficiency. The user is able to select as many start points as desired and is prompted to click through to the next page. After making selections, the user is prompted to select an endpoint node type (Plant, Chemical, Gene, Bacteria, Pathway, Phenotype, or Diet) where a query is run on pre-assembled paths from the initial node type. This returns all of the possible paths in a table, the ability to download the results to a csv file, and a query that can be used to generate those results in their own neo4j interface.

3.6 Application

3.6.1 Phytochemicals, microbiota, and phenotypes

Phytochemicals have long been touted as semi-magic pills for all kinds of maladies, cancers, and diet-related diseases [91, 92]. For example, plants in the *Brassica* family (broccoli, mustard, and brussels sprouts) contain glucosinolates, a type of phytochemical linked with anti-cancer, antioxidant, and anti-inflammatory benefits (Figure 3.4) [93]. However, there are aspects of this benefit that remain unknown, namely the molecular mechanisms [94]. In the ABCkb (Figure 3.5), several genes are linked with glucosinolates, one of which is the cytochrome p-450 gene, CYP4F3 [95]. This gene is one of three key branch points in the arachidonic acid metabolism pathway [96]. Arachidonic acid is released due to stress, injury, and is also found from dietary sources like red-meat, fish, eggs, and poultry [97, 98]. The cytochrome p-450 mediated arachidonic acid pathway has been shown to produce cardioprotective epoxyeicosatrienoic acids[96]. Thus, consumption of glucosinolate-rich foods to influence the breakdown of arachidonic acid through the cytochrome p-450 pathway is of great nutritive importance.

However, glucosinolates like glucoraphanin are not activated unless they are transformed by the enzyme myrosinase [99]. This enzyme is found in these *Brassica* vegetables in varying amounts, but their availability is affected by multiple factors including cooking method, crop-breeding, and anatomical plant location [93]. Therefore, it is likely that the actual amounts of bio-available sulforaphane from myrosinase activity alone is lower than expected.

The level of bio-available glucosinolates is further complicated by the interaction between the microbiome and phytochemicals. A search within the ABCkb was performed between the chemical node, glucosinolates, and bacteria (Figure 3.6). *Bifidobacterium*, a common probiotic genus, is shown in the ABCkb to affect glucosinolates. *Bifidobacterium* have a similar effect on glucosinolates like myrosinase but with

a propensity to metabolize further[100]. Additionally, a bacterium recently linked to digestive disorders, *Alistipes putredinis* is also shown to metabolize glucosinolates in the gut from the ABCkb [101]. The *Alistipes* genus was only recently described in 2003 and had formerly been under the *Bacteroides* classification [24]. Abundances of *A. putredinis* bacterium are correlated with ulcerative colitis disease activity and several other phenotypic conditions [102].

To explore connections between *A. putredinis* and phenotypes, a knowledgebase query was constructed in the ABCkb from *Alistipes putredinis* to phenotype (Figure 3.7). Connections from the Disbiome database indicate manually curated connections from *Alistipes* to phenotype and text-mining revealed an additional connection to obesity. This connection was from a study tracking weight loss patients over time and found that *Alistipes* abundance is negatively correlated with obesity[103]. Given the disbiome connection between *A. putredinis* and reduction of liver cirrhosis, the link between this bacterium and obesity is logical. Non-Alcoholic Fatty Liver Disease (NAFLD) is linked with incidences of obesity along with many metabolic disorders like hypertension, type 2 diabetes, and hypertension [104].

What then would be the molecular mechanisms driving the interaction between *A. putredinis*, obesity, and connected metabolic disorders such as cirrhosis of the liver? The previously explored path between glucosinolates, the cytochrome P450 gene, and arachidonic acid metabolism provides a possible explanation between this link. Hydroxyeicosatetraenoic acids produced from the arachidonic acid metabolic pathway, not the pathway attributed to glucosinolates is linked with obesity [105]. Therefore it is plausible that gut microbiota perform the activation step, like myrosinase, necessary to yield the protective benefits and explains the connection between *A. putredinis* and obesity. This should be further explored.

3.6.2 Resveratrol, Akkermansia, and diabetes

Abundances in the gut of bacterium in the Akkermansia genus have been linked to many favorable health outcomes (Figure 3.8) [106]. Specifically metabolic-related health outcomes, such as weight management and diabetes, are found to be inversely linked to abundances of *Akkermansia muciniphila*[107]. This bacterium is gram-negative, oval-shaped, and lives in the mucosal layer of the gut along with several other places in the human body like the small intestine, human milk, and the mouth [108]. *Akkermansia* degrades mucin within the gut and produces Short Chain Fatty Acids (SCFAs), compounds often linked with gut health[109, 110]. While this may sound negative, and was previously considered detrimental, it is in fact a positive, normal process[111, 112]. One SFCA in particular, butyrate has been shown to upregulate goblet cell mucin production in the gut through interaction with MUC genes [113]. In addition, butyrate and other SFCAs are linked to the increase of other beneficial bacteria in the gut, and the gut-brain axis[114, 115]. Therefore, it is likely that increasing *Akkermansia* would subsequently increase SFCA production in the gut which would lead to more mucin production, and ultimately a healthier gut microflora. However, the molecular mechanism by which *Akkermansia* affects human health is still an avenue of exploration. In the ABCkb, Akkermansia is linked to 54 different Phenotype nodes. Of those 54 nodes, 4 are related to diabetes and metabolic disorders (Figure. 3.9). Links between *Akkermansia* and polyphenol-rich diets have been established[116].

The foods that contain polyphenols include grapes, blueberries, coffee, tea, and cocoa powder[117]. These foods with their polyphenolic compounds are attributed to a host of health effects[118, 119]. In the ABCkb, Akkermansia is linked to 74 Chemicals, of which polyphenols are prevalent (Figure. 3.10). One specific polyphenol, resveratrol found primarily in berries, is released in response to plant stressors like pests and infections[120].

Resveratrol is the primary compound that gives red wine its claimed health benefits. In the ABCkb, resveratrol is linked to 3,866 genes, 93 more than the Comparative Toxicogenomics Database has available [15]. There are 13 pathways related to glucose and insulin with over 1500 genes between (Table 3.2).

Table 3.2: The pathways connected to resveratrol related to glucose and insulin with the gene counts in-between.

Gene Count	Pathway Name	Pathway ID
899	Insulin Secretion	GO:0030073
426	Glucose Metabolic Process	GO:0006006
382	Glucose Import	GO:0046323
295	Glucose Homeostasis	GO:0042593
216	Gluconeogenesis	GO:0006094
136	Insulin Receptor Signaling Pathway	GO:0008286
24	Insulin Catabolic Process	GO:1901143
23	Insulin Metabolic Process	GO:1901142
18	Glucocorticoid Secretion	GO:0035933
10	Insulin Processing	GO:003070
10	Positive Regulation of Glucokinase Activity	GO:0033133
6	Renal Glucose Absorption	GO:0035623
4	Insulin Receptor Internalization	GO:003816

The generalized links between diabetes and resveratrol may be fully explained by the interaction between *Akkermansia* and resveratrol. This connection should be further explored to elucidate the effect of dietary polyphenols on *Akkermansia*.

3.7 Discussion

Generating these connections manually through intense literature reviews adds significant time burdens to overburdened researchers. This knowledgebase however

provides the nutritive connections in papers in a graphical, browsable format, with references for clear knowledge of where the associations are derived, and where to go to explore the study that produced the association. In addition, the size of the knowledgebase is a manageable 2.8 Gigabytes, over the 50 Gigabyte version original.

There are a few limitations to this knowledgebase and method of creation. The first is the availability, and the significant time component of obtaining permissions to mine through full text articles. During our quest to obtain access for full-text mining, publishers often expressed concern that a knowledgebase like this would reduce the number of journal readers. We propose that this knowledgebase and others like it would in fact increase journal readers as they discover the nutritive connections within articles and desire to probe further. Methods sections often house additional information pertinent to experimental design not contained within the knowledgebase. The ABCkb merely expedites the literature review process and can point to more relevant results. Queries for *Alistipes* in the NCBI pubmed repository return several results irrelevant to nutrition. Custom MeSH queries reduce the overall papers returned but the researcher is left to read each one of the papers to find the associations within. With a restriction of papers, there is a question of how much text is too much. In the *Alistipes* to obesity example, the text-mining result was in fact found in a paper examining different methods for analyzing microbiome count data which referenced the original study that we did not have access to text-mine [121]. In a case where not all texts are available to mine through, assertions can be found from texts that reference other texts. Future work could expand on this knowledgebase to rank associations and articles of interest based on similar reference sections. Abstracts and references are often freely browsable. Mining through abstracts to extract associations and then following up with a ranking method to rank related articles could provide researchers with additional articles to read for further information.

Another limitation to this knowledgebase is the method by which the results were

obtained. Rule and dictionary approaches work well for mining through literature sources where the dictionaries are well-defined, however microbiome literature is a metaphorical wild-west. When bacteria are taxonomically reclassified, connecting old bacteria names to the reclassified name is a significant challenge. In addition, bacteria are referred to in some studies with an informal moniker to designate multiple strains. Connecting these associations may be possible through some complicated regular expression checking or a true machine learning approach may be required.

Finally the desire to probe drug compounds and their relation to the plant molecular mechanism pathway has been expressed. Nutrichem 2.0 included drugs and some of their effects in their database [14]. One common example is the ability for grapefruit to reduce the efficacy of some heart medications [122]. Similar effects with different drugs have been seen in people with varying microbiota abundances.

3.8 Conclusion

This knowledgebase, the ABCkb 2.0, further expands on the original knowledgebase with the addition of microbiome connections and dietary connections. Text-mining full-text articles extracted more associations than abstracts, but also requires more strict filtering. Two knowledgebase analyses were produced to demonstrate the effectiveness of the ABCkb and the potential to generate testable hypotheses. Further work on this should validate the associations within through physical assays.

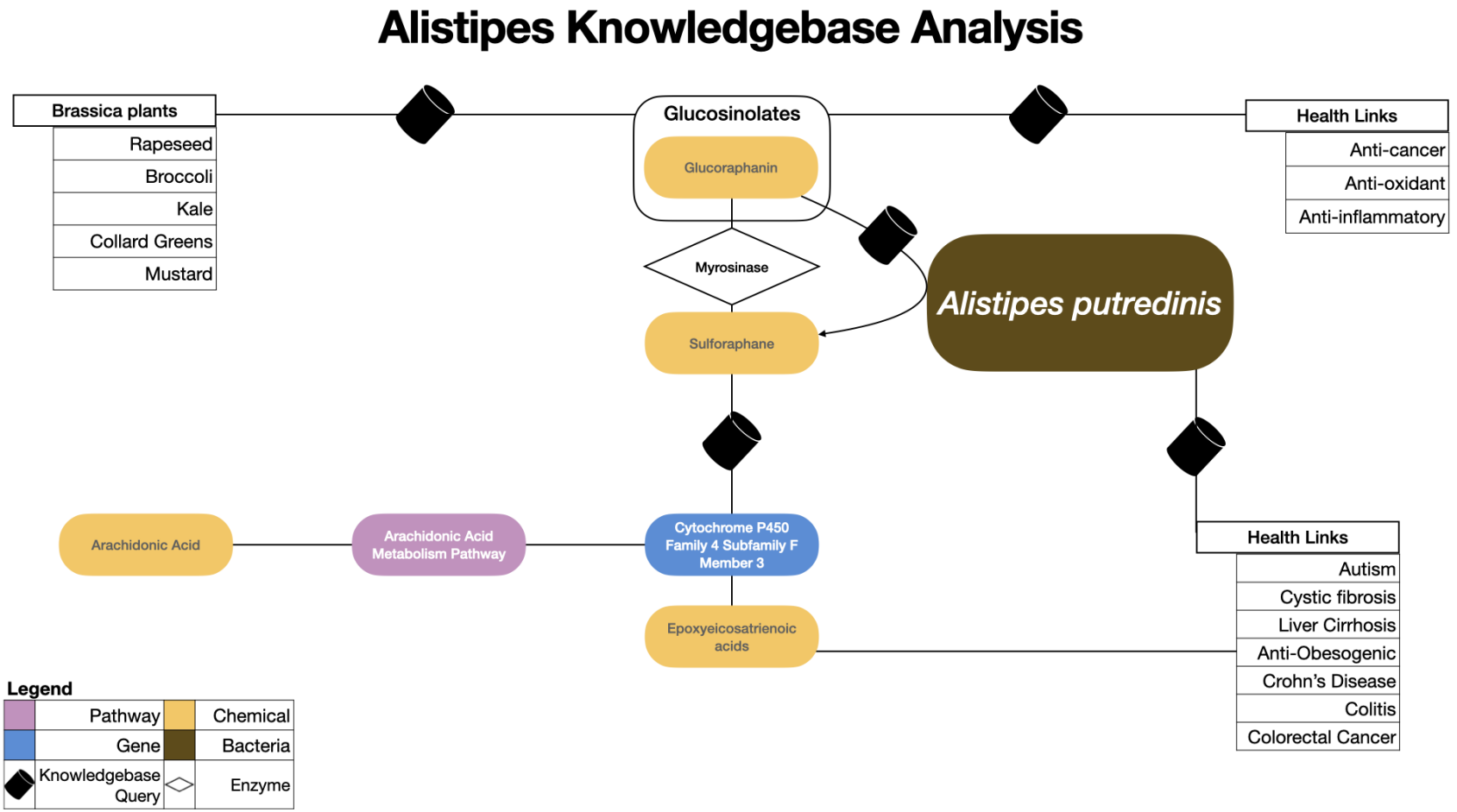

Figure 3.4: The overview of the knowledgebase analysis explaining the connection between *Akkermansia spp.*, broccoli, and the claimed health benefits of glucosinolates

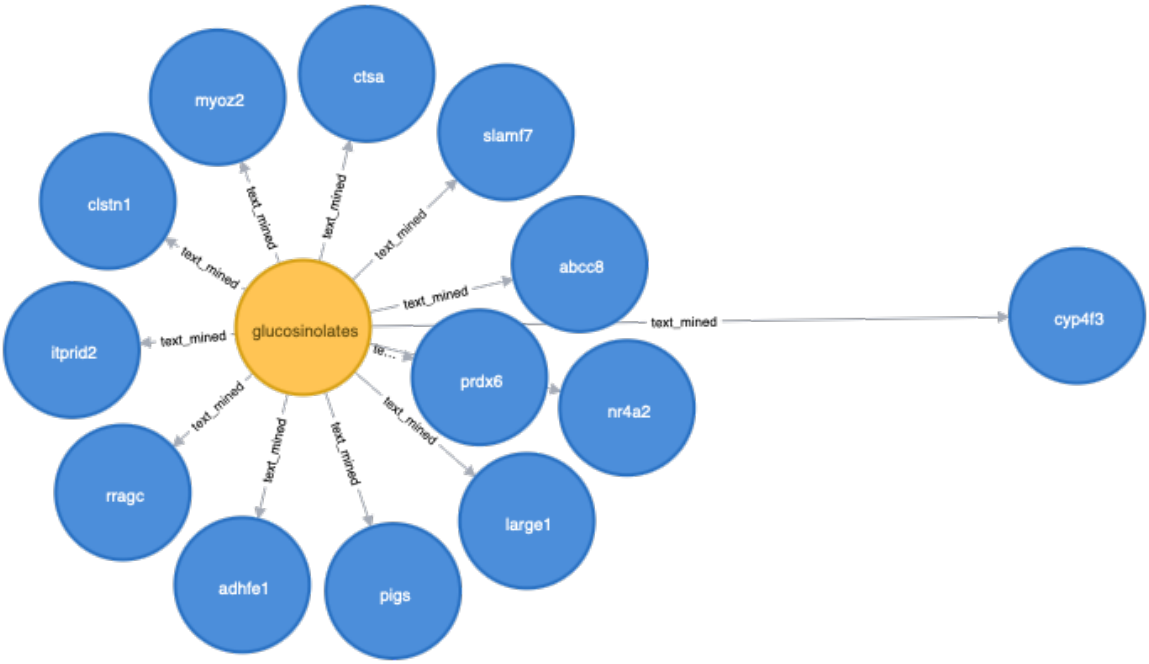

Figure 3.5: A graph from the ABCkb 2.0 showing the connectivity from glucosinolates to genes. One gene in particular has been singled out, Cytochrome p-450.

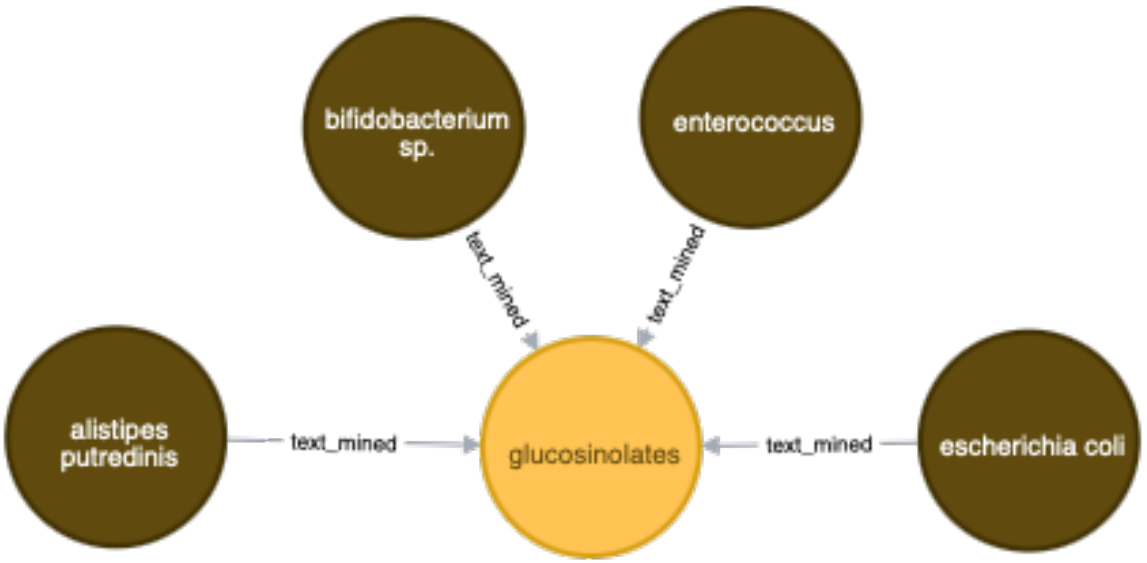

Figure 3.6: A graph from the ABCkb 2.0 showing the connectivity from glucosinolates to bacteria.

Figure 3.7: A graph from the ABCkb 2.0 showing the connectivity from glucosinolates to bacteria.

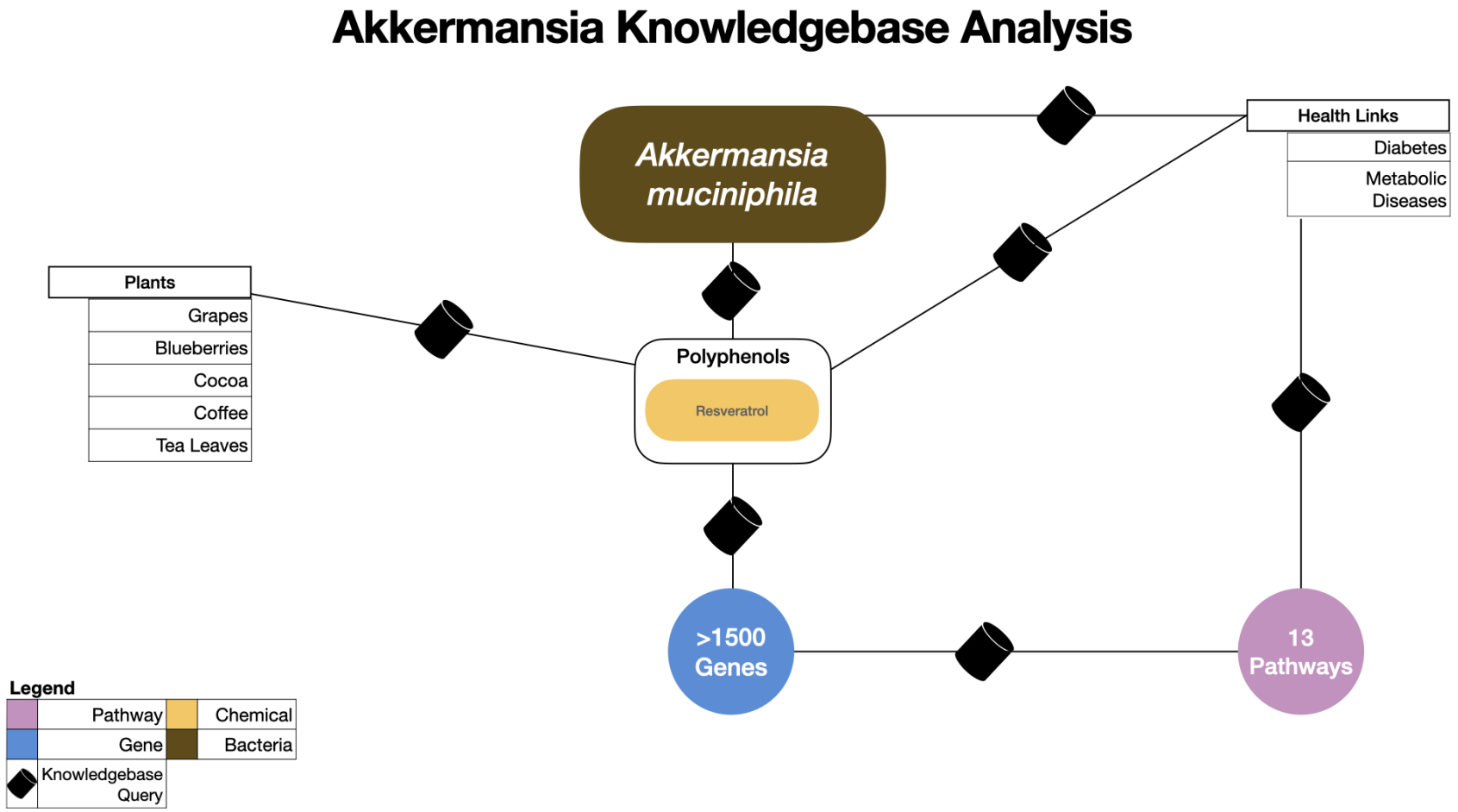

Figure 3.8: The overview of the knowledgebase analysis explaining the connection between *Akkermansia muciniphila*, diabetes, and metabolic disorders

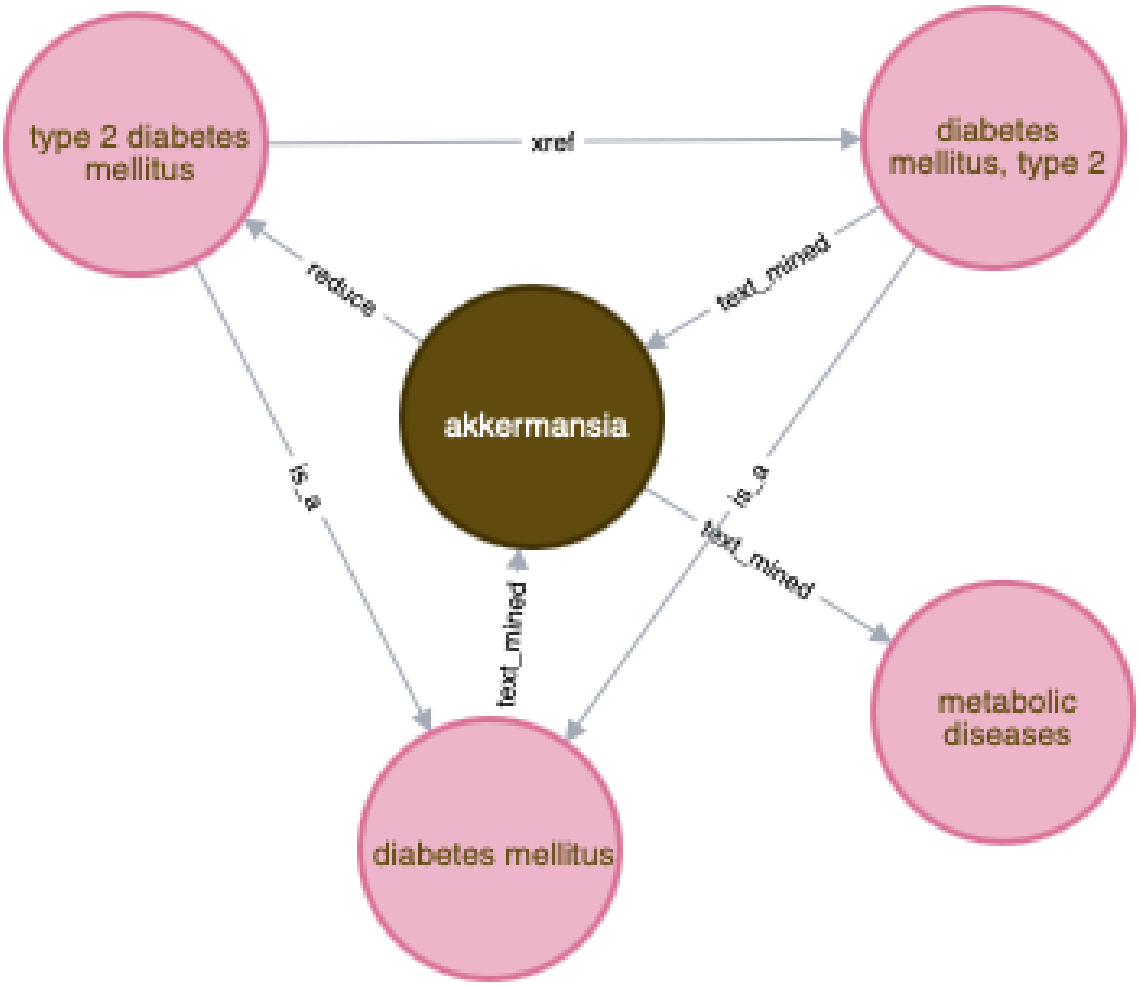

Figure 3.9: A graph from the ABCkb 2.0 showing the connectivity from Akkermansia to metabolic disorders and type 2 diabetes.

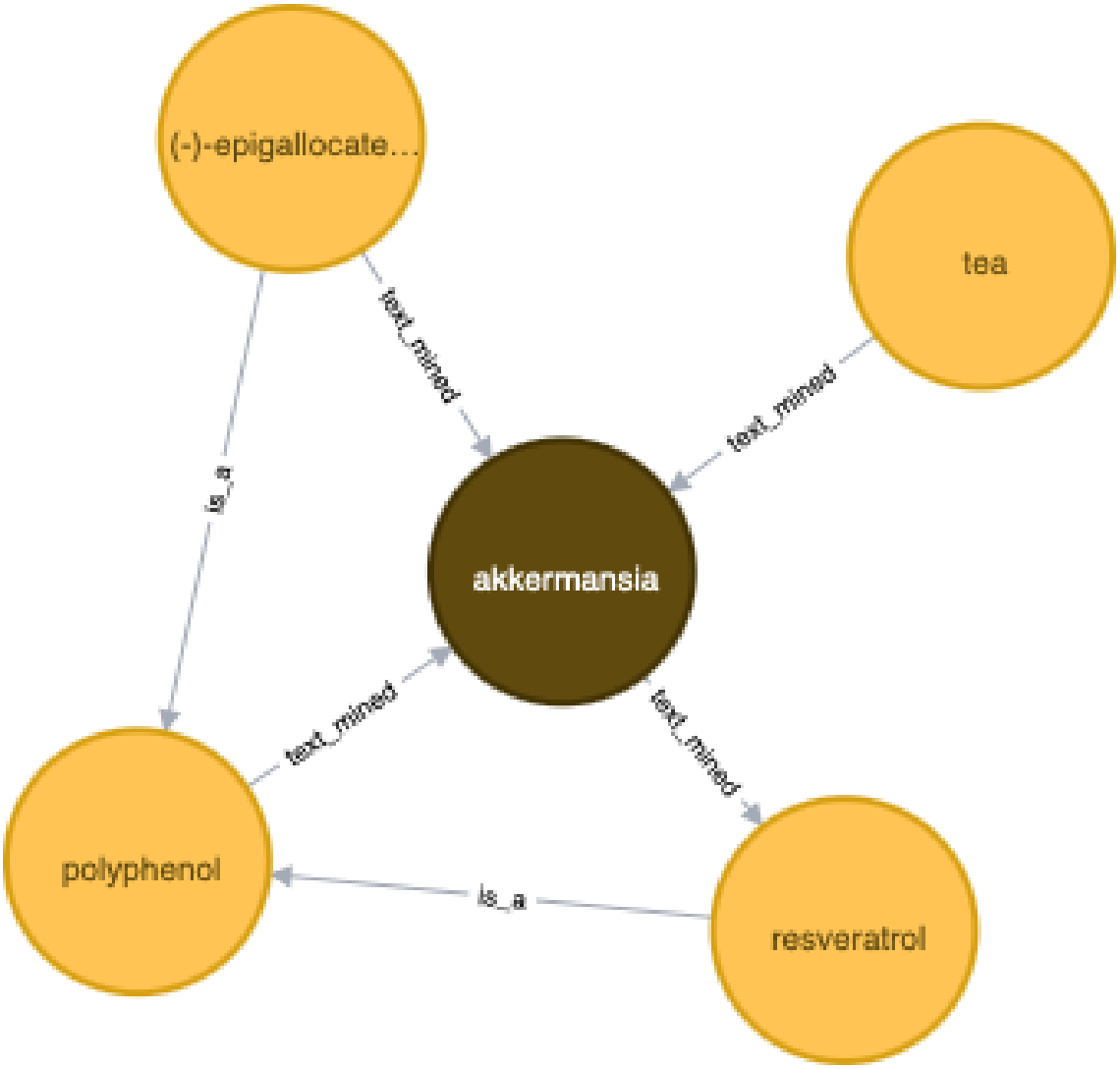

Figure 3.10: A graph from the ABCkb 2.0 showing the connectivity from Akkermansia to polyphenols.

CHAPTER 4: APPLICATION OF KNOWLEDGEBASED-DISCOVERY TO THE SARS-CoV-2 PANDEMIC AND DEVELOPMENT OF THE COVID TO DIET KB (CDkb)

Coronaviruses are prevalent in nature and are responsible for various respiratory infections, many of which are non-life threatening [123]. They spread among bat populations and the first significant coronavirus infection with a high mortality rate was first identified in 2003 after it spread in a hotel [124]. This coronavirus was named SARS-CoV-1 which stands for Severe Acute Respiratory Syndrome CoronaVirus 1. It took a significant effort to achieve sequencing and many hours of work [125, 126]. Fortunately the spread of the virus was mitigated, however experts continued to study the virus and the effect on the human body. In 2008 another coronavirus appeared in the middle-east and was named MERS-CoV. Finally, in late 2019 a coronavirus began to spread in the Wuhan province of China which reached pandemic proportions within a few months [123]. Technological advances led to a quick identification of the virus, and SARS-CoV-2 was the culprit. While our previous knowledgebases have been applied to noncommunicable diseases caused by dietary or genetic factors, communicable diseases of pandemic proportion will benefit from this method.

4.1 Background

Coronaviruses are positive-sense RNA genome, circular viruses that exist in abundance in nature [123]. The SARS-CoV-2 genome is around 30 kb which makes it one of the largest in the class of RNA viruses[124]. Covering the exterior of coronaviruses is a lipid bilayer envelope with three protein types (membrane, envelope, and spike) protruding out from the surface[127]. In SARS-CoV-2, viral infection is facilitated

through the spike protein binding to ACE2 receptors on host cells[128]. Infection is further enhanced through its host immune system evasive detection measures through utilization of glycan molecules that coat the spike proteins[128]. This coating limits host T-cell activation. While some general pathology was known prior to the pandemic, from other coronavirus studies, much of this information was discovered and broadly communicated through preprint articles[129].

The acceptance of pre-print article servers and repositories is still somewhat controversial [130]. Traditional scientists dislike pre-print articles due to the fact that peer-review is the stage-gate that reduces the spread of bad science. The process of review identifies and removes faulty experiment methods, unfounded results, and poor writing, which overall increases the quality of published literature [131]. In contrast, modern scientists argue that the large volume of papers being submitted for publication overwhelms the peer-review process and delays scientific progress. The process can be expedited but that requires reviewers pre-selected for journal topics that are in demand or of focus. It is argued that preprinted articles are able to be reviewed by a wider panel of scientists in the community [132]. Additionally, peer-reviewing is not as strict of a stage-gate as believed. A study found that articles can be passed through peer-review using language generated from computational methods [133]. Pre-print servers then exist to spread scientific information in an expedited manner while still honoring the process of peer-review.

The utility of pre-print articles was seen early in the pandemic. An average of 39.5 pre-print articles were circulated per day during the pandemic and by 6 months, over 6,000 articles were available[134, 129]. An additional factor to consider in the flow of information is the technological progress that has been made since the previous coronaviruses. The sequence for SARS-CoV-2 was available much sooner than for SARS-CoV-1, and MERS-CoV (Figure 4.1).

A weekly updated preprint repository was set up specifically for SARS-CoV-2 pa-

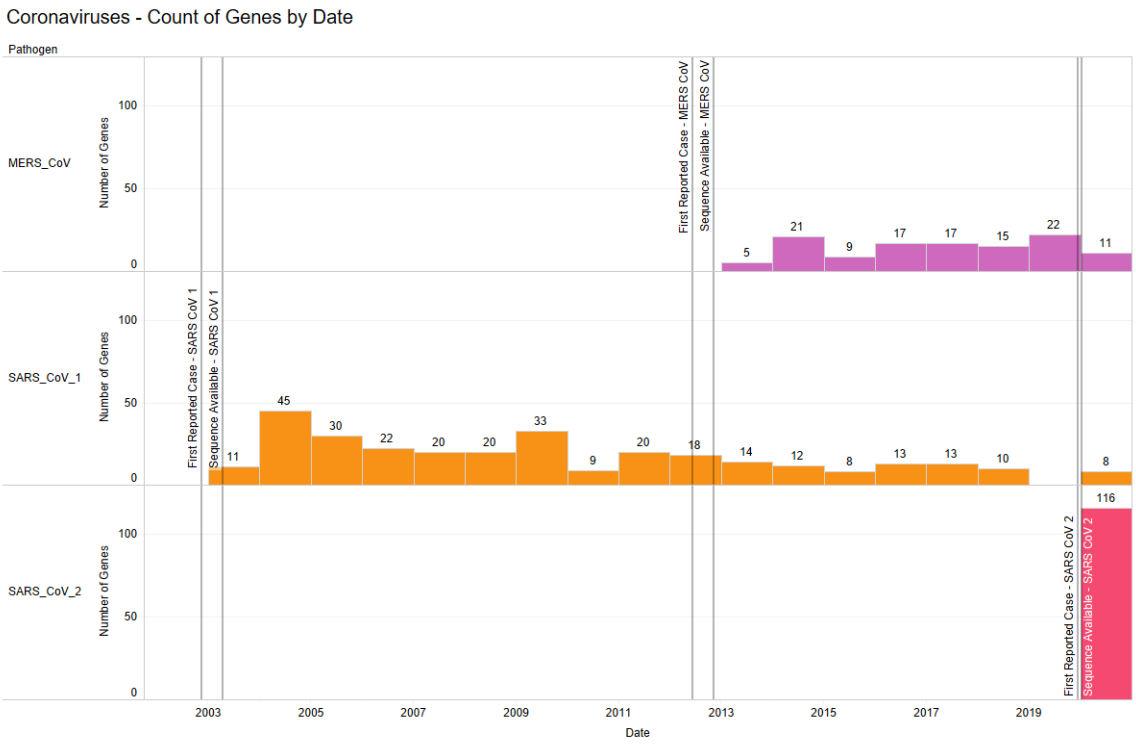

Figure 4.1: A comparison between the three major coronavirus contagion sequence availabilities and amount of genes linked to each individual pathogen by year, generated 6 months after the start of the pandemic.

pers, along with the peer-reviewed papers on the relative coronaviruses [135]. The pre-printed articles in this repository also came with reviewer comments on the articles for researchers to identify any areas of concern and mitigate any false information. All documents in the repository were provided in JavaScript Object Notation for convenient processing by text-mining tools. As the pandemic progressed, one symptom seen in a notable portion of infected patients is a prolonged infection of SARS-CoV-2.

One of the challenging parts of the pandemic was the broad range of symptoms present in infected individuals [127]. Some patients had symptoms similar to a cold, where others presented more severe symptoms. Infections in many cases led to a condition called COVID-pneumonia which resulted in patients requiring supplemental oxygen. In severe circumstances patients had to be sedated and placed on a ventilator. Once patients conquered the initial infection, many were left with a condition

called long-covid [136]. This condition persists for several months following the initial infection and presents with myalgia-like symptoms along with mild symptoms present in the initial infection: anosmia, ageusia, and difficulty breathing. Survival rates of SARS-CoV-2 infections were linked to noncommunicable patient comorbidities like type 2 diabetes (T2D), heart disease, and obesity[137]. These comorbidities are all diet-related. Therefore, diet alteration and subsequent reduction of patient comorbidities is of great interest. This knowledgebase was built to explore and build hypotheses around the molecular mechanisms between pathogen and plant phytochemical metabolic pathways.

4.2 Building the CDkb

The Covid to Diet knowledgebase (CDkb) was built from the original ABCkb 1.0 with the addition of associations from the CORD-19 dataset and drug information from the drugbank database (Figure 4.2) [135, 138].

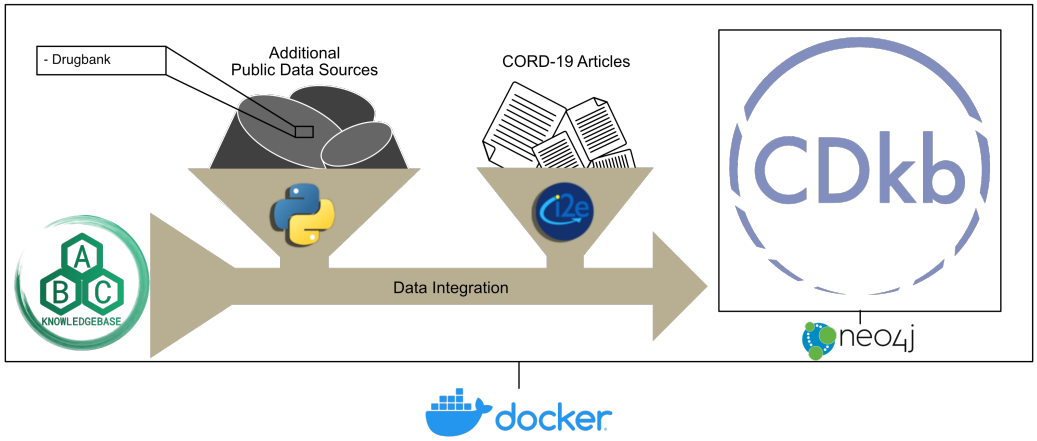

Figure 4.2: The process of creating the CDkb starts with the original knowledgebase and adding drugbank information with text-mined information from the CORD-19 dataset.

4.2.1 Structured data sources

Data in the CDkb comes from the original 11 sources from the ABCkb 1.0 with the addition of drug information. Though drugs are technically a subset of chemicals,

within this knowledgebase they are marked as a separate set of nodes. This allows for more efficient queries for open-discovery with drugs as the starting point. Drugs from the drugbank database were parsed and added to the knowledgebase with connections to genes from the original ABCkb[138]. Then text mined associations from I2E were added to the knowledgebase.

4.2.2 Natural Language Processing Relationships between Viruses and Human genes

A subset of literature from the CORD-19 dataset updated in July, 2021 with preprinted research, papers from Elsevier corpii and other Pubmed Central articles was downloaded and converted from jsonified text to xml for indexing in I2E[135]. This conversion was done through a custom python script and extracted sections where they were available within the original json header. When sections were unavailable, the text was labeled "general." The preprinted subset contained articles with reviewer comments, which was not useful for our immediate use and increased duplicate hits. To reduce unnecessary duplicate text-mined results, matches from the files containing reviewer comments were negated by removing any Additionally, any matches from methods and reference sections were negated. Articles were grouped into batches and uploaded to the server with the I2E interface.

After indexing, text-mining queries from I2E from the "class affects class" subset were run and refined to identify hits where SARS-CoV-1, SARS-CoV-2, or MERS-CoV affect human genes. An ontology of predicates provided by Linguamatics was used to link the Coronavirae to human genes in the literature. All of this information was inserted into a Neo4j graph database in a Docker container (Figure 4.3).

4.2.3 Projecting genes on pathways

Pathview is a tool produced to visually display heatmap gene expression values on KEGG pathways [139]. We used Pathview to map text-mined gene hits between

strains to determine top pathways identified and visualize gene hits on pathway maps. We also compared two methods: using the raw number of hits identified from text-mining, as gene expression values, or calculating a modified TF/IDF score. Raw hit scores are computed as the sum of gene hits per strain. Limitations of using raw hits as a metric for comparing gene hits across strains is the obvious discrepancy in corpus size given a short time-frame for papers on SARS-CoV-2 vs MERS-CoV and SARS-CoV-1. A TF/IDF score is commonly used to mitigate this limitation. The TF/IDF score is calculated from two separate values: Term Frequency, and Inverse Document Frequency. TF(T,D) is calculated as the number of times the term (T) is found in a document (D) 4.1. IDF(T) is calculated from taking the logarithm of the number of documents N, divided by the sum of documents containing term (T) (4.2). These two values are multiplied together to produce a score indicating importance, irrespective of frequency. Pathview allows us to map identified genes in pathway maps from the Kyoto Encylopedia of Genes and Genomes (KEGG) and visualize the strains together to determine the most relevant pathways, given the preliminary research.

$$TF(T, D) = f_{T,D} \tag{4.1}$$

$$IDF(T) = log\frac{N}{|\{d \in D : T \in d\}|} \tag{4.2}$$

$$TFIDF = TF * IDF \tag{4.3}$$

4.2.4 Exploring nutritive associations

Nutritive associations built in the ABCkb provide a link from plants to their phenotypes. After mapping COVID-19 associations, we probed possible molecular mechanisms linked with viral infections, along with long-covid infections. Additionally, dietary habits may exacerbate viral infections by up-regulating or down-regulating

genes necessary for mounting an immune response. Dietary phytochemical connections were also explored to produce a knowledgebase analysis.

4.3 Results

4.3.1 Text-Mined Results

The I2E text-mining query extracted 232,595 associations between the three Coronaviruses and host factors (Genes). Text-mining revealed over 4,000 host factors between these 3 coronaviruses with 1500 common host factors between SARS-CoV-1 and SARS-CoV-2 (Figure 4.4). This is over two-times the host factors mined from SARS-CoV-2 and MERS-CoV. These associations were found in over 49,000 articles with an average of 10 documents to each gene (Figure 4.5).

4.3.2 Knowledgebased discovering molecular mechanisms of long-covid

Long-COVID, or post-COVID syndrome is a range of symptoms experienced by patients four months beyond the initial SARS-CoV-2 infection[140]. The current range of symptoms of post-covid symptoms includes many of the symptoms experienced in the initial infection with additional mental and cognitive symptoms, along with symptoms that increase with physical exertion[141]. A cursory glance at the literature reveals many pathways that may result in experiencing these symptoms after infection.

There are several possible routes to generate a hypothesis around long-covid molecular mechanisms. As the symptoms are broad and seemingly disconnected, we began the knowledgebase search with known symptoms unique to long-covid (Figure 4.6. This search revealed a set of 29 phenotypic nodes from 9 symptoms (Appendix Table C.1). These phenotypic nodes alone are linked with 218 pathways and 36,369 gene nodes. There are 1,453 genes in the intersection of the text-mined host factors linked to SARS-CoV-2 in the CDkb and the genes connected with these 29 phenotypic nodes.

With the filtered genes from the knowledgebase, we performed a Pathview path-

way analysis with the calculated TF*IDF scores from text-mining [139]. Using automatic pathway selection which filters KEGG pathways to the most relevant pathways between the 3 Coronaviridae types. This analysis produced three KEGG pathway maps: Ubiquitin Mediated Proteolysis (Appendix Figure D.1), Protein Processing in Endoplasmic Reticulum (Figure 4.7), and RIG-I-Like Receptor Signaling Pathway (Appendix Figure D.2). Each map indicates the genes identified through text-mining with the TF*IDF scores providing color. Additionally, gene blocks are segmented by coronaviridae in order (left to right): SARS-CoV-1, SARS-CoV-2, MERS-CoV. One specific pathway of interest from the Pathview analysis is the Endoplasmic Reticulum Protein Processing pathway. This pathway is involved in protein folding within the cell.

Synthesized proteins are exported through the Endoplasmic Reticulum (ER) for cutting, folding, shaping and other modifications [142]. These modifications are facilitated by a set of molecular chaperones and enzymes that ensure correct protein folding and export the protein to the golgi body for any additional post-translational modifications [143]. The ER employs a series of accuracy checks before exporting the protein which ensures that the protein has been folded, folded properly, or needs to be degraded. ER stress is caused by a multiplicity of factors. Under stress, cascading pathways lead to either the adjustment of ER processing, stimulation of endoribonuclease activity, attenuation of cell translation, or signaling apoptosis of the cell [144]. From the Pathview analysis, there are several genes in these cascading pathways highlighted under protein stress, which is expected given the mechanism of viral production. Three genes of interest are eukaryotic translation initiation factor 2 alpha kinase 3 (PERK), endoplasmic reticulum to nucleus signaling 1 (IRE1), and activating transcription factor 6 (ATF6).

Traditional myalgias include Chronic Widespread Pain (CWP) and Fibromyalgia Syndrome FMS [145]. These are considered syndromes due to there being no clear dis-

ease pathway identified and the broad spectrum of symptoms that are selectively experienced by patients. Many of the symptoms experienced by patients with myalgias are also experienced in long-covid patients. In one study, myalgia patients have been found to experience severe pain hypersensitivity which is linked to widespread production of advanced oxidative protein products, formed in response to oxidative stress that lead to cellular apoptosis [146]. One specific gene drawn out from that study is the mitogen-activated protein kinase 8 (JNK) gene, highlighted in the graph and downstream from the IRE1 mediated cascade. Links between myalgias, blood coagulation pathways, and immunity pathways have also been shown [147]. Other evidence shows that microclots are found in long-covid patients, which mirrors what is known of apoptotic cells and the affect on procoagulation pathways[148, 149]. Therefore, it is possible that a high level of infection, leading to ER stress and widespread apoptosis, facilitated by increased IRE1 gene activation, causes long-term symptomatic periods for patients following SARS-CoV-2 infection. As syndromes like CWP and FMS have no definitive test to confirm diagnoses, testing for increases in clotting factors or microclots may provide a path forward for diagnosing patients.

4.3.3 Knowledgebased discovery of nutritional myalgia-like symptom mediation in long-covid patients

Mediation and reduction of long-covid symptoms that mirror myalgia symptoms often follows a course of Selective Serotonin Reuptake Inhibitor (SSRI) drugs in conjunction with pain reduction drugs like Tylenol or Ibu-profen [150]. However there may be nutritive pathways that alleviate the underlying molecular mechanisms driving the experience of symptoms. Using the identified genes in the ER protein processing pathway as initial start points, a knowledgebase query was developed to identify possible nutriceutical phytochemicals.

There are over 1,700 phytochemicals linked to over 6,900 plants in the Covid to Diet Knowledgebase. As the links between leafy green vegetables, and members of

the brassica family, to heart health is well established, we started with those. Two of the genes in this pathway that lead to apoptosis are linked with glucosinolates, and sulforaphane (Figure 4.8). These compounds are found in vegetables in the brassica family like Broccoli, Collard Greens, and Cabbage. Sulforaphane has been found to affect seratonin release, which increases the strength of this connection [151]. If widespread systemic cellular ER stress causes experience of myalgia symptoms in long-covid patients, the CDkb indicates that increased consumption of glucosinolate vegetables may provide relief through the mentioned molecular mechanisms. This connection should be explored further.

4.4 Discussion

Knowledge based discovery has many application pathways. Here we demonstrated one pathway using the ABCkb 1.0 with text-mining from preprinted articles to extract knowledge and produce a new knowledgebase, CDkb. From this, we provide hypotheses around molecular mechanisms of possible nutriceutical phytochemicals to alleviate myalgia symptoms in long-covid patients. We also explored using a common tool, Pathview, to filter out pathways and view gene hits between the three coronaviridae strains. There are a few limitations to this approach.

This approach assumes all text-mined data is of the same quality. Mentioned previously, one of the drawbacks of utilizing pre-print articles is that some erroneous conclusions may come through, which would likely be identified in the peer-review process. Those utilizing this method would need to fully understand where the data was coming from, investigate, and deal with any spurious conclusions. Our weighting metric allows us to calculate TF*IDF scores to mitigate any of these conclusions. In addition, this method assumes that journals are willingly allowing text-mining of their content. Acquiring text-mining permissions can be a significant challenge to overcome, which may be a hurdle for future studies. If journals are unwilling to allow full-text mining, asking permission from the author is a secondary method to

consider.

Additional information that may prove useful is the incorporation of Single Nucleotide Polymorphisms (SNP). These positions in patients have provided insights otherwise unavailable through traditional experiment methods. Many of these SNPs found in GWAS studies have been found in non-coding regions [152]. The collection of SNP to disease data reveals a path for future nutriceutical studies to examine and explain the molecular mechanisms behind complex diseases with no singular cause. The CDkb does not incorporate this data as it was out of the scope of this research.

4.5 Conclusion

As the world grows more interconnected, pandemics will likely become more frequent. By using preprinted articles with knowledgebased discovery, the coordinated efforts of scientists around the globe can provide insights leading to faster hypothesis generation, and hope for expedited discovery of new and better treatments.

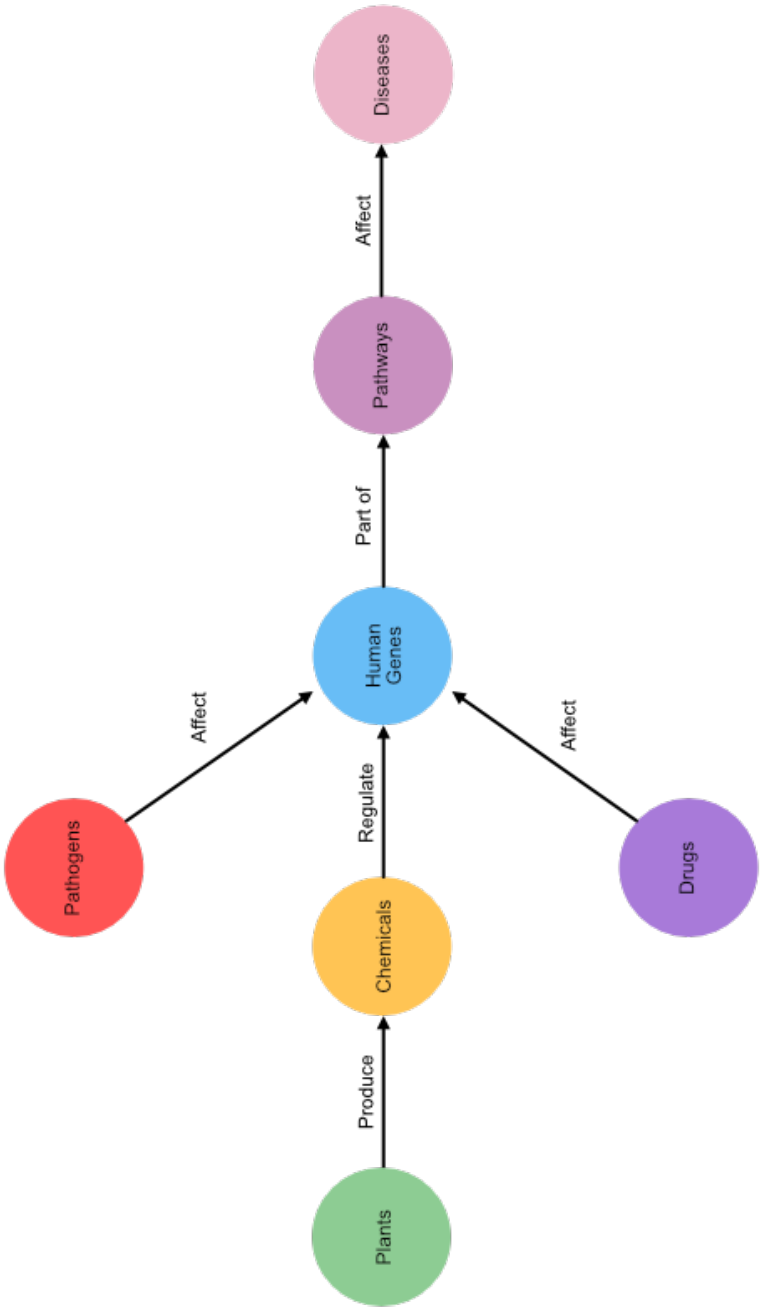

Figure 4.3: The schema for the CDkb showing connectivity to the elements from the ABCkb 1.0

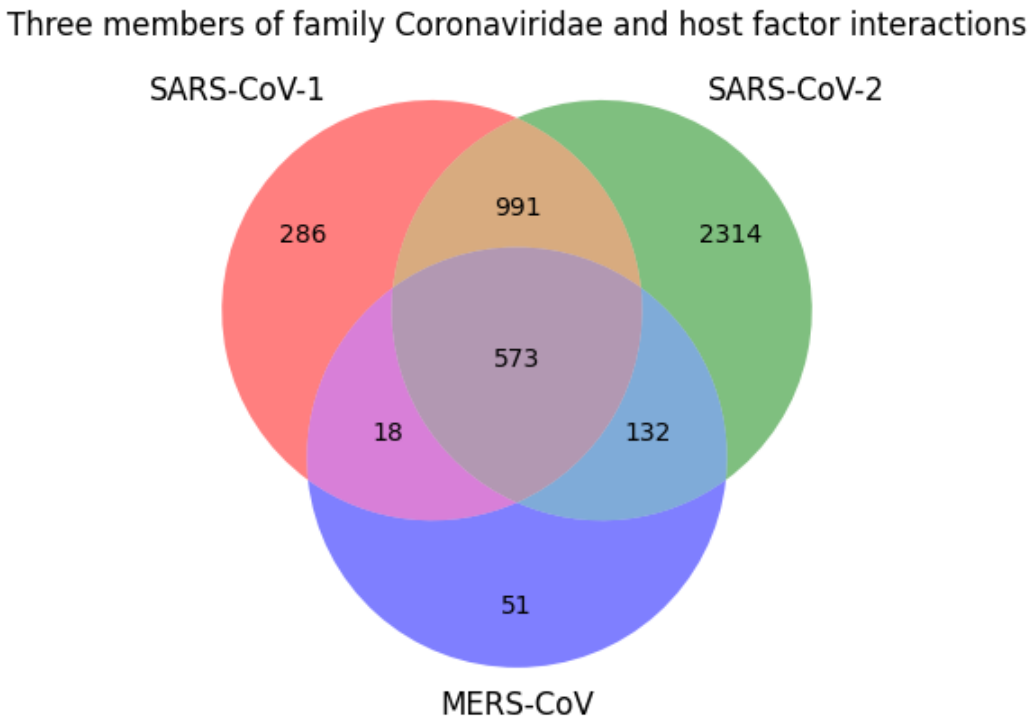

Figure 4.4: The intersection between 3 members of the family Coronaviridae and the text-mined host factor interactions from CORD-19

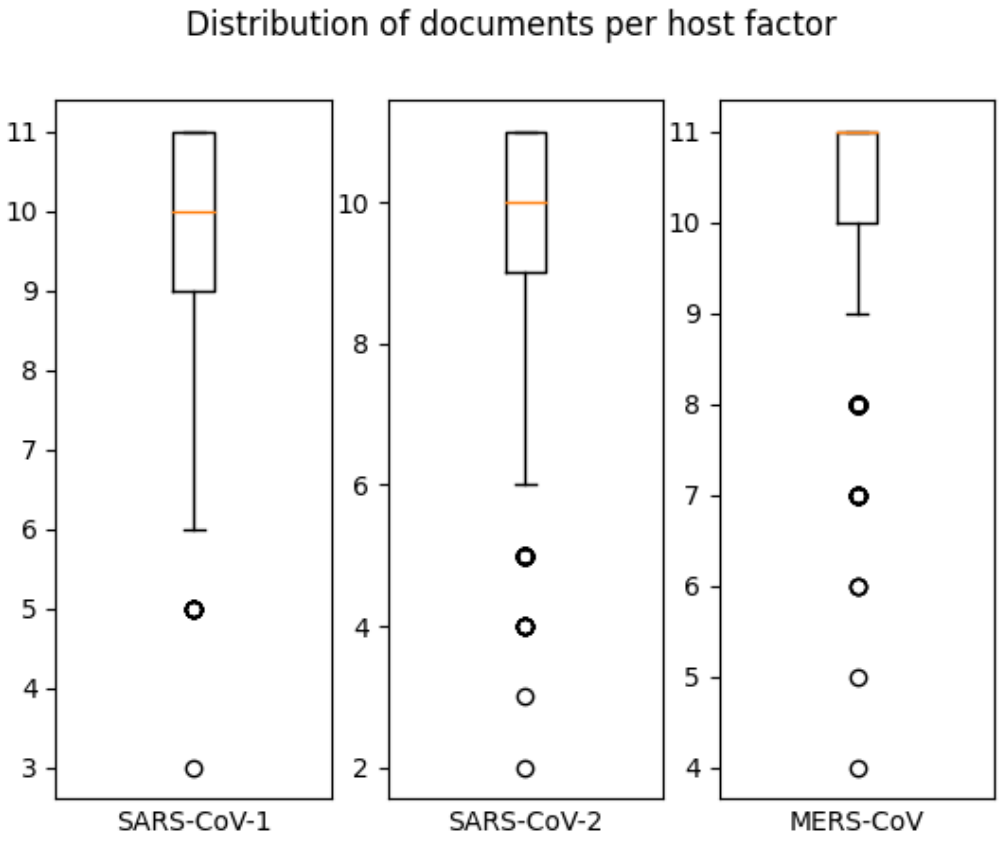

Figure 4.5: The distribution of documents to host factor identified through I2E text-mining with an average of 10 documents per host factor.

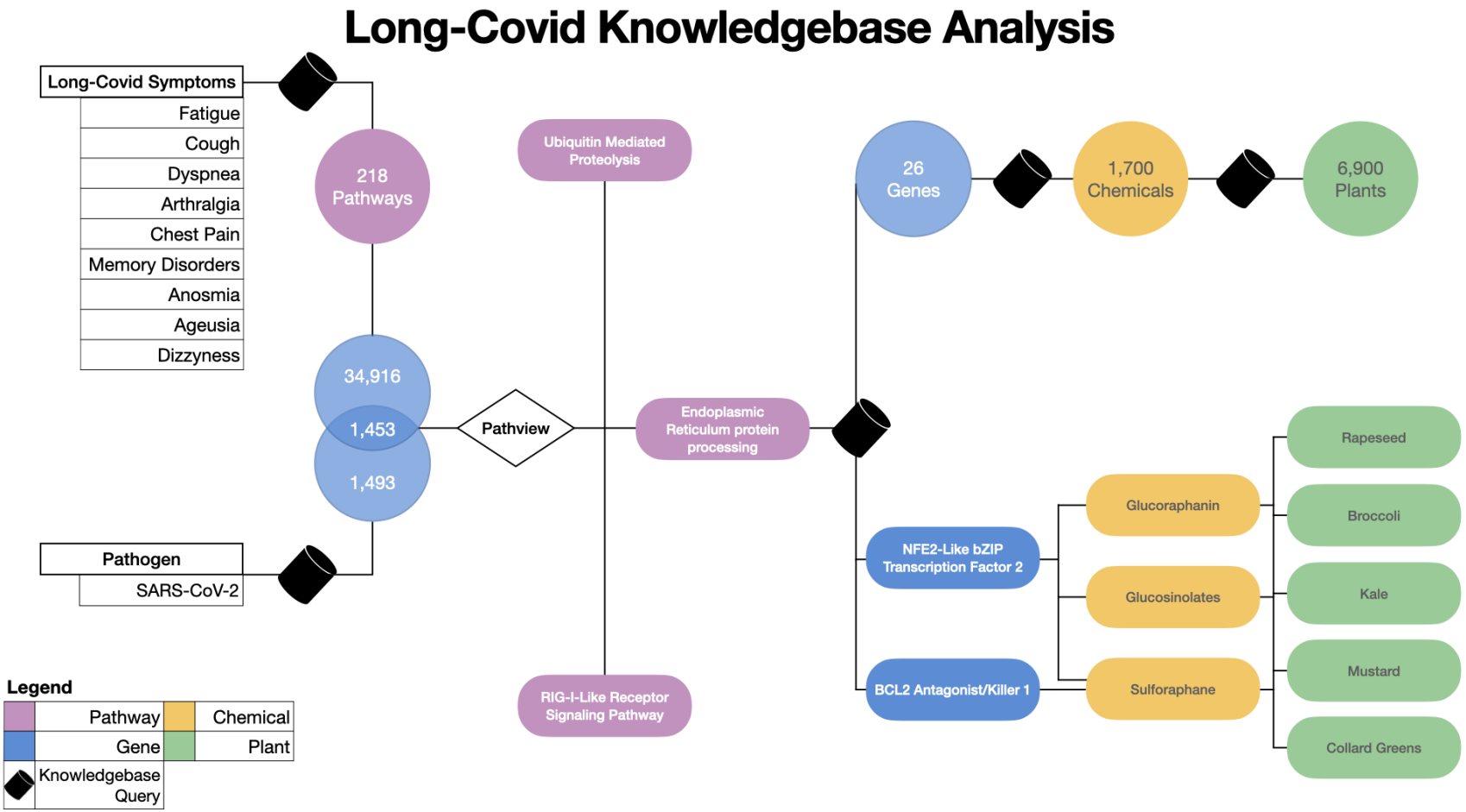

Figure 4.6: The high-level overview of the knowledgebase analysis around long-covid and the hypothesised plants and phytochemicals with possible alleviation

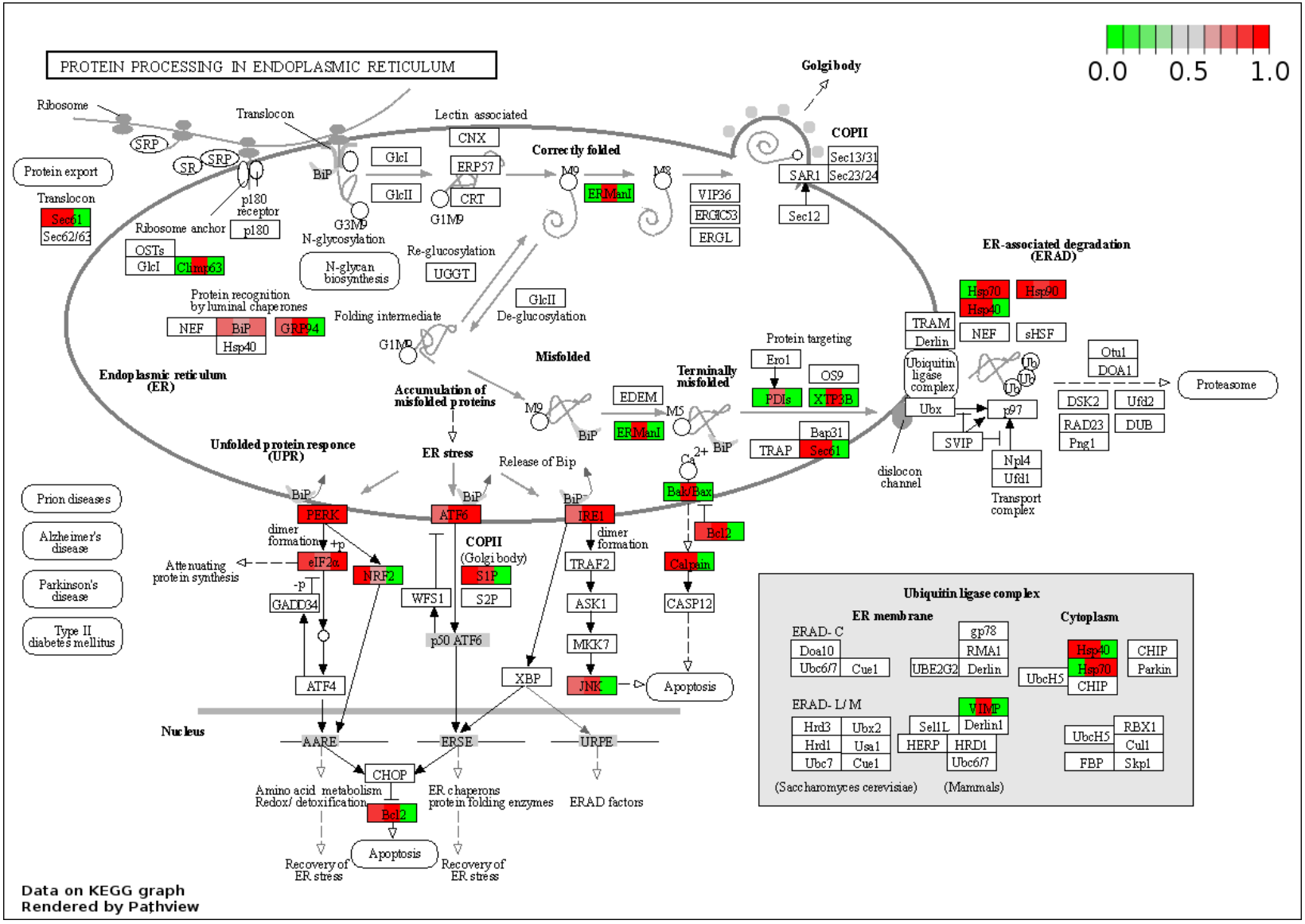

Figure 4.7: The hits from the pathview analysis with TF*IDF scores providing color values with the 3 coronaviridae. Colored gene boxes from left to right are: SARS-CoV-1,SARS-CoV-2,MERS-CoV

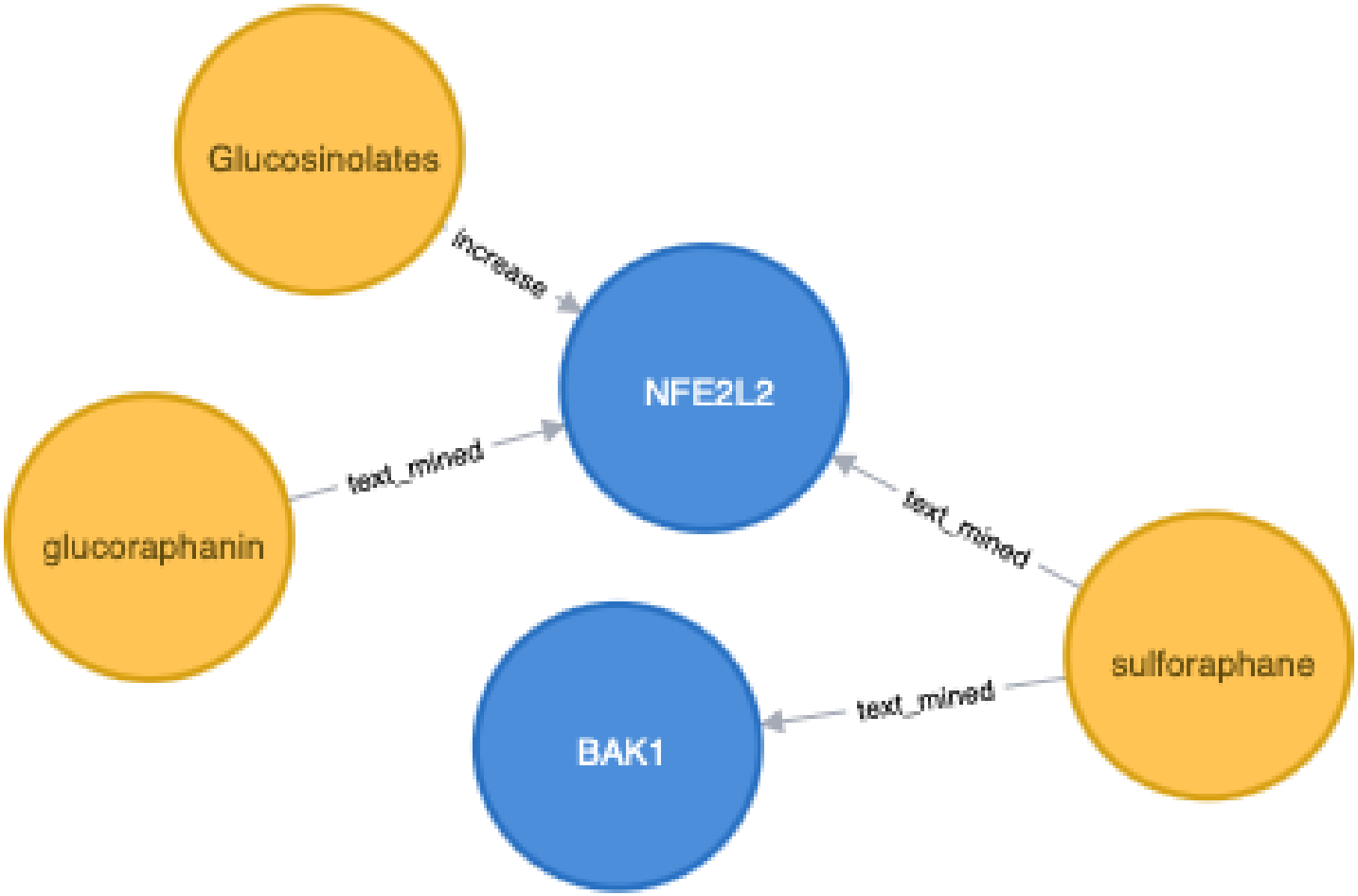

Figure 4.8: The hits from the pathview analysis with TF*IDF scores providing color values with the 3 coronaviridae. Colored gene boxes from left to right are: SARS-CoV-1,SARS-CoV-2,MERS-CoV

CHAPTER 5: Conclusions

There are three knowledgebases that have been generated as a result of this work. The first knowledgebase developed, ABCkb 1.0 is a cleaned up, slimmed down, updated version of the knowledgebase produced by Dr. Richard Linchangco [18]. Additionally, a simple interface was created to browse the contents and perform open-discovery without the use of Cypher Query Language. The second knowledgebase (ABCkb 2.0) adds microbiome connections from full-text text-mining along with duplicate node collapsing from separate sources. Two applications of discovery are provided, and the interface was updated to include the microbiome information. Finally, the third knowledgebase applies this method to a pathogen of pandemic proportions and utilizes knowledge-discovery to provide possible links to alleviate long-covid symptoms from dietary phytochemical sources. This knowledgebase is ready to be deployed to a team of individuals and move from the theoretical exploratory phase, to full-scale production.

5.1 Build a Support Team

The ABCkb is now large enough that it requires a multidisciplinary team to support the future development and use of the resource. There are several areas of support that will be required. An IT support team, web developers, machine learning and linguistics specialists. Additionally it would greatly benefit the future of this knowledgebase to partner with a nutrition lab, industrial or academic, for examining produced nutritive connections.

5.1.1 IT Support

This knowledgebase will require the support of IT professionals experienced in managing large databases. On Amazon Web Services (AWS), this knowledgebase is considered as a large storage bucket because of the size of the input data. Future IT work on this resource should consider building a custom server for management and migration away from AWS. Additionally, a hybrid approach between graph and relational databases may provide a more performant resource.

5.1.2 Web Interface

The web interface built for this knowledgebase is a simple resource to browse the data without users needing to learn a new query language. More work is left to do on the interface. There are three primary considerations for the future interface: a closed-discovery method, links to external resources, and graph building.

5.1.2.1 Closed-Discovery Browsing

Currently the only type of discovery supported by the interface is open-discovery. This is defined as the process of graph exploration from a known start node to any number of end nodes. This can be defined as a number of jumps, or to an end type. The goal is primarily to explore the endpoints and the connections between. In contrast, closed-discovery is primarily concerned with finding connections between specific start and end nodes. The interface should support selection of a number of end nodes and then provide the pathways in-between

5.1.2.2 External Resource Linking

As of now, the ABCkb 2.0 browser provides node names, but links are stored as properties of the individual nodes. A fully-supported interface should provide links to external resources for further knowledge discovery. This would enable researchers to connect additional information from the connected resources and further elucidate diet to disease molecular mechanisms.

5.1.2.3 Visualization of Results

The future interface should support integrated graph building for visualization and sharing of data. D3 is the Javascript library used by Neo4J in the provided GUI, however this library can become problematic. Often the spring constant that gives a bounce effect to nodes in a graph can crash a browser window and cause it to become non-responsive as the number of nodes increases. This greatly hinders graph exploration. Another option that should be explored is graphviz [153]. This library is a well made graph application built in C with support for python, command-line, and many other languages. Graphviz can create static graphs as necessary with custom parameters. This would enhance the visibility of the ABCkb by enabling graph sharing in publications, presentations, and posters. Visualization of search results is necessary for the ABCkb to proceed into the next chapter.

5.1.3 Machine Learning and Linguistics

Currently, with the ABCkb 2.0, I2E is the sole generator of text-mined diet to disease connections. I2E is a rule and dictionary based method that works well to index and mine through large literature sources. However, as language changes, the rules and dictionaries must also change. There are several Machine Learning modeling options for generating these connections that should be explored. Researchers using both supervised and unsupervised mathematical models have seen success in vectorizing the text and discovering connections using these vectors, specifically in chemical compound detection and property prediction [154, 155]. I2E indexes can take a week or longer to complete and as science is constantly updated, one week is a long time. The expediency of query results is negated by index time length.

5.2 Ranking Algorithm

Future work on the KB should improve the efficiency for researchers developing testable hypotheses; therefore, a knowledgebase utilized for this purpose will only

be as useful as the results returned by a query. Many implementations for ranking graphs have been devised and utilized in other domains such as PageRank, HITS, HeteSim, and HetERel [156, 50, 157]. These algorithms derive their strengths from utilizing links with an assumption that an increase in links indicates an increase in accuracy or relevance. A drawback of these methods is that citation counts do not infer the accuracy of an article. For instance, according to the Web of Science, the infamous redacted Andrew Wakefield MMR vaccine paper has been cited over 1,300 times and is ranked in the top 0.04% of papers published in 1998 when ordered by number of citations. An algorithm may incorrectly judge the accuracy of relationships extracted from this paper if based on the number of citations. There are several ways to improve the output of a query including removing false positive matches, limiting returned nodes, and ordering by importance or relevance.

5.2.1 Previous ranking methods

Significant work has been done in the past to produce algorithms calculating semantic value of predicates between two nodes in the context of many domains. HeteSim was developed as a measure to calculate similarity scores using a pairwise random walk strategy based on the theory that relevant objects reference each other[157]. Richard Linchangco, who generated the original diet to disease KB, also developed an algorithm to calculate an edge-adjusted relevance of heterogeneous objects, HetERel, based on the same theory[18]. There are three steps to the HetERel algorithm. The first is Kulczynski Product Edge Weight (KPEW) calculation, which calculates the value of each predicate connected to each node extracted from text mining. Then the object relationships are converted to weighted adjacency matrices, with KPEW scores as weights, and are normalized across row vectors to produce Transition Probability Matrices (TPM). Finally a normalized score is calculated which provides a metric for relevance between nodes. The formula to calculate a normalized HetERel score is defined in equation 5.1 where a is the start node, c is the target node and

$U_P(n,:)$ is the n'th row in the transition probability matrix. Though HetERel calculates relevance between nodes within heterogeneous networks, it does not provide ranking metrics for query results in the scope of hypothesis generation and must be run at query time which increases the time and computational resource cost for each user.

$$HetERel(a,c|P) = \frac{U_{AB}(a,:) * U_{BC}(c,:)}{||U_{AB}(a,:)||_2 * ||U_{BC}(c,:)||_2} \tag{5.1}$$

5.3 Heuristically Evaluating Weighting Metrics

Successful implementation of a weighting metric will accomplish a few goals. As previously mentioned, the metric will need to be calculated at knowledgebase build time to reduce the time complexity of knowledgebase queries. In addition, a bounded metric between 0 and 1 will expedite relationship weight calculation. Finally, a query should reduce generic, vague, or arbitrary results. To determine if the metric is successful, standard test queries should be utilized and run against current ranking standards, the HetERel and HetESim metrics. Ideally, they should encompass jumps of various lengths to ensure that the algorithm is successfully portable as the knowledgebase grows and incorporates more data types.

5.3.1 Items for Algorithmic Consideration

5.3.1.1 Calculating Accuracy of Extracted Relationships

The accuracy of a text mined knowledgebase query depends on two factors, the extracted relationship from the source, and the accuracy of the source material. Evaluating the accuracy of extracted relationships through NLP queries are performed through manual efforts or automatic systems. Limitations of manual evaluations include the subjectivity of what is classified a "good" result by each individual, user background experience, and time[158]. Automatic methods derive their strength from good training and test data sets. NLP is by nature a classification problem for which

multiple forms of accuracy detection have been developed and utilized in other domains. The standard accuracy measure for text mining queries is a F-measure, which is the harmonic mean of precision and recall. Weighting precision and recall equally however is not always an optimal solution. To remedy this, the F-measure can be adjusted by a factor β to weight precision to recall accordingly to improve the evaluation of a NLP query as seen in equation 5.2. Let $\beta > 1$ for a higher emphasis on recall, and $\beta < 1$ for an emphasis on precision.

$$F_\beta = (1 + \beta) * \frac{PPV * TPR}{(\beta * PPV) + TPR} \tag{5.2}$$

With sample data shown in table 5.1, the calculated F_1-measure is: 0.84, the $F_{0.5}$-measure is: 0.85, and the F_2-measure is: 0.82. The threshold for an acceptable F-measure is 0.85. Object identification, text mining nodes should place a greater emphasis on precision and relationship extraction should place a greater emphasis on recall. This will yield more overall relationships which can be filtered for quality later on through a confidence value.

Table 5.1: Sample confusion matrix

	Condition Positive	Condition False
Predicted Positive	85	20
Predicted Negative	12	50

A confidence value for extracted relationships is traditionally defined in the context of graphs as a weight. Identifying which weight is appropriate for the graph can be performed heuristically by comparing current algorithms that have been applied in various domains. An important feature to consider is calculating weights at knowledgebase build time rather than at query time, as mentioned previously. This will greatly expedite hypothesis generation and increase the research value of the

knowledgebase and will lead into the generation of a question answering system.

5.4 Question Answering Systems

The final stage of future development for the knowledgebase is a question-answer system. The largest barrier to incorporating public databases in hypothesis generation is learning each query language. Structured Query Language (SQL) databases are the standard choice to house data. Access through an interface where users can explore the database contents is a common method with prebuilt, fill-in-the-blank queries. An example of this is the Comparative Toxicogenomics Database[15]. The drawback to this method are the jumps required to connect diet with disease through multiple resources. This poses two main problems, reproducibility and productivity. Reproducibility is difficult to maintain when interacting with various interfaces, as the underlying data changes which affects outputted query results. This has largely been solved in other domains, namely computer science with version control, and these solutions can be incorporated into an interface. The other issue is productivity, which is ultimately affected by the need to jump from databases and resources to generate a hypothesis from a research question. A well designed interface can solve these problems by merging databases and developing a Query Development system (QD) to write cypher queries, and lead to a Question Answering system (QandA).

The most common example of a QandA system is Watson from IBM which is an open-domain QandA system[159]. Open-domain systems are intended to produce answers to questions formed in Natural Language about any topic, in contrast to closed-domain systems which provide answers to limited domains. Watson was designed to take questions in Natural Language form, determine the entities requested, and produce an answer. Both open and closed domain QandA systems rely on ontologies to detect requested entities, similar to the ontologies used for text mining in I2E. Other examples of QandA systems include many of the personal assistants in phones and call-center answering machines which have an additional layer of abstrac-

tion to interpret vocal sounds. The first step in any QandA system is to interpret the question asked. Limiting the system to answering questions in English narrows the problem space.

5.4.1 Develop a Question Answering System

Two primary components of interpreting questions are answer type determination, and the type of entity asked, which Watson denotes Lexical Answer Type (LAT)[160]. Determining the type of answer required is more complicated than interrogating the initial word in the question and will require the use of a lexical dictionary such as, WordNet. An open-source WordNet solution is available with the Natural Language ToolKit in Python. Entity recognition is a large task in NLP systems and requires the use of ontologies. Many of the ontologies utilized for text mining should be used for this task. An additional step for interpreting questions is entity recognition. AllenAI has several machine learning models available for entity recognition [161]. The complete library is available in python and is open source (https://github.com/allenai/allennlp).

5.4.2 Provide Answers

Answers should be produced by querying the KB utilizing the interpretation of each question. Queries can either be prewritten with substitution from detected entities or developed dynamically. A drawback to prewritten queries is the limitation to answerable questions. In this case, the questions asked must fit into a limited range of queries. In contrast, dynamically developed queries may prove more useful as the KB expands to cover a wider range of data.

5.5 Conclusion

This work was motivated by a desire to provide molecular mechanisms and discover the bidirectional connections between microbiota and diet. There are many directions available for the knowledgebase to expand and increase utility of the tool.

One hindrance has been the lack of a testable use-case from hypothesis generation to a clinical trial or bench-top study. Any further improvements to the knowledgebase will require the collaboration of partners willing to participate in the analysis of the generated hypotheses. In addition, the knowledgebase has grown to a point where a multidisciplinary support team for maintaining and updating source information is required. As the rates of non-communicable, diet-related diseases rise, the knowledgebase will facilitate development of nutritive solutions with molecular mechanistic explanations. This will lead to novel solutions that are nonsynthetic, natural, reduce the medical burdens of patients suffering from these diseases, and improve the quality of life for these patients.

REFERENCES

[1] C. Canavan, J. West, and T. Card, "The epidemiology of irritable bowel syndrome," *Clinical Epidemiology*, vol. 6, pp. 71–80, Feb. 2014.

[2] World Health Organization, "Non communicable diseases," Apr. 2021.

[3] H. Greenberg, "Diet and Non-Communicable 105 Diseases: An urgent need for new paradigms," p. 14.

[4] F. A. Olatona, O. O. Onabanjo, R. N. Ugbaja, K. E. Nnoaham, and D. A. Adelekan, "Dietary habits and metabolic risk factors for non-communicable diseases in a university undergraduate population," *Journal of Health, Population and Nutrition*, vol. 37, p. 21, Aug. 2018.

[5] J. Hunter-Adams and J. Battersby, "Health care providers' perspectives of diet-related non-communicable disease in South Africa," *BMC Public Health*, vol. 20, p. 262, Feb. 2020.

[6] A. Agus, K. Clement, and H. Sokol, "Gut microbiota-derived metabolites as central regulators in metabolic disorders," *Gut*, vol. 70, pp. 1174–1182, June 2021. Publisher: BMJ Publishing Group Section: Recent advances in basic science.

[7] K. Oliphant and E. Allen-Vercoe, "Macronutrient metabolism by the human gut microbiome: major fermentation by-products and their impact on host health," *Microbiome*, vol. 7, p. 91, June 2019.

[8] N. R. Smalheiser, "Rediscovering Don Swanson:The Past, Present and Future of Literature-based Discovery," *Journal of Data and Information Science*, vol. 2, pp. 43–64, Dec. 2017.

[9] D. R. Swanson, "Fish Oil, Raynaud's Syndrome, and Undiscovered Public Knowledge," *Perspectives in Biology and Medicine*, vol. 30, pp. 7–18, Jan. 2015.

[10] E. W. Sayers, J. Beck, E. E. Bolton, D. Bourexis, J. R. Brister, K. Canese, D. C. Comeau, K. Funk, S. Kim, W. Klimke, A. Marchler-Bauer, M. Landrum, S. Lathrop, Z. Lu, T. L. Madden, N. O'Leary, L. Phan, S. H. Rangwala, V. A. Schneider, Y. Skripchenko, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, and S. T. Sherry, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Research*, vol. 49, pp. D10–D17, Jan. 2021.

[11] D. J. Rigden and X. M. Fernendez, "The 2021 Nucleic Acids Research database issue and the online molecular biology database collection," *Nucleic Acids Research*, vol. 49, pp. D1–D9, Jan. 2021.

[12] A. Trautman, R. Linchangco, R. Walstead, J. J. Jay, and C. Brouwer, "The Aliment to Bodily Condition knowledgebase (ABCkb): A database connecting plants and human health," tech. rep., Mar. 2021. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[13] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki, "NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods," *Nucleic Acids Research*, vol. 43, pp. D940–D945, Jan. 2015. Publisher: Oxford Academic.

[14] Y. Ni, K. Jensen, I. Kouskoumvekaki, and G. Panagiotou, "NutriChem 2.0: exploring the effect of plant-based foods on human health and drug efficacy," *Database: The Journal of Biological Databases and Curation*, vol. 2017, p. bax044, June 2017.

[15] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers, and C. J. Mattingly, "The Comparative Toxicogenomics Database: update 2019," *Nucleic Acids Research*, vol. 47, pp. D948–D954, Jan. 2019. Publisher: Oxford Academic.

[16] J. A. Rothwell, J. Perez-Jimenez, V. Neveu, A. Medina Remon, N. M'Hiri, P. Garcia-Lobato, C. Manach, C. Knox, R. Eisner, D. S. Wishart, and A. Scalbert, "Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content," *Database*, vol. 2013, Jan. 2013.

[17] "FooDB."

[18] R. V. Linchangco, *The Semantics of Diet And Health: Knowledge Based Discovery Through Data Integration, Text Mining, and Network Analysis*. Ph.D., The University of North Carolina at Charlotte, United States – North Carolina, 2018.

[19] L. K. Ursell, J. L. Metcalf, L. W. Parfrey, and R. Knight, "Defining the Human Microbiome," *Nutrition reviews*, vol. 70, pp. S38–S44, Aug. 2012.

[20] S. L. Prescott, "History of medicine: Origin of the term microbiome and why it matters," *Human Microbiome Journal*, vol. 4, pp. 24–25, June 2017.

[21] R. I. Aminov, "A Brief History of the Antibiotic Era: Lessons Learned and Challenges for the Future," *Frontiers in Microbiology*, vol. 1, Dec. 2010.

[22] S. T. Shulman, H. C. Friedmann, and R. H. Sims, "Theodor Escherich: The First Pediatric Infectious Diseases Physician?," *Clinical Infectious Diseases*, vol. 45, pp. 1025–1029, Oct. 2007.

[23] J. Lloyd-Price, G. Abu-Ali, and C. Huttenhower, "The healthy human microbiome," *Genome Medicine*, vol. 8, Apr. 2016.

[24] M. Rautio, E. Eerola, M.-L. Väisänen-Tunkelrott, D. Molitoris, P. Lawson, M. D. Collins, and H. Jousimies-Somer, "Reclassification of Bacteroides putredinis (Weinberg et al., 1937) in a New Genus Alistipes gen. nov., as Alistipes putredinis comb. nov., and Description of Alistipes finegoldii sp. nov., from Human Sources," *Systematic and Applied Microbiology*, vol. 26, pp. 182–188, Jan. 2003.

[25] J. S. Johnson, D. J. Spakowicz, B.-Y. Hong, L. M. Petersen, P. Demkowicz, L. Chen, S. R. Leopold, B. M. Hanson, H. O. Agresta, M. Gerstein, E. Sodergren, and G. M. Weinstock, "Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis," *Nature Communications*, vol. 10, p. 5029, Nov. 2019.

[26] T. J. Sharpton, "An introduction to the analysis of shotgun metagenomic data," *Frontiers in Plant Science*, vol. 5, p. 209, June 2014.

[27] M. Shakya, C.-C. Lo, and P. S. G. Chain, "Advances and Challenges in Metatranscriptomic Analysis," *Frontiers in Genetics*, vol. 10, p. 904, 2019.

[28] J. M. Janda and S. L. Abbott, "16S rRNA Gene Sequencing for Bacterial Identification in the Diagnostic Laboratory: Pluses, Perils, and Pitfalls," *Journal of Clinical Microbiology*, vol. 45, pp. 2761–2764, Sept. 2007. Publisher: American Society for Microbiology.

[29] M. Pichler, O. K. Coskun, A. Ortega-Arbulu, N. Conci, G. Worheide, S. Vargas, and W. D. Orsi, "A 16S rRNA gene sequencing and analysis protocol for the Illumina MiniSeq platform," *MicrobiologyOpen*, vol. 7, p. e00611, Mar. 2018.

[30] "16S Sequencing vs Shotgun Metagenomic Sequencing."

[31] R. Bharti and D. G. Grimm, "Current challenges and best-practice protocols for microbiome analysis," *Briefings in Bioinformatics*, vol. 22, pp. 178–193, Jan. 2021.

[32] M. O. Press, A. H. Wiser, Z. N. Kronenberg, K. W. Langford, M. Shakya, C.-C. Lo, K. A. Mueller, S. T. Sullivan, P. S. G. Chain, and I. Liachko, "Hi-C deconvolution of a human gut microbiome yields high-quality draft genomes and reveals plasmid-genome interactions," tech. rep., Oct. 2017. Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article.

[33] S. Bahmani, N. Azarpira, and E. Moazamian, "Anti-colon cancer activity of Bifidobacterium metabolites on colon cancer cell line SW742," *The Turkish Journal of Gastroenterology: The Official Journal of Turkish Society of Gastroenterology*, vol. 30, pp. 835–842, Sept. 2019.

[34] D. Y. Graham, "History of Helicobacter pylori, duodenal ulcer, gastric ulcer and gastric cancer," *World Journal of Gastroenterology : WJG*, vol. 20, pp. 5191–5204, May 2014.

[35] P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, and J. I. Gordon, "The Human Microbiome Project," *Nature*, vol. 449, pp. 804–810, Oct. 2007.

[36] P. Markowiak and K. Śliżewska, "Effects of Probiotics, Prebiotics, and Synbiotics on Human Health," *Nutrients*, vol. 9, p. E1021, Sept. 2017.

[37] E. National Academies of Sciences, D. o. E. a. L. Studies, B. o. L. Sciences, B. o. E. S. Toxicology, , and C. o. A. U. o. t. I. o. E.-C. I. w. t. H. Microbiome, *Microbiome Variation*. National Academies Press (US), Dec. 2017. Publication Title: Environmental Chemicals, the Human Microbiome, and Health Risk: A Research Strategy.

[38] L. B. Bindels, N. M. Delzenne, P. D. Cani, and J. Walter, "Towards a more comprehensive concept for prebiotics," *Nature Reviews. Gastroenterology & Hepatology*, vol. 12, pp. 303–310, May 2015.

[39] C. Grootaert, J. A. Delcour, C. M. Courtin, W. F. Broekaert, W. Verstraete, and T. Van de Wiele, "Microbial metabolism and prebiotic potency of arabinoxylan oligosaccharides in the human intestine," *Trends in Food Science & Technology*, vol. 18, pp. 64–71, Feb. 2007.

[40] S. B. R. d. Prado, V. C. Castro-Alves, G. F. Ferreira, and J. P. Fabi, "Ingestion of Non-digestible Carbohydrates From Plant-Source Foods and Decreased Risk of Colorectal Cancer: A Review on the Biological Effects and the Mechanisms of Action," *Frontiers in Nutrition*, vol. 6, 2019. Publisher: Frontiers.

[41] H. H. Creasy, V. Felix, J. Aluvathingal, J. Crabtree, O. Ifeonu, J. Matsumura, C. McCracken, L. Nickel, J. Orvis, M. Schor, M. Giglio, A. Mahurkar, and O. White, "HMPDACC: a Human Microbiome Project Multi-omic data resource," *Nucleic Acids Research*, vol. 49, pp. D734–D742, Jan. 2021.

[42] Y. Janssens, J. Nielandt, A. Bronselaer, N. Debunne, F. Verbeke, E. Wynendaele, F. Van Immerseel, Y.-P. Vandewynckel, G. De Tré, and B. De Spiegeleer, "Disbiome database: linking the microbiome to disease," *BMC Microbiology*, vol. 18, p. 50, June 2018.

[43] L. Cheng, C. Qi, H. Zhuang, T. Fu, and X. Zhang, "gutMDisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions," *Nucleic Acids Research*, vol. 48, pp. D554–D560, Jan. 2020.

[44] S. Leonelli, "The challenges of big data biology," *eLife*, vol. 8.

[45] E. Birney and M. Clamp, "Biological database design and implementation," *Briefings in Bioinformatics*, vol. 5, pp. 31–38, Mar. 2004.

[46] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "The role of ontologies in biological and biomedical research: a functional perspective," *Briefings in Bioinformatics*, vol. 16, pp. 1069–1080, Nov. 2015.

[47] L. M. Schriml, E. Mitraka, J. Munro, B. Tauber, M. Schor, L. Nickle, V. Felix, L. Jeng, C. Bearer, R. Lichenstein, K. Bisordi, N. Campion, B. Hyman, D. Kurland, C. P. Oates, S. Kibbey, P. Sreekumar, C. Le, M. Giglio, and C. Greene, "Human Disease Ontology 2018 update: classification, content and workflow expansion," *Nucleic Acids Research*, vol. 47, no. D1, pp. D955–D962, 2019.

[48] S. Kohler, L. Carmody, N. Vasilevsky, J. O. B. Jacobsen, D. Danis, J.-P. Gourdine, M. Gargano, N. L. Harris, N. Matentzoglu, J. A. McMurry, D. Osumi-Sutherland, V. Cipriani, J. P. Balhoff, T. Conlin, H. Blau, G. Baynam, R. Palmer, D. Gratian, H. Dawkins, M. Segal, A. C. Jansen, A. Muaz, W. H. Chang, J. Bergerson, S. J. F. Laulederkind, Z. Yüksel, S. Beltran, A. F. Freeman, P. I. Sergouniotis, D. Durkin, A. L. Storm, M. Hanauer, M. Brudno, S. M. Bello, M. Sincan, K. Rageth, M. T. Wheeler, R. Oegema, H. Lourghi, M. G. Della Rocca, R. Thompson, F. Castellanos, J. Priest, C. Cunningham-Rundles, A. Hegde, R. C. Lovering, C. Hajek, A. Olry, L. Notarangelo, M. Similuk, X. A. Zhang, D. Gomez-Andres, H. Lochmuller, H. Dollfus, S. Rosenzweig, S. Marwaha, A. Rath, K. Sullivan, C. Smith, J. D. Milner, D. Leroux, C. F. Boerkoel, A. Klion, M. C. Carter, T. Groza, D. Smedley, M. A. Haendel, C. Mungall, and P. N. Robinson, "Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources," *Nucleic Acids Research*, vol. 47, pp. D1018–D1027, Jan. 2019. Publisher: Oxford Academic.

[49] "Medical Subject Headings - Home Page." Library Catalog: www.nlm.nih.gov Publisher: U.S. National Library of Medicine.

[50] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. 1998.

[51] E. J. Yates and L. C. Dixon, "PageRank as a method to rank biomedical literature by importance," *Source Code for Biology and Medicine*, vol. 10, p. 16, Dec. 2015.

[52] C. Engstrom, "PageRank in Evolving Networks and Applications of Graphs in Natural Language Processing and Biology," 2016.

[53] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," July 2004.

[54] S. Bolasco and P. Pavone, "Automatic Dictionary- and Rule-Based Systems for Extracting Information from Text," in *Data Analysis and Classification* (F. Palumbo, C. N. Lauro, and M. J. Greenacre, eds.), Studies in Classification, Data Analysis, and Knowledge Organization, (Berlin, Heidelberg), pp. 189–198, Springer, 2010.

[55] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Research*, vol. 41, pp. W518–W522, July 2013.

[56] J. Laster and L. A. Frame, "Beyond the Calories—Is the Problem in the Processing?," *Current Treatment Options in Gastroenterology*, vol. 17, pp. 577–586, Dec. 2019.

[57] B. M. Popkin, "Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases," *The American Journal of Clinical Nutrition*, vol. 84, pp. 289–298, Aug. 2006. Publisher: Oxford Academic.

[58] M. B. Schulze, M. A. Martínez-González, T. T. Fung, A. H. Lichtenstein, and N. G. Forouhi, "Food based dietary patterns and chronic disease prevention," *BMJ*, vol. 361, June 2018.

[59] R. W. Reid, C. R. Brouwer, E. W. Jackson, and M. A. Lila, "A need for a transdisciplinary environment: the Plant Pathways Elucidation Project," *Trends in Plant Science*, vol. 19, pp. 485–487, Aug. 2014. Publisher: Elsevier.

[60] J. Hastings, G. Owen, A. Dekker, M. Ennis, N. Kale, V. Muthukrishnan, S. Turner, N. Swainston, P. Mendes, and C. Steinbeck, "ChEBI in 2016: Improved services and an expanding collection of metabolites," *Nucleic Acids Research*, vol. 44, pp. D1214–D1219, Jan. 2016. Publisher: Oxford Academic.

[61] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene Ontology: tool for the unification of biology," *Nature genetics*, vol. 25, pp. 25–29, May 2000.

[62] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still GOing strong," *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, 2019.

[63] C. J. Mungall, J. A. McMurry, S. Kohler, J. P. Balhoff, C. Borromeo, M. Brush, S. Carbon, T. Conlin, N. Dunn, M. Engelstad, E. Foster, J. P. Gourdine, J. O. B. Jacobsen, D. Keith, B. Laraway, S. E. Lewis, J. NguyenXuan, K. Shefchek, N. Vasilevsky, Z. Yuan, N. Washington, H. Hochheiser, T. Groza, D. Smedley, P. N. Robinson, and M. A. Haendel, "The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic Acids Research*, vol. 45, pp. D712–D722, Jan. 2017. Publisher: Oxford Academic.

[64] "Agricultural Thesaurus and Glossary Home Page."

[65] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Research*, vol. 40, pp. D136–143, Jan. 2012.

[66] E. W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K. D. Pruitt, and I. Karsch-Mizrachi, "GenBank," *Nucleic Acids Research*, vol. 47, no. D1, pp. D94–D99, 2019.

[67] G. R. Brown, V. Hem, K. S. Katz, M. Ovetsky, C. Wallin, O. Ermolaeva, I. Tolstoy, T. Tatusova, K. D. Pruitt, D. R. Maglott, and T. D. Murphy, "Gene: a gene-centered information resource at NCBI," *Nucleic Acids Research*, vol. 43, pp. D36–D42, Jan. 2015.

[68] M. Halavi, D. Maglott, V. Gorelenkov, and W. Rubinstein, *MedGen*. National Center for Biotechnology Information (US), Dec. 2018. Publication Title: The NCBI Handbook [Internet]. 2nd edition.

[69] T. M. S. Wolever and D. R. a. R. Duss, "Oat B-Glucan Reduces Serum LDL Cholesterol in Humans with Serum LDL Cholesterol $<$ 160mg/dL," May 2016.

[70] L. Guo, L.-T. Tong, L. Liu, K. Zhong, J. Qiu, and S. Zhou, "The cholesterol-lowering effects of oat varieties based on their difference in the composition of proteins and lipids," *Lipids in Health and Disease*, vol. 13, Dec. 2014.

[71] B. Verges, "New insight into the pathophysiology of lipid abnormalities in type 2 diabetes," *Diabetes & Metabolism*, vol. 31, pp. 429–439, Nov. 2005.

[72] D. R. Pokharel, D. Khadka, M. Sigdel, N. K. Yadav, S. Acharya, R. Kafle, R. M. Sapkota, and T. Sigdel, "Prevalence and pattern of dyslipidemia in Nepalese individuals with type 2 diabetes," *BMC research notes*, vol. 10, p. 146, Apr. 2017.

[73] M. J. Shahwan, A. A. Jairoun, A. Farajallah, and S. Shanabli, "Prevalence of dyslipidemia and factors affecting lipid profile in patients with type 2 diabetes," *Diabetes & Metabolic Syndrome*, vol. 13, pp. 2387–2392, Aug. 2019.

[74] S. K. Paulsen, S. B. Pedersen, S. Fisker, and B. Richelsen, "11Beta-HSD type 1 expression in human adipose tissue: impact of gender, obesity, and fat localization," *Obesity (Silver Spring, Md.)*, vol. 15, pp. 1954–1960, Aug. 2007.

[75] C. Dammann, C. Stapelfeld, and E. Maser, "Expression and activity of the cortisol-activating enzyme 11B-hydroxysteroid dehydrogenase type 1 is tissue and species-specific," *Chemico-Biological Interactions*, vol. 303, pp. 57–61, Apr. 2019.

[76] L. V. Turek, N. Leite, R. L. Rodrigues Souza, J. K. Lima, G. E. Milano, L. d. S. Timossi, A. C. V. Osiecki, R. Osiecki, and L. F. Alle, "Gender-dependent association of HSD11B1 single nucleotide polymorphisms with glucose and HDL-C levels," *Genetics and Molecular Biology*, vol. 37, pp. 490–495, Sept. 2014.

[77] S. Nair, Y. H. Lee, R. S. Lindsay, B. R. Walker, P. A. Tataranni, C. Bogardus, L. J. Baier, and P. A. Permana, "11beta-Hydroxysteroid dehydrogenase Type 1: genetic polymorphisms are associated with Type 2 diabetes in Pima Indians independently of obesity and expression in adipocyte and muscle," *Diabetologia*, vol. 47, pp. 1088–1095, June 2004.

[78] A. Gambineri, F. Tomassoni, A. Munarini, R. H. Stimson, R. Mioni, U. Pagotto, K. E. Chapman, R. Andrew, V. Mantovani, R. Pasquali, and B. R. Walker, "A combination of polymorphisms in HSD11B1 associates with in vivo 11{beta}-HSD1 activity and metabolic syndrome in women with and without polycystic ovary syndrome," *European Journal of Endocrinology*, vol. 165, pp. 283–292, Aug. 2011.

[79] D. S. Freedman, B. A. Bowman, S. R. Srinivasan, G. S. Berenson, and J. D. Otvos, "Distribution and correlates of high-density lipoprotein subclasses among children and adolescents," *Metabolism: Clinical and Experimental*, vol. 50, pp. 370–376, Mar. 2001.

[80] D. C. Goff, R. B. D'Agostino, S. M. Haffner, and J. D. Otvos, "Insulin resistance and adiposity influence lipoprotein size and subclass concentrations. Results from the Insulin Resistance Atherosclerosis Study," *Metabolism: Clinical and Experimental*, vol. 54, pp. 264–270, Feb. 2005.

[81] R. Abdul Rani, R. A. Raja Ali, and Y. Y. Lee, "Irritable bowel syndrome and inflammatory bowel disease overlap syndrome: pieces of the puzzle are falling into place," *Intestinal Research*, vol. 14, pp. 297–304, Oct. 2016.

[82] Z. D. Wallen, M. Appah, M. N. Dean, C. L. Sesler, S. A. Factor, E. Molho, C. P. Zabetian, D. G. Standaert, and H. Payami, "Characterizing dysbiosis of gut microbiome in PD: evidence for overabundance of opportunistic pathogens," *npj Parkinson's Disease*, vol. 6, pp. 1–12, June 2020. Number: 1 Publisher: Nature Publishing Group.

[83] D. R. Mack, "Probiotics," *Canadian Family Physician*, vol. 51, pp. 1455–1457, Nov. 2005.

[84] D. Davani-Davari, M. Negahdaripour, I. Karimzadeh, M. Seifan, M. Mohkam, S. J. Masoumi, A. Berenjian, and Y. Ghasemi, "Prebiotics: Definition, Types, Sources, Mechanisms, and Clinical Applications," *Foods*, vol. 8, p. 92, Mar. 2019.

[85] E. P. Halmos, C. T. Christophersen, A. R. Bird, S. J. Shepherd, P. R. Gibson, and J. G. Muir, "Diets that differ in their FODMAP content alter the colonic luminal microenvironment," *Gut*, vol. 64, pp. 93–100, Jan. 2015.

[86] R. K. Singh, H.-W. Chang, D. Yan, K. M. Lee, D. Ucmak, K. Wong, M. Abrouk, B. Farahnik, M. Nakamura, T. H. Zhu, T. Bhutani, and W. Liao, "Influence of diet on the gut microbiome and implications for human health," *Journal of Translational Medicine*, vol. 15, p. 73, Apr. 2017.

[87] S. Goertz, A. B. d. Menezes, R. J. Birtles, J. Fenn, A. E. Lowe, A. D. C. MacColl, B. Poulin, S. Young, J. E. Bradley, and C. H. Taylor, "Geographical location influences the composition of the gut microbiota in wild house

mice (Mus musculus domesticus) at a fine spatial scale," *PLOS ONE*, vol. 14, p. e0222501, Sept. 2019. Publisher: Public Library of Science.

[88] V. K. Gupta, S. Paul, and C. Dutta, "Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity," *Frontiers in Microbiology*, vol. 8, p. 1162, June 2017.

[89] D. Westergaard, H.-H. Staerfeldt, C. Tonsberg, L. J. Jensen, and S. Brunak, "Text mining of 15 million full-text scientific articles," *bioRxiv*, p. 162099, July 2017.

[90] 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 Computational Biology*, vol. 14, p. e1005962, Feb. 2018. Publisher: Public Library of Science.

[91] F. Khan, M. M. R. Sarker, L. C. Ming, I. N. Mohamed, C. Zhao, B. Y. Sheikh, H. F. Tsong, and M. A. Rashid, "Comprehensive Review on Phytochemicals, Pharmacological and Clinical Potentials of Gymnema sylvestre," *Frontiers in Pharmacology*, vol. 10, 2019.

[92] R. Guan, Q. Van Le, H. Yang, D. Zhang, H. Gu, Y. Yang, C. Sonne, S. S. Lam, J. Zhong, Z. Jianguang, R. Liu, and W. Peng, "A review of dietary phytochemicals and their relation to oxidative stress and human diseases," *Chemosphere*, vol. 271, p. 129499, May 2021.

[93] R. Verkerk, M. Schreiner, A. Krumbein, E. Ciska, B. Holst, I. Rowland, R. De Schrijver, M. Hansen, C. Gerhäuser, R. Mithen, and M. Dekker, "Glucosinolates in Brassica vegetables: the influence of the food supply chain on intake, bioavailability and human health," *Molecular Nutrition & Food Research*, vol. 53 Suppl 2, p. S219, Sept. 2009.

[94] A. Mazumder, A. Dwivedi, and J. du Plessis, "Sinigrin and Its Therapeutic Benefits," *Molecules (Basel, Switzerland)*, vol. 21, p. 416, Mar. 2016.

[95] O. Vang, J. Mortensen, and O. Andersen, "Biochemical effects of dietary intake of different broccoli samples. II. Multivariate analysis of contributions of specific glucosinolates in modulating cytochrome P-450 and antioxidant defense enzyme activities," *Metabolism: Clinical and Experimental*, vol. 50, pp. 1130–1135, Oct. 2001.

[96] B. Wang, L. Wu, J. Chen, L. Dong, C. Chen, Z. Wen, J. Hu, I. Fleming, and D. W. Wang, "Metabolism pathways of arachidonic acids: mechanisms and potential therapeutic targets," *Signal Transduction and Targeted Therapy*, vol. 6, pp. 1–30, Feb. 2021. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Cancer;Cardiovascular diseases Subject_term_id: cancer;cardiovascular-diseases.

[97] A. J. Higgins and P. Lees, "The acute inflammatory process, arachidonic acid metabolism and the mode of action of anti-inflammatory drugs," *Equine Veterinary Journal*, vol. 16, pp. 163–175, May 1984.

[98] H. Tallima and R. El Ridi, "Arachidonic acid: Physiological roles and potential health benefits – A review," *Journal of Advanced Research*, vol. 11, pp. 33–41, Nov. 2017.

[99] J. W. Fahey, W. D. Holtzclaw, S. L. Wehage, K. L. Wade, K. K. Stephenson, and P. Talalay, "Sulforaphane Bioavailability from Glucoraphanin-Rich Broccoli: Control by Active Endogenous Myrosinase," *PLoS ONE*, vol. 10, p. e0140963, Nov. 2015.

[100] D.-L. Cheng, K. Hashimoto, and Y. Uda, "In vitro digestion of sinigrin and glucotropaeolin by single strains of Bifidobacterium and identification of the digestive products," *Food and Chemical Toxicology: An International Journal Published for the British Industrial Biological Research Association*, vol. 42, pp. 351–357, Mar. 2004.

[101] B. J. Parker, P. A. Wearsch, A. C. M. Veloo, and A. Rodriguez-Palacios, "The Genus Alistipes: Gut Bacteria With Emerging Implications to Inflammation, Cancer, and Mental Health," *Frontiers in Immunology*, vol. 11, 2020.

[102] K. Nomura, D. Ishikawa, K. Okahara, S. Ito, K. Haga, M. Takahashi, A. Arakawa, T. Shibuya, T. Osada, K. Kuwahara-Arai, T. Kirikae, and A. Nagahara, "Bacteroidetes Species Are Correlated with Disease Activity in Ulcerative Colitis," *Journal of Clinical Medicine*, vol. 10, p. 1749, Apr. 2021.

[103] S. Louis, R.-M. Tappu, A. Damms-Machado, D. H. Huson, and S. C. Bischoff, "Characterization of the Gut Microbial Community of Obese Patients Following a Weight-Loss Intervention Using Whole Metagenome Shotgun Sequencing," *PLoS ONE*, vol. 11, p. e0149564, Feb. 2016.

[104] R. Sarwar, N. Pierce, and S. Koppe, "Obesity and nonalcoholic fatty liver disease: current perspectives," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 11, pp. 533–542, Sept. 2018.

[105] C. A. Pickens, L. M. Sordillo, C. Zhang, and J. I. Fenton, "Obesity is positively associated with arachidonic acid-derived 5- and 11-hydroxyeicosatetraenoic acid (HETE)," *Metabolism: Clinical and Experimental*, vol. 70, pp. 177–191, May 2017.

[106] Y. Naito, K. Uchiyama, and T. Takagi, "A next-generation beneficial microbe: Akkermansia muciniphila," *Journal of Clinical Biochemistry and Nutrition*, vol. 63, pp. 33–35, July 2018.

[107] M. C. Dao, A. Everard, J. Aron-Wisnewsky, N. Sokolovska, E. Prifti, E. O. Verger, B. D. Kayser, F. Levenez, J. Chilloux, L. Hoyles, M.-O. Consortium,

M.-E. Dumas, S. W. Rizkalla, J. Doré, P. D. Cani, and K. Clément, "Akkermansia muciniphila and improved metabolic health during a dietary intervention in obesity: relationship with gut microbiome richness and ecology," *Gut*, vol. 65, pp. 426–436, Mar. 2016. Publisher: BMJ Publishing Group Section: Gut microbiota.

[108] S. Y. Geerlings, I. Kostopoulos, W. M. de Vos, and C. Belzer, "Akkermansia muciniphila in the Human Gastrointestinal Tract: When, Where, and How?," *Microorganisms*, vol. 6, p. 75, July 2018.

[109] E. E. Blaak, E. E. Canfora, S. Theis, G. Frost, A. K. Groen, G. Mithieux, A. Nauta, K. Scott, B. Stahl, J. van Harsselaar, R. van Tol, E. E. Vaughan, and K. Verbeke, "Short chain fatty acids in human gut and metabolic health," *Beneficial Microbes*, vol. 11, pp. 411–455, Sept. 2020.

[110] M. Derrien, E. E. Vaughan, C. M. Plugge, and W. M. de Vos, "Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium," *International Journal of Systematic and Evolutionary Microbiology*, vol. 54, pp. 1469–1476, Sept. 2004.

[111] F. Van Herreweghen, K. De Paepe, H. Roume, F.-M. Kerckhof, and T. Van de Wiele, "Mucin degradation niche as a driver of microbiome composition and Akkermansia muciniphila abundance in a dynamic gut model is donor independent," *FEMS Microbiology Ecology*, vol. 94, p. fiy186, Dec. 2018.

[112] K. E. Norin, B. E. Gustafsson, B. S. Lindblad, and T. Midtvedt, "The establishment of some microflora associated biochemical characteristics in feces from children during the first years of life," *Acta Paediatrica Scandinavica*, vol. 74, pp. 207–212, Mar. 1985.

[113] E. Gaudier, A. Jarry, H. M. Blottière, P. de Coppet, M. P. Buisine, J. P. Aubert, C. Laboisse, C. Cherbut, and C. Hoebler, "Butyrate specifically modulates MUC gene expression in intestinal epithelial goblet cells deprived of glucose," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 287, pp. G1168–G1174, Dec. 2004. Publisher: American Physiological Society.

[114] J. A. Jiminez, T. C. Uwiera, D. W. Abbott, R. R. E. Uwiera, and G. D. Inglis, "Butyrate Supplementation at High Concentrations Alters Enteric Bacterial Communities and Reduces Intestinal Inflammation in Mice Infected with Citrobacter rodentium," *mSphere*, vol. 2, Aug. 2017.

[115] Y. P. Silva, A. Bernardi, and R. L. Frozza, "The Role of Short-Chain Fatty Acids From Gut Microbiota in Gut-Brain Communication," *Frontiers in Endocrinology*, vol. 11, 2020.

[116] F. F. Anhê, G. Pilon, D. Roy, Y. Desjardins, E. Levy, and A. Marette, "Triggering Akkermansia with dietary polyphenols: A new weapon to combat the metabolic syndrome?," *Gut Microbes*, vol. 7, no. 2, pp. 146–153, 2016.

[117] J. Pérez-Jiménez, V. Neveu, F. Vos, and A. Scalbert, "Identification of the 100 richest dietary sources of polyphenols: an application of the Phenol-Explorer database," *European Journal of Clinical Nutrition*, vol. 64, pp. S112–S120, Nov. 2010. Number: 3 Publisher: Nature Publishing Group.

[118] J. Joven, V. Micol, A. Segura-Carretero, C. Alonso-Villaverde, J. A. Menéndez, and Bioactive Food Components Platform, "Polyphenols and the modulation of gene expression pathways: can we eat our way out of the danger of chronic disease?," *Critical Reviews in Food Science and Nutrition*, vol. 54, no. 8, pp. 985–1001, 2014.

[119] H. Cory, S. Passarelli, J. Szeto, M. Tamez, and J. Mattei, "The Role of Polyphenols in Human Health and Food Systems: A Mini-Review," *Frontiers in Nutrition*, vol. 5, p. 87, Sept. 2018.

[120] P. Jeandet, C. Clément, and S. Cordelier, "Regulation of resveratrol biosynthesis in grapevine: new approaches for disease resistance?," *Journal of Experimental Botany*, vol. 70, pp. 375–378, Jan. 2019.

[121] X. Zhang, H. Mallick, Z. Tang, L. Zhang, X. Cui, A. K. Benson, and N. Yi, "Negative binomial mixed models for analyzing microbiome count data," *BMC bioinformatics*, vol. 18, p. 4, Jan. 2017.

[122] D. G. Bailey and G. K. Dresser, "Interactions between grapefruit juice and cardiovascular drugs," *American Journal of Cardiovascular Drugs: Drugs, Devices, and Other Interventions*, vol. 4, no. 5, pp. 281–297, 2004.

[123] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, and V. Thiel, "Coronavirus biology and replication: implications for SARS-CoV-2," *Nature Reviews Microbiology*, vol. 19, pp. 155–170, Mar. 2021. Number: 3 Publisher: Nature Publishing Group.

[124] M. Rastogi, N. Pandey, A. Shukla, and S. K. Singh, "SARS coronavirus 2: from genome to infectome," *Respiratory Research*, vol. 21, p. 318, Dec. 2020.

[125] M. A. Marra, S. J. M. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. N. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham,

M. Krajden, M. Petric, D. M. Skowronski, C. Upton, and R. L. Roper, "The Genome sequence of the SARS-associated coronavirus," *Science (New York, N.Y.)*, vol. 300, pp. 1399–1404, May 2003.

[126] "CDC Media Relations: Press Release."

[127] K. Yuki, M. Fujiogi, and S. Koutsogiannaki, "COVID-19 pathophysiology: A review," *Clinical Immunology (Orlando, Fla.)*, vol. 215, p. 108427, June 2020.

[128] M. Scudellari, "How the coronavirus infects cells — and why Delta is so dangerous," *Nature*, vol. 595, pp. 640–644, July 2021. Bandiera_abtest: a Cg_type: News Feature Number: 7869 Publisher: Nature Publishing Group Subject_term: SARS-CoV-2, Virology.

[129] N. Fraser, L. Brierley, G. Dey, J. K. Polka, M. Pálfy, F. Nanni, and J. A. Coates, "The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape," *PLOS Biology*, vol. 19, p. e3000959, Apr. 2021. Publisher: Public Library of Science.

[130] T. Sheldon, "Preprints could promote confusion and distortion," *Nature*, vol. 559, pp. 445–445, July 2018. Bandiera_abtest: a Cg_type: World View Number: 7715 Publisher: Nature Publishing Group Subject_term: Publishing, Society, Communication.

[131] C. F. D. Carneiro, V. G. S. Queiroz, T. C. Moulin, C. A. M. Carvalho, C. B. Haas, D. Rayêe, D. E. Henshall, E. A. De-Souza, F. E. Amorim, F. Z. Boos, G. D. Guercio, I. R. Costa, K. L. Hajdu, L. van Egmond, M. Modrák, P. B. Tan, R. J. Abdill, S. J. Burgess, S. F. S. Guerra, V. T. Bortoluzzi, and O. B. Amaral, "Comparing quality of reporting between preprints and peer-reviewed articles in the biomedical literature," *Research Integrity and Peer Review*, vol. 5, p. 16, Dec. 2020.

[132] N. K. Fry, H. Marshall, and T. Mellins-Cohen, "In praise of preprints," *Microbial Genomics*, vol. 5, p. e000259, Apr. 2019.

[133] R. Van Noorden, "Publishers withdraw more than 120 gibberish papers," *Nature*, Feb. 2014. Publisher: Nature Publishing Group.

[134] L. Brierley, "Lessons from the influx of preprints during the early COVID-19 pandemic," *The Lancet Planetary Health*, vol. 5, pp. e115–e117, Mar. 2021. Publisher: Elsevier.

[135] L. Lu Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The Covid-19 Open Research Dataset," *ArXiv*, p. arXiv:2004.10706v2, Apr. 2020.

[136] H. C. Maltezou, A. Pavli, and A. Tsakris, "Post-COVID Syndrome: An Insight on Its Pathogenesis," *Vaccines*, vol. 9, p. 497, May 2021.

[137] I. Djaharuddin, S. Munawwarah, A. Nurulita, M. Ilyas, N. A. Tabri, and N. Lihawa, "Comorbidities and mortality in COVID-19 patients," *Gaceta Sanitaria*, vol. 35 Suppl 2, pp. S530–S532, 2021.

[138] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: a knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Research*, vol. 36, pp. D901–906, Jan. 2008.

[139] W. Luo and C. Brouwer, "Pathview: an R/Bioconductor package for pathway-based data integration and visualization," *Bioinformatics*, vol. 29, pp. 1830–1831, July 2013.

[140] A. Raveendran, R. Jayadevan, and S. Sashidharan, "Long COVID: An overview," *Diabetes & Metabolic Syndrome*, vol. 15, no. 3, pp. 869–875, 2021.

[141] O. L. Aiyegbusi, S. E. Hughes, G. Turner, S. C. Rivera, C. McMullan, J. S. Chandan, S. Haroon, G. Price, E. H. Davies, K. Nirantharakumar, E. Sapey, and M. J. Calvert, "Symptoms, complications and management of long COVID: a review," *Journal of the Royal Society of Medicine*, vol. 114, pp. 428–442, Sept. 2021.

[142] F. J. Stevens and Y. Argon, "Protein folding in the ER," *Seminars in Cell & Developmental Biology*, vol. 10, pp. 443–454, Oct. 1999.

[143] I. Braakman and D. N. Hebert, "Protein Folding in the Endoplasmic Reticulum," *Cold Spring Harbor Perspectives in Biology*, vol. 5, p. a013201, May 2013. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[144] C. J. Adams, M. C. Kopp, N. Larburu, P. R. Nowak, and M. M. U. Ali, "Structure and Molecular Mechanism of ER Stress Signaling by the Unfolded Protein Response Signal Activator IRE1," *Frontiers in Molecular Biosciences*, vol. 6, 2019.

[145] P. Olausson, B. Gerdle, N. Ghafouri, D. Sjöström, E. Blixt, and B. Ghafouri, "Protein alterations in women with chronic widespread pain – An explorative proteomic study of the trapezius muscle," *Scientific Reports*, vol. 5, p. 11894, July 2015. Number: 1 Publisher: Nature Publishing Group.

[146] R. Ding, B. Sun, Z. Liu, X. Yao, H. Wang, X. Shen, H. Jiang, and J. Chen, "Advanced Oxidative Protein Products Cause Pain Hypersensitivity in Rats by Inducing Dorsal Root Ganglion Neurons Apoptosis via NADPH Oxidase 4/c-Jun N-terminal Kinase Pathways," *Frontiers in Molecular Neuroscience*, vol. 10, 2017.

[147] K. Wåhlén, B. Ghafouri, N. Ghafouri, and B. Gerdle, "Plasma Protein Pattern Correlates With Pain Intensity and Psychological Distress in Women With Chronic Widespread Pain," *Frontiers in Psychology*, vol. 9, p. 2400, Nov. 2018.

[148] A. Yang, F. Chen, C. He, J. Zhou, Y. Lu, J. Dai, R. B. Birge, and Y. Wu, "The Procoagulant Activity of Apoptotic Cells Is Mediated by Interaction with Factor XII," *Frontiers in Immunology*, vol. 8, p. 1188, Sept. 2017.

[149] E. Pretorius, M. Vlok, C. Venter, J. A. Bezuidenhout, G. J. Laubscher, J. Steenkamp, and D. B. Kell, "Persistent clotting protein pathology in Long COVID/Post-Acute Sequelae of COVID-19 (PASC) is accompanied by increased levels of antiplasmin," *Cardiovascular Diabetology*, vol. 20, p. 172, Aug. 2021.

[150] T. Oskotsky, I. Marić, A. Tang, B. Oskotsky, R. J. Wong, N. Aghaeepour, M. Sirota, and D. K. Stevenson, "Mortality Risk Among Patients With COVID-19 Prescribed Selective Serotonin Reuptake Inhibitor Antidepressants," *JAMA Network Open*, vol. 4, p. e2133090, Nov. 2021.

[151] L. Mastrangelo, A. Cassidy, F. Mulholland, W. Wang, and Y. Bao, "Serotonin receptors, novel targets of sulforaphane identified by proteomic analysis in Caco-2 cells," *Cancer Research*, vol. 68, pp. 5487–5491, July 2008.

[152] E. Cano-Gamez and G. Trynka, "From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases," *Frontiers in Genetics*, vol. 11, 2020.

[153] J. Ellson, E. R. Gansner, E. Koutsofios, S. C. North, and G. Woodhull, "Graphviz and Dynagraph — Static and Dynamic Graph Drawing Tools," in *Graph Drawing Software* (M. Jünger and P. Mutzel, eds.), pp. 127–148, Berlin, Heidelberg: Springer, 2004.

[154] M. Galushka, C. Swain, F. Browne, M. D. Mulvenna, R. Bond, and D. Gray, "Prediction of chemical compounds properties using a deep learning model," *Neural Computing and Applications*, vol. 33, pp. 13345–13366, Oct. 2021.

[155] S. Jaeger, S. Fulle, and S. Turk, "Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition," *Journal of Chemical Information and Modeling*, vol. 58, pp. 27–35, Jan. 2018.

[156] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods," in *Computing and Combinatorics* (T. Asano, H. Imai, D. T. Lee, S.-i. Nakano, and T. Tokuyama, eds.), Lecture Notes in Computer Science, (Berlin, Heidelberg), pp. 1–17, Springer, 1999.

[157] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu, "HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks," *arXiv:1309.7393 [cs]*, Sept. 2013. arXiv: 1309.7393.

[158] A. Clark, C. Fox, and S. Lappin, eds., *The handbook of computational linguistics and natural language processing*. Blackwell handbooks in linguistics, Chichester, West Sussex ; Malden, MA: Wiley-Blackwell, 2010. OCLC: ocn500823419.

[159] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: Beyond Jeopardy!," *Artificial Intelligence*, vol. 199-200, pp. 93–105, June 2013.

[160] C. Derici, K. Celik, E. Kutbay, Y. Aydin, T. Gungor, A. Ozgur, and G. Kartal, "Question Analysis for a Closed Domain Question Answering System," in *Computational Linguistics and Intelligent Text Processing* (A. Gelbukh, ed.), vol. 9042, pp. 468–482, Cham: Springer International Publishing, 2015.

[161] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. Liu, M. Peters, M. Schmitz, and L. Zettlemoyer, "AllenNLP: A Deep Semantic Natural Language Processing Platform," *arXiv:1803.07640 [cs]*, May 2018. arXiv: 1803.07640.

APPENDIX A: KB Code

This project utilizes Python heavily to Extract, Transform, and Load source data. Linguamatics queries were developed to mine through abstracts and full-text articles. Docker was used to build and run instances of the Knowledgebase and interface. This allows for standardizing systems regardless of user hardware. All code written for this project can be found at the following url: https://github.com/atrautm1/Dissertation.

APPENDIX B: Research Timeline

Table B.1: The timeline for all research aims.

Aim	Task	Est. Duration (days)
1.1	Develop a method to Auto Download Source Data	90
1.2	Update source data	120
1.3	Containerize the knowledgebase	90
1.4	Load Data into KB	60
1.5	Develop an interface for exploration	60
2.1	Identify relevant Full Text (FT) articles	30
2.2	Acquire permissions for text mining FT articles	45
2.3	Download FT articles	45
2.4	Index FT articles in I2E	45
2.5	Develop Text Mining Queries	90
2.6	Source structured microbiome data	90
2.6	Connect ABCkb 2.0 to interface	90
3.1	Get quarantined	60
3.2	Download Cord-19 subset	45
3.3	Index Cord-19 subset	45
3.4	Develop Text mining queries	45
3.5	Connect text mined data to ABCkb	30
3.5	Pathview analysis	30

APPENDIX C: CDkb Long-Covid Nodes

Table C.1: These are the phenotype nodes from the CDkb that are connected with 9 common symptoms of long-covid.

Source ID	Node Name	Long Covid Symptom
D005221	Fatigue	Fatigue
HP:0012378	Fatigue	Fatigue
D005222	Mental Fatigue	Fatigue
HP:0012432	Chronic fatigue	Fatigue
DOID:8544	chronic fatigue syndrome	Fatigue
NAL:918	cough	Cough
D003371	Cough	Cough
HP:0012735	Cough	Cough
HP:0031246	Nonproductive cough	Cough
D004417	Dyspnea	Dyspnea
NAL:919	Dyspnea	Dyspnea
HP:0002829	Arthralgia	Arthralgia
D018771	Arthralgia	Arthralgia
D002637	Chest Pain	Chest Pain
HP:0100749	Chest Pain	Chest Pain
NAL:204102	memory disorders	Memory Disorders
HP:0002354	Memory impairment	Memory Disorders
D008569	Memory Disorders	Memory Disorders
HP:0000458	Anosmia	Anosmia
MONDO:0010528	anosmia (disease)	Anosmia
HP:0012247	Specific anosmia	Anosmia

Continued on next page

Table C.1 – continued from previous page

Source ID	Node Name	Long Covid Symptom
HP:0010633	Partial anosmia	Anosmia
HP:0041051	Ageusia	Ageusia
HP:0031249	Parageusia	Ageusia
D004408	Dysgeusia	Ageusia
D000370	Ageusia	Ageusia
D004244	Dizziness	Dizziness
D012678	Sensation Disorders	Dizziness
HP:0002321	Vertigo	Dizziness

APPENDIX D: Pathview Pathway Analysis with 3 Coronaviridae and Text-mining

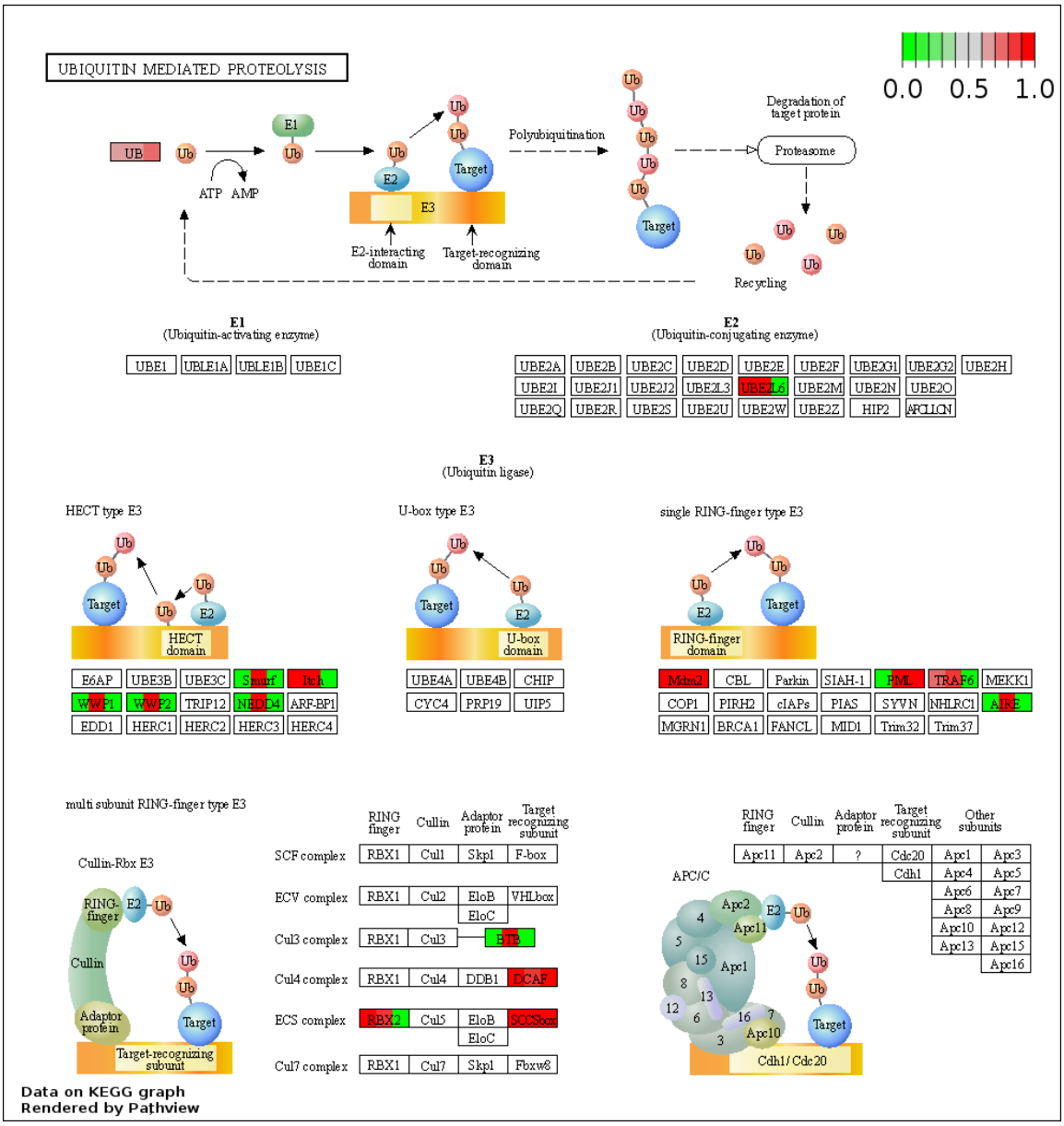

Figure D.1: The hits from the pathview analysis with TF*IDF scores providing color values with the 3 coronaviridae. Colored gene boxes from left to right are: SARS-CoV-1,SARS-CoV-2,MERS-CoV

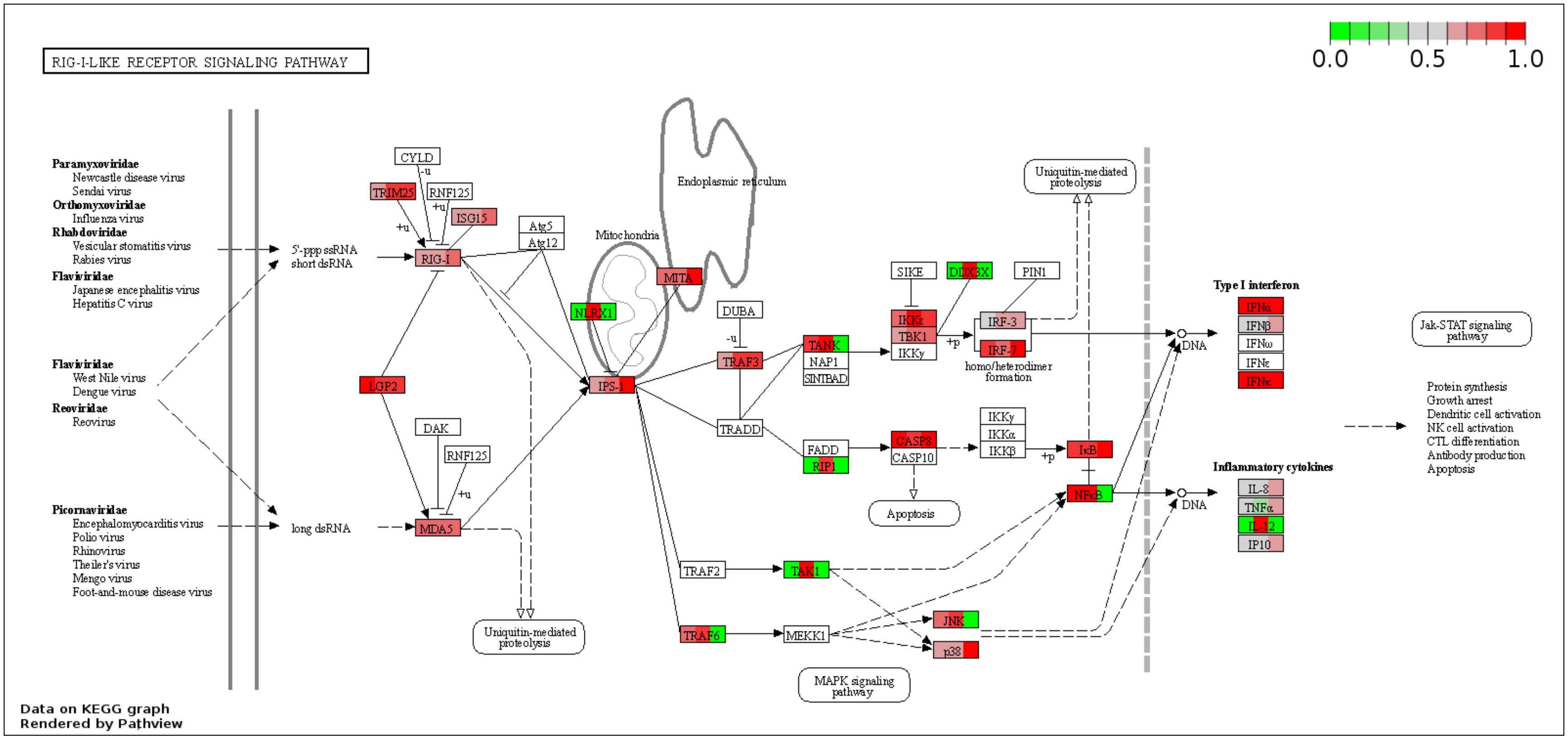

Figure D.2: The hits from the pathview analysis with TF*IDF scores providing color values with the 3 coronaviridae. Colored gene boxes from left to right are: SARS-CoV-1,SARS-CoV-2,MERS-CoV