THE SEMANTICS OF DIET AND HEALTH: KNOWLEDGE BASED DISCOVERY THROUGH DATA INTEGRATION, TEXT MINING, AND NETWORK ANALYSIS

by

Richard V. Linchangco

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 Genomics

Charlotte

2018

Approved by:

_____
Dr. Cory Brouwer

_____
Dr. Jessica Schlueter

_____
Dr. Jennifer Weller

_____
Dr. Xinghua Shi

_____
Dr. Michael Turner

ABSTRACT

RICHARD V. LINCHANGCO. The Semantics of Diet and Health: Knowledge Based Discovery through Data Integration, Text Mining, and Network Analysis. (Under the direction of DR. CORY BROUWER)

Consumption of fruits and vegetables has been linked to a reduced risk of cancer and other chronic diseases, but the molecular mechanisms supporting these connections remain largely unknown. A wealth of association data linking components of a plant-based diet, human genes, biological pathways, and phenotypes can be found in public databases and scientific literature. However, this massive amount of data is distributed across disparate sources, presenting a significant barrier to the investigation of the effects that plant-based diets impart on human health.

This dissertation describes an integrated association network composed of existing curated and text-mined relationships which connect the agricultural and biomedical entities that define diet and disease. This research also describes HetERel, a meta path-based relevance ranking method for extracting highly relevant relationships between different types of entities in this network, such as a plant and the chemicals it produces. HetERel is tested on a network of chemical-disease association data for validation and performance. The method is then applied to the full-scale, integrated diet-disease network to discover distant, indirect links between plant-based chemicals and human phenotypes.

The integrated diet-disease association network provides a foundational resource that connects plants and human health. Paired with the relevance search and prioritization method, HetERel, these methods empower researchers to generate hypotheses

which elucidate the molecular mechanisms between plants and human disease.

ACKNOWLEDGMENTS

This entire process has been an enlightening endeavour. I would like to thank those who supported me with their intellect, constructive criticism, and emotional support.

Dr. Cory Brouwer provided me with the opportunity to pursue this research and constantly encouraged me throughout the arduous journey. He introduced me to a distinct, yet integrated area of research in bioinformatics. Dr. Jeremy Jay was my biggest critic, but in retrospect, it was to better my research. Many foolish algorithmic ideas and programming questions were directed to him for rationalization. Dr. Robert Reid provided collaborative ideas and direction for many of the biological cases of interest for my research. His upbeat attitude also kept me going through the process. Steven Blanchard provided all the necessary computational resources I needed at all times.

The numerous Ph.D. students and interns of the Plant Pathways Elucidation Project learned and grew with me throughout the research process. I would also like to thank Dr. Eric Jackson who inspired and helped fund the project. The UNC GA was instrumental in providing the funding for research opportunities.

My family who supported me and listened to my constant streams of ideas. My sisters who supported me with their human health use case ideas. My brother who underwent the trials and tribulations of school and research with me.

Last, but not least, Bertsie Castillo, my rock and enforcer. Your passion to pursue Veterinary Medicine and patience with me through graduate school has motivated me time and time again!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: DISCOVERING LINKS BETWEEN DIET AND HUMAN HEALTH AT THE MOLECULAR LEVEL

Three of the leading causes of death in the United States, cardiovascular disease, diabetes, and cancer, are significantly linked to lifestyle choices. These chronic diseases have reached epidemic proportions, prompting researchers to investigate contributing factors such as physical activity, genetics, and diet [29]. Traditionally, diet and human health have been studied at the phenotypic and observational level through epidemiological studies [148]. These types of studies create a gap in knowledge of the mechanisms by which dietary components affect human health.

High throughput (HT) studies have begun to characterize the composition of foods and the genetic variations between healthy and diseased individuals [127]. Nutritional genomics uses HT techniques to study effects elicited from the interaction of dietary components with the human genome [118]. Different combinations of HT techniques have associated specific compounds in foods with altered gene activity and changes in human health phenotypes [54]. The challenge lies in not only identifying associations between dietary components and human health, but also in the discovery of molecular mechanisms supporting such associations. However, these associations are distributed across separate heterogeneous databases and buried within the text of scientific publications. Analysis of these associations requires the assimilation of massive amounts of data from myriad disparate resources such as association databases and

the scientific literature.

The knowledge-based identification of diet and human health associations provides insight into the importance of the molecular effects of diet on disease prevention. An understanding of these molecular mechanisms can provide scientific support to filter the deluge of fictitious health claims commonly propagated in the media today. A similar problem exists in pharmacogenomics when exploring drug response in various genotypes and phenotypes. Pharmacogenomics researchers have applied data mining techniques to process the immense volume and heterogeneity of data from databases and the literature on drug response [69]. Data mining techniques offer a comprehensive view of existing knowledge and require significantly less time and resources than expert manual curation or benchtop assays [13, 69, 150]. These techniques can be implemented to build upon and complement research in pharmacogenomics in order to identify associations between dietary compounds and disease related genes. The identification of these associations aid in explicating the molecular mechanisms behind diet and human health.

This research describes a novel method to elucidate the molecular mechanisms by which plant-based foods influence human health through the identification and prioritization of linked information from a large relationship graph of agricultural and biomedical associations. First, the difficulties of procuring, extracting, and integrating data from many diverse sources are discussed. Relevant entities and relationships between plants, chemicals, genes, pathways, and disease are extracted from biomedical and agricultural databases, as well as scientific literature using a host of mining techniques. Second, a novel, quantitative metric for searching and ranking the relevance

of entities within a heterogeneous information network is designed, implemented, and tested on a chemical-disease network. Finally, the ranking method and metric are applied to biological use cases in a diet-disease network for the discovery of testable biological hypotheses. This approach consolidates current knowledge from a variety of public resources, expedites hypothesis development through entity prioritization, and aids in the discovery, at the molecular level, of the relationship between diet and human health.

1.1 Integrating Agricultural and Biomedical Association Data

Advances in HT technologies have generated massive volumes of data available for nutrition research. The most disruptive of these technologies has been high throughput sequencing. Sequencing was instrumental to the release of the Human Genome Project, which supplied a reference genome for researchers to investigate human phenotypes at the gene level [35, 36]. Agricultural research makes use of HT technologies, such as metabolomics, to determine the biochemical composition of plants and sequencing to understand the basis for the biochemical products of plants [46]. These and other types of agricultural and biomedical data are collected, stored, and interpreted by different research groups utilizing various standards. The problem of identifying associations that connect agricultural and biomedical entities for prioritization is exacerbated by the multitude of uniquely defined standards.

To overcome this obstacle, diverse datasets must be aggregated, integrated, and made easily accessible. These datasets are stored in disparate sources of varying types, such as databases and structured vocabularies. In order to harness the wealth

of data for hypothesis generation through candidate prioritization, the complexity of data from agricultural and biomedical research sources need to be understood.

1.1.1 Public Resources of Curated Data

Only machine accessible data can be used in computational algorithms for the prioritization of molecular mechanisms between diet and human health. There are numerous public databases that span the domains of plants, chemicals, genes, biological pathways, and human phenotypes. These resources store entity and association data in a number of different formats, including association databases and structured vocabularies, made accessible as a database or other parseable flat files. **Table 1** provides an overview of collective data sources that encompass biomedical and agricultural entities. Association databases contain entities, their associations, including cross references to other sources, and metadata (data that describes the data). Structured vocabularies contain hierarchically related entities used to describe the semantic skeleton of a domain. The quality, reusability, throughput, and accurate representation of biological domain were the criteria for selecting publicly available sources for the diet-disease network.

Table 1: The listed data sources are comprehensive sources with multiple databases and structured vocabularies that contain entity and association data describing the agricultural and biomedical domains. These sources are discussed in further detail in **Chapter 2**.

Data Source	**Entity Types**	**References**
National Center for Biotechnology Information	Species, Chemicals, Genes, Phenotypes	[151]
European Bioinformatics Institute	Chemicals, Genes, Phenotypes	[89]
National Center for Biomedical Ontology	Species, Chemicals, Pathways, Phenotypes	[164]
USDA National Agricultural Library	Species, Chemicals, Genes, Phenotypes	[124]

Each source contains unique entity identifiers (EIDs), preferred terms (PT), and

associations for agricultural and biomedical entities. EIDs are distinct references to entities within a data source, such as chemicals or genes. PTs are lexical descriptors of entities, such as the term glucoraphanin which describes a phytochemical. Associations are explicit connections between entities within a source and cross references to external sources. A semantic label generically explains associations, such as *is_a* which indicates a subsuming association. In database instances, semantic labels are often not specified. Associations in structured vocabularies, ontologies specifically, provide semantic labels that help organize entities hierarchically. Associations may also have metadata that provide information, such as the source or confidence value, for the association. Association types must be independently quantified based on quality, scope, and detail, which will be discussed in detail in **Chapter 2**.

1.1.1.1 National Center for Biotechnology Information Sources

The National Center for Biotechnology Information (NCBI) maintains publicly accessible data and tools for computational analysis in the biomedical domain [2]. There are 39 databases for biomedical research, called Entrez, in NCBI. NCBI's E-Utilities web service was developed as a query interface system for programmatic access to the Entrez databases. This research integrates data from the following four resources of the NCBI.

The Taxonomy Database, hosted by NCBI, contains a classification of all species represented in the Entrez sequence databases [20]. It represents a curated, hierarchically organized nomenclature for almost 550,000 taxa, including all land plants, Embryophyta. Embryophyta encompass 170,650 agriculturally relevant plants. Data

from the NCBI Taxonomy Database can be downloaded from NCBI's FTP site or accessed with E-Utilities.

The Entrez Gene Database stores gene information for over 20.5 million species [106]. Gene information from Entrez Gene can include nomenclature, sequences, membership in pathways, associations to phenotypes, and cross references to other relevant databases. Entrez Gene serves as the most commonly used reference for gene information in biomedical research. Entrez Gene data can be downloaded as flat files from NCBI's FTP site or through E-Utilities.

PubMed is an indexed, freely available citation database containing over 27 million biomedical publication abstracts [60]. PubMed contains citation metadata such as author lists, journals, publication dates, and keyword indices. **MEDLINE** is a subset of PubMed that contains over 24 million abstracts from biomedical and life sciences research [59]. MEDLINE does not contain in-process or "Ahead of Print" citations. It is the most widely used body of literature in text mining for biomedical research. PubMed and MEDLINE share cross references to many other NCBI hosted databases and resources, such as Entrez Gene and the Medical Subject Headings. The annual baseline set of citations for PubMed and MEDLINE can be downloaded in XML format from the NCBI FTP site.

The Medical Subject Headings (MeSH) resource is a structured vocabulary maintained by NCBI that defines a hierarchically organized, standard terminology and provides synonyms for biological and medical terms [100]. MeSH is used as a means of indexing the articles found in PubMed and MEDLINE. These articles are annotated with MeSH terms by expert curators which aids in filtering PubMed searches.

MeSH files can be downloaded from the National Library of Medicine MeSH FTP site.

The Online Mendelian Inheritance in Man Database (OMIM) is a highly curated, online resource that holds detailed descriptions and associations between human genes and phenotypes [70]. The database focuses on hereditary diseases and draws on expert curation from publications which associate genes and disease. OMIM is regarded as a g̈old standardf̈or gene and disease associations, based on its extensive scope in human disease. OMIM is accessible through the OMIM API and as flat files from the OMIM FTP site.

1.1.1.2 United States Department of Agriculture Sources

The United States Department of Agriculture (USDA) oversees the National Agricultural Library (NAL) which maintains a physical and electronic library of resources related to agriculture [124]. The NAL hosts numerous services for agricultural research, such as repositories, and information centers for food safety and nutrition.

NAL hosts the Agricultural Online Access (AGRICOLA), a public citation database devoted to agricultural publications [123]. AGRICOLA contains over 6 million publication records from scientific journals, books, and government reports. AGRICOLA can be accessed through the NAL Catalog site `https://agricola.nal.usda.gov/`. In its entirety, AGRICOLA is available by lease from the National Technical Information Service as a data file.

The NAL Thesaurus is a structured vocabulary that includes descriptions, synonyms, and hierarchical relationships between agriculture related terms [125]. NAL

Thesaurus terms are used to index citations in AGRICOLA. Curators annotate publication records using the NAL Thesaurus terms to assist in filtering the output of AGRICOLA article searches. The NAL Thesaurus is available for download in multiple file formats from the USDA NAL site.

The USDA National Nutrient Database for Standard Reference (NDB) stores nutrient information for raw and processed foods commonly consumed in the United States [182]. The NDB connects over 8,000 foods and 150 nutrients. Food and nutrient data from the NDB include association sources, descriptions, food weights, and nutrient values. The NDB can be downloaded as a set of flat files from the Agricultural Research Service website.

1.1.1.3 Open Biological and Biomedical Ontology Sources

The Open Biological and Biomedical Ontologies (OBO) Foundry is a repository of biomedical controlled vocabularies, called ontologies. The OBO Foundry was developed with the goal of fostering interoperability through the use of a standardized, flat file format called OBO [164]. These ontologies contain term definitions, synonyms, relationships, and cross references to other ontologies and databases.

The Gene Ontology (GO) is a structured vocabulary that standardizes terminology which describes the role of genes and proteins in cells [10]. The GO specifies and associates concepts that explain gene function, location, and pathway membership across all species. It is comprised of three ontologies, molecular function, cellular component, and biological process. The GO is the most widely known and used ontology in biomedical research and has been used to annotate and categorize results

in tens of thousands of publications. The Gene Ontology can be downloaded directly from the project's FTP site.

The Plant Ontology (PO) adheres to the OBO format to provide a structured vocabulary describing plant anatomy, morphology, and developmental stages [12]. It contains cross references to external ontologies such as the Gene Ontology and the Chemicals of Biological Interest. The PO is available from the Planteome project FTP site.

The Human Phenotype Ontology (HPO) is a structured vocabulary representing phenotypic anomalies of diseases in humans [145]. The HPO has cross references to many ontologies used in this research. It also aids in providing structural depth where other resources, such as MeSH, are lacking. The HPO and phenotype to gene annotations are available from the HPO GitHub reposistory.

The Mammalian Phenotype Ontology (MPO) is similar to the HPO but contains terminology specific to rat and mouse models of human biology and disease [165]. The MPO is available from the Mouse Genome Informatics website.

The Disease Ontology (DO) is a publicly available, structured vocabulary that assists in integrating biomedical data associated with human disease [152]. It provides definitions of human disease terms, phenotypic characteristics, and medical vocabulary concepts. The DO has extensive mappings to MeSH and OMIM. The DO can be downloaded from OBO Foundry repository.

1.1.1.4 European Molecular Biology Laboratory and European Bioinformatics Institute Sources

The European Molecular Biology Laboratory (EMBL) and European Bioinformatics Institute (EBI) are the main public resources for molecular biology information in Europe [89]. They support data repositories and analysis tools for molecular and computational biologists.

The Chemical Entities of Biological Interest (ChEBI) is a structured vocabulary for small molecular entities produced in nature or synthesized to affect pathways in living organisms [72]. ChEBI serves as a database and an ontology, storing relevant information such as molecular formula, structure, charge, and mass. The ChEBI ontology and taxonomic origins of compounds are available from the EBI FTP site.

1.1.1.5 Chemical, Gene, and Disease Association Databases

Secondary association databases gather data from primary sources in order to tease out interesting patterns and investigate newly generated hypotheses. These association databases map interactions between genes, chemicals, gene products, and human disease.

The Comparative Toxicogenomics Database (CTD) contains associations of chemicals, genes, and diseases [42]. The CTD is a publicly available association database that aids in the investigation of how chemicals influence human disease. Associations in the CTD are highly curated and map entities between biomedical resources, such as MeSH, GO and Entrez Gene. Association data from the CTD is accessible via the CTD's project website as flat files.

As evidenced by the information sources described, biomedicine and agriculture continue to be active areas of independent research. The identification of genetic contributors to disease focuses on the molecular function of genes and gene products. Research in agriculture focuses on the benefit of bioactive chemicals for stress response in plants. However, an understanding behind the molecular mechanisms linking phytochemicals with their effects on human genes is still largely missing.

Although a deluge of data from agricultural and biomedical research exists, it is distributed across numerous public repositories and shares sparse relationships that connect the different domains. There are monolithic sources that contain cross references to link various databases in the biomedical domain, such as the National Center for Biotechnology Information Sources or the European Bioinformatics Institute [89, 151]. Despite these monolithic resources, cross domain associations that link entities, such as phytochemicals and human genes, are few in number. There is a dearth of sources that associate plants and their phytochemical products to the effects of those phytochemicals on human genes. Scarce associations can be aggregated through the integration of publicly available sources and supplemented with the addition of relationships mined from scientific literature.

1.1.2 Augmenting Public Resources of Curated Data

The sources for extensively curated diet-disease relationships are scarce and generally disjointed. A minimum of four domains (*plants - chemicals - genes - phenotype*) must be traversed to link plant-based diets to disease. These four domains span the agricultural and biomedical research space, yet do not frequently overlap. Sparse

overlap creates a need to connect these domains to explain the relationship between diet and disease. This need can be addressed by extracting overlapping associations from the scientific literature. A wealth of agricultural and biomedical knowledge describing the interaction of phytochemicals with the human genome is available in the scientific literature.

Unlike the structured data found in databases, scientific literature follows a free-text format, easily understood by humans but unavailable for computation by computers. Compounding the problem is the size of the scientific literature. MEDLINE, the most prominent citation database in biomedicine, contains over 24 million abstracts to date with a growth rate over one million articles per year [59]. At this pace, scientists struggle to remain current, even within their specific field of research. In addition, in order to study plant-based diets it is necessary to include literature from agricultural research. AGRICOLA is the most widely used citation database for agricultural research, storing over 6 million abstracts [123]. The volume of literature across biomedicine and agriculture, totaling over 30 million abstracts, cannot be assimilated by humans alone. High throughput methods, such as text mining, are required to extract associations buried in these massive sets of literature [5]. Extracting the latent associations from both agricultural and biomedical literature provides the supplemental associations needed to link plant-based diets with human health and disease.

1.2 Literature Based Discovery Methods for Relevance Search

Literature based discovery (LBD) is the process of using collections of scientific literature to infer implicit relationships from existing knowledge, resulting in data driven, testable hypotheses [189]. Swanson was the first to propose a literature driven hypothesis, proposing that dietary fish oils (DFOs) could prevent and treat the effects of Raynaud's Syndrome (RS) [174]. From all available literature about DFOs, Swanson deduced that DFOs lower blood viscosity, platelet aggregability, and vascular reactivity, providing the potential to improve blood circulation. In all available literature relevant to RS, Swanson found high blood viscosity, platelet aggregability, and vasoconstriction to be associated with the circulatory disorder RS. At the time, the existing knowledge for DFOs and RS shared common attributes but were found in non-interacting sets of literature. Swanson combined observations from these non-interacting sets of literature, leading to the novel postulation that DFOs might ameliorate or prevent RS.

Swanson realized that by aggregating scattered bits of research literature, a reader can infer implicit connections. The logical connections were generalized into the ABC model of discovery [174]. The model states that two concepts, A and C, may be connected by transitive property if an intermediate set of concepts, B, connects both A and C. To illustrate Swanson's DFO-RS hypothesis, let A represent DFOs and B represent reduced blood viscosity, platelet aggregability, and vasoconstriction, and let C represent RS amelioration, as in **Figure 3**. The literature shows DFOs (A) cause reduced blood viscosity (B). The literature also shows reduced blood viscosity

(B) causes RS amelioration (C). According to Swanson's model, because A causes B and B causes C, it can be inferred that A may cause C, or that DFOs may cause RS amelioration. This hypothesis, along with others, was clinically validated in later experiments [161–163, 174–176].

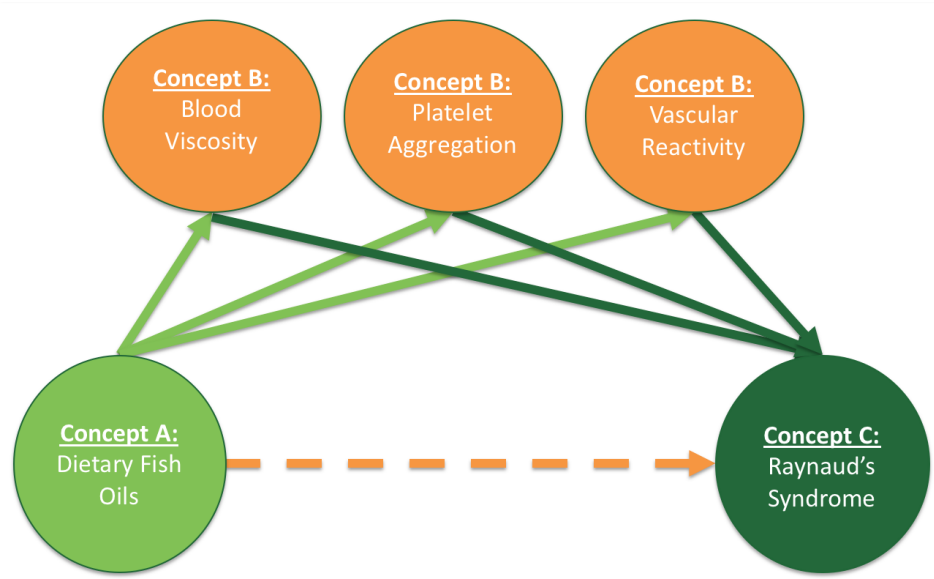


Figure 1: ABC Model of Discovery illustrates the inference between Dietary Fish Oils and Raynaud's Syndrome. Solid lines represent the existing links from concepts A and C to the set of B concepts. The dashed line represents the inferred link between concepts A and C due to the overlap of intermediate set of B concepts.

The ABC model takes two forms, called closed discovery and open discovery. In the closed discovery process, both concepts A and C are known, restricting the candidate concepts of interest to the set B, common to both A and C. To illustrate, an observed association in nutrition research states that broccoli, concept A, reduces chronic inflammation, concept C [185]. Closed discovery attempts to find the connection between broccoli(A) and chronic inflammation(C) via a transitive relationship between A-B and B-C. Oxidative stress is a shared concept associated to both broccoli and chronic inflammation that provides such a transitive relationship, acting as

a concept B [199]. In the open discovery process, only the starting concept, A, is known. Following the previous example, if A was broccoli, concepts B and C could be any biomedical concepts of interest, such as chemicals or metabolic pathways. Open discovery is a computationally difficult problem, as it presents a combinatorial explosion of candidate concepts for investigation. This research focuses on the more tractable form of closed discovery, with the modularity to extend the method to open discovery in the future.

LBD methods are composed of three major tasks that include text mining, association mining for ranking, and network analysis [101]. Text mining methods range from simple co-occurrence within documents to the use of natural language processing techniques for high throughput extraction of interesting entities and their relationships. Co-occurrence methods provide non-descript associations, while advanced text mining methods provide semantic relationships between entities. The vast possibility of associations from text mining qualify the need for methods to filter and prioritize candidate hypotheses inferred by LBD, such as interestingness measures and network analysis techniques.

1.2.1 Text Mining

Text mining is the process of extracting useful information from document collections through the identification and exploration of interesting patterns in unstructured textual data [53]. The value of text mining comes from automation and throughput. As previously mentioned, text mining is essential in LBD for identifying and extracting associations from the vast volume of scientific literature. Its application in

pharmacogenomics, mining knowledge between drug and human gene interactions, parallels its potential in the field of nutrition, where interactions of phytochemicals and human genes are the focus [69].

A document corpora, such as abstracts from MEDLINE, is the input for text mining. The output varies based on the goals of mining but is generally organized and results in machine-readable data structures containing entities and associations of interest. In the case of this research, text mining output consists of tables of agricultural and biomedical entities and the semantic relationships between them. Text mining can be broken down into two subtasks, information retrieval and information extraction.

Information retrieval (IR) is the storage, organization, and access to information found in informational objects such as documents or web pages [1]. The basic function of information retrieval systems is to rapidly and accurately reduce the search space of a given information collection by returning information objects relevant to keyword-based queries. An IR system allows researchers to quickly access the publications in PubMed based on keyword searches. Keywords from a user are mapped to MeSH, the vocabulary used to index all publications in PubMed. The IR system finds all documents indexed with the mapped keywords and returns them to the user in seconds. For example, a researcher interested in publications relevant to broccoli and glucosinolates would input the two terms into PubMed's IR system. The system maps the terms to their closest matches in MeSH. For instance, broccoli would map to *Brassica oleracea* var. *italica*. All publications indexed with *Brassica oleracea* var. *italica* and glucosinolates would then be returned to the researcher. IR quickly

narrows the corpora of literature to a subset of the most pertinent, user desired information. Smaller subsets can still be unmanageable, such as the broccoli and glucosinolates query which returns over one thousand publications. The next subtask of text mining, information extraction, assists in automatically structuring information from the free text within the literature from IR for computational analysis.

Information extraction (IE) transforms the data from unstructured text into relevant, machine-readable information [53]. Information extraction is the most important subtask in text mining as the structured information it produces ultimately represents the output of text mining. IE can be divided into two major subtasks, named entity recognition and relation extraction, that identify entities and the relationships between them from unstructured text. Modern approaches combine the subtasks using a combination of named entity recognition techniques.

Named entity recognition (NER) identifies entities of interest, such as agricultural or biomedical terms, from the documents returned by IR [96]. There are three main categories of NER techniques: dictionary-based approaches, rule-based approaches, and machine learning approaches.

Dictionary based NER is the simplest and most accurate approach. In the context of NER, a dictionary is a collection of terms that represent entities of interest [96]. In this approach, exact matches of dictionary terms are found in the text, ensuring precision but to the detriment of recall. Dictionary based approaches also lack the flexibility to recognize undefined terms, such as synonyms or alternate spellings, not found in the dictionary. For example, if b̈roccoliïs the only explicitly defined term for the plant in the dictionary, other instances, such as B̈roccoliör *Brassica oleracea* var.

italica, will not be recognized. Inexact or fuzzy matching techniques can generate spelling variants of terms to increase matches. Integrating ontologies and controlled vocabularies as dictionaries supplies synonyms and other representations of terms for increased recall in dictionary based NER methods. In the Whatizit system, all dictionaries are generated from biomedical databases and ontologies to maximize recall [138]. The added benefit to using database and ontology terms in these systems is the ease of interoperability of sources when mining across domains.

Rule-based NER approaches rely on rules or patterns derived from prior knowledge [34, 96]. Early approaches used rules taken from biomedical nomenclature to identify entity classes. For example, protein and gene symbols tend to be single ẅordsc̈onsisting of upper-case letters, numbers, and hyphens, such as the protein interleukin-1 symbol, IL-1 [61]. More advanced systems define specific patterns as rules, such as "chemical entity" *regulates the expression of* "gene entity". The most sophisticated rule-based systems employ natural language processing techniques for NER.

Natural language processing (NLP) makes use of linguistic concepts, such as parts of speech and grammatical structure, to parse and represent free text [85]. An early example of an NLP text mining tool is the freely available, web-based Chilibot system that implements NLP techniques for NER and rule extraction [28]. The requirement of prior knowledge and the potential time investment for manual deduction of rules are drawbacks of rule-based NER approaches. Domain specific rules can also make it difficult to scale these methods to larger, diverse document collections.

Machine learning (ML) and statistical based approaches treat NER as a classifica-

tion problem [34, 96]. Based on the value of defined features, these approaches can tag entities and parts of speech, or classify entire documents into different categories. ML approaches require a comprehensive and manually annotated set of data to train classifiers. The three most implemented machine learning techniques for NER are naive Bayes, hidden Markov models, and support vector machines.

Naive Bayes classifiers commonly represent a class as a vector of feature variables. In NER, these feature variables can be words, phrases, or other characters that define a class entity. Each feature is assumed to be independent of any other feature. A supervised learning stage involves training the classifier with a manually annotated dataset. After training, the classifier is fed the test dataset to classify. The conditional probability of a certain class or category existing in a document given the features of that document, is calculated and evaluated against a defined threshold.

Support vector machines (SVMs) are similar in concept to naive Bayes classifiers. The difference lies in the assumption of independence of features in naive Bayes. In SVMs, linear combinations of features, called support vectors, are found from a set of positive and negative examples. These linear combinations separate the feature space into either positive or negative, classifying text based on which side of the linear combination they fall [68]. SVMs perform best when the assumption of independence between features does not hold, according to the data and problem being solved.

Hidden Markov models (HMMs) build on feature identification and add another level of complexity to NER. HMMs take into account the sequence of features (words or phrases in NER) as they appear in text. Then HMMs use statistical information from annotated examples which predict the most probable sequence of features and

if that exists within the text being searched [140]. As with the other ML techniques, HMMs tend to require larger, highly curated training datasets for the best results. Also, the more specific the features are to a domain, the less they can be reused as features for other domains.

NER is best performed with hybrid approaches that draw from dictionary, rule, and machine learning based methods. Current approaches utilize ontologies, controlled vocabularies, and database entries to define dictionaries for NER. These dictionaries are combined with grammatical and common pattern rules to annotate training datasets to be used in machine learning classifiers for NER.

Relation extraction for information extraction is built on either co-occurrence or NLP methods. Co-occurrence counts the existence of two entities of interest found within the same text as a relation. The definition of same text can differ, based on what boundaries are set. Co-occurrence can exist within the entire document, within a specific section of a document (such as the results section of a publication), or within a sentence. Co-occurrence based text mining makes use of dictionary-based methods of NER. NLP based approaches are more advanced, implementing rule and ML based algorithms that were previously described. The resulting output of IR and IE are relations and entities in a structured format for use in downstream computational analyses. [77]

The main goal of text mining in this work is to augment existing associations for knowledge discovery: the realization of new information through identification of implicit associations between relevant entities [69].

1.2.2 Association Mining for Ranking

Association mining is the probabilistic determination of associations between co-occurring items within a collection, expressed as association rules of the form ($X \rightarrow Y$) [179]. Association rules signify that whenever items in X are present in the collection, items in Y are also likely to be present. In the context of knowledge based discovery, association mining discovers the probability of a pair of entities sharing an association or relationship within a collection, such as an integrated knowledge base.

Association mining treats a dataset of associated items as a series of event instances, where each instance contains a set of co-occurring items called an itemset. For example, purchase data from a grocery store can be thought of as a series of transactions where each transaction contains a set of purchased items, such as milk and eggs. Association mining can be succinctly described in two general steps. First, unique itemsets are defined across the collection of event instances. The occurrence of unique itemsets are then counted. Those itemsets that meet or exceed a user-defined minimum probability threshold are called frequent itemsets. Frequent itemsets are used in the next step of rule generation. In rule generation, frequent itemsets are ranked based on an interestingness measure. Highly ranked frequent itemsets produce the most likely candidate association rules for that dataset [177]. In many cases, confidence is used as the default interestingness measure and those candidate rules with high confidence are returned as strong association rules.

Interestingess measures used in biomedical association mining are split into two categories, objective and subjective. Objective measures use statistics that describe the

data and a user-defined threshold to filter uninteresting associations. These measures are domain independent and data driven with little reliance on prior user knowledge or input. In addition to the dataset, subjective measures depend on user input and expert knowledge. As this work crosses the domains of agriculture and biomedicine, we focus exclusively on objective measures.

Support and confidence were the first formally defined interestingness measures [4]. Support is the probability that an event contains items of both itemsets X and Y. Confidence is the probability that an event contains the items of Y given those in X. The support-confidence framework has limitations for both support and confidence. Support eliminates low probability items that may create interesting patterns while confidence ignores the support of the consequent itemset, Y in $(X \Rightarrow Y)$. In doing so, confidence generates rules where an item is highly likely to occur on its own, regardless of the presence of other items. Lift is an interestingness measure that was developed to overcome the drawback of confidence by including the support of the consequent in calculating the interestingess of a rule [137]. It is essentially a test of independence, where a value of 1 indicates independence between X and Y. Other common objective interestingness measures that test independence are chi-squared (χ^2) and Fisher's exact test. These correlation measures face the limitation of being influenced by proportional changes to the sample size. In many biological databases, a specific association between two entities, such as a chemical interaction with a gene, is often a small probability event in relation to the total number of associations within the database. Within such a database, many associations will not include that chemical and gene, and are deemed null events with respect to that chemical or

gene. Interestingness measures affected by null events have been found to perform poorly in large databases [195]. The importance of the null-invariance property in interestingness measures has been investigated in a number of studies [126, 178, 194].

Null-invariant measures account for small probability events, where a particular item in an itemset may not occur frequently given the total number of events. These low probability events are of potential interest for developing novel hypotheses between bioactive components in food and human health phenotypes. Null-invariant measures, including the Jaccard similarity coefficient, cosine similarity, and Kulczynski measure, have been reviewed in different rule mining studies [62, 177, 194]. Null-invariant interestingness measures have been extensively studied and the selection criteria for the best applicable measure has been determined to be data dependent [195]. Null-invariant measures follow an inherent ordering and follow four properties. The first property is that each measure follows a range from 0, representing no co-occurrence, to 1, indicating two entities always co-occur. The second property states that more co-occurrences leads to a higher interestingness value, while the inverse is also true. The third property states that the measure is symmetric under event permutations. The final property of these measures is that they are invariant to scaling, where multiplying the support values by a scaling factor makes no difference in the overall value [195]. These properties are favorable to association analysis within large databases, thus warranting their comparison for use in assessing the interestingness of relationships in a diet-disease network. Furthermore, the design of a new relevance ranking measure, based on a null-invariance measure in combination with heterogeneous information network analysis techniques, will be discussed in

great detail in **Chapter 3** of this work.

1.2.3 Information Network Analysis

Data in numerous research fields, including the social sciences, library science, and biology, are increasingly modeled as large, highly linked collections of interrelated objects, called information networks [73, 109, 186]. Information networks contain objects, which represent concepts, connected by links that represent the relationships between concepts. When the objects and links in an information network all share a single type, it is considered a homogeneous network. The objects and links in heterogeneous networks have different types which convey subtle semantic information. Heterogeneous networks are more representative of current data, such as the interaction between chemicals, genes, and disease. Consequently, many different methods have been developed for analyzing homogeneous and heterogeneous networks.

Methods for the quantification of similarity between two objects within an information network are of particular interest in information network analysis for LBD. An application of similarity search in the biological domain is the identification and ranking of genes associated to disease. Given a network of gene-disease associations, a similarity search can identify and prioritize genes based on their similarity to a specific disease. The similarity of two objects can be quantified and evaluated by a similarity measure. Traditional similarity measures, such as the Jaccard coefficient or PageRank, were developed for homogeneous networks, unable to capture the semantic information of heterogeneous networks [98, 128].

Similarity measures designed for heterogeneous networks, such as PathSim and

Path Constrained Random Walk, utilize the concept of meta paths to distinguish the semantics of paths connecting differently typed objects [91, 169]. A meta path is a sequence of relations between different object types, which defines a composite relation between the first and last object types of the path. These meta path based methods perform well in cases where similarity is measured between objects of the same type, such as calculating the similarity of two genes associated to a disease. However, cases exist where the relationship between objects of different types is of interest. For example, a plant researcher would be interested in determining the relevance between plants and chemicals to generate a plant's chemical profile.

The distinct task of measuring the relevance of differently typed objects is less studied than measuring the similarity of same typed objects. Straight forward measures, such as path count (PC) and pairwise random walk (PRW), calculate relevance but have biases [169]. Path count measures relevance as the number of instances of a meta path between a start and end object. It favors objects with high link counts. Himmelstein proposed a relevance measure that extends path count, implementing a down weighting factor to overcome high count bias [75]. Pairwise random walk begins by splitting the meta path into two even paths. Then it calculates the probability of two random walks, originating from the start and end objects of the meta path, reaching the same middle object. PRW is biased because it values densely linked middle objects. HeteSim is an extension of PRW that accounts for bias by normalizing the random walk probability for each step composite relation in a meta path [156]. The HeteSim measure exhibits a pair of beneficial properties. It is a symmetric measure, which provides a single value to compare the relative relatedness between pairs

of differently typed objects. It also has the property of a self-maximum, meaning HeteSim values are constrained to a range between 0.0 and 1.0. This allows for easy comparison to other measures with a self-maximum.

1.3 Challenges and Limitations of Current Relevance Search Methods

Current relevance measures are applied to small, bibliographic networks or sparse, disease specific networks generated from subsets of data found in monolithic sources. The sparsity of linked data between agricultural and biomedical research is exacerbated by the isolation of existing associations in domain specific data silos with low interoperability, such as those hosted by the USDA. This work overcomes the sparsity of association data connecting agriculture and biomedicine through integration of current data and augmentation with text mined relationships extracted from scientific literature. The combination of these techniques results in the development of a large, semantically rich diet-disease network.

Recent network analysis methods focus on similarity search within homogeneous networks. These measures are based on incomplete models of the complex networks they represent and ignore the latent semantic information in association data. Most measures developed for heterogeneous networks search for similarity between objects of the same type, neglecting the potential value of quantifying the relevance between differently typed objects. Measures designed for relevance search fail to incorporate all possible semantic detail. The heterogeneity of link types extracted from text mining and curated sources can provide detailed semantics which affect the relatedness of two objects. This work considers link types mined from the literature and curated

sources to better inform the relevance between objects.

1.4 Improving LBD with Semantically Informed Heterogeneous Network Analysis

It is difficult to draw meaningful conclusions from the sparse data available linking plant-based diets with human health. Current methods attempt to overcome data sparsity by integrating existing curated data, but curated associations connecting agricultural and biomedical research exhibit low interoperability and are generally not publicly available. My method integrates data from public sources of agricultural and biomedical associations. My method also augments this data with text mined relationships that provide support for low probability associations and bridge the gap in association data between plants, chemicals, and human health. The extraction of text mined relationships add heterogeneity and subtle semantic information that can be incorporated into heterogeneous information network analysis tasks, such as relevance search. My semantically informed technique helps generate phytochemical profiles, hypotheses for diet and gene interaction, and enables the elucidation of molecular mechanisms by which plants affect human health phenotypes.

This research produces a diet-disease network that supplements curated data with text mining and a semantically informed relevance search measure for ranking related agricultural and biomedical entities. This work investigates three separate, but related, topics: data integration and augmentation methods for disparate data sources, a heterogeneous network analysis method for relevance search, and a large-scale relevance analysis of agricultural and biomedical entities in an integrated diet-disease network. **Figure 2** shows the overall workflow implemented in this work to investigate

the molecular mechanisms behind plant-based diet and human health phenotypes.

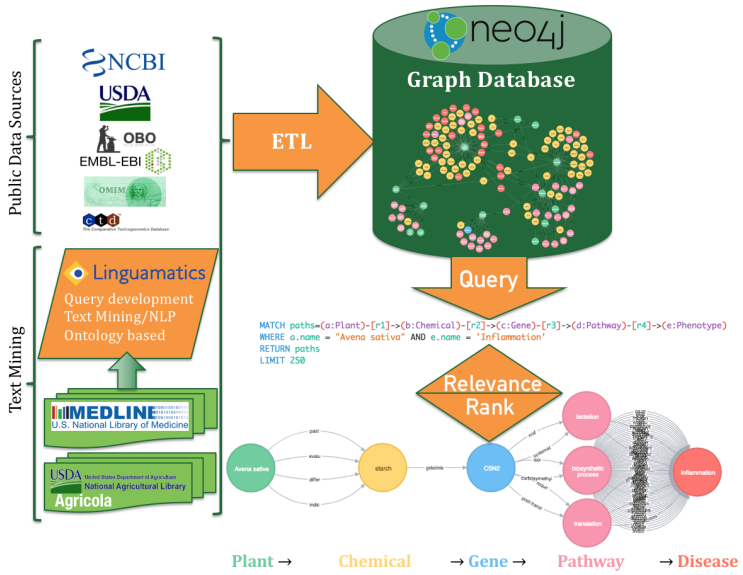


Figure 2: This is the general workflow used in this research. The curated sources and text mined results from scientific literature are identified and subjected to several data integration techniques. The association data is then transformed for storage in a Neo4j graph database. Relevance calculations are performed on results from database queries and prioritized objects and paths are returned as tables. Ranked results can be visualized in the Neo4j Browser.

Chapter 2 discusses the development of a novel diet-disease network, stored in a graph database. The sources, format, and types of data for the network, both curated and mined, are described in detail. Concepts of graph theory and applicability to the computational problem of linking multiple domains are also explained.

Chapter 3 describes the design, development, and implementation of a meta path based relevance ranking measure for generating candidate hypotheses. The identification and quantification of relevant objects from a heterogeneous network of curated and text mined data is a difficult task. It is necessary to filter the noise from the explosion of information available in agriculture and biomedicine. Recently, the relevance search task is more commonly approached with machine learning techniques

that require large, curated training datasets which are difficult to generate and specific a research domain. The novel ranking measure HetRel is semi-supervised, requiring little input from the user. HetRel is evaluated and compared to existing relevance measures using gold standard datasets, ranked by each measure.

Chapter 4 investigates the biological use cases of the novel diet-disease network. It also implements the HetRel ranking measure described in the previous chapters. This chapter showcases the utility of the diet-disease network and the HetRel measure when used in concert.

Chapter 5 briefly describes limitations of this work and the future improvements and goals which can overcome them. It outlines the applicability of the modular framework for many other domain linking problems relevant to research areas in human health phenotypes.

CHAPTER 2: DEVELOPMENT OF A DIET-DISEASE ASSOCIATION NETWORK

The advent of HT technology has produced vast amounts of data in biomedical research that allows a broader view of biological interactions at the systems level. The rise in read length and depth, coupled with the falling cost of genome sequencing has afforded new insight into plant products and function. The last decade has seen a growing trend where contextual evidence is incorporated across domains to draw meaningful conclusions from the influx of data, as evidenced by the field of nutrigenomics that diverged from nutrition science [117]. Elucidating the molecular mechanisms behind the health benefits of plant-based foods requires the aggregation, integration, and curation of data from relevant domains.

One of the main obstacles to identifying the mechanisms by which diet affects disease is that a comprehensive resource which integrates knowledge from the literature with relevant contextual data does not exist. Present databases describing diet and disease entities limit their scope to specific domains, exhibit sparse associations, and lack descriptive relationships for their associations. Additively, these issues impede the discovery of plausible molecular mechanisms of diet and disease. We developed a novel diet-disease network that integrates heterogeneous biological domains and enriches current associations to address these obstacles.

Developing a diet-disease network requires the identification, aggregation, and inte-

gration of data from relevant information domains. Data from available sources need to be understood for proper integration and subsequent analyses. The data should follow accepted standards (such as those outlined by the OBO Foundry [164]), contain cross references to other pertinent sources, be comprehensive, and properly represent molecular biology. A great example is the Gene Ontology (GO) which provides definitions and relationships for concepts describing gene and gene product function across species in both OBO and OWL formats [10]. The GO has been adopted for gene annotation by all major genomic repositories, generating associations across many prevalent databases. Data that meet these criteria facilitate the integration process. Numerous public sources contain such data, yet a dearth of associations and relationships linking different domains still exists.

Within the published scientific literature are relationships that link otherwise separate domains of knowledge. Recent studies have showcased the power of text mining, particularly in the field of biomedicine, for integrating knowledge from literature and other relevant data sources [13,30,69,153]. Text mining is a high throughput means of aggregating information from the literature, but returns a higher error rate than manually curated associations. The combination of curated associations from databases with text mined associations has been found to result in a larger collection of results, while maintaining an adequate error rate [81, 135]. In this work, text mining is used to extract relationships from titles and abstracts in PubMed and Agricola to augment the associations found in and between the agricultural and biomedical domains.

A comprehensive dataset that integrates agricultural and biomedical entities provides a foundation for studying the relationship between plant-based foods and human

disease. It is costly from a time and resources standpoint to design, perform, and analyze large random assays to determine phytochemical profiles or molecular interactions of plant-based foods. It is more feasible to assimilate current knowledge of plant products and their effects on the human genome to generate more focused, data-driven hypotheses for further inquiry. A heterogeneous network that integrates domains in agriculture and biomedicine broadens the resources for understanding the molecular mechanisms of diet and disease.

Proper data integration is crucial for downstream analysis in LBD and can affect downstream results if done poorly. Integration methods that broadly match entities can lead to an overabundance of associations and relationships between domains which creates excess noise within the dataset. On the other hand, overly specific integration methods that match few entities can suffer from a sparsity of associations and relationships which reduces the signal from the dataset. To that end, this chapter discusses solutions to overcome challenges posed by the integration of biological entities that describe diet and disease. Biological entities included in the integration process include plant species, chemicals and compounds, genes and gene products, biological pathways, and human phenotypes with a focus on disease.

2.1 Data Sources and Types

There are many agricultural and biomedical data sources for the elucidation of molecular mechanisms linking plant-based diets and disease. These available sources can be classified into three broad data formats: 1)controlled vocabularies, 2)association databases, and 3)text mined associations and relationships. Controlled vocabu-

laries provide a list of explicitly defined, unambiguous, and non-redundant terms to which describe a domain. Secondary databases consist of data derived from analyses of experimental data found in primary databases. These databases store associations and relationships determined from experimental studies that were gathered into a domain specific resource. Text mined associations and relationships are extracted from the literature, as previously described in **Chapter 1**, to quickly capture existing knowledge that may not have been assimilated into other sources. When integrated, the data from these three sources form a consistent representation of plant-based diets and disease from which hypotheses can be generated.

Determining which sources and data to include within the network was essential to the investigating the effects of diet on disease. Each source has distinct properties, influenced by objectives and standards adopted by that particular project, which dictate the source's utility in achieving project goals. The goals of the project direct considerations of pertinence for entities and associations of sources. For example, the Gene Ontology project provides a structured vocabulary that defines gene and gene product functions, cellular locations, and involvement in biological processes. However, the Gene Ontology does not contain genome and gene-specific information such as gene names and the species genes are found in. Entrez Gene stores genomes and gene-specific information, but must cross reference other repositories, such as the Gene Ontology, to associate genes with function.

These sources span various domains and are built on different assumptions and datasets that produce dissimilar formats, making interoperability difficult. **Table 2** catalogs the curated sources incorporated into the diet-disease network by their data

formats and lists the data types extracted from each. To understand the complexity involved in integrating heterogeneous data, the data types of all sources are described in greater detail.

Table 2: Data sources integrated into the Diet-Disease Network are listed along with their format and entity types. Data formats: **CV**- Controlled vocabularies, **Asc**- Association database, **Citation**- Citation database. Data types: **EID**- entity identifier, **EN**- preferred entity name, **Assoc**- entity associations/relationships

Data Source	**Data Format**	**Data Types**	**Entity Identifier**	**Entity Name**	**References**
NCBI Taxonomy	CV	EIDs,ENs,Assocs	36774	*Brassica oleracea var. italica*	[52]
Medical Subject Headings	CV	EIDs,ENs,Assocs	D015179	Colorectal Neoplasms	[100]
Chemical Entities of Biological Interest	CV	EIDs,ENs,Assocs	CHEBI:24279	glucosinolate	[72]
National Agricultural Thesaurus	CV	EIDs,ENs,Assocs	6949	alpha-tocopherol	[125]
USDA Nutrient Database	CV	EIDs,ENs,Assocs	341	Tocopherol, beta	[182]
Gene Ontology	CV	EIDs,ENs,Assocs	GO:0034599	cellular response to oxidative stress	[10]
Disease Ontology	CV	EIDs,ENs,Assocs	DOID:11981	morbid obesity	[152]
Human Phenotype Ontology	CV	EIDs,ENs,Assocs	HP:0001022	Albinism	[145]
Mammalian Phenotype Ontology	CV	EIDs,ENs,Assocs	MP:0014159	stomach fibrosis	[165]
Plant Ontology	CV	EIDs,ENs,Assocs	PO:0005849	primary xylem	[12]
Entrez Gene	Asc	EIDs,ENs,Assocs	4780	NFE2L2	[106]
Online Mendelian Inheritance in Man Database	Asc	EIDs,ENs,Assocs	614594	Olmsted syndrome	[70]
Comparative Toxicogenomics Database	Asc	Assocs			[42]
PubMed	Citation	Assocs			[60]
AGRICOLA	Citation	Assocs			[123]

2.1.1 Controlled Vocabularies

The main purpose of a controlled vocabulary is to express the knowledge of a domain in a standardized, reusable format to resolve ambiguity between concepts. A controlled vocabulary is a defined list of terms for use within a domain. Further information, such as associations, can be added to a controlled vocabulary to provide it with structure. Controlled vocabularies can be categorized by the type of data contained and format followed. Three categories of controlled vocabularies commonly utilized in integration analyses are taxonomies, thesauri, and ontologies.

Taxonomies are controlled vocabularies that contain parent-child relationships that form a hierarchical structure. The NCBI Taxonomy is a prime example of a taxonomy, with parent-child relationships structured by biological taxonomic rank [52]. Taxonomies contain, at minimum, three types of data that include an entity identifier, parent-child relationships, and metadata, which in this case included preferred term names and possible synonyms. Aside from describing taxonomic homology amongst species, a taxonomy alone provides little in the way of investigating the effects of plant based foods on human health.

Thesauri also contain relationships that form a structured vocabulary. They include declarative links of varying expressivity between entities. Relationships such as "broader term" and "narrower term" further specify parent-child associations, like those expressed in taxonomies. A thesaurus also stores additional metadata such as synonyms, cross references to other structured vocabularies and databases, descriptions, and dates of discovery. The prominent thesaurus in biomedical research is the

Medical Subject Headings Thesaurus (MeSH) from the NCBI National Library of Medicine [100]. In agriculture, the most widely used thesaurus is the National Agricultural Library Thesaurus from the United States Department of Agriculture [125]. Synonyms, cross references, and descriptions were extracted from these two thesauri for integration into the diet-disease network. Thesauri cover a broad range of entities but, in doing so, may lack depth and detail.

Ontologies are hierarchically structured vocabularies with relationships and attributes specific to a particular domain. The purpose of an ontology is to create a reliable semantic specification to promote interoperability and communication within a domain. Relationships in ontologies are the most expressive links found in structured vocabularies. The Gene Ontology (GO), ubiquitous in life science research as evidenced by tens of thousands of article citations, serves as a great example of expressive relationships in ontologies [10]. There are 11 relation types of varying detail used to describe the links between gene products within the GO. A relation provides semantic meaning to a relationship far better than a simple parent-child association, like those of thesauri and taxonomies. Relations, such as *negatively_regulates* or *positively_regulates*, yield further semantic detail and add negative or positive directionality to the link. Other ontologies contain distinctive relations for phenotypes and disease (*has_symptom*), chemicals (*is_conjugate_base_of*), and plants (*isolated_from_germplasm*). Data extracted from ontologies consisted of entity identifiers, preferred terms, synonyms, descriptions, and relation types. Ontologies provide a wealth of detail for specific domains, but suffer from sparsity of entity links across domains.

Controlled vocabularies are made available in numerous data formats. The NCBI Taxonomy is separated into a number of files that contain the species identifiers, preferred terms, and synonyms. These files are structured into rows and columns, delimited by specific characters. In the case of the NCBI Taxonomy, columns are separated with a combination of tabs and pipes, as such: [*column1Value* | *column2Value*]. Thesauri and ontologies are commonly self-contained within single files that are structured into stanzas for each entity definition. A stanza is a related group of lines that consist of the entity identifier and all other properties of that entity, such as preferred term, synonyms, and associations. The OBO format serves as a standard for biological ontologies. The GO and all other ontologies used in this work followed the OBO format for ease of parsing and extensibility. Links to parsers for the controlled vocabularies used to populate the diet-disease network can be found in **Appendix A**.

2.1.2 Association Databases

NCBI and EBI host biomedical databases that act as reference libraries to researchers across all domains of the biological sciences. The main purpose of these databases is to classify and provide access to biomedical data. Inherently, classifying the data involves standardization, which aids in downstream interoperability and ease of access for these databases.

Primary and secondary data can be stored in association databases. Experimentally obtained data in primary databases is submitted directly from scientists into primary databases. Secondary databases assimilate primary data into a collection to

investigate broader patterns through data mining. Databases serve the dual purpose of defining biological entities, such as genes in Entrez Gene, and linking these entities across multiple sources. A thesaurus or ontology acts as a foundation for databases, defining concept classes that classify instances of data. When based on a thesaurus or ontology, a database will inherit entity identifiers, preferred terms, synonyms, descriptions and associations from that structured vocabulary. Databases build upon an ontology by incorporating curated associations between internal entities and cross references derived from external databases and structured vocabularies. The amount and variety of metadata and associations available in association databases imparts greater density to the skeletal structure of structured vocabularies. The wealth of associations, both internal and external, from databases add significant detail and interoperability to the diet-disease network described in this work.

The Entrez Gene Database is an example of a primary database for gene information where researchers directly deposit experimentally derived data. Entrez Gene entity identifiers and links to species, pathways, and phenotypes were extracted as definitions of genes and their associations with other pertinent entities involved in diet to disease interactions. OMIM and CTD represent secondary association databases for chemical, gene, pathway, and phenotype information amalgamated from various sources. These secondary databases draw information from MeSH, NCBI Taxonomy, GO, and other ontologies to associate chemicals, genes, and disease. OMIM contains detailed text summaries of curated instances of associations from the scientific literature. As previously mentioned, developers of biomedical databases use standard ontologies and thesauri to not only classify, but to foster the sharing of data and

information.

Data from biomedical databases can be accessed in variety of ways in part due to an ability to be indexed from the structure of ontologies and thesauri. NCBI has developed web portals for all resources it hosts, including NCBI Taxonomy, MeSH, OMIM, and Entrez Gene. Web portals such as these provide quick single query access to data within these resources. For larger, multi-step and batch queries, biomedical repositories allow programmatic access via application programming interface tools, such as NCBI E-Utilities. Studies that require the entirety of data in these databases are able to download database or flat files directly from their respective ftp sites.

Data from biomedical repositories are commonly available in two file formats, flat file records or as character delimited files. The database files used in this work were downloaded from FTP sites in the form of multiple character delimited files per database. These delimited files required entity mapping to one another, similar to that of tables within a relational database. This concept of linking tables with key fields in relational databases will be explained in greater detail in the following section. Links to parsers for data extraction from databases used within the diet-disease network can be found in **Appendix A**.

2.1.3 Data Types

Data from controlled vocabularies and association databases must be parsed from original source files for use in computational analyses. The foremost consideration in data integration is deciding which data types best describe and connect agricultural and biomedical entities. The data types that define agricultural and biomedical con-

cepts are entity identifiers, associations that connect entities, and metadata. All data sources utilized to develop the diet-disease network contain, at minimum, these three data types.

Entity identifiers (EIDs) for biological concepts are the most prevalent type of data in agricultural and biomedical sources. EIDs act as unique references to specific agricultural or biomedical concepts, such as a species, chemical, or gene. They can be characters, numbers, letters, or a combination of those. **Table 2** displays examples of entity identifiers from each of the three data source formats. The degree of uniqueness of an identifier varies by data source. NCBI Taxonomy and Entrez Gene employ arbitrary EIDs to species and genes. The EIDs are a series of sequential numbers assigned to entities based on the order that data was input into the database. As such, these EIDs are only unique to specific entities within their respective databases. Ontologies, such as the GO, use a combination of letters, characters, and numbers to ensure uniqueness and opacity (when an identifier provides no metadata, such as the order of data entry). The GO has standardized EIDs to follow this particular format of a prefix, "GO:", followed by a seven digit number with leading zeroes. For example, the EID for *chronic inflammatory response* is *GO:0002544*. Opaque identifiers are critical for combining and integrating data from disparate sources.

Entity associations connect agricultural and biomedical concepts within sources and provide cross references to other sources. These associations are stored in structured vocabularies and databases. Internal associations, between entities within a source, occur frequently in structured vocabularies. Association databases consist mainly of cross reference associations that connect various external sources.

Table 3: The relationship types extracted from various data sources and formats are listed. When relationships had no explicit type: Internal relationships were issued the *is_a* type, External relationships were issued the *xref* type

Data Source	**Data Format**	**Relationship Types**	**References**
Gene Ontology	CV	*occurs_in*	[10]
Chemical Entities of Biological Interest	CV	*has_role*	[72]
Disease Ontology	CV	*complicated_by*	[152]
Plant Ontology	CV	*isolated_from_germplasm*	[12]
NCBI Entrez Gene	Asc	*xref*	[106]
PubMed	Citation	*produc*	[60]

Connections between entities have varying levels of detail. Generally, associations are expressed as mapped pairs of EIDs. An association identifies the existence of a simple link between two entities. Building upon that, a relationship assigns a descriptive type to links between entities. Descriptive types are standardized by structured vocabularies, namely ontologies such as the Gene Ontology. Examples of relationship types are provided in **Table 3**. In this case, a relationship forms a triple of the format *(EID, relationship type, EID)*. The addition of relationship types aids in standardizing complex associations.

The associations and relationships available from the sources previously described are highly curated. Manual curation demands high costs in both time and labor, which limits the scope and quantity of associations from these sources. The diet-disease network presented here integrates resources using unique internal identifiers and expands the scope of data through the aggregation of associations across domains. It provides a single access point to dense amounts of highly curated associations in a standard format.

Metadata includes information relevant to the association, such as data version or the publication identifiers from which data was retrieved. It is crucial in integration

and reproducibility of analyses to maintain the history and provenance of source data. For instance, every updated version of a structured vocabulary will add new entities and associations. If the versions of two integrated sources are out of sync, adding associations from an entity in the updated source to a non-existent entity in the out of date source would fail. This work has recorded versions, dates of access, and noted manual alterations to data acquired from aforementioned sources in order to track metadata.

Optional data types contained by these sources include entity definitions and synonym lists. Entity definitions are text explanations of the given entity. Synonym lists reduce entity ambiguity by mapping common terms sharing the same meaning to a single, preferred term, such as broccoli, a synonym falling under the preferred entity term *Brassica oleracea var. italica*.

This compilation of entities is essential for developing text mining queries, which extract valuable information from scientific literature. Volumes of literature are available to create phytochemical profiles for plants found in the human diet. Fortunately, the data aggregated from publicly available sources including ontologies, taxonomies, food composition sources, and chemical databases aids text mining methods in the identification of entities and their associations in the literature.

2.2 Text Mining Workflow

Even with the integration of curated sources boasting millions of associations, there exists a sparsity of computationally accessible connections between plant-based foods and human health. A lack of associations, particularly between plants, chemicals,

and human genes, causes a disconnect between plant-based foods and their effects on human health.

The published scientific literature is a valuable resource for augmenting existing associations in agriculture and biomedicine. Reading and assimilating the information and knowledge within the vast amount of literature is not humanly possible. In order to access such knowledge, a high throughput method capable of mining the literature for relationships between relevant entities and extracting the data in a machine readable format is necessary. Text mining methods, as were discussed in **Chapter 1**, are able to perform these tasks.

Text mining the scientific literature serves three purposes within knowledge based discovery. The extracted knowledge from text mining creates a review of the current research in agriculture and biomedicine, adds semantic detail to entity associations, and augments sparse associations from current repositories.

Mining scientific articles provides a cursory view of the multiple facets in agricultural and biomedical research. Raw results from text mining uncover popular research trends in specific plants and phytochemicals with health benefits. Associations of phytochemicals to agronomic plants of interest extracted from the scientific literature communicate the disparity in research for different plant-based foods. **Figure 3** displays the top 10 plants with the highest citation count of associations with chemicals, which include corn, rice, wheat, soybean, and spinach. Cruciferous vegetables (*Brassicaceae*) show far fewer phytochemical associations from the scientific literature, which suggests the potential for further research.

A drawback of simple associations is the constricted expressivity of associations

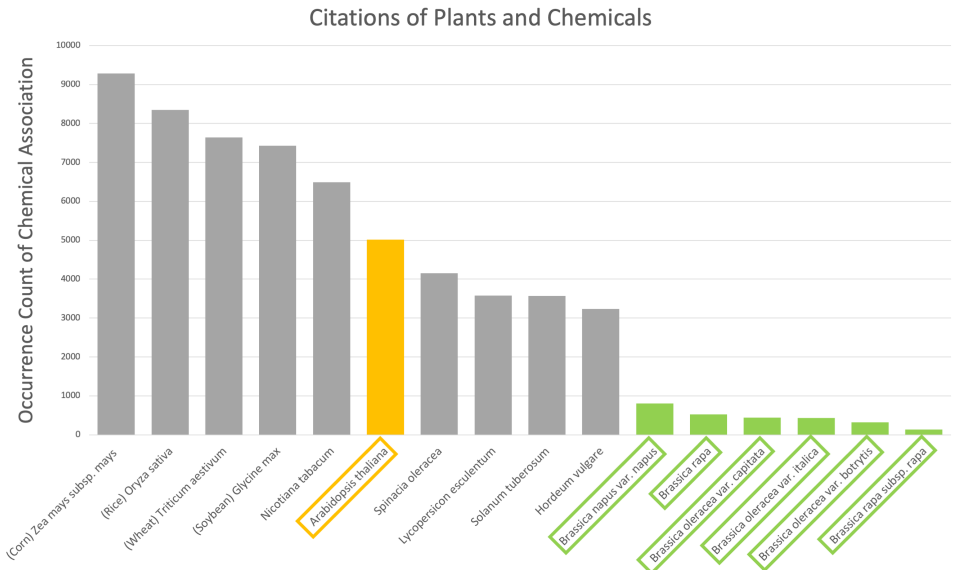


Figure 3: Bar plot depicts the state of research for agriculture. The first 10 plants have the highest citation count for plant to chemical associations in articles from PubMed and Agricola. The remaining 6 plants in the plot represent research in the cruciferous vegetable family *Brassicaceae*. The juxtaposition of the two types highlights the gap in knowledge in agricultural research.

between entities. The vast majority of text mining studies for knowledge based discovery rely on co-occurrence methods of simple associations [75,77,81]. Co-occurrence mining methods exhibit high recall, returning large numbers of associations. These methods return instances where two entities occur within a text. The definition of text differs between implementations, where text could be a document, title, abstract, article section (e.g. Results), paragraph, or sentence. Many studies establish a text as an entire abstract in order to reduce computation during mining and improve recall. In many instances, these associations are completely unrelated aside from being mentioned within the same abstract. The relationship of the co-occurring entities is unknown without reading the context of the association, impeding the speed of interpretation.

NLP based text mining methods begin with co-occurrence techniques and refine

results based on grammatical and semantic features surrounding the co-occurrence. The benefit of NLP methods is their ability to identify and extract extra information from the co-occurrence, such as grammatical verbs and actions (semantic predicates or relationship types). Semantic predicates provide meaning to associations between entities, as opposed to manually identifying their meaning from co-occurrence or concept derived associations [78,143]. For example, extracting the semantic predicate *(dietary fish oils INHIBITS platelet aggregation)* indicates the specific relationship between DFOs and platelet aggregation, rather than a simple association between the entities. Semantic predicates extracted by text mining will be overlaid on the diet-disease network generated from ontologies and association databases. The extraction of relationship types will return a smaller number of associated entities because, within these extracted relationships, co-occurrences are known to exist but many are filtered out due to a strict set of rules in text mining queries. An example of a rule defines the boundary of a text, where a text mining result is returned if two entities co-occur within the same sentence or phrase, as opposed to the more lax boundaries of the entire abstract.

Text mining is capable of programmatically and comprehensively adding entity connections to current research databases. The benefit of text mining published literature in agriculture and biomedicine is that these relationships are accepted as truths, having been published in peer-reviewed journals in their respective fields. With current sources of agriculture and biomedical research, it is impossible to find a connection linking plants, their chemicals, and human genes, pathways, and phenotypes. Explicitly mining for relationships between each set of these entities provides

a large collection of associations to augment current sources, enabling researchers to make that connection.

The methods applied in this work were chosen to take advantage of the benefits of text mining just described. These methods consist of defining the document corpus, designing or identifying text mining softwares and pipelines, query development and result extraction, and an overview of text mining results.

2.2.1 Document Corpus

The main input of text mining is the collection of documents for information extraction. The three requirements for selecting a document corpus for text mining are a defined scope, evaluation of information quality, and ease of accessibility.

Studies which utilize large text corpuses, such as Wikipedia [114, 198], attempt to identify general trends and information. The problem with using general texts as the source input for text mining is that the information, although vast and comprehensive, do not define a specific scope, which introduces noise in mined results. Also, mining irrelevant text unnecessarily increases the need for computational time and resources. In addition, such large text sources are not always curated, which leads to variability in information quality. However, sources such as Wikipedia are publicly available and easily accessible via download.

A combination of the citation databases Medline and Agricola were selected to fulfill the selection requirements for the document corpus in this work. This study encompasses the fields of biomedical and agricultural research. Medline is the standard text corpus for biomedical text mining and is used exclusively in the majority

of studies [13, 24, 30, 69, 154]. The agricultural equivalent to Medline is Agricola. To date, this work is the first to text mine the Agricola bibliographic database. As the scope of the project includes agriculture, it is necessary to incorporate relevant agricultural research texts from a reputable source. Together, these biographical databases comprehensively represent current knowledge in the scope of biomedicine and agriculture. The millions of citations in these databases are amassed from thousands of peer-reviewed journals, theses, and book chapters which ensures high quality information. Medline and AGRICOLA are also publicly available in common file formats. In biomedical text mining, bibliographic records are formatted using available metadata, such as journal, publication date, and authors, based on the text mining method or software used.

2.2.2 Text Mining Tools

Following document corpus selection and formatting, the next step in text mining is designing or implementing a text mining tool. The important features of text mining tools are the underlying text mining methods and licensing availability.

As previously discussed in **Chapter 1**, many text mining methods exist such as co-occurrence, statistical, machine learning, and NLP based methods [34, 85, 200]. Early tools relied heavily on co-occurrence as the primary method of text mining, generating high recall at the expense of precision. Many associations extracted by these methods returned false positives, requiring filtering and curation. Studies employed these tools and methods for their simplicity, high recall, and open source licensing [34]. More recently, hybrid approaches have been found to manage the shortfalls of previous

techniques by combining text mining methods, resulting in higher precision with a reasonable cost to recall [53, 56, 81]. Higher precision, hybrid approaches have become prevalent in current open source and proprietary text mining tools used in biomedical research. A major difference between open source and proprietary tools is the level of support, in the form of documentation and developer wikis, available to researchers.

Several open source and proprietary tools are available for text mining in the context of biomedical research. Some of these tools are discussed in the following section.

2.2.2.1 Open Source Tools

The foundation of many open source text mining tools is the Stanford Natural Language Toolkit (NLTK) [25, 31, 147]. The NLTK is a suite of tools for developing text mining programs in the Python programming language. It includes modules for text parsing, classification, natural language processing algorithms, and visualizations [21]. The corpora and controlled vocabularies included in the NLTK are generic, meant for broader mining projects with input text usually unrelated to agriculture or biomedicine. The NLTK provides a platform for biomedical text mining tools to be built on, but is purposely non-specific for use across any domain. As a result, it is the most widely used open source tool with vast and constantly updated documentation and community support. The major limitation of the NLTK is the programming language it supports. Python has a shallow learning curve that reduces the barrier to entry for new users, has a rapid development and testing cycle, is fairly human readable, and is object oriented [104]. However, it is an arduous task to develop multi-threaded programs in Python, making it difficult to scale NLTK-based

text mining tools for big data [149]. The document corpus in this project is over 30 million documents, which can prove computationally intensive when there is an inability to run text mining processes in parallel.

The tm package is a framework for developing text mining applications within the R programming language. This package in R is capable of preprocessing text data, performing association analyses, document clustering, concept summarization, and text classification [111]. R makes use of co-occurrence and statistics methods to perform clustering and association analyses. The advantage of the tm package is that it can extend other R packages such as OpenNLP and lsa [14, 190]. Text mining with the tm package provides access to advanced statistical methods, such as latent semantic analysis, and open source NLP algorithms. The statistical methods in R are invaluable for text and data mining, but the R programming language is not able to scale to handle extremely large datasets. With a document corpus of over 30 million documents, running advanced statistical methods on complex sentences within this many documents would be not be feasible in a reasonable amount of time.

WhatIzIt is a suite of open-source text mining modules hosted by the EBI for biomedical literature, such as the EBI installation of Medline. This tool has modules specific to named entity types, such as organisms, chemicals, and diseases [138]. WhatIzIt is specific to biomedical literature, drawing from controlled vocabularies such as ChEBI, GO, and NCBI Taxonomy. It takes into account morphological variability and includes ambiguous acronyms for named entities. The tool is hosted as a webservice to scale with the size of literature in Medline and number of controlled vocabularies available for named entity recognition. However, as a webservice there

are limitations to batch processing and ability to configure custom functionalities.

MetaMap is a highly configurable text mining tool that maps concepts from the highly curated unified medical language system metathesaurus to biomedical literature. This tool can be accessed as a webservice or installed to run on a local machine, allowing highly configurable batch processing. MetaMap is heavily computational and based on symbolic, NLP, and linguistic techniques [8]. The drawback of MetaMap's exhaustive thoroughness in concept mapping is its relative slowness, making it inappropriate for real-time use and iterative refinement. Complex sentences can generate hundreds of thousands of candidate concept mappings which increases the problem of ambiguity, requiring several hours for MetaMap to complete [9].

2.2.2.2 Proprietary Tools

Statistical Analysis Text Miner, a product of SAS, implements text mining techniques to identify patterns found within the entirety of a document collection. SAS differentiates the definition of text mining from NLP and knowledge extraction, citing the single document specificity of the former compared to the overall collection in the latter [3]. The Text Miner preprocesses and parses free-format text with text mining techniques to transform it into structured data and make it available to data mining algorithms. The SAS Text Miner is capable of text clustering, classification, and identifying essential concepts in the context of an entire text collection. To that end, SAS Text Miner has been implemented in sentiment analyses, business intelligence, and the development of drug efficacy models through prediction. This work requires not only a summarization of trends within a collection, but also document specific

information extraction including relation extraction. For this reason, SAS Text Miner does not fulfill the requirements of a text mining tool for this project.

I2E is an NLP based text mining platform developed by Linguamatics for the task of information extraction from large document collections across numerous domains [115]. I2E provides an interactive information extraction tool that allows users to identify relationships between entities of interest using user-defined ontologies, taxonomies, and thesauri in combination with rule-based pattern matching, NLP, and linguistic algorithms to help define the context of a query. This platform supports input from external domain sources to help define context. Queries are built accessing classes of concepts defined by ontologies combined with linguistic patterns. A beneficial feature is that users can develop and refine queries in real time over millions of documents. A simple query against an indexed Medline collection (over 24 million citations) for eleven micronutrients in the Brassica family and ten potential molecular interactions in mammals takes under 28 seconds. Results from I2E queries are returned in configurable, machine-readable, structured formats for further analysis. The efficiency, advanced NLP mining methods, and extensive configurability of the I2E system, coupled with great support from Linguamatics subject matter experts and an active user community, fulfilled the needs of this project.

I2E has successfully been implemented in domains of the life sciences for various studies. Bandy *et al.* extracted protein-protein interactions to investigate the associations between a set of 50 genes of interest [15]. Liu *et al.* utilized I2E results to train an algorithm that automatically categorized pneumonia diagnoses from chest x-ray reports [103]. Tari *et al.* incorporated I2E into a drug target and biomarker discovery

pipeline. Linguamatics I2E will be used in this work to extract relationships between entities describing the effects of plant based diets and disease from the abstracts of over 30 million scientific articles in Agricola and PubMed.

2.2.3 Text Mining with Linguamatics I2E

The text mining workflow in I2E consists of three major steps that are conceptualized in **Figure 4**. It starts with the acquisition, preprocessing, and indexing of the document corpus to be mined. In parallel, controlled vocabularies and ontologies are also acquired and indexed to form "classes" which represent named entities. This step generally occurs once per corpus and controlled vocabulary update. Indexing is followed by specific query development, which produces query results. Query development and results are involved in an iterative loop of refinement that continues until the results meet the user's standards. The details of controlled vocabulary selection and query development and refinement for investigating diet and disease relationships are discussed in this section.

2.2.3.1 Named Entities Indexed for Text Mining

I2E was developed as an ontology-based interactive information extraction system, providing background knowledge through the incorporation of domain specific ontologies. Entities from these ontologies are treated as classes in I2E, retaining relationship structure and sets of synonyms for each class. This inherited knowledge from ontologies enables querying at a conceptual level, requiring little to no knowledge of all the synonyms or subsumed (included under a broader definition) entities of a class. It also allows for queries between a specific entity or a family of entities. For example, if one

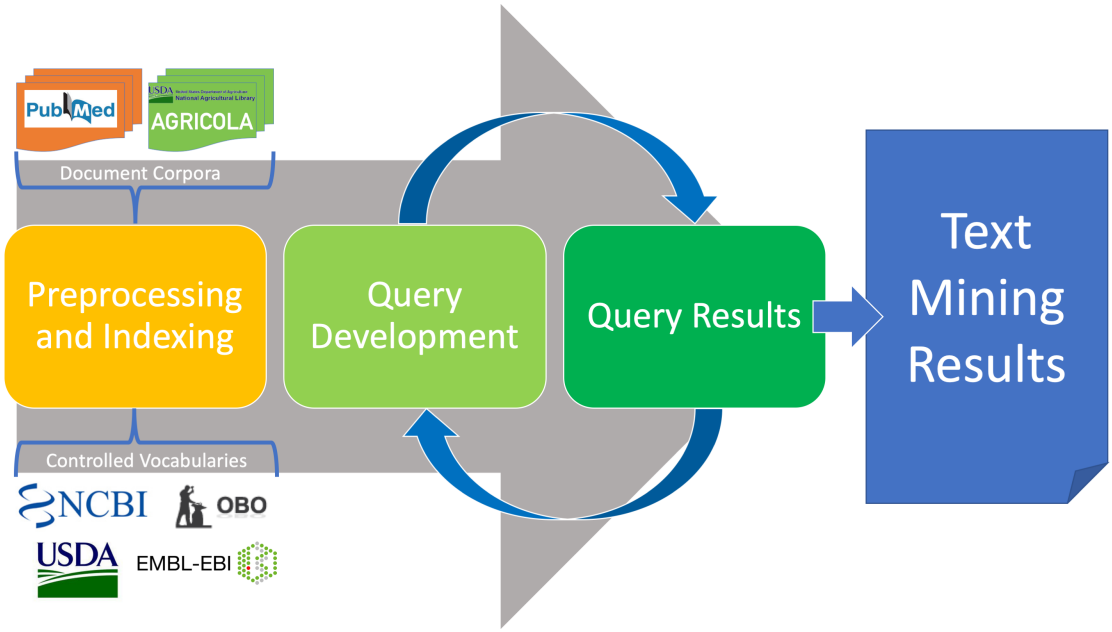


Figure 4: Overview of three step text mining workflow. Workflow begins with induction, preprocessing, and indexing of document corpora and controlled vocabularies. Named entities from controlled vocabularies are used in query development and iterative refinement of query results. Following result refinement, text mining results are exported for further use.

were interested in mining the literature for the phytochemical profile of broccoli, they could select the specific class of broccoli from the NCBI Taxonomy and the more general class of chemical entities from ChEBI. I2E would search the corpus for all synonyms of *Brassica oleracea var. italica*, the preferred term for broccoli, and their occurrences with any chemical subsumed by the chemical entity class of ChEBI, such as glucosinolate or glucoraphanin. Without background knowledge from ontologies, a query such as this would prove arduous, requiring manual compilation of broccoli synonyms and list of chemicals.

To elucidate the molecular mechanisms of diet on disease, six structured vocabularies describing plants, chemicals, genes, biological pathways, and human health phenotypes (described in **Chapter 1**) are indexed as named entity classes in I2E.

As MeSH and the NAL Thesaurus index the entirety of the corpus, named entities from these structured vocabularies are heavily utilized in text mining for this project. **Table 4** separates named entity classes into distinct types based on domain and identifies the structured vocabularies that define them within the context of this text mining project.

Table 4: The entity types of nodes in the diet-disease network and their sources.

Entity Type	**Data Sources**
Plant	NCBI Taxonomy(Embryophyta)
	MeSH(Embryophyta)
	NALT(Plantae)
Chemical	ChEBI(chemical entity)
	MeSH(Chemicals and Drugs Branch)
	NALT(Biochemical compounds)
Gene	Entrez Gene(*Homo sapiens*)
Pathway	GO(Biological Process)
	MeSH(Diseases Branch)
Phenotype	Disease ontology
	NALT(diseases and disorders)
	MeSH(Diseases Branch)

2.2.3.2 Query Development

In this work, the overall goal of text mining is to augment the relationships linking the domains of diet and disease at the molecular level. This goal dictates entities of interest and the structure of the relationships that connect them. The biological inquiry can be broken down into four more manageable questions: 1) What phytochemicals are associated to plants? 2) What effect do phytochemicals have on human gene expression? 3) How does altered gene expression affect biological pathways in humans? and 4) How do perturbations in biological pathways affect human phenotypes? Four sets of text mining queries, connecting five domains, were developed to answer these questions.

As previously mentioned, the format of extracted entity relationships in this work follows that of RDF triples. Triples add detail to associations found in the literature by providing semantic predicates for the extracted relationship. Additionally, the extraction of triples acts as a filter by restricting entity associations to those connected by a semantic predicate within the boundary of a sentence. For example, if broccoli and a pesticide are mentioned within the same article abstract, the pair would not be extracted as an association. If broccoli and isothiocyanates are mentioned within the same sentence and are connected by a predicate, such as contains or produces, then the triple *(broccoli, contains, isothiocyanates)* would be returned as a mined result. With this in mind, the iterative process of query development can be described.

Each set of queries begins with a combination of named entity classes defined in **Table 4**. The classes chosen to represent the five domains are broad and subsume a multitude of classes that can introduce spurious results if not carefully excluded. Universal exceptions in text mining exist to reduce noise introduced by common terms and phrases. The natural language group at Stanford introduced a stop word list that was included within each query developed in this project. Otherwise, exclusions varied based on the context of the biological question, as well as specificity of the selected classes. The most biologically relevant exclusions were in plant, chemical, and phenotype entity types. In order to be comprehensive, all green land plants are considered to be edible and were included in this study. In the NCBI Taxonomy and MeSH, Embryophyta, also referred to as Plantae in the NAL Thesaurus, is the classification of terrestrial plant species indexed for text mining for this analysis, excluding Chlorophyta, green algae. In chemical classes, ChEBI specifically, fertilizers,

pesticides, and chemical solutions were excluded from text mining queries. The focus of this study was to identify naturally occurring chemicals within plants, not those added or supplemented by humans. Excluded phenotype classes included sprains, strains, wounds, and injuries. Changes in these phenotypes are unrelated to diet and are excluded to reduce computational time and resources in text mining.

Preliminary queries start by mining for the co-occurrence of entity classes. The entire article is set as the initial boundary for the co-occurrence of entities. After reviewing cursory co-occurrence mining results, the next iteration of query development sets stricter boundaries, such as within the title or abstract, down to a single sentence of an article. Upon manual review of successive co-occurrence results, the sentence structure of entity mentions are abstracted to linguistic patterns to refine the query. Linguistic patterns differ given the set of mined entities classes being queried. Again, each query is driven by one of the four biological questions. For example, a linguistic pattern between a plant and chemical entity is structured as, *plant X contains/produces/synthesizes chemical Y*. Alternatively, another linguistic pattern of plants and chemicals is, *chemical Y isolated/extracted/derived from plant X*. With every iteration, precision is increased while being mindful of drastic decreases in recall. Semantic predicates mentioned in the linguistic patterns, such as *contains* and *isolate*, were not limited to previously defined verbs or phrases in order to preserve the semantic detail of relationships found in the literature. Instead, semantic predicates were discovered by way of a parts of speech class that identifies verbs and phrases between specified entities implemented in I2E.

I2E incorporates a series of linguistic constraints that were also used in query devel-

opment to increase precision in text mining. An option in I2E allows for the selection of boundaries for matching a query entity class. The three available settings are linguistic, exact, or strict, ordered by the level of specificity of the boundary. The linguistic boundary matches any part of an identified linguistic unit. A linguistic unit is a phrase that contains a queried entity class. For example, a query for broccoli, a verb, and a chemical could return the linguistic unit, "Broccoli sprouts contain health-promoting glucosinolates". The linguistic units for broccoli and chemical would be, "Broccoli sprouts" and "health-promoting glucosinolates". The linguistic setting would match any part of the linguistic unit for the queried entity class while the strict setting would not match broccoli in "Broccoli sprouts" because it is part of a larger linguistic unit. Another linguistic constraint imposed in text mining queries was a disambiguation filter determined by a Linguamatics calculated confidence metric. Disambiguation attempts to resolve the multiple meanings of words mined in the literature. Over the course of query refinement, it was determined that the threshold for the confidence metric should be set to 50.

Query results were manually reviewed until each query returned 80% accuracy for 1000 randomly selected text mined triples. Results were output once these requirements were satisfied. Queries, formatted as .i2q files, are available in **Appendix A**.

2.2.4 Text Mining Results

Text mining output included entity properties, relationship metadata, and relationship evidence for text mined results. Entity properties, for both entities extracted

in the relationship triple, consist of the entity's original source identifier, preferred term, and term identified within the literature. Relationship metadata includes the semantic predicate and location, title or abstract, where the relationship was identified. The evidence extracted is the exact sentence from the literature containing the relationship. Results were output in a tab delimited format for ease of organization and parsing.

Text mining results augmented curated connections between the five different domains spanning from plant species to human phenotypes. **Table 5** summarize the results returned from the four sets of queries between plants, chemicals, genes, biological pathways, and human health phenotypes. 3,249,155 relationships were returned across all query sets. Of these over three million relationships, there were 72,470 distinct semantic predicates and 54,293 individual entities. These relationships were extracted from 103,723 Agricola and 821,777 Medline citations. The addition of all extracted relationships could enrich the investigation but, realistically, confidence in text mining results is variable. It is important to consider the proper balance between the recall of text mining and it's error rate.

Table 5: Text mined results separated by query. Result statistics include total overall and unique results of relationships and relationship types.

Entity Pair Query	**Overall Mined**	**Unique Mined**	**Overall Predicates**	**Unique Predicates**
Plant, Chemical	622,559	281,277	14,015	4,010
Chemical, Gene	910,605	409,833	23,080	3,975
Gene, Pathway	599,622	385,357	26,550	3,582
Pathway, Phenotype	1,116,369	415,907	33,419	4,502
All Pairs	3,249,155	1,492,371	72,470	8,946

Manual curation of millions of relationships is an onerous task. To this end, dif-

ferent methods of filtering were performed to assist in curation. First, only results from articles written in English were included as viable results. Relationships extracted from articles in 14 other languages, all found within the Agricola citation database, were excluded from the results. The highly heterogeneous number of semantic predicates introduced noise, with many low and single occurrence instances, while providing little extra semantic detail. To retain semantic detail while reducing heterogeneity in relationships, predicates were collapsed into higher, subsuming relationships. Predicate phrases were tokenized and the longest single term was selected from these tokens to represent the semantic detail of the phrase. All single term predicates were then stemmed to remove morphological affixes using the Porter-Stemmer implementation from the NLTK. The combination of selecting the longest term in predicate phrases and stemming predicates significantly reduced the number of unique predicates from 72,470 to 8,946. After applying these filters, the resulting number of distinct, overall relationships from text mining was 1,492,371.

Entities, associations, and relationships from curated sources and text mining discussed in this Chapter are used to populate the diet-disease network described in the following sections.

2.3 Data Munging and Integration

Generally, data munging is the process of cleansing and transforming data into a usable form for computational analysis. Cleansing data is a crucial step in data munging and overall analysis that consumes a majority of the time of a data project due to many considerations. A researcher must contemplate what data to include,

map, and establish provenance during the cleansing component of data munging [50]. This crucial stage facilitates data integration and heavily influences all downstream analyses.

2.3.1 Cleansing Data from Diet-Disease Sources

Data cleansing begins with a detailed understanding of the data sources and types needed for a project. The data types included in the development of the Diet-Disease Network, described in **Section 2.1**, are entity identifiers, associations, and metadata. Entity identifiers (EIDs), previously described, provide distinct points of reference for mapping entity data, associations, and metadata. Once EIDs are parsed from a source, a researcher must consider what supporting data for entities, associations, and metadata to include in the data project.

Each data type has a set of features that can be used in this data project. The inclusion of features is determined by the goals of the project and the discretion of the researcher. EIDs feature synonym lists, entity definitions, and a preferred term. These features provide labels to entities for visualization and pattern recognition. They also act as a means of pattern matching that eases the burden of knowing a specific entity name while querying the data. Association features consist of the source and evidence of an association. Association sources fall into 3 categories, experimentally validated, curated from publications, or inferred by computational means. Evidence for associations is dependent on the source of the association. Evidence for experimentally validated associations may include instrument measurements. Associations curated from publications store publication identifiers of articles mentioning the associations.

Evidence supporting computationally inferred associations may be a qualitative metric, such as sequence similarity between genes of related species, or a text excerpt of the association from text mining.

The selected data features are then parsed to be mapped to EIDs. At this juncture, the quality of data is taken into account, particularly associations with metric-based evidence features. Acceptance thresholds, specific to the association evidence, are useful for cleaning up a dataset by excluding associations that can introduce noise and provide false positives in the analysis. For instance, text mined relationships in this project were excluded based on their number of occurrences within the collection of publications. A text mined relationship between two entities was accepted into the dataset if it occurred at least twice within the publication collection. This occurrence threshold prevents the introduction of excess noise from low confidence relationships.

Data provenance describes the lineage of how data was generated, processed, and modified. Provenance keeps a record of metadata for data, such as version and source information. Data provenance can occur at different levels, such as a database as a whole, files within a database, or single entries in files [183]. Establishing data provenance within data projects is important to ensure compatibility when mapping datasets to one another, especially when integrating data from multiple sources into a searchable data warehouse.

2.3.2 Data Integration

Data integration is the task of combining data from heterogeneous sources to provide a unified, systematic view of these data [95]. The aggregation and integration of

data from heterogeneous domain sources is a powerful way to overcome data sparsity. It accumulates stronger evidence through consolidating scarce data into a more complete dataset. It also helps to add semantic structure and detail that would otherwise not exist in a single data source. The model framework for data integration is composed of two major components, a set of data sources and a global schema that maps and reconciles the data. Irrespective of the integrative mapping approach, discrete identifiers are necessary to properly map the associations and relationships between entities.

2.3.2.1 Resolving Ambiguous Identifiers

As described in **Section 2.1.3**, some sources assign arbitrary identifiers to entities. The identifiers are unique within those sources but upon integration with multiple sources they may be ambiguous and result in possible entity identifier collisions. Listed in **Table 6**are the five sources in this study that utilize arbitrary EIDs. Discrete, internal identifiers were created to circumvent EID collisions within the diet-disease network. The designated format follows a similar pattern to those used in ontologies, such as the Gene Ontology. The pattern prefix includes the source name, with words separated by periods, followed by *.id:*. EIDs from the original source are appended to this prefix to generate discrete, internal identifiers. An example of a discrete EID from Entrez Gene is *nih.nlm.ncbi.gene.id:3586* for the IL10 gene in *Homo sapiens*. Mappings from sources that refer to arbitrary EIDs from the original sources are redirected to the unique internal EIDs for those entities. After resolving the ambiguity between source EIDs in an integrated resource, one must specify the

approach for mapping data. The two common approaches are known as the Global As View and the Local As View.

Table 6: Five data sources with arbitrary entity identifiers were integrated into the diet-disease network. Internal entity identifiers were assigned to these sources to ensure uniqueness.

Data Source	**Source EID**	**Internal EID**
NCBI Taxonomy	36774	nih.nlm.ncbi.taxonomy.id:36774
NCBI Entrez Gene	4780	nih.nlm.ncbi.gene.id:4780
National Agricultural Library Thesaurus	6949	usda.nal.thesaurus.id:6949
USDA Nutrient Database	341	usda.ndb.id:341
OMIM	614594	omim.disease.id:614594

2.3.2.2 Entity Association Mapping

The Global As View (GAV) attempts to map entities of the global schema to those found in the original data sources. The GAV approach models the integrated data such that structuring queries is simplified, needing only to create queries based on the global schema. This results in efficiency for query development and execution. The main disadvantage of the GAV is the requirement of explicitly specifying the mapping and merging of entities between multiple sources into a global schema [95]. This requirement does not allow the addition of a new data source independently of other sources. It stymies the incorporation of new data by necessitating a remapping of the global schema for new sources.

The Local As View (LAV) follows the opposite approach of the GAV, mapping entities from the local schemas of original data sources to the global schema. The LAV approach addresses the disadvantage of the GAV, allowing new data sources to be integrated independently of existing sources. However, the approach introduces complexity to query development, requiring more sophisticated queries to capture

data not explicitly defined in the global schema [95]. In this project, the LAV mapping approach was utilized for data integration to allow efficient addition of new data sources for. The development of more complex queries for the integrated data was a logical trade-off for the flexibility of incorporating new data independently of the global schema. This modularity enables scalability of the project in the future, given the exponential growth of biological data available.

This project incorporates associations that link entities across a variety of sources to enable integration. Many of the sources provide associations linking chemical, gene, pathway, and phenotype entities described by the NCBI, ChEBI, and Gene Ontology. Sparse data in single sources is supplemented in the integrated Diet-Disease Network by aggregating and mapping entity associations from the multitude of sources. The scarcity of data connecting plant species with their phytochemical profiles was augmenting with relationships extracted via ontology-based text mining, which used structured vocabularies from sources such as the NCBI and EBI.

2.3.3 Data Formatting

In order to use the aggregated and integrated data as a single, unified resource, it must to be formatted into a form suitable for data management systems. Database management systems (DBMSs) are responsible for defining, creating, creating, querying, and updating a structured collection of information. DBMSs are capable of accepting many data format types as input, the most popular being the delimiter separated value format.

The delimiter separated value format stores data as a two dimensional array of

columns and rows. This is achieved by separating column values in each row with specific delimiter characters. A delimiter can be any character, with the stipulation that the character does not appear in any value of the data. Commas, tabs, colons, spaces, and the vertical bar, also known as a pipe, are the most widely used delimiters due to their rarity in most data. This format is simple to interpret and parse, making it highly versatile and widely accepted as a data input format. The data of columns and rows is decided by the database systems that best meets the needs of a data project.

2.4 Data Storage

Traditionally, the standard method of storing and querying data in the biological sciences has been relational databases. Relational databases (RDBs) are capable of storing, organizing, and providing access to large amounts of association data. They also efficiently capture metadata for biological concepts. However, when faced with the challenge of storing and querying large, heterogeneous data from multiple domains, RDBs are met with computational limitations.

2.4.1 Canonical Database Limitations

The most prominent limitation of RDBs is the normalization of data into tables. Normalization organizes data into varying levels of a normal form, essential to reducing data redundancy and maintaining data integrity within a database. The issue with normalized RDBs arises when complex queries require a series of aggregation and retrieval inquiries (called by the JOIN method in the SQL query language) mapping primary keys and foreign keys from multiple, normalized tables. JOINs are com-

putationally intensive and when many-to-many relationships exist between tables, require the generation of associative tables which raises computational costs exponentially [43]. For example, a table of genes and a table of biological pathways share many-to-many relationships. A biological pathway can involve many genes and a gene can be involved in many biological pathways. De-normalizing data to decrease the number of joins offset computational costs from complex queries in RDBs. However, these techniques can only go so far and also disregard the strict design of RDBs.

Another limitation of RDBs stems from their strict modeling of data, again brought about due to normalization. Strict data models are great for reducing redundancy and preserving data integrity particularly when assimilating updated data. An issue arises when new data types are added to an RDB. The introduction of new data types requires a redesign of the database schema to accommodate the structure of new associations. Database schema redesign is an intricate process, accounting for previous considerations that improved query efficiency while following logical constraints for data normalization, to incorporate new data [17]. To remain current with the expanding breadth of data in the biological sciences, RDBs will inevitably require extensive schema redesigns. This task, in addition to the complications from complex JOIN queries, introduces a formidable hurdle to scalability for storing and querying across molecular biology concepts.

Within the context of identifying the molecular mechanisms behind diet and disease, RDBs are not the best option for creating a diet-disease network. The amount, heterogeneity, and interconnected nature of data aggregated to describe the effects of diet on disease test the limitations found in RDBs. Recently, graph databases (GDBs)

have been adopted in various domains of computational biology because of their efficiency in traversing highly inter-connected data. In the remainder of this section, we discuss the benefits and development of a graph-based diet-disease network.

2.4.2 Applicability of Graphs for Diet-Disease Network

Graphs are frequently used in various domains for fraud detection, social networks, and recommendation engines [6, 142]. More recently, graphs have been applied in computational biology to study protein-protein interactions, gene clustering, and metabolic networks [64, 113, 130, 191]. Modeling biological entities and interactions as a graph has gained popularity because of the highly connected nature of the data. Graph databases are capable of overcoming the limitations of complex query efficiency, schema rigidity, and scalability found in relational databases.

Native graph databases are efficient in traversal operations, especially with highly connected data. On the other hand, traversal operations in relational databases require computationally expensive and complicated join queries which can exponentially degrade query performance. Traversal operations are efficient in graph databases because the data is formatted as a graph where both entities and relationships between entities are stored. In contrast, relational databases only store data as tables and infer relationships through multiple join operations. Explicitly storing the relationships between entities circumvents expensive join operations which leads to significantly faster queries when comparing graph and relational databases. Graph databases will not fall prey to expensive many-to-many join operations slowdowns.

Graph databases are schema-less which allows for scalability and data intake and

integration. In relational databases, all data is required to conform to a strict, normalized schema structured as tables. The rigid schema of a relational database must be redesigned when introducing new data types. In schema-less graph databases, new data types can be incorporated quickly without significantly affecting existing data. Although schema-less, graph databases generally model data for efficient query performance. Large scale applications of graphs, such as massive and dynamic social networks, are capable of efficiently storing and querying across millions of entities and relationships. This demonstrates the capability of graphs to encompass the deluge of data in the biological domain.

The application of graphs for biological research was introduced as early as 1994 [65], as a means of storing and querying the growing data of the human genome. The ease of data integration, scalability, and complex query efficiency were key factors for proposing the implementation of a graph database for the human genome. The efficacy of traversal queries in graphs drove Wilkowski to introduce graph theory to literature-based discovery. He investigated the pathophysiology of depressive disorder by implementing "discovery chains" of proteins, pathways, and phenotype relationships as paths in a graph [191]. Graph traversals have also been used in pharmalogical research to infer indirect connections between drugs and phenotypes. A combination of graph and linguistic theory was proposed to mathematically drive the development of drug repurposing hypotheses [64]. The principles of graph theory remain constant amongst these studies, but the implementation of graphs is variable. Graphs can be implemented as data models within relational databases, but can also exist as native graph databases.

2.4.3 Graph Database Management Systems

Native graph databases store, manage, and analyze relationship data. They improve performance, scalability, and flexibility when handling highly interconnected data compared to relational databases. Open-source native graph database management systems include Neo4j, OrientDB, and Titan. Currently, Neo4j is one of the most popular graph databases and has outperformed OrientDB and Titan [19].

Neo4j is an open-source, ACID compliant transactional database with native graph storage, processing, and analytics [119]. It is managed by the Neo4j graph platform which includes well supported application programming interfaces and drivers, a declarative query language called Cypher, built-in graph algorithms, and an interactive user interface that provides basic visualization. Neo4j also supports plugins and extensions for added functionality, such as performing graph queries with the Gremlin query language and visualizing results in Cytoscape. An advantage of Neo4j being open-source is that it has the largest active graph database community, which contributes to comprehensive documentation and current development support.

Many biological databases implementing graphs for storage and querying have adopted Neo4j. The most prominent biological database using Neo4j is the Reactome Knowledgebase. Reactome provides curated, molecular details of biological processes including gene, protein, and pathway information [51]. Another database built on Neo4j is biochem4j, which aggregates chemical, biochemical, and other biology resources for expanding research in systems biology [173]. Hetionet utilized the Neo4j graph database management system to encode compounds, diseases, genes,

and pathways to develop an edge prediction method for drug repurposing [76]. The applicability of the Neo4j graph database management system is made obvious by these and other published projects.

2.4.4 Graph Structure and Modeling

Graphs help conceptualize data by abstracting diverse concepts as nodes and making connections between them using edges. This conceptualization is beneficial for visualization and pattern recognition as it parallels how the human mind organizes information. Formally, a graph G is defined as a pair *(V,E)* where V is a set of vertices (referred to as nodes in this text) and E is a set of edges, denoted as G = *(V,E)*. Nodes represent concepts within a domain, such as entities describing plant-based diets and human health phenotypes. Edges represent associations and relationships between these concepts. Edges are defined as $E = (i, j)|i, j \in V$ where each member of E is a single connection between the nodes i and j [37]. Different variations of graphs are based on the properties exhibited by their edges.

In directed graphs, an edge E = *(i,j)* has direction from i to j. Each edge in E is mapped to an ordered pair of nodes in V and are called directed edges. Directed graphs are most applicable for modeling biological pathways or sequential processes. Directed acyclic graphs (DAGs) are directed graphs not containing edges that connect a node to itself, called cyclic edges [37]. DAGs provide semantic structure with directed, acyclic parent-child relationships. In the biological contest, DAGs are implemented for organizing data in ontologies and taxonomies. A DAG is the natural structure for integrating data from structured vocabularies into a single graph re-

source.

In weighted graphs, each edge is assigned a weight, given by a weight function $w : E \Rightarrow R$ where R is the set of all real numbers. The weight *w(i,j)* of an edge $(i, j) \in E$ represents the confidence in or relevance of an association. Weighted edges aid the efficiency of graph traversal algorithms by prioritizing edges between nodes. Performing graph traversals in large graphs with millions of nodes and edges is computationally demanding but can be executed by graph algorithms that utilize weights to quickly disregard low confidence edges. The diet-disease network described in this Chapter is structured as a weighted directed acyclic graph and is implemented as a Neo4j graph database instance.

Traditionally, graphs have been treated as homogeneous networks where nodes and edges are considered to belong to a single type, such as a protein-protein interaction network. In the last decade, researches have realized that graphs effectively model real world networks that contain objects and relationships of multiple, different types, such as social or bibliographic networks. These heterogeneous graphs have various types, or labels, that provide semantic information about nodes and the relationships between them. The computational difference between homogeneous and heterogeneous networks will be discussed in detail in **Chapter 3**.

2.4.4.1 Graph model

Neo4j databases are modeled as labeled property graphs. As with any graph, a labeled property graph (LPG) consists of nodes and edges. Nodes represent entities, such as a plant or a chemical. Edges represent associations or relationships between

entities. LPGs are unique in that their nodes and edges have internal structure in the form of stored attributes. Each node and edge in an LPG stores an identifier, a set of key-value pairs called properties, and types that are called labels. Nodes and edges may have any number of properties that describe the entity or association. Nodes can also be assigned labels that identify the type of node and edges can be assigned relationship types which define associations [188]. The Diet-Disease Network follows the labeled property graph model, as evidenced by the graph schema shown in **Figure 5**.

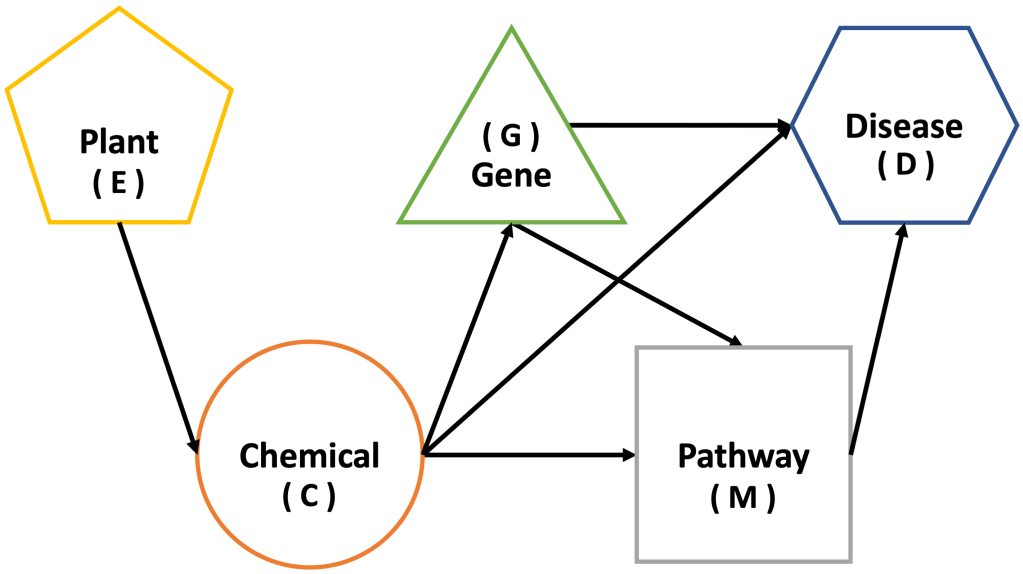


Figure 5: Diet-Disease Network schema which includes five node types and numerous edge types.

In this Diet-Disease Network, five entity labels are used to logically partition nodes for organization and query efficiency. The five entity types are: Plant, Chemical, Gene, Pathway, and Phenotype. In Neo4j, multiple node labels can be assigned, allowing for a source label for each of the 13 sources aggregated into the data warehouse. Retaining the source of each entity assists in data provenance and integrity by

way of validating entity and association numbers throughout the extract, transform, and load process of database population. There is a minimum requirement of data for each node. It includes the entity's source identifier and preferred term or name. When available, a synonym list and the definition of the entity is also included as node properties. Nodes are connected by edges in a property graph, which also have sets of properties and labels.

The most defining characteristic of edges in LPGs is the definition of edge types. Edge types provide semantic meaning to associations between entities. The advantage of assigning types to edges is the ability to uniquely identify instances of a relationship. Compared to other storage graph models, such as RDF triple stores, edges in LPGs can have multiple instances of the same relationship type. In RDF triple stores, connections of the same type between the same node pair cannot occur because it would represent the same triple with no added information. The ability to uniquely identify instances of the same edge type creates a more compact graph. It also allows an accurate count of edge types for use in graph traversal algorithms and ranking metrics. In the Diet-Disease Network, relationship types are derived from specified associations from ingested curated sources and predicates extracted from text mining, resulting in a highly heterogeneous set. Heterogeneity of relationship types is beneficial for adding semantic detail to associations, commonly ignored by most studies who utilize text mined data [139]. However, high heterogeneity means high disparity in instance counts for common and rare relationship types. Instance counts are an important factor in the weighting function for edge weights, influencing downstream graph querying and analyses. This disadvantage is resolved through the

weighting scheme and ranking metric described in **Chapter 3**.

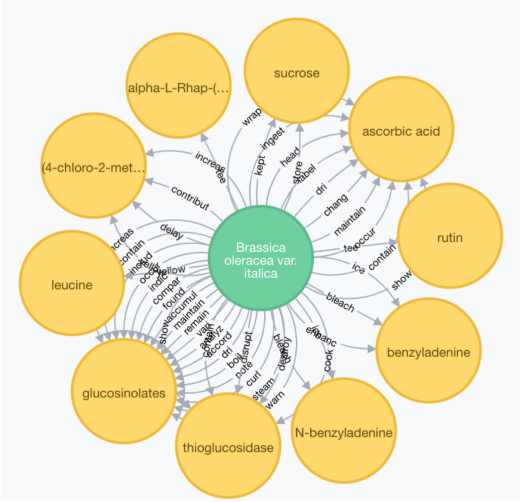
In addition to relationship type, each edge stores the source and evidence of the relationship as properties. Edges are sourced from curated structured vocabularies, association databases, and text mined from scientific publications. Text mined edges also store evidence for each instance of an edge as an array of tilde (˜) separated strings. Each string has six fields that explain instances of that edge type. Each evidence string includes an article identifier, the publication section the relationship was mined (either the title or abstract), the *(subject,predicate,object)* triplet, and the sentence of text the relationship was extracted from. Evidence, specifically the section of the publication the relationship was found, is used as part of the weighting scheme and ranking metric developed in **Chapter 3**.

The main features of the labeled property graph model are labels for nodes, relationship types for edges, and key-value properties that are assigned to both nodes and edges. An understanding of the Diet-Disease Network graph model is necessary to translate biological questions about data in the graph into queries that return plausible answers.

2.4.5 Graph Querying

The most beneficial feature of Neo4j is the graph-centric Cypher query language. Cypher is an expressive, declarative query language designed exclusively for Neo4j's native, labeled property graphs. It conceptually visualizes nodes and their respective edges as the textual queries themselves. **Figure 6** displays a Cypher query alongside its result, visualized by the Neo4j web interface. The *MATCH* clause can be seen

as a visualization of a specific relationship type connecting two nodes, denoted by closed parentheses. Cypher consists of clauses, keywords, and expressions that are similar to those in SQL, the universal query language for relational databases. Basic SQL statements are easily translated into Cypher statements because they share the same basic format and logic. Cypher has also become open-source as part of the openCypher project with the goal of becoming a modern graph query language supported by many graph databases [121]. As an open-source project, Cypher is under constant development with efficiency and support from a large community. It is familiar and intuitive, allowing a researcher to focus on developing more complex queries to uncover insights from the data.




```
MATCH (e:Plant)-[r]->(c:Chemical)
WHERE e.name = 'Brassica oleracea var. italica'
RETURN e,r,c
```

Figure 6: Example cypher query to determine the chemicals associated with the plant broccoli (*Brassica oleracea var. italica*). Results are truncated for brevity.

The efficiency of the Cypher language is a direct result of graph theory and algorithms. To that end, practical graph algorithms for searching a graph and computing

shortest paths between vertices have been directly implemented as clauses in Cypher. In addition to these canonical graph algorithms, contributors to openCypher have implemented in Java a series of user-defined procedures named Awesome Procedures On Cypher (APOC) [120]. The APOC package includes algorithms for computing PageRank, closeness and betweenness centrality, and community detection. The ability to call these procedures directly from Cypher strengthens the languages versatility and efficiency in graph querying. It also allows for "real-time" visualization of query results.

The Neo4j Server provides a web browser-based graphical user interface that allows access to the graph database. The Neo4j Browser supplies a Cypher console for interacting with the database, data profiling tools, query templates, and customizable query result visualization. Exploratory queries can be run in the Cypher console and the results visualized immediately. The visualization can be customized to better illustrate patterns and anomalies in the results. Data profiling tools determine quantities and characteristics of graph data, such as the number of nodes, relationships, and relationship types. Query templates and can be utilized in combination with result visualization for iterative refinement during query development. Programmatic access to Neo4j graphs is available through drivers in Python, Java, JavaScript, and C#. This project makes use of the Python driver for programmatic access to the Diet-Disease Network and the Neo4j Browser for result visualizations.

2.5 Integration and Augmentation of Data Sources to Span Diet-Disease Domains

To augment the available relationships describing the connection between diet and disease, the presented methods of data integration, storage, and querying were applied to the data types and sources discussed in **Chapters 1.1 and 2.1**. The application of these methods culminated in a unified data warehouse called the Diet-Disease Network that serves as a foundational resource to answer the questions: 1) What is the phytochemical profile of plant-based foods? 2) How do these phytochemicals affect human genes? and 3) How do these effects on human genes influence human health phenotypes, such as disease?

The aggregated data provides a vast, comprehensive dataset for determining the molecular mechanisms that drive the effects of diet on human health. **Table 7** exemplifies the sparsity of singular data sources when compared to the aggregated dataset compiled. After the inclusion of closure inferences and consolidation of non-unique relationships, the integrated dataset stores over 732,094 entities connected by over 460 million relationships and associations.

2.5.1 Data Source Integration

Data describing the components of diet and their relation to human phenotypes is sparse and scattered in various resources. In particular, comprehensive sources of connections between plants and their chemicals are few and exhibit little integration with human genomic and metabolic pathway data. Separately, no single source is capable of connecting the entities linking plant-based foods to human health phenotypes at the molecular level. Monolithic repositories, such as NCBI, contain and link

Table 7: Entity and association count statistics for all data sources included in the diet-disease network. Note that 42 of the 52 million associations from CTD are gene to disease associations.

Data Source	**Total Entities**	**Total Associations**
NCBI Taxonomy	170,651	203,002
Medical Subject Headings	252,626	406,477,330
Chemical Entities of Biological Interest	102,425	236,589
National Agricultural Thesaurus	58,113	150,251
USDA Nutrient Database	8,768	9,908
Gene Ontology	30,366	95,143
Disease Ontology	9,678	14,379
Human Phenotype Ontology	11,940	16,675
Mammalian Phenotype Ontology	14,309	11,864
Plant Ontology	1,730	2,893
Entrez Gene	59,599	324,029
Online Mendelian Inheritance in Man Database	20,914	26,312
Comparative Toxicogenomics Database		52,088,758
TOTALS	741,119	459,592,356

genomic, pathway, and phenotype data but lack significant connections to chemicals from plant-based foods.

The integration of diverse, reliable sources adds substantially more associations that increase connectivity amongst diet and disease entities. Structured vocabularies contribute semantic details while association databases connect the entities of different domains.

2.5.2 Augmentation of Data Sources To Traverse Diet-Disease Path

The aggregation and integration of 13 data sources creates a unified data warehouse for querying and discovery browsing. However, even with the integration of multiple sources, a dearth of associations connecting edible plants, their chemicals, and their effect on human gene expression exists. The diet-disease network composed of the 13 integrated data sources has only 3537 relationships between plants and chemicals.

To augment the current data, text mining was used to extract latent knowledge

from agricultural and biomedical literature. The introduction of text mined associations to the diet-disease network increased connectivity fifteen-fold between plants and chemicals alone to 55,651. The combination of integrated sources and text-mined relationships better represents the current domain knowledge while providing more relationships as evidence to discern important entities. Additionally, the integration and augmentation of data describing the components of diet and disease facilitates reachability across the entirety of the diet-disease network. For example, without augmenting the 13 data sources with text-mined relationships, a query between a plant, such as broccoli, and a biological pathway, such as inflammation, would not return a result. Text mining provides connections between domains that do not exist in the 13 data sources. This improved reachability allows researchers to ask questions such as, "What effects do plant X have on human health phenotypes?". The challenges of querying, filtering, and determining important paths from a large integrated dataset such as the diet-disease network will be discussed and resolved in great detail in the next chapter.

CHAPTER 3: A META PATH BASED RELEVANCE SEARCH AND RANKING METHOD

The surge of big data in genomics and metabolomics furnishes the evidence necessary to explain the molecular effects of plant-based foods on human phenotypes. These biological datasets are enormous, containing millions of highly connected objects and links of multiple types. Existing studies show a recent trend in modeling biological data as information networks, stored in graph data structures, to emphasize the relationships between molecular entities [27, 51, 75, 82, 173, 187, 202]. Information networks enable the discovery of knowledge by facilitating integration, scalability, and efficient mining of large amounts of interrelated data. A fundamental task of the discovery process is the evaluation of similarity or relevance between two entities within an information network. In the context of diet and disease, thousands of potential associations between phytochemicals of plant based foods and human health phenotypes must be identified and prioritized to develop data-driven hypotheses that explain the molecular mechanisms behind the effects of diet on human health.

Traditionally, the similarity between two objects in information network analysis has focused on their similarity, determined by node or edge based methods in homogeneous information networks [82, 98]. Node based methods, such as common neighbors and the Jaccard Coefficient, follow the notion that two nodes are similar if they share a large overlap of adjacent nodes. Edge based methods, such as PageRank and Sim-

Rank, compute similarity based on random walk algorithms between the nodes of interest [80,128]. These methods perform well on homogeneous information networks but ignore the inherent information stored in the heterogeneity of object and link types. The consideration of object and link types hinders the direct application of homogeneous network analysis techniques in heterogeneous information networks.

Heterogeneous Information Networks (HINs) encapsulate the semantic information provided by unique types of objects and links that exist in complex networks, such as drug-target interaction and gene-disease association networks [27, 75]. HINs are graphs consisting of nodes and edges of multiple types that allow for a scalable and detailed expression of semantics within a dataset. HIN analysis methods utilize the concepts of meta paths and meta structures to include the information from different object and link types [75, 76, 79, 97, 157, 168–170]. Meta paths are sequences of differing object types connected by various link types. Current techniques for biological network integration and gene prioritization utilize the concept of meta paths to rank the relevance of objects in HINs [75, 76, 97, 169]. Recently, Huang introduced the concept of meta structure, a directed acyclic graph of multiple types of objects and links, to represent the relationship of two objects in a graph [79]. However, many of these studies measure the similarity of objects of the same type, such as the similarity of a gene with another gene. Searching for the similarity of differently typed objects seems counterintuitive, yet many query applications require a measure to evaluate objects with different types. For instance, researchers who study drug targets would like to measure the relatedness of a chemical to a disease in order to rank which diseases would be affected by a drug. In this work, these HIN analysis methodolo-

gies are extended to rank the relevance between differently typed objects in a large, heterogeneous diet-disease network.

Heterogeneous networks integrate data from different domains to provide context and encompass the complexity of systems being investigated, reiterating the purpose of the diet-disease network described in **Chapter 2**. The integration of numerous diet and disease data sources provides aggregated support for low confidence relationships recorded in single resources. Aggregated support increases the signal to noise ratio for such entity relationships and allows them to be identified as novel, evidence-driven candidates for further investigation. Entity and relationship types in the diet-disease network are defined by predicates extracted from the literature and ontology associations. These types provide a semi-structured network schema able to accommodate new data sources and types of any size.

The link mining task of link based object ranking in heterogeneous networks requires a general and extensible method to accommodate the continued exponential growth of biological datasets from current and new molecular techniques. A ranking method should also account for the latent semantic information present within a heterogeneous network. This chapter describes a meta path based method to rank objects by their connectivity in a heterogeneous diet-disease network. These rankings help to elucidate the molecular mechanisms between plant-based foods and human health. The method utilizes a novel relevance measure described and evaluated in **Section 3.2**.

3.1 Data Mining in Heterogeneous Networks of Nutritional Systems Biology

In functional genomics and pharmacogenomics, data mining methods employing measures of similarity and interestingness have been used to evaluate similarity and prioritize relevant relationships between entities in a network, such as chemicals or genes to disease [22, 64, 80, 83, 99, 102, 128, 129, 131, 132, 141, 187]. Many of these network based methods were developed for homogeneous networks, which ignore the heterogeneity of entities and relationships in real world information networks, such as biological systems. Link mining is an area of research that investigates data mining techniques which explicitly consider heterogeneous, linked data [62]. Motivated by the relevance search problem in heterogeneous information networks, this work develops a new method for measuring the relevance of and ranking objects with different types in a heterogeneous network. The method also incorporates null invariant measures from association mining to account for novel, low probability events.

3.1.1 Heterogeneous Network Definitions and Concepts

Sun proposed a distinct formalization of Information Networks to distinguish the difference between homogeneous and heterogeneous information networks [170]. Sun introduced essential concepts of heterogeneous information analysis, such as network schema, meta paths, and a novel similarity measure called PathSim. A more encompassing notion of meta paths, called meta structures, was recently proposed by Huang to further capture underlying semantic information within heterogeneous networks [79]. In this section, these concepts are defined to lay the foundation for the design of the novel relevance measure presented in this work.

Formally, an **Information Network** is defined as a directed graph $G = (V,E)$ with an object type mapping function $\phi: V \rightarrow A$ and link type mapping function $\psi: E \rightarrow R$, where each object $v \in V$ belongs to one particular object type $\phi(v) \in A$, and each link $e \in E$ belongs to a particular relation type $\psi(e) \in R$. This is a unique definition of an information network as it distinctly identifies object and link types within the network [169]. Information networks are classified as heterogeneous information network when object types $|A| > 1$ or link types $|R| > 1$ and is considered a homogeneous information network for all other instances.

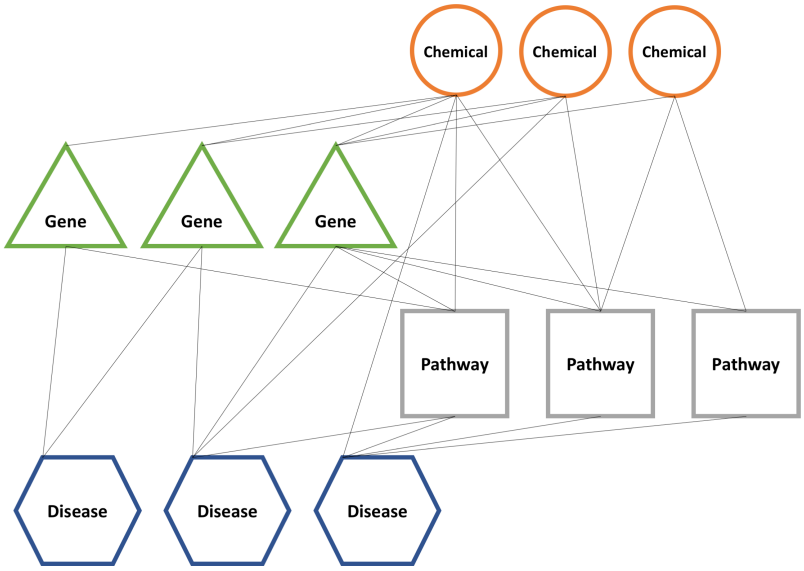
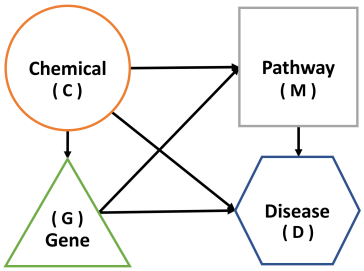


Figure 7: This representation of a heterogeneous information network contains four biological object types, denoted by different shapes and labels. Object types are linked to one another by arrows.

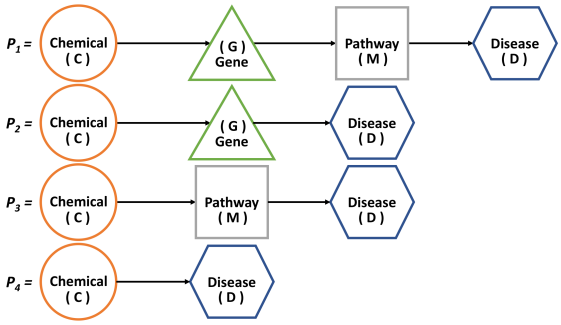
Figure 7 displays an example of a heterogeneous information network in biology that stores objects of the following types: Chemical compounds (C), Genes (G), Biological Pathways (M), and Diseases (D). Each disease $d \in D$ has links to a set of biological pathways, genes, and chemicals, while each chemical compound $c \in C$ has links to a set of plants. These relationships represent the various link types in the

biological information network. HINs can become complex, requiring an abstract, or meta, view of the network to better comprehend their semantic relationships. Heterogeneous networks follow a network schema which provides a meta-structure that aids in search, mining, and analysis of the network [168]. Cognizant of this, the concept of a Network Schema is defined.

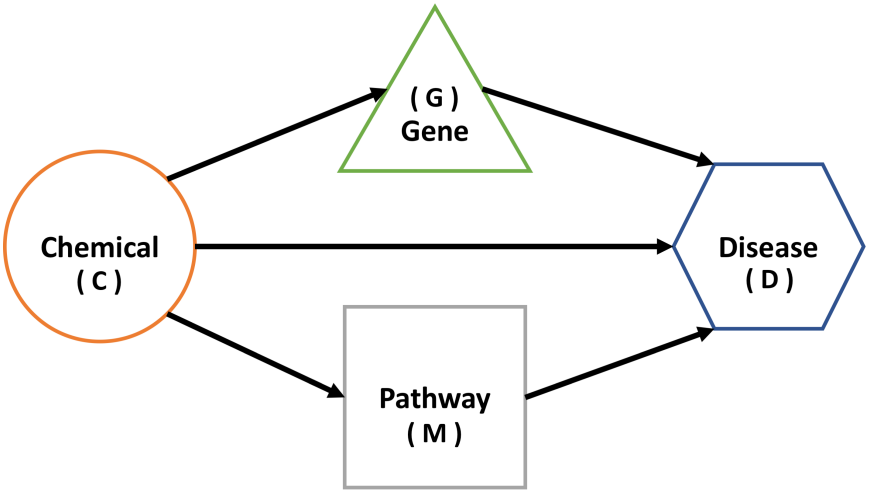
A **Network Schema** is a meta view for a heterogeneous network *G = (V,E)* with object type mapping function $\phi: V \rightarrow A$ and link type mapping function $\psi: E \rightarrow R$. It is a directed graph $T_G = (A, R)$, defined over object types A with link types R. A network schema acts as a blueprint by identifying all object types and link types connecting a HIN. **Figure 8a** provides a network schema for the biological information network example in **Figure 7**. Network schemas in HINs are conceptually similar to Entity Relationship models in current relational databases. The difference is that network schemas are more abstract in that they only model entity and relationship types, disregarding entity type attributes. This property allows a more general framework with the capability to model unstructured, non-normalized data, while also facilitating graph theoretic network analysis methods, such as the quantification of paths between objects.



(a) Network Schema of Biological Heterogeneous Information Network



(b) Meta Paths in Biological Heterogeneous Information Network



(c) Meta Structure in Biological Heterogeneous Information Network

Figure 8: Visualization of key concepts that describe a heterogeneous information network, such as those found in biology. **8a** is a representation of a heterogeneous information network contains four biological object types, denoted by different shapes and labels. Object types are linked to one another by arrows. **8b** displays four possible meta paths from **8a** between the Chemical and Disease object types are displayed, each with a different number of steps. **8c** is the meta structure from Chemical to Disease object types derived from the network schema in **8a**

.

As evidenced in network schemas, two objects can be linked by more than one path. In graph theory, a path is defined as a sequence of adjacent, distinct vertices connected by distinct edges. A meta path is an extrapolation of this concept. Formally, a meta path P is a path defined on the graph of an HIN network schema $T_G = (A, R)$, and is

expressed in the form $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between types A_1 to A_{l+1}, where $\circ$ is the composition operator on relations. For brevity, meta paths can be denoted as a sequence of adjacent object types $P = (A_1, A_2, \ldots, A_l)$. For example, given the network schema in **Figure 8a**, the meta path $P_1 = (C, G, M, D)$ in **Figure 8b** describes the relationship of a chemical compound (source) and a disease (target) that share associations with sets of genes and biological pathways. A path $p = (a_1, a_2, \ldots, a_{l+1})$ in network G is a path instance of a meta path $p \in P$ if $\forall i, \phi(a_i) = A_i$, and each link $e_i =< a_i a_{i+1} >$ belongs to each relation $R_i \in P$. Two meta paths $P_1 = (A_1, A_2, \ldots, A_l)$ and $P_2 = (A'_1, A'_2, \ldots, A'_k)$ are considered concatenable if and only if $A_l = A'_1$. The concatenated path is denoted as $P = (P_1, P_2)$, equivalent to $P = (A_1, A_2, \ldots, A_l, A'_2, \ldots, A'_k)$. A simple concatenated path example can be made between *(G,M)* and *(M,D)*, denoted as *(G,M,D)*, which describes the relationship between genes and diseases by way of common biological pathway associations.

A further abstraction of meta paths, called meta structures, was proposed by Huang for complex relationships between source and target objects that meta paths are unable to express [79]. For example, illustrated in **Figure 8c** is a complex relationship S between a chemical compound and a disease that cannot be expressed by a single meta path. One way to capture this relationship would be to calculate the linear combination of relevances of the meta paths derived from the meta structure. This method loses the information gained from a node with edges to different types within a meta structure. Formally, a meta structure S is a directed acyclic graph with a single

source node n_s with in-degree 0, and a single target node n_t with out-degree 0, defined on an HIN schema $T_G = (L, R)$. A meta structure is denoted as $S = (N, M, n_s, n_t)$, where N is a set of nodes and M is a set of edges. For any node $x \in N, x \in L$; for any link $(x, y) \in M, (x, y) \in R$. The concept of a layer for meta structure is essential for meta structure based relevance measures. Given a meta structure $S = (N, M, n_s, n_t)$, the nodes can be partitioned by their topological order in S. The nodes of the *i-th* layer can be denoted as $S[i] \subseteq N$ and $S[i\colon j](1 \leq i \leq j)$ as the nodes from the *i-th* layer to the *j-th* layer. The number of layers is denoted by d_S, meaning $S[1\colon d_S] = N$. To illustrate, the number of layers in the meta structure in **Figure 8c** is $d_S = 3$, where *S[i]* for $1 \leq i \leq 3$ are $\{C\}, \{G, M\}$, and $\{D\}$, respectively.

The concept of meta paths has been applied in numerous data and link mining tasks, such as ranking, top-k search, link prediction, and clustering [75, 91, 97, 156, 169–171]. Recent relevance search methods also use meta paths and meta structures to measure and rank heterogeneous objects [75, 79, 91, 156]. With the concepts of heterogeneous networks defined, similarity and relevance search methods can be reviewed for their influences in the development of a new meta path based relevance measure.

3.1.2 Similarity Search in Heterogeneous Networks

The most equivalent data mining task to relevance search is similarity search. Similarity search is extensively studied and its methods can be separated into node or edge based approaches. Node based approaches measure the overlap between objects to assess similarity between objects. These approaches include cosine similarity, the

Jaccard Coefficient, and the Rand Index [98]. Another class of node based methods are derived from association rule mining. Association rule mining methods, particularly null-invariant measures, are capable of searching for rare, unusual associations and relationships [195].

Edge based approaches make use of the connectivity of objects within a network to measure object similarity. The most well-known edge based similarity search method, Personalized PageRank, calculates the likelihood of reaching a target object from a starting object via a random walk with restart algorithm [128]. SimRank is another edge based approach that evaluates object similarity through the similarity of neighboring objects. The intuition behind SimRank is "two objects are similar if similar objects reference them" [80]. These node and edge based similarity search methods are grounded in the assumption that all nodes and edges are the same type, ignoring the semantics of paths with objects of different types.

In the last decade, the formalization of heterogeneous networks has encouraged the design of various different similarity search methods. Many of these methods are based on the concept of meta paths, previously described in **Section 3.1.1**. Sun defined heterogeneous information networks and introduced four meta path based similarity measures that laid the foundation for similarity and relevance measures [169]. PathCount, **Equation 1**, is the most straightforward measure and evaluates similarity by counting the number of path instances p between x and y following a meta path P. It should be noted that PathCount is unbounded, making it difficult to compare to bounded measures.

$$PathCount(x, y) = |\{p\colon p \in P\}| \tag{1}$$

Random walk, **Equation 2**, calculates the probability of a random walk that starts at x and ends at y following meta path P. The similarity measure is the sum of probabilities for all path instances $p \in P$ between x and y.

$$RandomWalk(x, y) = \sum_{p \in P} Prob(p) \tag{2}$$

Pairwise random walk, **Equation 3**, decomposes a meta path P into two equal length meta paths $P = (P_1, P_2)$. It measures similarity by the probability of two random walks that start from x and y which terminate at the same middle object.

$$PairwiseRandomWalk(x, y) = \sum_{(p_1 p_2) \in (P_1 P_2)} Prob(p_1) Prob(p_2^{-1}) \tag{3}$$

These count based measures are biased towards highly visible (degree centrality) or concentrated (betweenness centrality) objects. Path count and random walk reward paths of objects with more connections, while pairwise random walk favors paths with middle objects that act as bridges between start and target objects. The PathSim metric normalizes the PathCount measure to avoid the bias of high visibility. Given a symmetric meta path P, PathSim computes similarity between two objects of the same type x and y as shown in **Equation 4**.

$$PathSim(x, y) = \frac{2 * |\{p_{x \to y}\colon p_{x \to y} \in P\}|}{|\{p_{x \to x}\colon p_{x \to x} \in P\}| + |\{p_{x \to y}\colon p_{x \to y} \in P\}|} \tag{4}$$

In **Equation 4**, $p_{x \to y}$ is a path instance between x and y, $p_{x \to x}$ is an instance be-

tween x and itself, and $p_{y \to y}$ is an instance between y and itself. These heterogeneous network based measures have been shown to perform better than similarity measures that disregard the semantics in paths of multiple object and link types. However, they focus on symmetric meta paths where the start object and target object share the same type. For instance, given the biological heterogeneous network in **Figure 7**, a symmetric path could involve a path $P = (G, M, G)$ where the similarity of two genes is the focus, based on shared links to a biological pathway. Asymmetric paths are equally interesting and applicable in various domains. Within the biological heterogeneous network in **Figure 7**, a biologically interesting asymmetric path $P = (G, M, D)$ searches for the influence of genes on disease via their affects in biological pathways. Recently, more robust methods have been proposed to handle both symmetric and asymmetric paths in heterogeneous networks in search of relevance, as opposed to similarity.

3.1.3 Relevance Search in Heterogeneous Networks

Relevance and similarity search methods are built on the same concepts in heterogeneous network analysis. However, relevance search measures the relevance of objects with different types within a heterogeneous network, as opposed to the similarity of objects of the same type. In information retrieval, relevance denotes how pertinent a returned document, or set of documents, is to the information needs of a query [92]. In the context of heterogeneous networks, relevance search defines the relatedness of objects with different types based on their connectedness in the network.

Many of the methods developed for relevance search between differently typed ob-

jects relied on earlier similarity search methods, such as those proposed by Sun [169]. Degree Weighted Path Count (DWPC) is a measure based on the meta path based measure of PathCount. DWPC forgoes a normalized PathCount (NPC), **Equation 5**, in favor of distinct degree adjustments for each meta path instance [167].

$$NPC_m(s,t) = \frac{PC_m(s,t)}{\sum_{t_i \in T_m} PC_m(s,t_i) + \sum_{s_i \in S_m} PC_m(s_i,t)} \tag{5}$$

The denominator of the NPC favors paths of high degree nodes over more specific, lower degree paths. To combat this bias, a Path Degree Product (PDP) is calculated by downweighting each relation in a path instance by raising both the in-degrees and out-degrees of each node in a path to the -w power, where $w \geq 0$, and multiplying all degrees **Equation 6**. The DWPC, **Equation 7**, is the sum of all PDPs between the source s and target t objects.

$$PDP(path) = \prod_{d \in D_{path}} d^{-w} \tag{6}$$

$$DWPC_m(s,t) = \sum_{path \in Paths_m(s,t)} PDP(path) \tag{7}$$

As simple as the computation is, DWPC is a supervised method that requires learning for optimization for the damping exponent w parameter. Another supervised relevance measure for ranking objects of different types was the Path Constrained Random Walk (PCRW) method, introduced by Lao and Cohen. The PCRW method searches labeled, directed graph networks following a random walk algorithm that assigns weights for each edge label as proximities between objects. The random walk algorithm is constrained by a defined sequence of objects, referred to as "path

experts" [91]. The labeled, directed graph can be thought of as a network schema and "path experts" sequences as meta paths defined on the schema. The PCRW is a supervised learning model which calculates the probability that a random walk, restricted on a path P that starts from an object o_s will arrive at an ending object o_t. Although PCRW can be used as a relevance method, Shi stressed a possible weakness due to it being an asymmetric measure.

Shi argued that a symmetric measure allows for the comparison of relatedness for pairs of heterogeneous objects. To that end, Shi proposed a symmetric relevance measure called HeteSim. HeteSim follows the basis of the SimRank measure in homogeneous networks that objects are more likely to be related if they are referenced by other similar objects [156]. For example, a gene is more relevant to a disease if the gene is associated with the biological pathways that affect the disease. Given a relevance path $P = R_1 \circ R_2 \circ \ldots \circ R_l$, HeteSim between two objects s, $s \in R_1.S$ and t, $t \in R_l.T$ is calculated by **Equation 8**.

$$HeteSim(s,t|R_1 \circ R_2 \circ \ldots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j(t|Rl)|R_2 \circ \ldots \circ R_{l-1}) \tag{8}$$

In **Equation 8** $O(s|R_1)$ is the out-neighbors of s given relation R_1 and $I(t|R_l)$ is the in-neighbors of t given relation R_l. HeteSim simplifies the complexity of SimRank by constraining a pairwise random walk to a specific relevance path, while being able to determine relevance between heterogeneous objects of arbitrary path length. The properties and efficiency of HeteSim make it useful in a variety of applications such as

ranking, clustering, and filtering. All relevance measures described are based on the notion of meta paths and efficiently quantify the relatedness between heterogeneous objects of different types. However, meta paths are not always capable of capturing more complex relationships.

Most recently, a meta structure based relevance measure was proposed by Huang. A meta structure, formally described in **Section 3.1.1**, is a directed acyclic graph of differently typed objects connected by differently typed edges describing the relationships between them. Huang developed two meta structure based relevance measures that are combined into a unified measure called the Biased Structure Constrained Subgraph Expansion (BSCSE) [79]. Given a heterogeneous information network G = *(V,E)*, a meta structure S, a source object $o_s \in V$, and a target object $o_t \in V$, the relevance of an *i-th* layer subgraph $g \subseteq G$ is defined in **Equation 9**.

$$BSCSE(g, i|S, o_t) = \frac{\sum_{g' \in \sigma(g,i|S,G)} BSCSE(g', i+1|S, o_t)}{|\sigma(g, i|S, G)|^{\alpha}} \tag{9}$$

In **Equation 9**, $\sigma(g, i|S, G)$ represents the *(i+1)-th* layer's instances expanded from $g \in S[1\colon i]$ on G and α is a bias factor between [0,1]. A smaller α will bias the measure towards higher visibility subgraphs, while a larger α favors the probability of random expansions reaching the target object. The BSCSE measure performs subgraph matching over a heterogenous network which can become computationally intensive given the size of a typical heterogeneous network. In large datasets, this method provides minimal gain in the task of ranking at the high expense of computational resources and time.

3.1.4 Integrated Heterogeneous Information Network for Link Mining Analysis

This chapter describes a semi supervised, semantically rich, scalable method for relevance ranking analysis in a large heterogeneous information network. This semi supervised method negates the need to identify true positive data sets for training purposes, as is required by other learn to rank, machine learning methods. The definition and inclusion of both object and link types within the heterogeneous information network captures semantic information that is lost in single typed, homogeneous networks. Lastly, the method is applied to a data structure capable of easily storing, integrating, and querying large datasets, regardless of data type.

The data for relevance analysis is stored in a graph framework which follows the description of a heterogeneous information network outlined in **Section 3.1.1**. It utilizes a labeled property graph model, stored, queried, and modified with the Neo4j graph database. Nodes, representing biological entities, are connected by edges, the relationships extracted from the data sources and literature. Formally, the graph *G = (V,E)* maps the set of nodes to four different entity types, $v : V \rightarrow \{A_1, A_2, A_3, A_4\}$, and the set of edges to six different relationship types, $e : E \rightarrow \{R_1, R_2, R_3, R4, R_5, R_6\}$. **Figure 8a** visually defines the schema of the heterogeneous network. The network schema abstracts the connectivity in the network while capturing latent semantic information held within its topology. This analysis also accounts for specific edge labels provided by curated and text mined sources. Curated sources link entities with a set of types defined by curators, such as *ïs_a¨* or *p̈art_of¨* from the Gene Ontology. Link types from text mining are the predicates extracted with a pair

of entities from the published literature. These link types can be highly variable but add semantic detail to text mined relationships. The definition of meta paths on the schema simplify the heterogeneity of link types by aggregating them as an abstract type, based on the entity types they connect. **Figure 8b** illustrates four meta paths derived from the network schema that provide support for quantifying and ranking the relevance between a chemical and disease. Meta paths embody the semantics and structure of a heterogeneous network.

In the development of any analysis method, it is necessary to possess an intimate understanding of the data to be analyzed. This relevance analysis employs a combination of databases and structured vocabularies that describe and connect chemicals, genes, biological pathways, and phenotypes. The National Center for Biotechnology Information (NCBI) hosts numerous linked databases and structured vocabularies encompassing a majority of these biological entities. Publicly available NCBI databases such as Entrez Gene and OMIM contain cross references to entities in structured vocabularies such as the Medical Subject Headings, Gene Ontology, and Disease Ontology. The Comparative Toxicogenomics Database (CTD) contains curated links for entities stored in the data sources of the NCBI. Incorporating NCBI databases and structured vocabularies with association information from the CTD creates a highly connected, heterogeneous dataset perfectly suited to test a novel relevance ranking measure. The data for genes, biological pathways, and phenotypes were restricted to human only for the purposes of testing and evaluation. The data repositories and their statistics are described in **Chapter 2** while Python scripts for data munging and database loading are available in **Appendix A**.

3.2 Design and Development of A Meta Path Based Relevance Measure

Several design objectives and considerations must be observed in the development of a meta path based relevance measure. Most importantly, the relevance measure must be applicable to heterogeneous information networks, which are ubiquitous data representations across many research domains, such as biology. Second, designing a semi supervised method to measure relatedness reduces the barrier to acceptance and use. A final design objective for a relevance measure is the ease by which it can be interpreted and compared to other measures. Although beneficial, these design objectives introduce issues that must be accounted for in the development of a meta path based relevance measure.

3.2.1 Objectives and Considerations for a Meta Path Based Relevance Ranking Method

Biological systems have been modeled as networks for numerous link mining tasks, such as similarity search and link prediction. Many methods have been developed to measure the similarity between connected entities within these biological networks [80, 131, 132]. However, early network analysis methods conflated the naturally heterogeneous node and edge types found in complex networks into single, homogeneous types. Inherently, biological networks are composed of different entity types connected by relationships that convey subtle semantic meaning as paths.

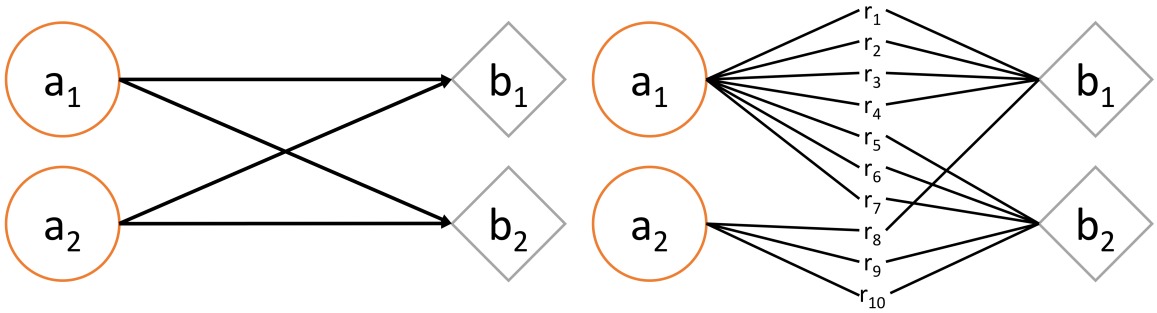
Meta path based methods were developed to quantify and evaluate the similarity of objects while considering the topological structure of the heterogeneous network. These methods impose a path constraint on network queries which can reduce the

search space, and therefore computational time in calculating similarity [157,168,169]. Many meta path based similarity methods were designed to find similar objects of the same type. For example, given a bibliographic network one can compare the similarity of an author to another author, based on which conferences they published papers in. Meta path based relevance search methods extend the concept of similarity to measure relatedness between objects of the same or different types. For example, within that same bibliographic network, one may be interested in the relatedness of an author and a research domain, based on terms included in their published papers. Within the context of a biological network, a relevance search between a gene and a disease provides insight into genetic factors or molecular drug targets for a disease.

An issue introduced by meta path based methods is the oversimplification of link types. Curated sources provide link types which add specific semantic detail to direct relationships between entities. The most common link type in ontologies is the parent-child relationship (is_a). Text mining can also produce link types which detail the relationship between extracted entities. Link types extracted from text mining are generally predicates which describe or modify the concepts within text, such as the link type (*produces*) between plant and chemical type objects. Current methods aggregate these heterogeneous link types into abstract types defined by the pairs of entity types they connect.

Including this semantic detail reduces the bias from high visibility paths. **Figure 9** illustrates this concept between two entity types A and B connected by various link types r. The number of links between A_1 to B_2 is 3 (r_5, r_6, r_7) and A_2 to B_2 is 2 (r_9, r_{10}). Although A_1 has a higher path count to B_2 than A_2, the majority of links

from A_2 (2/3) connect to B_2 while a smaller ratio of links from A_1 (3/7) connect to B_2. From this illustration it can be concluded that A_2 and B_2 have a stronger association than A_1 and B_2.



(a) Conflated $A \xrightarrow{R} B$ Relation (b) Expanded $A \xrightarrow{R} B$ where $r \in R$

Figure 9: Semantic value exists in the inclusion of weights derived from the detailed connections available from controlled vocabularies and the literature.

The second design objective considers the input requirements of supervised and semi supervised methods. Supervised methods are used to predict outcomes based on validated, true positive training data. This input requirement places the onerous task of identifying large sets of training data on end users. In many research fields, a sparsity of large, publicly available, true positive datasets for make it difficult to adequately train supervised machine learning models. In contrast, semi supervised methods require little to no human input which reduces the barrier to implementation by removing the onus of locating true positive training data. Less or no human input also improves the reproducibility of results. Semi supervised, meta path based relevance measures only require the definition of meta paths.

The design of any measure should consider the interpretability of the resulting values. Unbounded measures, such as PathCount, produce values that are difficult to

interpret without context [169]. Simple methods of normalization, such as finding the arithmetic mean of a value, scale unbounded values into an easily interpretable range, commonly between 0-1.0. Consequently, many similarity and relevance measures are designed to produce values that fall within these bounds, which creates a common interface for comparing the efficacy of measures. Therefore, a meta path based relevance measure for ranking should be bounded to aid in result comprehension and evaluation against existing measures.

There are two major considerations in designing a meta path based relevance measure for heterogeneous object ranking. The first is the selection of a search method for quantifying the relevance between objects in a heterogeneous network. The second consideration is the determination of technique for defining meta paths in a heterogeneous network. These selected methods should adhere to the design objectives previously outlined.

The relevance search method that best satisfies the criteria outlined is the HeteSim measure [156]. Firstly, the HeteSim measure provides a uniform framework to calculate the relatedness of same or differently typed objects for arbitrary paths in a heterogeneous network. This flexibility permits the use of the measure in numerous data mining tasks, such as object profiling and clustering. HeteSim scores are normalized via a cosine function that scales them into the easily interpretable and comparable range of 0-1.0. In addition, the HeteSim measure also exhibits the property of symmetry, which allows for comparisons of the relatedness between heterogeneous object pairs to show their relative importance. Lastly, the HeteSim measure is semi supervised, with input limited to interesting meta paths.

Meta paths used in relevance measures can be defined by a user based on domain knowledge, experience, and research question. They can also be automatically selected by supervised learning methods that calculate the importance of meta paths [90]. The HeteSim measures selects paths based on the knowledge and requirements of users. This enables users to tailor relevance searches to specific, pertinent research inquiries. For example, given a biological network of plants, chemicals, genes, biological pathways, and diseases, a nutrition researcher may be interested in measuring the relevance of certain plant based foods to human diseases and define a meta path between plants, chemicals, and human diseases. Based on these design objectives and considerations, the next section describes the development of a novel, meta path based relevance measure for ranking differently typed objects within a heterogeneous network.

3.2.2 A Meta Path Based Relevance Metric

The heterogeneity of detailed link types within a meta path provides latent semantic value. Considering the design objectives and considerations described, I define a novel metric called the Kulczynski Product Edge Weight (KPEW) that quantifies the semantic value of variable link types between node types in a meta path. Given two differently typed objects a and b, where $a \in A, b \in B$, and the link type between them R specified by a meta path $P = R_1 \circ R_2 \circ \ldots \circ R_l$, the Kulczynski Product Edge Weight is defined by

$$KPEW(a, b, R) = 1 - \sqrt{\frac{\sum_{r \in (R|P)} KRP(a, r, b)^2}{n(R)}} \tag{10}$$

The variable r represents detailed, heterogeneous types that belong to the broader link type R that connects objects of type A and B, from the given meta path P. Including the distinct r types in KPEW incorporates the subtle semantic value of predicates extracted from ontologies and text mining.

KPEW implements a sum of squares based mean of Kulczynski Relationship Products (KRP) between objects a of type A and b of type B to aggregate and normalize object-link-object probabilities as a single, quantified weight connecting a and b. A sum of squares average approach also upweights the low probability events calculated by the KRP and downweights high probability events.

The Kulczynski Relationship Product (KRP) is the product of the probabilities that object a of type A and b of type B are associated through link r of link type R. KRP is unique to each step, $R_1 \circ R_2 \circ \ldots \circ R_l$ in a meta path P. This allows each r between different object type pairs to provide a unique semantic value. For example, the predicate *affects* could exist between the object types of chemical-gene and gene-pathway. The probability that *affects* links a chemical to a gene is different than a gene to a pathway based on the distribution of associations of the two steps. KRP utilizes the Kulczynski 2 Index to determine these conditional probabilities, as shown in **Equation 11**.

$$KRP(a, r_i, b) = Kulc2(a, r_i) * Kulc2(b, r_i) \tag{11}$$

$$Kulc2(v_i, r_j) = \frac{(P(v_i|r_j) + P(r_j|v_i))}{2} \tag{12}$$

In **Equation 11**, the average conditional probability between object and detailed

link types *a,r* and *b,r* is calculated with the Kulczynski 2 Index (Kulc). Kulc is a null-invariant measure used in assessing the likelihood of small probability events. Small probability events are discretized events that occur a small number of times with respect to the total number of interactions, such as the existence of an association between an object, a or b, and a detailed link type, r. **Equation 12** defines the Kulczynski 2 Index between an object v and a detailed link type r.

Based on the concept that relevant objects are referenced by other relevant objects, KPEW is directly applied to HeteSim to produce a novel, meta path based relevance metric I have named HetERel (Heterogeneous Edge-adjusted Relevance), to parallel HeteSim.

$$HetERel(s,t|R_1 \circ R_2 \circ \ldots \circ R_l) = \frac{\sum_{i=1}^{|O(KPEW(s,R_1,k)|R_1)|} \sum_{j=1}^{|I(KPEW(t,R_l,m)|R_l)|} HetERel(O_i(KPEW(s,R_1,k)|R_1), I_j(KPEW(t,R_l,m)|R_l)|R_2 \circ \ldots \circ R_{l-1})}{|O(KPEW(s,R_1,k)|R_1)||I(KPEW(t,R_l,m)|R_l)|} \quad (13)$$

Equation 13 defines HetERel, where $O(KPEW(s,R_1,k)|R_1)$ are the KPEW weighted out-neighbors of s given relation R_1, and $I(KPEW(t,R_l,m)|R_l)$ are the KPEW weighted in-neighbors of t given relation R_l. For instances where s has no out-neighbors ($O(KPEW(s,R_1,k)|R_1) = 0$) or t has no in-neighbors ($I(KPEW(t,R_l,m)|R_l) = 0$) following the given path, relevance scores are set to 0, as a random walk from start to target will never complete. In the case that s and t share the same object type and are equivalent (i.e. $s = t$), it should be assumed that the object has a self-relation to itself. Logic posits that an object would be most relevant to itself and warrant a relevance measure score of 1. However, this logic does not hold in the

equation of HetERel, requiring a normalization step.

In order to describe the normalization step, I first discuss the calculation of HetERel for any two objects that follow a given meta path of a heterogeneous information network. HetERel is computed in three phases: 1) KPEW calculation, 2) matrix multiplication, and 3) relevance computation. **Figure 10** provides an example of the HetERel calculation, based on the definition of a simple relation $A \xrightarrow{R} B$.

The first phase of HetERel, shown in **Figure 10a**, describes the relationship R between A and B as a set of detailed link types $r, r \in R$ known as predicates. The semantic value of each predicate r within link type R is quantified by the Kulczynski Relationship Product. As denoted in **Equation 11**, the KRP of each object pair and predicate *(a,r,b)* is calculated as the product of two independent Kulcznski 2 Index (**Equation 12**) calculations between *(a,r)* and *(b,r)*. The set of KRPs are input into **Equation 10** to quantify the semantic value of relation R between A and B. In **Equation 13** KPEW is calculated for each instance of relationship R between all objects in types A and B.

The second phase of computing HetERel represents object relationships as matrices. The quantified relationship R, between A and B can be expressed as a weighted adjacency matrix, $(W_{AB})_{n*m}$ where the KPEW value for each object pair *(A,B)* is used to represent adjacency. W_{AB} normalized along the row vector generates U_{AB}, the transition probability matrix (TPM) of $A \xrightarrow{R} B$. Formally, U_{AB} is determined by **Equation 14**. Normalizing W_{AB} along the column vector results in V_{AB}, which represents the TPM of $B \xrightarrow{R^{-1}} A$. Previously, it was proven by Shi that a TPM has the property $U_{AB} = V'_{BA}$ and $V_{AB} = U'_{BA}$ where V'_{BA} is the transpose of V_{BA} [156].

$$U_{AB}(i,j) = \frac{W_{AB}(i,j) * KPEW(i,j,R)}{\sum_{k=1}^{m} W_{AB}(i,k) * KPEW(k,j,R)} \tag{14}$$

The simple relationship $A \xrightarrow{R} B$ can be extended into a meta path $P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$ of length l, where R is a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$. The relationship from A_1 to A_{l+1} from path P can be expressed as a reachable probability matrix RPM_P defined in **Equation 15**, where $(RPM(i,j))$ is the probability that object $i \in A_1$ reaches object $j \in A_{l+1}$ under path P.

$$RPM_P = U_{A_1A_2}U_{A_2A_3}\ldots U_{A_lA_{l+1}} \tag{15}$$

Applying the TPM property proven by Shi to **Equation 15** produces the non-normalized equation of HetERel. It is the inner product of two probability distributions that A_1 reaches a middle object type along path P and A_{l+1} reaches the same middle object type against path P. The non-normalized equation for HetERel is derived by

$$HetERel(A_1, A_{l+1}|P) = HetERel(A_1, A_{l+1}|P_LP_R) = RPM_{P_L}RPM'_{P_R^{-1}} \tag{16}$$

where $RPM'_{P_R^{-1}}$ is the transpose of RPM_{P_R}. **Equation 16** separates a meta path into two equal parts. Paths of even length can be split into two paths, $P = P_LP_R$, where $P_L = A_1 \ldots A_{mid}$ and $P_R = A_{mid} \ldots A_{l+1}$, and $mid = \frac{l}{2} + 1$. For example, a meta path *P=(GMD)* has an even length of two. Objects of type G and D meet at the middle type object M. The path can easily split into two equal paths, $P_L = GM$

and $P_R = MD$.

Odd length meta paths have start and target object types that never meet at a middle object type. For instance, a path *P=(CGMD)* with start object type C and target object type D would meet at the intersection of object types GM, as opposed to either G or M. To account for this, Shi proposed a decomposition step for odd length paths that adds an intermediate middle object type E between the atomic relation of the two middle object types such that start and target objects would meet at E (e.g., *P=(CGMD)* becomes *P=(CGEMD)*). This method increases the computational complexity of the calculation which prevents application to large, highly connected datasets. HetERel handles odd length paths with a computationally efficient approach that creates two separate split paths and calculates their mean HetERel score. The split paths are as follows: 1) $P_L = A_1 \dots A_{\frac{l+1}{2}-1} A_{\frac{l+1}{2}}$ and $P_R = A_{\frac{l+1}{2}} A_{\frac{l+1}{2}+1} \dots A_{l+1}$, 2) $PL = A_1 \dots A_{\frac{l+1}{2}} A_{\frac{l+1}{2}+1}$ and $P_R = A_{\frac{l+1}{2}+1} \dots A_{l+1}$.

The final phase is the computation and normalization of the relevance score for two instances $a \in A_1$ and $b \in A_{l+1}$ given a meta path P. The vectors of a and b from their respective RPMs are multiplied and normalized over their cosine. Formally, the normalized HetERel score between two instances $a \in A_1$ and $b \in A_{l+1}$ given a meta path P is

$$HetERel(a, b|P) = \frac{RPM_{P_L}(a,:) * RPM'_{P_R^{-1}}(b:)}{\|RPM_{P_L}(a,:)\|_2 * \|RPM'_{P_R^{-1}}(b:)\|_2} \tag{17}$$

where $RPM_{P_L}(a,:)$ is the row vector of a across RPM_{P_L} and $RPM'_{P_{R^{-1}}}(b,:)$ is the row vector of b across $RPM'_{P_{R^{-1}}}$. HetERel measures the cosine of the weighted

probability distributions that a and b meet at a middle object type, with a following path P and b going against path P. In the case where the path length is odd, the scores for each separate pair of split paths are calculated and their mean is taken as the final set of scores.

Given the heterogenous information network in **Figure 10c**, we illustrate the phases for calculating HetERel scores between A and C type objects following the path P=ABC. First, the meta path P=ABC is split in two, where $P_L = AB$ and $P_R = BC$. **Figure 10b** illustrates the KPEW phase for a_1 to all instances of object type B, which is performed for all pairs of objects in P_L, P_R. The adjacency matrices of W_{AB} and W_{CB}, where $W_{CB} = (W_{BC})^T$ are weighted with their respective KPEW values, and their normalization along their row vectors into transition probability matrices U_{AB} and U_{CB} are calculated as follows:

Weighted Adjacency Matrices

$$W_{AB} = \begin{bmatrix} 0.85 & 0 & 0 & 0.67 \\ 0 & 0.65 & 0 & 0 \\ 0 & 0.55 & 0.77 & 0.74 \end{bmatrix}, W_{CB} = \begin{bmatrix} 0 & 0.68 & 0 & 0 \\ 0.55 & 0.74 & 0.80 & 0 \end{bmatrix}$$

Transition Probability Matrices

$$U_{AB} = \begin{bmatrix} 0.5592 & 0 & 0 & 0.4408 \\ 0 & 1 & 0 & 0 \\ 0 & 0.2670 & 0.3738 & 0.3592 \end{bmatrix}, U_{CB} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0.2632 & 0.3541 & 0.3828 & 0 \end{bmatrix}$$

In this instance, because the two separate paths $P_L = AB$ and $P_R = BC$ are both a single step, their reachable probability matrices RPM_R^{-1} are equal to their transition

probability matrices (i.e.- $RPM_L = U_{AB}$ and $RPM_R^{-1} = U_{CB}$. The final phase is calculating the normalized HetERel from the reachable probability matrices using

$$HetERel(a, c|ABC) = \frac{U_{AB}(a,:) * U_{CB}(c,:)}{\|U_{AB}(a,:)\|_2 * \|U_{CB}(c,:)\|_2}$$

where the start and target pair *(a,c)* can be substituted for each instance of object types A and C from the heterogeneous information network in **Figure 10c**. The global set of HetERel scores for this example are shown in the following matrix

$$HetERel(A, C|ABC) = \begin{bmatrix} & c_1 & c_2 \\ a_1 & 0 & 0.3539 \\ a_2 & 1 & 0.6062 \\ a_3 & 0.4579 & 0.6976 \end{bmatrix}$$

In contrast, a similar implementation where the adjacency matrices (W_{AB} and W_{CB}) have either a 1 when objects are adjacent or 0 when they are not, produces the following matrix of scores

$$HeteSim(A, C|ABC) = \begin{bmatrix} & c_1 & c_2 \\ a_1 & 0 & 0.4082 \\ a_2 & 1 & 0.5774 \\ a_3 & 0.5774 & 0.6667 \end{bmatrix}$$

HetERel is able to incorporate semantic information found in the predicates between objects and produce fine-tuned weights that resolve similar rankings, such as between (a_2, c_2) and (a_3, c_1).

The normalized version of HetERel is bounded, $HetERel(a, b|P) \in [0, 1]$, which allows ease of interpretation and comparison to other, bounded relevance metrics. The normalization step of HetERel also ensures a self-maximum, where $HetERel(a, b|P) = 1$ if, and only if, $RPM_{P_L}(a, :) = RPM_{P_{R^{-1}}}$. HetERel inherits a general symmetric property for both symmetric and asymmetric paths as it utilizes a path decomposition step for odd-length paths, akin to HeteSim [156]. The decomposition step provides modularity for HetERel scores to be calculated for any arbitrary path length, both even and odd.

3.2.3 Implementation of a Meta Path Based Relevance Analysis

As with any data analysis method, the input data underwent a series of preprocessing procedures prior to analysis by HetERel. Data sets were fetched from sources, parsed, and transformed into typed objects and links (nodes and edges, respectively) for input into the high-performance graph database platform, Neo4J, as previously described in **Chapter 2**. In parsing and loading the data, exclusions were made to restrict the scope of the analysis to knowledge pertinent to plants and human health. Neo4J provides interactive access to the data through a combination of the Cypher query language and numerous programming language drivers. The heterogeneous information network data preprocessing code is available in **Appendix A**. The implementation of HetERel follows the phases described in the previous section with the addition of an initial graph query phase at the beginning. The Neo4J graph instance is queried via Cypher for pairs of typed objects and their associated detailed links. Next, the KPEWs for each typed object pair (e.g., Plant to Chemical and Chemical

to Gene) are calculated for use in typed object pair adjacency matrices. The third phase generates weighted adjacency matrices W_{AB} between object type pairs, as determined by the given meta path. Weighted adjacency matrices are then normalized into transition probability matrices U_{AB} and undergo chain matrix multiplication to determine the two reachable probability matrices $RPM_L, RPM_{R^{-1}}$. Finally, HetERel is computed via matrix multiplication of RPM_L and RPM_R, and normalized over its cosine. **Algorithm 1** is an example of a Python implementation of KPEW and **Algorithm 2** is an example pseudo-code implementation of HetERel. Pairs of start and target objects are sorted by HetERel scores and written file for searching of specific instances.

Simple optimization methods were applied to the naive implementation of HetERel to attain a remarkable difference in runtime. The most significant speedup was achieved by adopting a reuse strategy for time-intensive tasks. The query, KPEW calculation, and HetERel calculation phases benefitted from implementing a reuse strategy that writes results to disk for later access. The query task requires communication with the Neo4J graph database that, when performed for each instance of an object type pair, accumulates unnecessary run time. The reuse strategy ensures that graph queries occur a single time and are made available on disk by subsequent equations of HetERel. The KPEW calculation phase of HetERel reuses saved graph query results to compute pairwise KPEW values once and save them to disk for the matrix multiplication and relevance computation phases. Finally, calculated HetERel scores are saved to file for fast look up between start and target objects of interest in *O(n)* time. The reuse strategy is applicable to HetERel for these tasks because the

data and calculations do not change for the duration of the analysis. The introduction of new data necessitates these time-intensive tasks to be rerun for new analyses.

Source code for HetERel computation and Cypher querying is available in **Appendix A**.

3.3 Evaluation Method for Comparing Meta Path Based Relevance Ranking

In this chapter, a heterogeneous information network that consists of chemicals, genes, pathways, and diseases was created to develop and test HetERel. From this network, HetERel can be used to identify and rank the relevance between chemicals and disease (3 steps away) and genes and disease (2 steps away). To evaluate the meta path based ranking capabilities of HetERel, it is necessary to aggregate "gold standard" sets of known associations for chemicals to disease and genes to disease.

3.3.1 Gold Standard Datasets

The Comparative Toxicogenomics Database (CTD) is a highly curated knowledgebase containing direct associations between chemicals and human health phenotypes [42]. The CTD is the foremost authority on chemical interaction with genomes of model organisms relevant to *Homo sapiens*. It contains millions of associations between chemicals, genes, and disease which are utilized within the heterogeneous information network described in **Chapter 2**. The CTD has previously been used as the baseline for chemical-disease association inference and analysis [11, 87, 180] and maps directly to MeSH term, Entrez Gene, and OMIM identifiers. In the CTD, direct chemical-disease associations can be supported by direct evidence from curated scientific articles or inferred through chemical-gene and gene-disease associations. The

Algorithm 1 KPEW Calculation Phase-Example Python Implementation

```
1: procedure KULCZYNSKISIM2(xr, x, r)
2:     term1 = xr/xr + x
3:     term2 = xr/xr + r
4:     ks2 = (term1 + term2)/2.0
5: procedure CREATECONTINGENCYTABLE(object, W_{XR_XY}, R_XY, kulcDict)
6:     for predicate in W_{XR_XY}[object] do
7:         XR = W_{XR_XY}[object][predicate]
8:         X = sum(W_{XR_XY}[object].values())
9:         R = R_XY[predicate]
10:        kulc = kulczynskiSim2(XR, X, R)
11:        kulcDict[object][predicate] = kulc
12: procedure KULCRELATIONSHIPPRODUCT(W_XY, kulcDict)
13:    krpDict = defaultdict()
14:    for xry in W_XY do
15:        objectX = xry[0]
16:        predicateR = xry[1]
17:        objectY = xry[2]
18:        xrk = kulcDict[x][r]
19:        yrk = kulcDict[y][r]
20:        krpDict ← xrk, yrk
21: procedure KULCPRODUCTEDGEWEIGHT(krpDict)
22:    kpewDict = defaultdict()
23:    for x in W_XY do
24:        for y in W_XY do
25:            totalN = 0
26:            kpewSum = 0
27:            for krp in W_XY do
28:                totalN+ = 1
29:                kpewSum+ = krp * krp
30:                kpew = 1 − sqrtkpewSum/totalN
31:                kpewDict[x][y] = kpew
32:    return kpewDict
33: procedure CALCULATEKPEW(R_AB, W_{AR_AB}, W_{BR_AB}, W_AB)
34:    kulcDict ← defaultdict()
35:    for a in W_{AR_AB} do
36:        createContingencyTable(a, W_{AR_AB}, R_AB, kulcDict)
37:    for b in W_{BR_AB} do
38:        createContingencyTable(b, W_{BR_AB}, R_AB, kulcDict)
39:    krpDict ← kulcRelationshipProduct(W_AB, kulcDict)
40:    kpewDict ← kulcProductEdgeWeight(krpDict)
41:    return kpewDict
```

Algorithm 2 HetERel Algorithm-Psuedocode Implementation

```
1: procedure calculateHalfPath(halfPath)
2:     for step in halfPath do
3:         associations ← parseSavedAssociations(step) return
4: procedure calculateHetERel(metaPath)
5:     if metaPath == even then
6:         leftPath ← metaPath/2
7:         rightPath ← metaPath/2
8:         calculateHalfPath(leftPath)
9:         calculateHalfPath(rightPath)
10:        for associationSet in halfPath do
11:            createAdjacencyMatrix(associationSet)
12:            normalizeAdjacencyMatrix(adjacencyMatrix)
13:        matrixMultiplication(leftMatrices)
14:        matrixMultiplication(rightMatrices)
15:        hetERelScores = cosineNormalization(leftMatrices, rightMatrices)
16:    if metaPath == odd then
17:        leftPath ← (metaPath/2) + 1
18:        rightPath ← (metaPath/2) − 1
19:        calculateHalfPath(leftPath)
20:        calculateHalfPath(rightPath)
21:        for pair of halfPaths do
22:            for associationSet in halfPath do
23:                createAdjacencyMatrix(associationSet)
24:                normalizeAdjacencyMatrix(adjacencyMatrix)
25:            matrixMultiplication(leftMatrices)
26:            matrixMultiplication(rightMatrices)
27:            hetERelScores2 ← cosineNormalization(leftMatrices, rightMatrices)
28:        finalScores = (hetERelScores + hetERelScores2)/2
```

CTD associates 8,840 chemicals to 3,075 diseases through direct evidence.

DisGeNET is a publicly available, comprehensive, integrated resource for gene and gene variant to disease associations [134]. It encompasses a variety of gene-disease association sources such as UniProt, the CTD, OMIM, and Genetics Association Database. DisGeNET includes gene-disease associations for human, rat, and mouse species. Associations in DisGeNET are organized into curated or predicted sets, with an all-inclusive set also available. In this evaluation, only curated gene-disease associations are considered. DisGeNET contains curated associations between 8,949 genes and 13,075 diseases.

Preprocessing of CTD and DisGeNET data was performed to compare HetERel scoring with chemical-disease and gene-disease gold standards. For the chemical-disease gold standard, the **CTD_chemicals_diseases.tsv** file from the CTD was parsed. All chemical and disease identifiers in the file are MeSH identifiers, facilitating the ease of comparison to results from HetERel. As previously mentioned, only associations with direct evidence were included as part of the gold standard set of chemical-disease associations, amounting to 88,359 curated chemical-disease associations from the CTD. The **curated_gene_disease_associations.tsv** file from DisGeNET was parsed to make the gene-disease gold standard. Entrez Gene identifiers are associated to diseases via Unified Medical Language System concept unique identifiers (CUIs). The **disease_mappings.tsv** file from DisGeNET maps UMLS CUIs to other controlled vocabularies, such as MeSH, the Disease Ontology, and the Human Phenotype Ontology. Only human genes and MeSH mapped diseases were parsed for the gold standard dataset. The association file from DisGeNET contained

130,821 curated gene-disease associations. The source code used to parse and preprocess the CTD and DisGeNET gold standards can be found in **Appendix A**.

3.3.2 Relevance Ranking Analyses

Global relevance analyses from genes and chemicals to disease concepts were performed on the test network described in **Section 3.1.4**. Ranked lists of genes and chemicals from these relevance analyses are compared to their respective gold standard sets to determine the total amount of true positives and false positives. True positives are genes or chemicals present in both the ranked list result and the gold standards set while false positives are genes or chemicals present only in the ranked list result. True positives and false positives are used to create a Receiver Operating Characteristic (ROC) curve to measure the performance of meta path based relevance ranking measures. An ROC curve plots the true positive rate (sensitivity) and false positive rate (1 - specificity) of results from the relevance measures. A perfect relevance measure would produce an ROC curve that follows the left hand border and top border of the plot, indicating a sensitivity and (1-specificity) of 1.

3.3.3 Comparison of Results to Existing Methods

Relevance ranking analysis were performed by HetERel and three existing meta path based relevance ranking methods for comparison. These meta path based methods are capable of determining similarity between objects of the same type or relevance between different types of objects in heterogeneous information networks. The methods tested, like HetERel, are semi-supervised and do not require a training data set for relevance ranking. Unlike these methods, HetERel includes predicate probabilities

between pairs of objects in its calculation for ranking.

HeteSim serves as the basis for two meta path based ranking methods compared in this work, HetERel and AvgSim. HeteSim is a pair-wise random walk method developed by Shi [156] that possesses metric properties, previously described, that provide a desirable foundation for numerous data mining tasks. However, HeteSim has been implemented in smaller, less connected heterogeneous information networks such as computer science conference citation networks, movie databases, and protein interaction networks [156,197]. The complexity of the decomposition step in HeteSim for odd length paths generates a combinatorial explosion of intermediate associations that hamper scalability to larger, more complex datasets.

AvgSim was developed to measure relatedness between objects of different types in large heterogeneous information networks. To circumvent the complexity in HeteSim's decomposition step for odd length paths, AvgSim implements a forward random walk from start to target object, and reverse random walk from target object to start object constrained to a given path. The AvgSim score is the arithmetic mean of these path constrained walks. Further optimization involves dynamic programming, to determine the most efficient order of matrix multiplication operations, and parallelization of matrix multiplication using the MapReduce algorithm [110]. AvgSim increases the computational efficiency of HeteSim for use in large networks but performance was only tested on the same datasets as HeteSim, with the addition of large, sparse, randomly generated data.

The Degree Weighted Path Count (DWPC) metric is the most dissimilar of the methods compared in this work. It is an extension of the path count algorithm

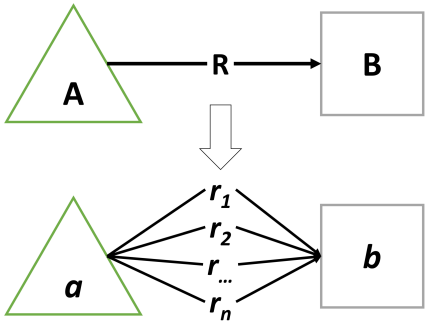
that reduces the bias introduced by general objects exhibiting high connectivity in a network. DWPC does so by introducing a damping exponent to the path count for each object in a meta path, thereby downweighting high degree objects along the meta path. DWPC has been used to quantify relationships for the prioritization of gene and disease, as well as drug and disease, associations [75,76]. Similar to HetERel, DWPC does assign predicates to relationships between objects but still summarizes them as a single predicate type.

11 well-known, diet related diseases were selected to compare the relevance ranking performance of HetERel, HeteSim, AvgSim, and DWPC in the large biological heterogeneous information network described in **Section 3.1.4**. The implementations for all methods can be found in **Appendix A**. For comparison, two meta paths of different length were selected, connecting chemicals to disease (meta path of length 3) and genes to disease (meta path of length 2). All possible pairs of chemicals or genes to disease were included and ranked by each method, when computationally possible. Due to the complexity and in-memory consumption of it's decomposition step, HeteSim could not calculate scores for the entirety of chemical to disease pairs. As such, the HeteSim decomposition step was substituted in favor of the HetERel decomposition step. This allowed HeteSim to perform a global analysis of the chemical to disease meta path.

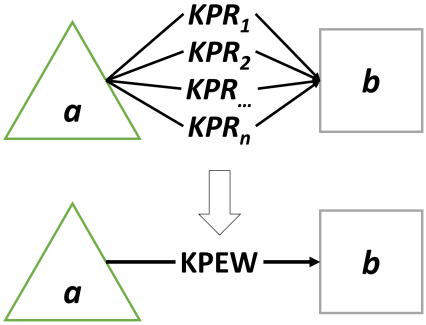
The results of relevance ranking from each method were saved in a pairwise, tab-separated format for ease of comparison. ROC analysis was performed on the results of the selected diseases and the highest ranking 2000 chemicals (**Figure 11**)and 1000 genes (**Figure 12**)were used to generate ROC plots and AUCs. HetERel out-

performs AvgSim and DWPC with significantly different (p ¡ 0.05) AUCs for all chemical to disease rankings except for Colonic Neoplasms, as calculated by implementation of a bootstrapped AUC comparison test [144]. The low overall AUC for all methods ranking chemicals and Colonic Neoplasms highlights the widespread effect of inflammation in cancers and the complex interaction of bacterial metabolites in the intestines [93, 105]. The performance between HeteSim and HetERel is only incrementally different in both chemical and gene to disease rankings, with HetERel producing similar or slightly larger AUCs.

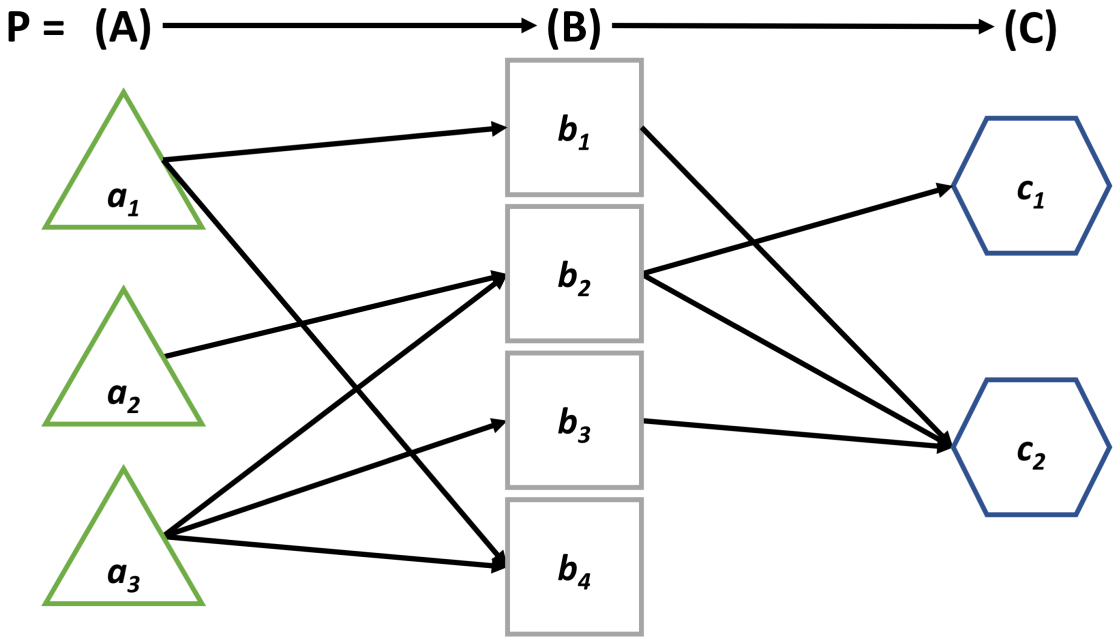
The similarity in performance can be attributed to the uniform framework implemented by both HeteSim and HetERel to quantify and search for relevance between objects of same or arbitrary types. HetERel weights the adjacency matrices using the KPEW while HeteSim favors a binary scoring system. The normalization step for each adjacency matrix relegates the influence of KPEW weighting on overall score calculations. However, the inclusion of predicate probability distributions in HetERel allows it to discern granular differences in pair-wise object ranking. This is exemplified in ROC plots (**Figure 13**)of the top 100 gene to disease rankings of **Figure 12** where HetERel outperforms all compared methods in each disease. Furthermore, the gene-disease rankings include well-studied diseases of dietary relevance, such as those affecting the cardiovascular and digestive systems. The ability of HetERel to provide a more accurate ranked list of the top 100 associated genes, or any biological concept, supplies a short, evidence-based set of biological candidates for experimental validation. This is extraordinarily functional in the biological domain that is inundated with thousands of proposed associations [166].



(a) Expansion of relation $A \xrightarrow{R} B$



(b) Calculation of KPEW between a and b.



(c) Example HIN for HetERel calculation

Figure 10: Breakdown and visualization of KPEW and HetERel calculations. In **10a** the simple relation $A \xrightarrow{R} B$ is expanded to further capture semantic information from detailed link types via $a \xrightarrow{r_1, r_2, \ldots, r_n} b$ where $r \in R$. **10b** visualizes the Kulczynski Relationship Products used to calculate the KPEW between A and B. In **10c** HetERel is calculated based on the heterogeneous information network, following the path $P{=}ABC$.

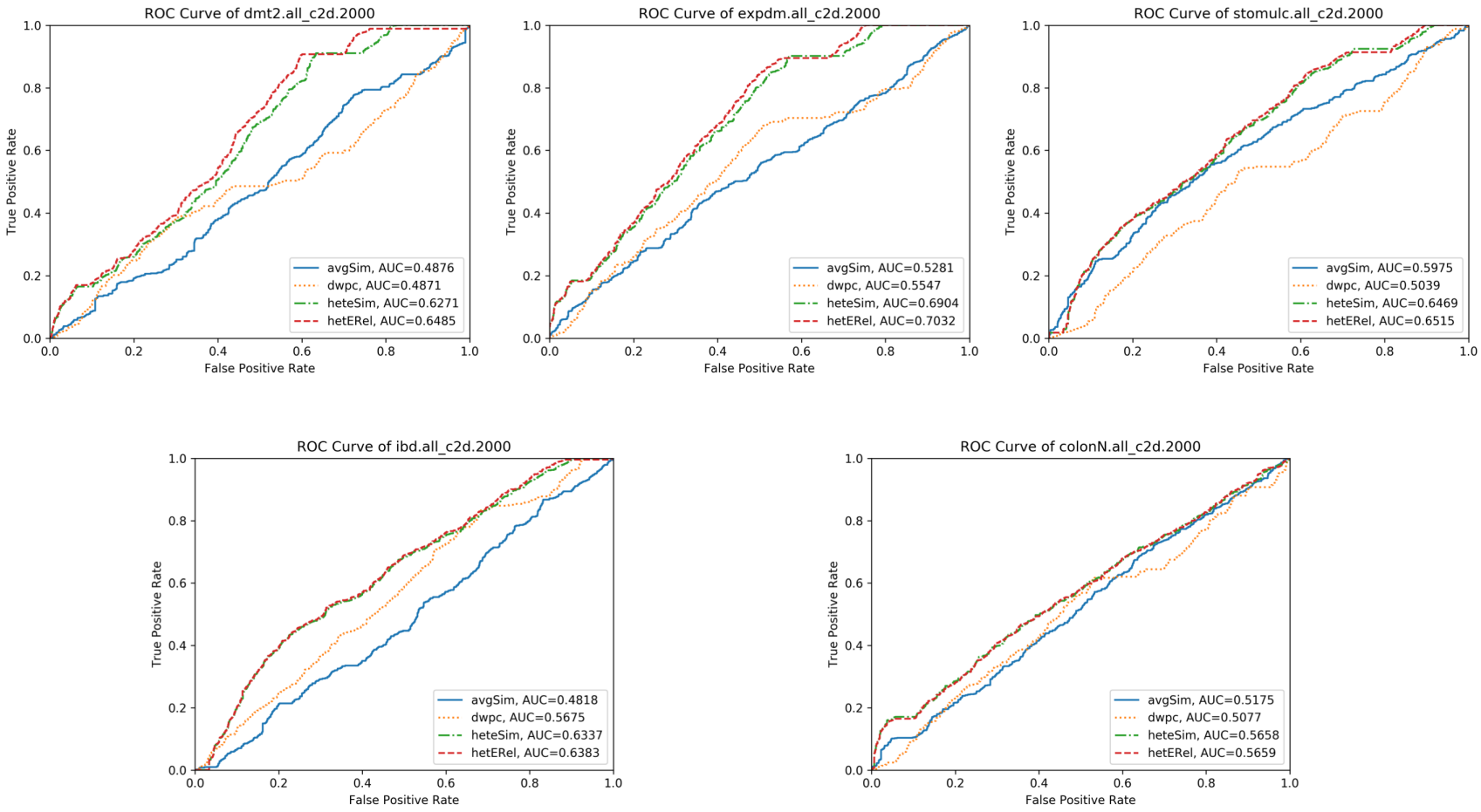


Figure 11: Relevance ranking performance comparisons- ROC plots for diet related digestive system diseases and their associated chemicals ranked by each relevance method. Diseases analyzed include: dmt2- Diabetes Mellitus, Type 2, expdm- Diabetes Mellitus, stomulc- Stomach Ulcers, and colonN- Colonic Neoplasms

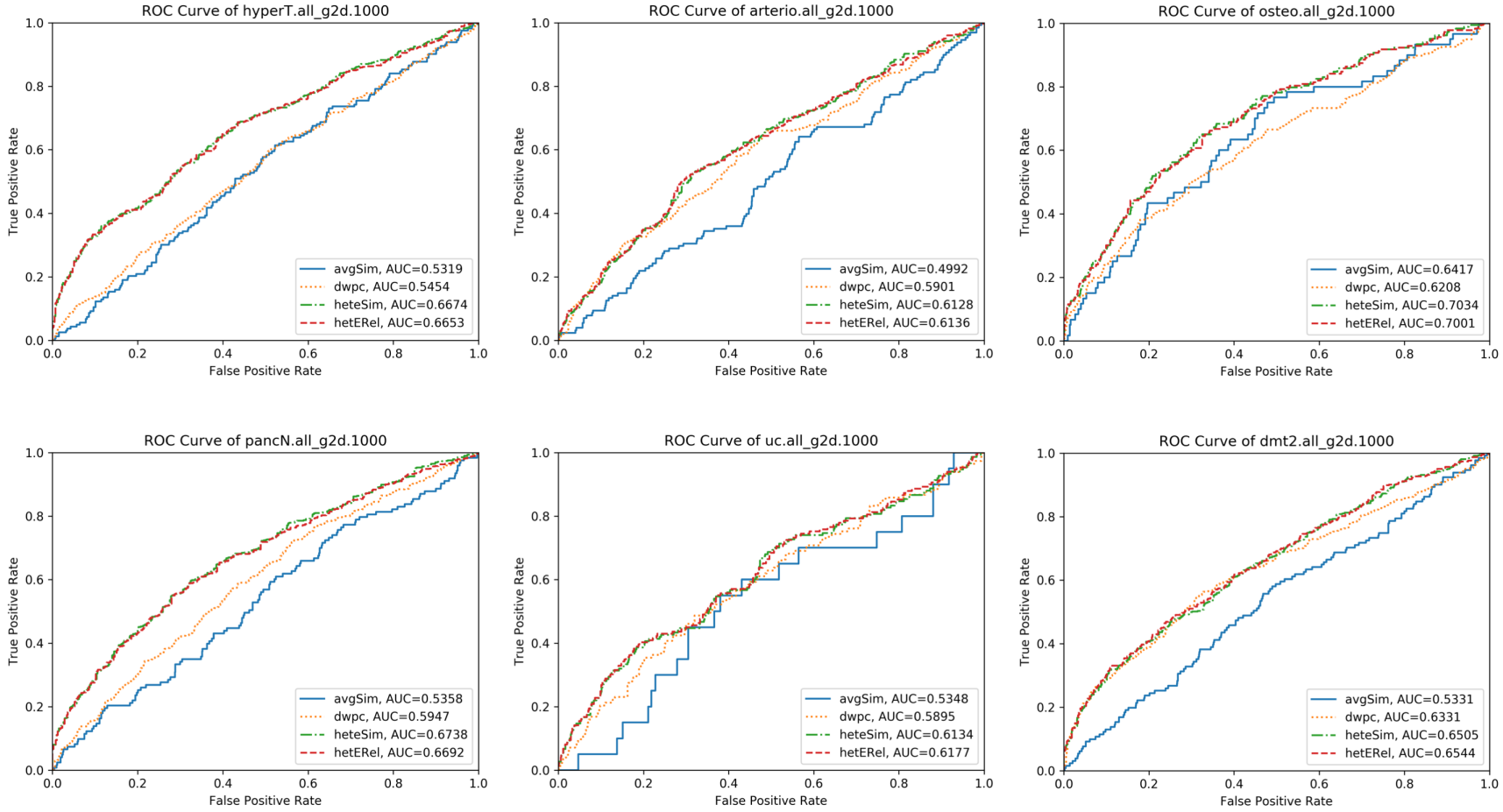


Figure 12: Relevance ranking performance comparisons- ROC plots for diet related cardiovascular and digestive system diseases and their associated genes ranked by each relevance method. Diseases analyzed include: hyperT- Hypertension, arterio- Arteriosclerosis, osteo- Osteoporosis, pancN- Pancreatic Neoplasms, uc- Ulcerative Colitis, and dmt2- Diabetes Mellitus, Type 2

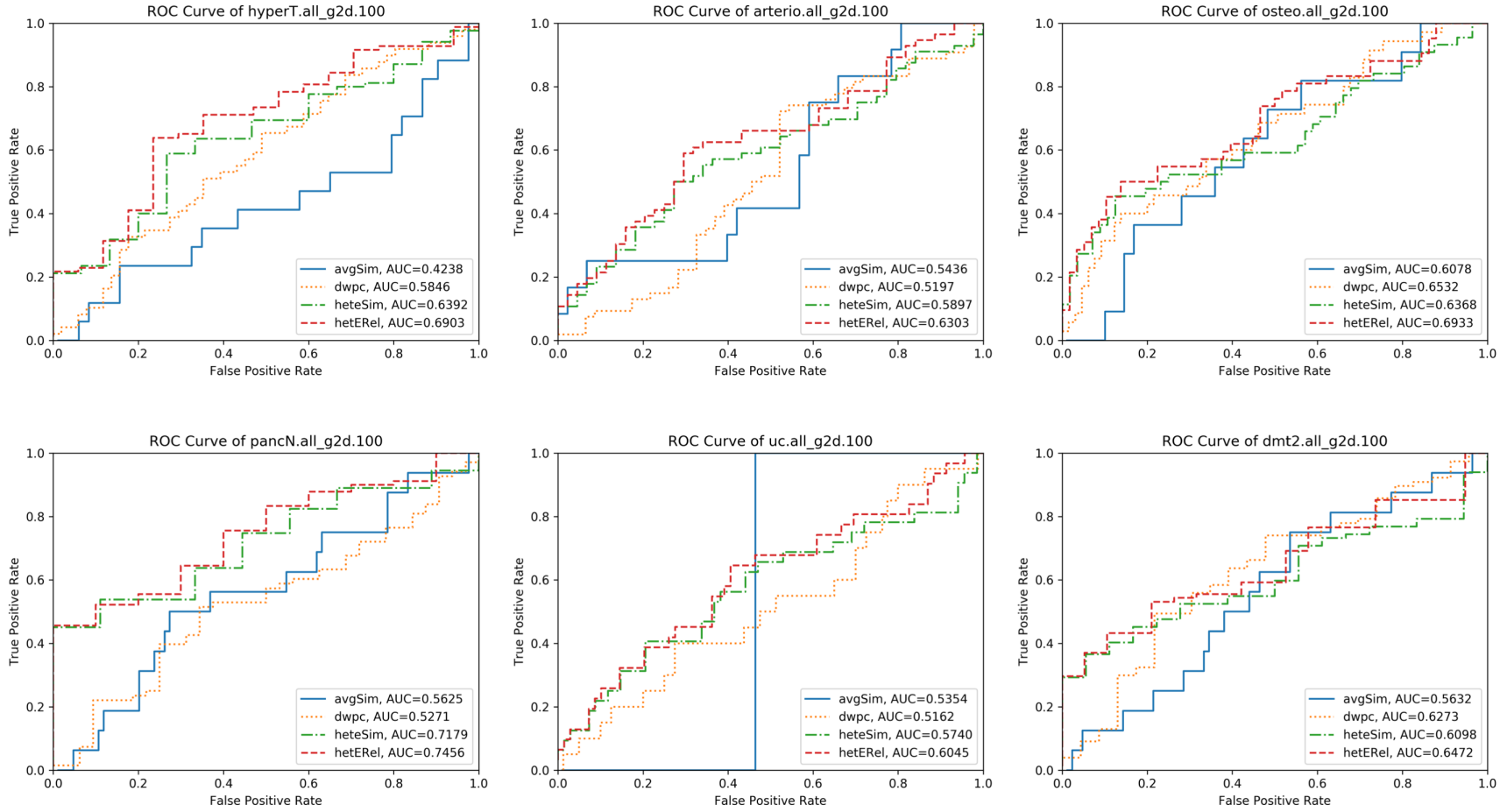


Figure 13: Relevance ranking performance comparisons- ROC plots for diet related cardiovascular and digestive system diseases and their associated genes ranked by each relevance method. Diseases analyzed include: hyperT- Hypertension, arterio- Arteriosclerosis, osteo- Osteoporosis, pancN- Pancreatic Neoplasms, uc- Ulcerative Colitis, and dmt2- Diabetes Mellitus, Type 2. HetERel outperforms each method as seen by the partial AUC calculated for the top 100 ranking genes in each disease

3.4 Conclusion

HetERel is a novel, data-agnostic metric which ranks the relevance between objects of arbitrary types in large, complex heterogeneous information networks. The HetERel metric incorporates specific predicates in object relationships to quantify the strength of relevance between objects. Developed as an extension of previous relevance metrics, HetERel inherits the metric properties of symmetry and self-maximum(scores bounded to [0-1]). In addition, HetERel exhibits a null-invariant property that allows recovery of low probability predicates which add semantic information and detail. The evaluation of HetERel displays the granular improvement in ranking gained from including the semantic value held in predicates mined from literature and extracted from structured vocabularies.

Meta path selection is an onerous task with numerous possibilities. Further investigation into the multitude of meta paths can improve the quantification and ranking of relevant objects within a heterogeneous network. For example, of the four meta paths presented in **Figure 8b** between chemicals and diseases in the test biological heterogeneous network, P_1 was tested for odd path length and variety of object types. The remaining meta paths may provide higher relevance rankings than P_1 based on the information they contain. The addition of connected data could also provide a better prioritization of relevant objects. The inclusion of curated and manually annotated relationships between objects would supplement existing relationships drawn from controlled vocabularies and text mining.

Relevance search methods, such as HetERel, are essential for mining large biolog-

ical knowledgebases for evidence-based, indirect associations and patterns that can elucidate the chemical interactions of food and the human genome.

CHAPTER 4: BIOLOGICAL USE CASES FOR THE DIET-DISEASE NETWORK AND META PATH BASED RANKING METHOD

Genetics and diet play crucial roles in the development of many cardiovascular, digestive, and metabolic diseases. Historically, biomedical research has emphasized the influence of genetic variance in the development of disease. As a result, thousands of associations connecting genes and disease have been identified and stored in various databases [48, 160, 184]. More recently, metabolomic studies have identified, beyond an observational level, associations between the components of diet and human health [23, 67]. However, the segmented and exponential growth of association data for diet, genetics, and disease outpaces the ability of researchers to analyze and prioritize interesting patterns.

Association data that describes the link between diet and disease is stored in numerous, domain specific databases. Associations between plants and chemicals are available from the UDSA Nutrient Database (sparse), while more specific chemical detail can be found in the Chemical entities of Biological Interest (CheBI) [72, 182]. The Comparative Toxicogenomics Database (CTD) links chemicals to the genes they interact with [42]. There are many gene to disease association databases which do not include the relationships from biological pathways to genes or disease [70, 134]. Genes and biological pathways are connected across many heterogeneous databases such as KEGG and the Gene Ontology [10, 84]. The data to connect the components

of plant-based diets with human disease phenotypes is separated amongst various data repositories, preventing general data exploration and high throughput hypothesis generation.

Currently, no single resource contains the data or a prioritization method capable of investigating the molecular effects of plant-based diet on human health. Existing resources tend to be disease-centric or focus on small molecule drugs as opposed to phytochemicals in diet [76, 82, 181]. A general resource that aggregates link data between biological and chemical concepts from plants to human health phenotypes can reduce the data sparsity found in specific, single resources. For example, when searching for genes relevant to a disease, such as type 2 diabetes mellitus, genes that affect biological pathways associated to diabetes would be returned, in addition to genes directly associated to diabetes. A method to mine this resource which provides a relevance-based prioritization of concepts of interest would allow researchers to efficiently investigate the molecular effects of phytochemicals on disease.

The development of an integrated diet-disease network encompassing all land plants and human health phenotypes was described in **Chapter 2**. In **Chapter 3**, the design and implementation of a meta path based relevance ranking framework was discussed. This chapter provides insights into the research trends of plants through exploratory analysis of the diet-disease network and amalgamates the meta path based ranking framework with the diet-disease network to determine the most relevant molecular interactions of dietary components and human disease.

4.1 Diet-Disease Network

The CTD is one of the most comprehensive resources connecting the components of plant-based diets with human disease phenotypes. It explicitly includes associations spanning from chemicals to genes, genes to biological pathways, and biological pathways to disease [42]. The CTD contains few associations between phytochemicals and their plants of origin. As a result, the relevance of most plants to chemicals, genes, and diseases cannot be calculated.

The Diet-Disease Network developed in **Chapter 2** extends and enhances the associations found in the CTD. The network aggregates multiple sources of association data and integrates their entities to ameliorate the sparsity of connections between plants, chemicals, and human genes. In addition, association density is augmented with text-mined relationships from the agricultural and biomedical literature. The diet-disease network also retains the relationship types extracted from text mining and curated association data. Specific relationship types provide semantic information that help to distinguish the relevance between two like objects. The integration and augmentation of data describing the components of diet and disease facilitates reachability across the entirety of the diet-disease network.

4.1.1 Overview of Diet-Disease Network

The diet-disease network consists of 5 types of agricultural, biological, and chemical entities integrated 13 different curated data sources. The 5 entity types are *Plant (E)*, *Chemical (C)*, *Gene (G)*, *Pathway (M)*, and *Disease (D)*. Overall, there are 732,094 unique nodes within the diet-disease network. These entity types are linked

by 9,109 unique relationship types. The high variation in relationship type results from the extraction of predicates from text mining. The heterogeneous information network schema in **Figure 14a** displays the connectivity between entities of the diet-disease network. From this network schema, several meta paths linking plants *(E)* to disease *(D)* can be deduced, as shown in **Figure 20b**. Meta paths which connect plants to all other entity types can be derived from those four meta paths. The semantics underlying each meta path portrays a distinct meaning between plants and disease. The biological inquiries posed by a researcher can be used to guide meta path selection.

4.1.2 Meta Path Based Exploratory Analysis of Diet-Disease Network

General inquiries of the diet-disease network provide insight into the interesting properties of plants with potential benefits for human health. The results from exploratory queries highlight plants with high connectivity to other concepts, such as chemicals and genes, summarizing the current state of agricultural and biomedical research. This entire work is premised on the question, "Why do certain plant-based foods confer benefits to human health?". In **Chapter 2.5**, that question is decomposed into three more manageable questions which guided the selection of meta paths for global exploratory analysis of the diet-disease network. The three biological questions are: 1)What is the phytochemical profile of plant-based foods? 2) How do these phytochemicals affect human genes? and 3) How do these effects on human genes influence human health phenotypes, such as disease?

The first question can be answered through analysis of the meta path $P = E \rightarrow C$.

This meta path, of length 1, consists of a single pair of associated objects linked by a single step. A unique count of plant to chemical associations quickly identifies the overall size of plant chemical profiles. The 10 plants with the most chemical associations from the diet-disease network are shown in **Figure 15a**, ranging from 2,180 down to 598 unique chemicals. Among them are common agricultural crops one would expect to be well characterized, such as soybeans (*Glycine max*), corn (*Zea mays subsp. mays*), spinach (*Spinacia oleracea*), coffee (*Coffea arabica*), and oats (*Avena sativa*). There is a clear difference in the number of unique chemicals associated to the top 10 plants in the diet-disease network compared to the top 10 chemically characterized plants from the CTD (**Figure 15b**). This indicates that the aggregation and integration technique used in developing the network successfully augmented the existing, curated data. It is interesting to note that two of the top ten plants, *Nicotiana tabacum* and *Arabidopsis thaliana*, are not edible plants, but are highly characterized for their economic and research value, respectively [196]. Plants with large chemical profiles exhibit a diverse pool of potential bioactive components which warrant further investigation. Specific plants and their associated chemicals will be discussed in further detail in the coming section, based on their relevance rankings.

The meta path $P = E \rightarrow C \rightarrow G$, of length 2, is composed of the three object types *Plant(E)*, *Chemical(C)*, and *Gene(G)*. This meta path answers the second biological question by determining the subset of human genes affected by chemicals associated to plants. **Figure 16a** displays the top 10 plants with unique associations to human genes, ranging from 19,161 (approximately a third of human genes) to 14,905 unique

genes. Four of the top ten plants associated to human genes did not have the largest phytochemical profiles. These exploratory results indicate that the quantity of chemical associations plays a role in but is not the driving factor for the overall interaction of a plant with the human genome. This trend continues as the top k number of plants is increased. When the reverse meta path is queried ($P = G \rightarrow C \rightarrow E$), the human genes most affected overall by plants is found. The top 10 human genes are shown in **Figure 16b**, interacting with between 3,326 and 2,427 plants. These top genes are involved hallmark biological processes which lead to chronic disease, including transcriptional regulation, inflammation, and oxidative stress [44, 86, 94]. Particular plants of interest and the human genes they affect will be discussed in detail in the following section.

The third question explores which biological pathways are perturbed by the interaction of phytochemicals and human genes. Querying the meta path $P = E \rightarrow C \rightarrow G \rightarrow M$ returns a list of biological pathways affected by each plant. To gain a general understanding of the most commonly affected biological pathways, the reverse meta path $P = M \rightarrow G \rightarrow C \rightarrow E$ is analyzed. The top 10 biological pathways affected by plants are shown in **Figure 17**, with counts of unique plants between 5,228 and 5,146. Plants affect many of the cellular processes in humans, such as signaling, biosynthesis, and apoptosis. Perturbation of these essential biological pathways can have detrimental effects on human health [16, 38, 107]. Noise can be introduced in meta path with longer lengths. Most search algorithms, such as Dijkstra, search for the shortest path to avoid the effects of noise. However, the implementation of a prioritization framework can account for this and identifies the signal from the noise. To

make the best use of the diet-disease network, we apply the meta path based ranking method developed in **Chapter 3**.

4.2 Application of Meta Path Based Ranking Framework in Diet-Disease Network

The exploratory analysis of the diet-disease network provides a cursory survey of plants and chemicals that elicit an effect on the human genome. From these insights, hypotheses can be developed to explain the molecular interactions between specific plants of interest and human disease. Hypothesis development is driven by the same biological questions that guide exploratory analysis. Each question guides the relevance analysis and informs a researcher about which meta paths best satisfy the biological inquiry. The application of the meta path based ranking framework in querying the diet-disease network prioritizes the molecular entities associated with the specific plant or chemical of interest. The top prioritized candidates help refine relevance results to generate evidence-based, testable hypotheses.

4.2.1 Phytochemical Profiling

Identifying the chemical profile of plants establishes the phytochemical space of bioactive candidates for human health. A global relevance analysis was performed on the diet-disease network along the meta path $P = E \rightarrow C$ which included all land plants (*Embryophyta*). The analysis generated phytochemical lists for each plant ranked by the HetERel relevance metric. Any *Embryophyta* can be quickly found and their ranked chemical profile returned. In this section, we investigate the profiles of three plants of dietary interest.

Broccoli (*Brassica oleracea var. italica*) is a commonly consumed vegetable as-

sociated with a reduced risk of numerous chronic diseases [159]. In the literature, this effect is attributed to derivatives of a group of chemicals called glucosinolates. Glucosinolates are found in the *Brassicaceae* family of cruciferous vegetables. They are broken down by the enzyme myrosinase through the process of mastication and release sulfur compounds called isothiocyanates. The most well-studied of these isothiocyanates is glucoraphanin and its breakdown products including sulforaphane [74]. The relevance analysis of broccoli returned 215 unique chemicals, which include many of the known derivatives of glucosinolates. **Table 8** displays the top 10 relevance ranked chemicals associated to broccoli in the diet-disease network. Many of these chemicals, such as 4-methoxyglucobrassicin and glucoiberin, have been found to have antioxidant properties linked to the reduction of inflammation in humans [40].

Oats (*Avena sativa*) are a global food staple, commonly eaten as oatmeal or cereal. The dietary fiber found in oats have been proven to lower cholesterol and reduce the risk of cardiovascular disease and obesity [49, 57]. Phytochemical groups such as avenanthramides and carbohydrates (mainly beta-glucans) are responsible for the antioxidant and LDL-cholesterol reducing effects found in oats [158]. These phytochemical groups are represented in the top 10 of 598 unique chemicals associated to oats (**Table 8**) in the diet-disease network. Beta-glucans, such as lichenin and xylotetraose, are well-known for their effects in cardiovascular health, but have recently shown potential antitumor properties as well [33, 201]. An interesting component of oat in the top 10 table, is the protein avenin. Avenin is a protein in oat that people, rarely, can be sensitive to, similar to that of gluten found in wheat or barley [71].

Coffee (*Coffea arabica*) is one of the most popular beverages worldwide. Coffee

is well-known for its caffeine content and effect on alertness [203]. Recently, meta-analyses of the correlation between coffee consumption and various health outcomes from a multitude of observational studies highlighted liver outcomes to have large and consistent effect sizes. Beneficial associations were also found for specific cancers and metabolic diseases, such as prostate cancer and type 2 diabetes mellitus [45, 66, 136]. Opposing studies have proposed that the presence of volatile compounds, such as pyrroles and 4-methylimidazole, have had detrimental, carcinogenic effects in in-vitro studies [39, 116, 133]. The top 10 chemicals associated with coffee (**Table 8**) are cited as key active compounds in a number of these studies. Cafestol is a bioactive diterpene proposed to decrease the mutagenic effect of multiple carcinogens through different mechanisms, such as antioxidant defense and inhibition of carcinogenic activation [26].

Defining the chemical profiles of plant-based foods reduces the search space for identifying potential bioactive compounds in the human diet. A compound has bioactive properties if it interacts with and produces an effect on human gene products. In order to identify these interactions, the meta path is extended another step to include associations between chemicals and human genes.

4.2.2 Gene Prioritization

Canonically, the purpose of gene prioritization has been to rank genes according to their association to a phenotype or disease. A wealth of data connecting human genes and disease exists in the biomedical domain for this task. Far less data exists linking plant-based foods and their chemical products to human genes. Text-mined associations integrated into the diet-disease network help bridge this gap in computer

readable data, allowing a researcher to apply the gene prioritization task to the domains of diet and nutrition. The meta path ($P = E \rightarrow C \rightarrow G$), of length 2, was used to analyze the relevance between plants and the human genes their chemical products interact with. In this section, the most relevant genes affected by broccoli(*Brassica oleracea var. italica*), oat(*Avena sativa*), and coffee(*Coffea arabica*) are discussed.

The compounds in the chemical profile of broccoli were associated to 11,507 unique human genes. To gain a general understanding of the functions of the gene set, the top ranked 1000 genes were subjected to a gene function analysis using the PANTHER classification system [112]. An overrepresentation analysis using a Fisher's Exact Test against the GO-slim molecular function annotation data set, with FDR correction $p < 0.05$, was performed to generate the bar plot in **Figure 18a**. The most overrepresented molecular functions of genes associated to broccoli include peroxidase activity, antioxidant activity, and oxidoreductase activity. It can be deduced from these results that the chemical constituents, namely isothiocyanates, of broccoli promote the reduction of peroxides and reactive oxygen species in the human body, as proposed in the literature [108]. This type of molecular function generally reduces inflammation and oxidative stress throughout the cardiovascular and digestive system via interaction with isothiocyanates that induce ARE-mediated pathways. [122, 193].

Oat related chemicals were found to be associated to 13,690 unique human genes. An overrepresentation analysis was also performed on the top 1000 ranked genes associated to Oat to characterize their molecular function. **Figure 18b** presents the most overrepresented molecular functions of oat associated genes. The most overrepresented functions are neuropeptide hormone activity at 8.97 fold enrichment, and

tumor necrosis factor receptor binding at 8.74 fold enrichment. The avenanthramides in oat have been found to have potent anti-inflammatory effects on the skin leading to reduced dermatological by blocking neurogenic inflammation from sensory receptors in the skin. This is accomplished through a mechanism of anti-inflammatory activity in human skin cells involving tumor necrosis factor (TNF) induced inhibition of the transcription factor NFKB [172]. Many pro-inflammatory genes from the interleukin family of genes are regulated by NFKB, meaning its inhibition by avenanthramides leads to a reduction in the production of interleukin gene products.

Coffee related chemicals in the diet-disease network were found to be associated with 13,249 unique human genes. The PANTHER molecular function analysis results (**Figure 18c**) show that the neuropeptide hormone activity, tumor necrosis factor receptor binding, and guanylate cyclase activity are the top 3 statistically overrepresented molecular functions of coffee associated genes. Cafestol, a chemical unique to coffee, has a suppressive effect on the expression of the COX2 gene which serves a crucial role in inflammation and carcinogenesis. It suppresses NFKB, which in turn suppresses TNF-mediated COX2 expression [155]. The caffeine found in coffee has been studied for mechanisms of reducing the risk of heart disease. Caffeine stimulates the reduction of calcium in the vascular smooth muscle cell, in turn increasing uptake in endothelial cells. It also stimulates the production of nitric oxide (NO). NO binds to the guanylate cyclase enzyme and activates it to create cyclic GMP which increases protein kinase activity, acting as a vasodilator and lowering blood pressure. At the same time, caffeine acts as a competitive inhibitor of phosphodiesterase enzymes that are meant to degrade cyclic GMP, leading to an accumulation of cyclic GMP and

increasing the effect of vasodilation [47].

4.2.3 Relevance of Plant-Based Diets and Human Health

The diet-disease network spans the agricultural, chemical, and biological domains to connect plants and human health. Building upon the relevance results from generating phytochemical profiles and prioritizing candidate genes, the indirect link between plant-based foods and human health can be calculated. In the diet-disease network, the meta path ($P = E \rightarrow C \rightarrow G \rightarrow M \rightarrow D$), of length 4, determines the relevance between plants and disease. This complete path sheds light on the molecular mechanisms by which edible plants affect human health.

Table 9 displays the top 10 diseases related to broccoli, oat, and coffee ranked by their HetERel relevance score. The most relevant diseases linked to broccoli are as expected, including 4 varieties of cancer, oxidative stress, and inflammation. Oat is most associated with chronic conditions of cardiovascular disease and precursors to diabetes, such as obesity and metabolic syndrome, as one would gather from current dietary guidelines and research. The top 10 diseases relevant to coffee in the diet-disease network are similar to oat, with the exception of brain damage and poisoning.

Upon cursory review of the literature, many published studies refer to the difficulty in performing the task of coffee preparation for those with brain damage [18,63]. Further research within the diet-disease network revealed that caffeine and caffeic acid in coffee inhibits the activation of TNF and the NFKB signaling pathway. This interaction provides vasodilative activity that reduces the risk of ischemia to not just the heart but also the brain, preventing brain damage**Figure 19**. Reference information

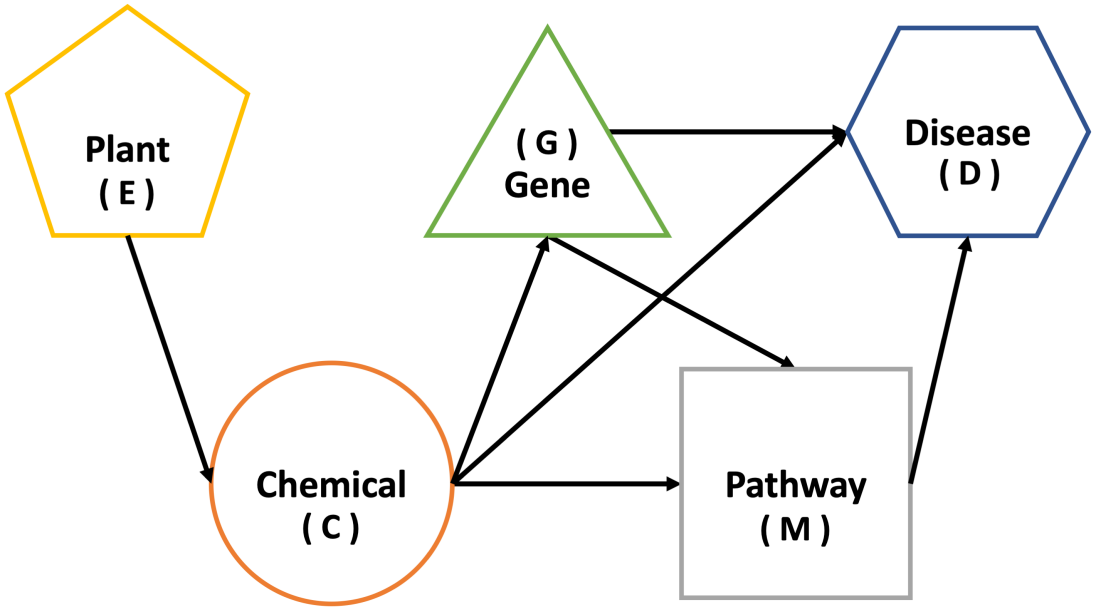
to the literature was readily available in the properties of each link in the diet-disease network. Uncommon inferences, such as the relevance between coffee and brain damage, can be easily investigated when the association data that links them is available in a single, integrated network. Hypotheses can be generated by traversing the diet-disease network along any desired meta path P previously described, each step guided by HetERel relevance scores. Two examples of hypotheses for the relevance between broccoli and disease are visualized with the Neo4J browser interface in **Figure 20**.

4.3 Conclusion

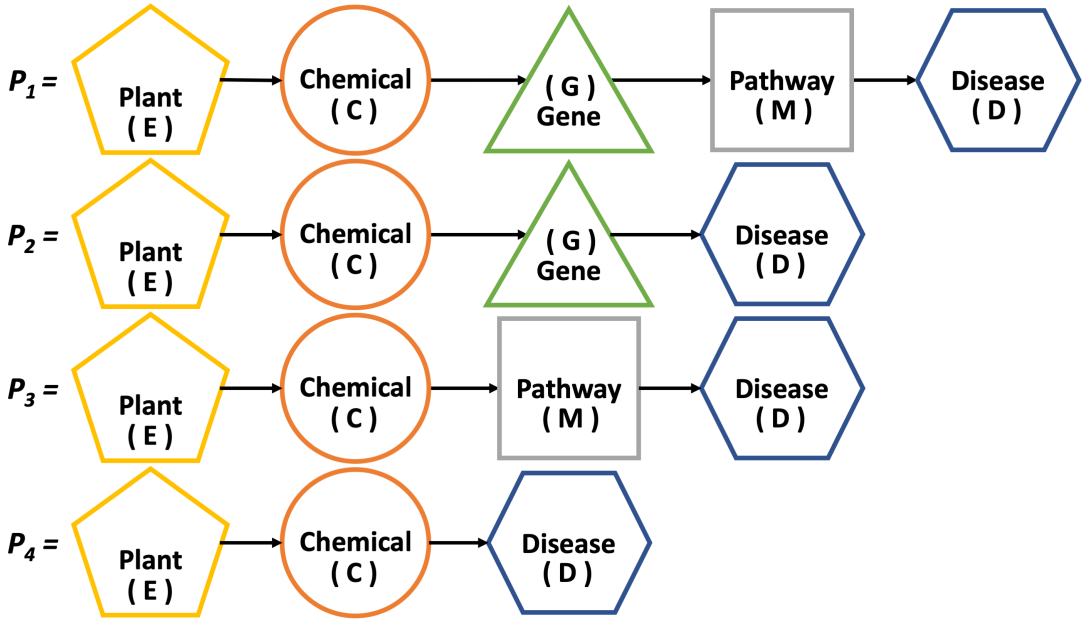
The ability to mine data linking plants, chemicals, human genes, biological pathways, and human health phenotypes in a high throughput manner is highly valuable for investigating the molecular effects of plant-based diets on human health. The relevance results and hypotheses generated in this chapter display the efficacy of the HetERel relevance ranking method in parallel with a comprehensive, integrated information network connecting plants and human disease.

Generating knowledge-based hypotheses that elucidate the molecular pathways leading to health effects from plant-based diets is an essential and time consuming task for an individual researcher. In combination, HetERel and the Diet-Disease Network provides prioritized, evidence-based lists of biologically relevant candidates for laboratory or clinical validation, with data provenance and predicate detail. The HetERel framework and the Diet-Disease Network enables agricultural and biomedical researchers to efficiently access the available knowledge in their fields to not only remain current, but develop data driven experiments in search of bioactive phyto-

chemicals.

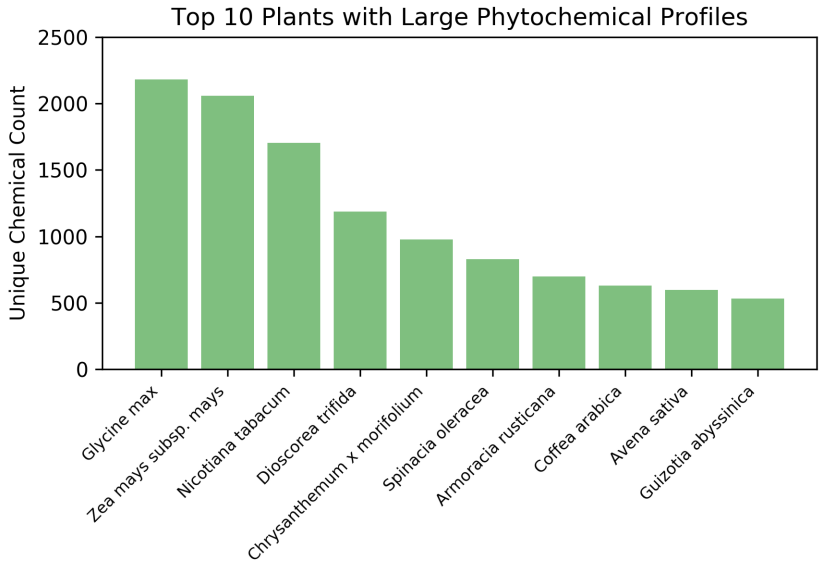


(a) Schema for Diet-Disease Network, as displayed in **Chapter 2**.

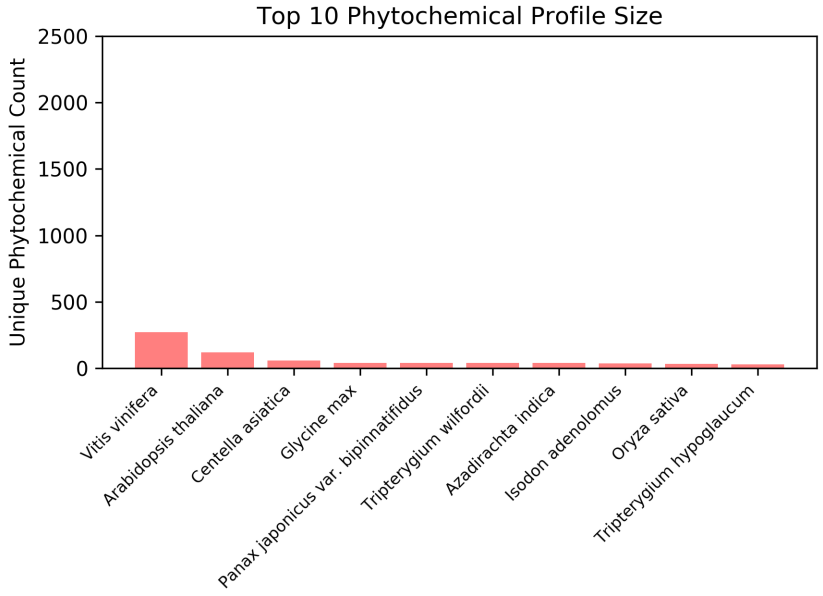


(b) All possible meta paths between objects of the types Plant (E) and Disease (D). These meta paths encompass all meta paths in the network.

Figure 14: The schema for the Diet-Disease Network and all possible meta paths between object types of Plant (E) and Disease (D)

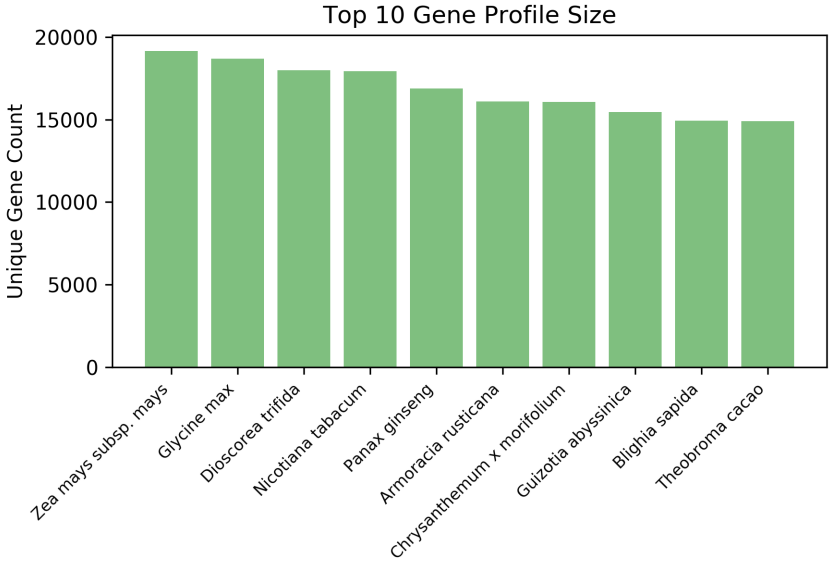


(a) Top 10 plants with unique phytochemical counts from diet-disease network

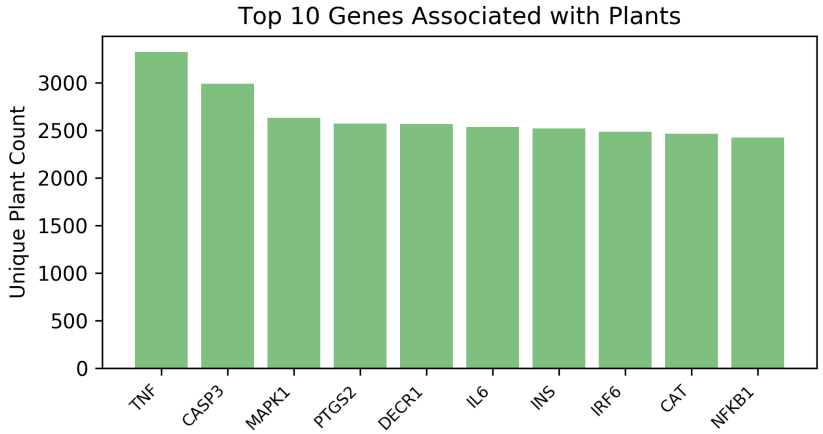


(b) Top 10 plants with unique phytochemical counts from the CTD

Figure 15: Comparison of the top 10 best characterized plants from the diet-disease network and the curated CTD following the meta path $P = E \rightarrow C$



(a) Top 10 Plants associated, through the meta path $P = E \rightarrow C \rightarrow G$. Highly connected plants are likely to influence disease development, progression, or amelioration



(b) Top 10 Genes with unique associations to plants through the meta path $P = G \rightarrow C \rightarrow E$

Figure 16: Overview of the top 10 plants that interact with human genes and the top 10 human genes that are affected by plants

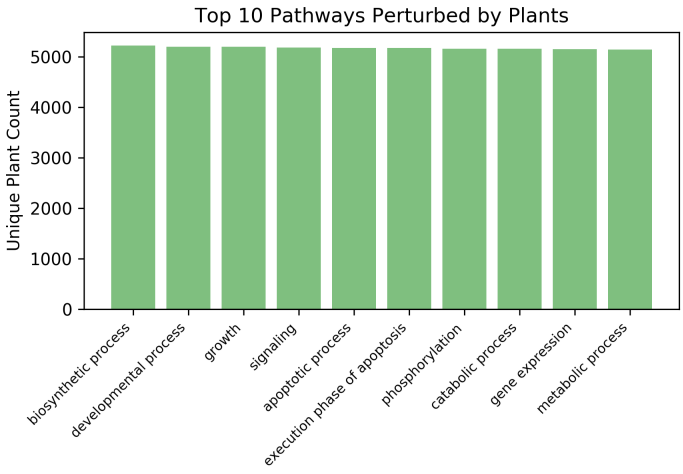
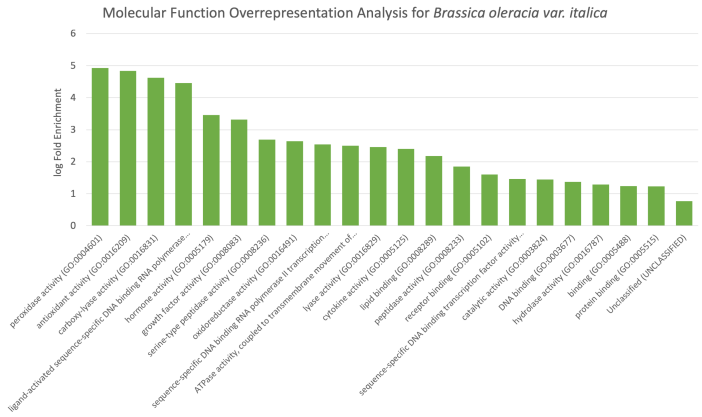


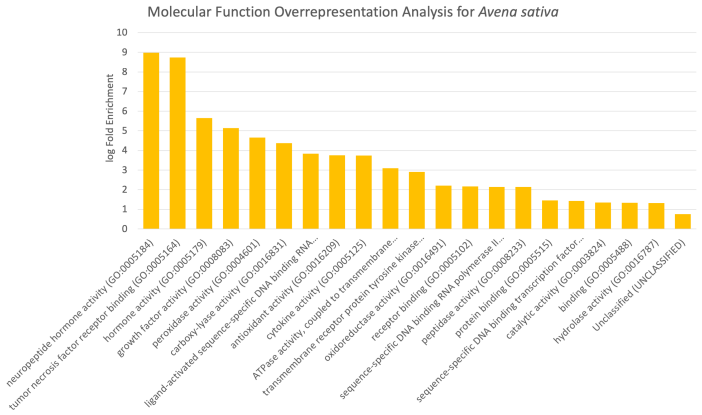
Figure 17: The top 10 biological pathways affected by plants

Table 8: Top 10 chemicals associated with *Brassica oleracea var. italica*, *Avena sativa*, and *Coffea arabica*.

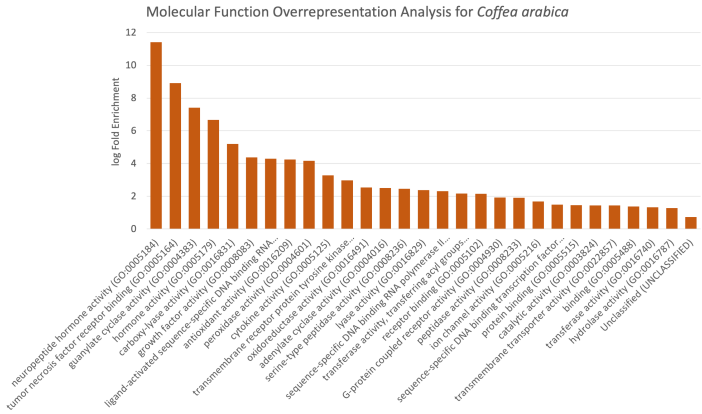
Plant	**Chemical**	**HetERel Score**
Brassica oleracea var. italica	glucoraphanin(1-)	0.066835
	3,3'-diindolylmethane	0.06670
	selenoproteins	0.066751
	4-methoxyglucobrassicin	0.066372
	glucoiberin	0.066369
	3H-1,2-dithiole-3-thione	0.066349
	(R)-sulforaphane	0.048000
	gluconasturtiin	0.047278
	glucotropeolin(1-)	0.047050
	neoglucobrassicin	0.046811
Avena sativa	vitexin 2"-O-beta-L-rhamnoside	0.040521
	lichenin	0.028638
	diferulic acid	0.028630
	2'-deoxymugineic acid	0.028628
	indol-3-ylacetaldehyde	0.028628
	methylpyrazine	0.028599
	hydroxyanthranilic acid	0.028570
	avenin	0.028560
	xylotetraose	0.028508
	gramine	0.028455
Coffea arabica	chalcogran	0.039570
	benzene-1,2,4-triol	0.039566
	isoamyl formate	0.039563
	delta-tocopherol	0.039537
	pyrroles	0.039432
	N-methylpyridinium	0.039349
	D-ribosylnicotinic acid	0.039338
	melanoidins	0.039304
	cafestol	0.039072
	4-methylimidazole	0.038269



(a) Overrepresented molecular function of genes associated to *Brassica oleracea var. italica*



(b) Overrepresented molecular function of genes associated to *Avena sativa*



(c) Overrepresented molecular function of genes associated to *Coffea arabica*

Figure 18: Fold enrichment for molecular function of genes associated to *Brassica oleracea var. italica*, *Avena sativa*, and *Coffea arabica*

Table 9: Top 10 Phenotypes associated with *Brassica oleracea var. italica*, *Avena sativa*, and *Coffea arabica*.

Plant	**Disease**	**HetERel Score**
Brassica oleracea var. italica	oxidative stress	0.565160
	liver neoplasms	0.535610
	breast neoplasms	0.528981
	inflammation	0.521293
	infarction	0.517824
	lung neoplasms	0.517176
	pneumonia	0.514951
	heat stress	0.513441
	colorectal neoplasms	0.511281
	rheumatoid arthritis	0.510549
Avena sativa	oxidative stress	0.684840
	hyperglycemia	0.658581
	metabolic syndrome	0.658245
	inflammation	0.655189
	breast neoplasms	0.654367
	myocardial infarction	0.645413
	myocardial ischemia	0.643975
	hypertrophy	0.639481
	rheumatoid arthritis	0.638790
	obesity	0.638421
Coffea arabica	oxidative stress	0.672117
	brain damage	0.633576
	infarction	0.629755
	inflammation	0.623943
	poisoning	0.621713
	liver neoplasms	0.615741
	hyperglycemia	0.615291
	myocardial ischemia	0.614992
	metabolic syndrome	0.614617
	stroke	0.612817

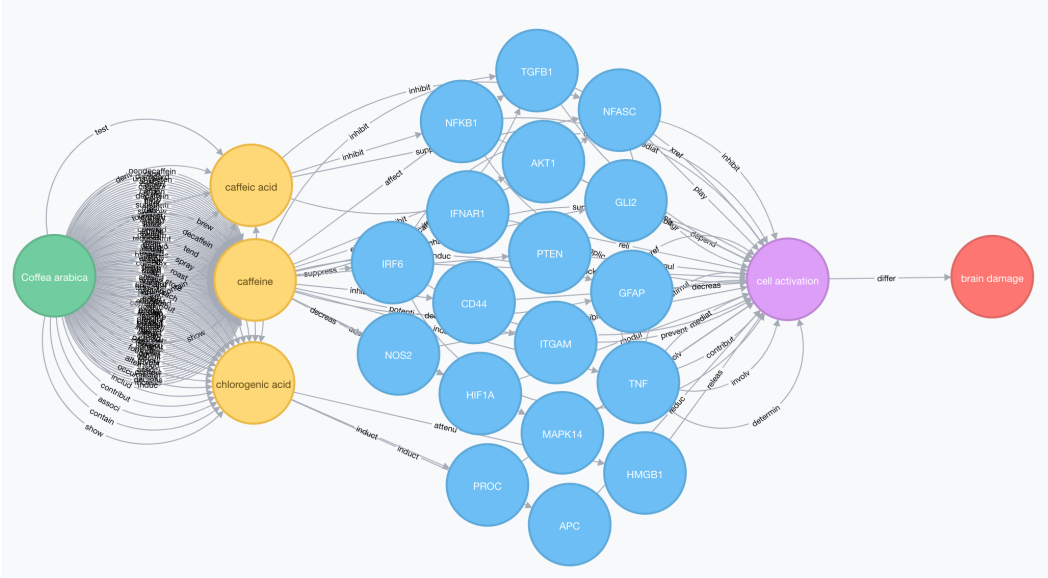
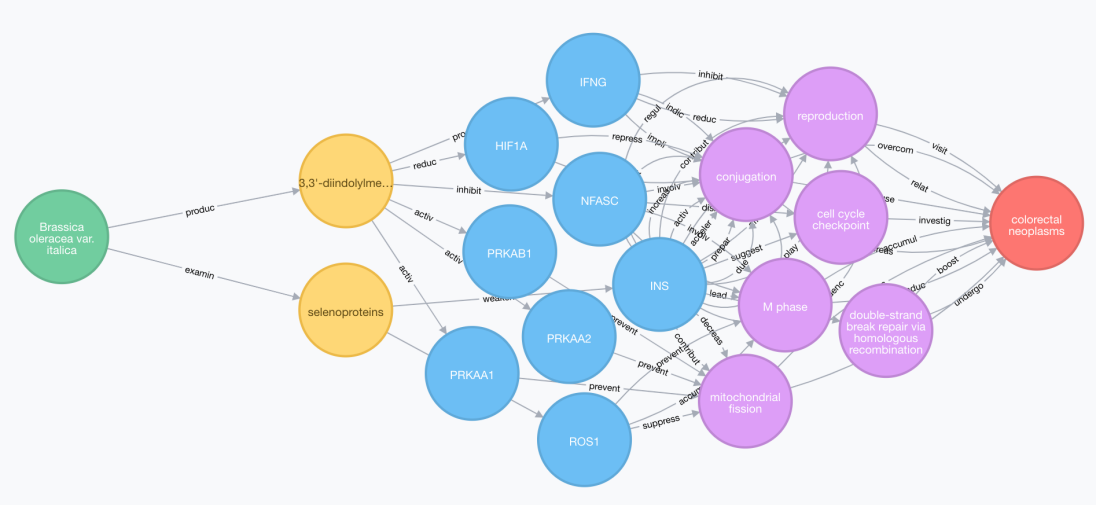
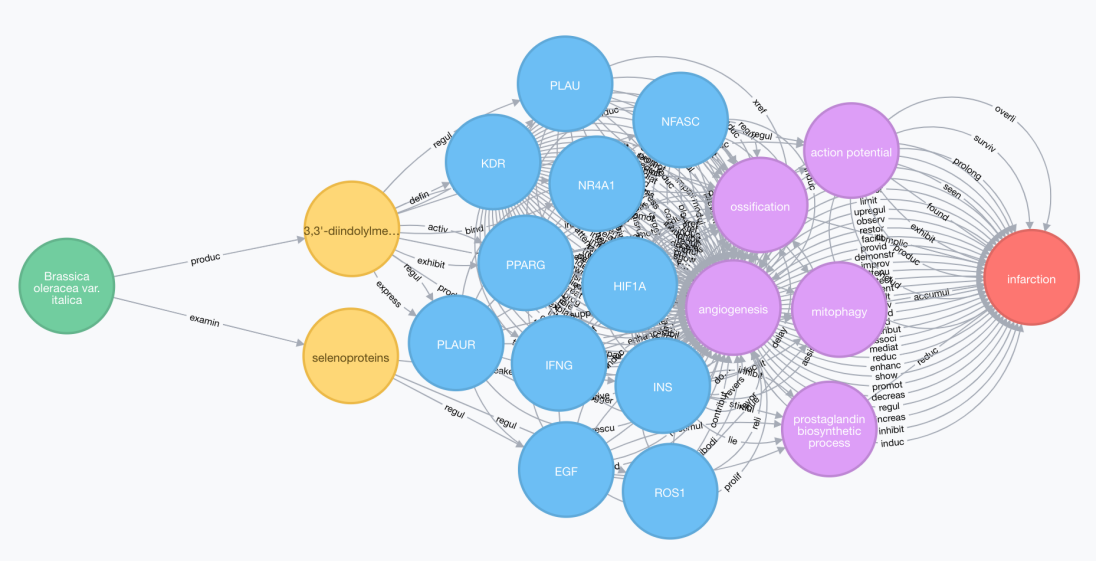


Figure 19: The Diet-Disease Network is capable of explaining inferred paths by providing access to the evidence for each link between objects. This visualization depicts the meta path (*P=E,C,G,M,D*) traversing the network from the plant *Coffea arabica*(green) to chemicals(yellow) to genes(blue) to pathways(purple) to phenotype(red).



(a) This visualization depicts the meta path (P=E,C,G,M,D). It generates multiple hypotheses explaining the relevance between broccoli and colorectal neoplasms.



(b) This visualization depicts the meta path (P=E,C,G,M,D). It generates multiple hypotheses explaining the relevance between broccoli and infarction.

Figure 20: Hypotheses can be generated when searching the diet-disease network with the HetERel meta path based relevance ranking framework.

CHAPTER 5: KNOWLEDGE BASED DISCOVERY THROUGH DATA MINING, INTEGRATION, AND SEMANTIC PRIORITIZATION

The advancements in high-throughput technologies has brought about an exponential increase in data for agricultural and biomedical research. The wealth of data continues to be studied, generating new knowledge and hypotheses. However, the rate of data collection has far outpaced current methods of data analysis. In addition, the data linking agricultural and biomedical research are distributed across disparate databases or hidden within the unstructured text of scientific literature. This necessitates the development of integrated knowledge bases which connect and encompass various research domains. The vast amount of data these knowledge bases amass requires an efficient framework capable of searching and determining the relevance of heterogeneous agricultural and biomedical concepts. The diet-disease network and meta path-based relevance ranking framework described in this work serve as a scalable foundation for linking distinct research domains and efficiently developing evidence-based insights from the current deluge of data.

5.1 Limitations and Considerations of the Diet-Disease Network and Relevance Ranking Method

The identification of limitations and considerations of a scientific work is important to address for the improvement of science and illuminate future avenues of research. In this work, the availability and access to data did impose certain limitations to

analysis.

Publicly available, curated association data connecting plants to chemicals was difficult to identify and procure, as opposed to the multitude of databases connecting chemicals, genes, pathways, and phenotypes. This gap in data was filled by the results of text mining the agricultural and biomedical literature. More recently, data resources, such as FooDB and PhenolExplorer, have become available that can augment the plant-chemical associations and extracted from text mining [146, 192].

Recently, studies in microbiome research have discovered the highly influential interaction between the gut microbiome, diet, and human health [32, 41]. This work does not analyze the effect of microbial communities in humans due to the minimal amount of microbiome data and knowledge available at the inception of this work, as well as the level of complexity it introduces. However, the modular and scalable design of the diet-disease network allows for the integration of microbial data, including bacterial taxonomy and chemical byproducts. The HetERel relevance ranking method is also data agnostic, meaning it is capable of analyzing new types of data, such as microbiome data.

The discussion of these limitations naturally leads into the various directions this work can lead to in the future. The considerations to the listed limitations provide an immediate goal of acquiring supplemental data to augment sparse data and refine the relevance metric.

5.2 Future Directions

In this work, an integrated network and relevance ranking framework were developed to investigate the interactions between dietary components and human health. The constant influx of new data and scientific publications presents an opportunity to expand the integrated network and apply relevance ranking to other diet-disease related domains of research.

As previously discussed, the gut microbiome has been found to play an impactful role in human health [32]. The integration of bacterial species, genes, and chemical byproducts into the diet-disease network would empower the development of hypotheses in the growing domain of microbiome research. Different biological inquiries could be investigated with the inclusion of microbiome data, such as the mitigated impact of diet on human health, based on the microbial composition of the human gut [7]. The variability of the human gut microbiome is attributed to various factors, such as a person's environment and their personal genetic makeup. The resulting hypotheses about the microbiome feed into other research fields, such as personalized nutrition.

The field of nutrigenomics combines nutrition and genome research to discover optimal diet and exercise plans based on an individual's genotypes and phenotypes [55]. The current integrated network and ranking framework could be extended to investigate the effects of dietary components on specific genotypes for personalized nutrition. Data from genetic variation databases and more general chemical databases, such as cosmic and ChEMBL, respectively, could be integrated into the network to provide insights into personalized nutrition [58, 88].

The development of the integrated diet-disease network and relevance ranking method enables researchers to explore new datasets and their connectivity with existing diet-disease knowledge in a scalable and efficient way.

5.3 Conclusion

The research described in this work addresses some of the recent issues brought about by big data faced by scientists across many research domains. The growing accumulation of agricultural and biomedical data and its distribution across numerous, disparate sources hampers knowledge-based discovery in the effects of diet on disease. This work provides a comprehensive analysis that leverages existing data and semantics to efficiently determine the most relevant results to explain the molecular paths of diet and disease interaction.

The diet-disease network is a foundational resource for investigating the mechanisms of action between dietary components and human health phenotypes. It is the first graph-based, heterogeneous information network to traverse the five distinct types of agricultural, chemical, and biological entities of diet and disease. The diet-disease network also incorporates text mined relationships as triples, including the predicates that define the association between entities. Text mined relationships extracted from two citation databases, PubMed and Agricola, provide overlap between agricultural and biomedical research.

The algorithmically simple weighting metric and meta path-based ranking method developed in this work discovers the molecular entities most relevant to describing the link between diet and disease. It takes into account the fine semantic detail in

relationships by calculating relevance with predicate probabilities. Simultaneously, the inclusion of predicates assigns evidence-based weight to the heterogeneous links within the diet-disease network. The evidence-based weights provide a more distinct measure of relevance compared to existing methods.

Together, the diet-disease network and novel meta path-based relevance ranking metric, HetERel, allow scientists to comprehensively investigate the associations proposed between diet and human health. The knowledge-based identification of diet and human health associations provides insight into the importance of the molecular effects of diet on disease prevention. An understanding of these molecular mechanisms can provide scientific support to filter the deluge of fictitious health claims commonly propagated in the media today.

REFERENCES

[1] R. Abilio, F. Morais, G. Vale, C. Oliveira, D. Pereira, and H. Costa. Applying information retrieval techniques to detect duplicates and to rank references in the preliminary phases of systematic: Literature reviews. *CLEI Electronic Journal*, 18(2):3–3, 2015.

[2] A. Acland, R. Agarwala, T. Barrett, J. Beck, D. A. Benson, C. Bollin, E. Bolton, S. H. Bryant, K. Canese, D. M. Church, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 42(D1):D7, 2014.

[3] C. B. Advantage. Sas® text miner.

[4] R. Agrawal, R. Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

[5] C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis. Literature mining, ontologies and information visualization for drug repurposing. *Briefings in bioinformatics*, 12(4):357–368, 2011.

[6] K. Anyanwu, A. Maduko, and A. Sheth. Semrank: ranking complex relationship search results on the semantic web. In *Proceedings of the 14th international conference on World Wide Web*, pages 117–127. ACM, 2005.

[7] J. Aron-Wisnewsky and K. Clément. The gut microbiome, diet, and links to cardiometabolic and chronic disorders. *Nature Reviews Nephrology*, 12(3):169, 2016.

[8] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[9] A. R. Aronson and F.-M. Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[10] 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, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[11] K. Audouze, S. Brunak, and P. Grandjean. A computational approach to chemical etiologies of diabetes. *Scientific reports*, 3:2712, 2013.

[12] S. Avraham, C.-W. Tung, K. Ilic, P. Jaiswal, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, et al. The plant ontology database: a community resource for plant structure and developmental stages controlled

vocabulary and annotations. *Nucleic acids research*, 36(suppl 1):D449–D454, 2008.

[13] N. C. Baker and B. M. Hemminger. Mining connections between chemicals, proteins, and diseases extracted from medline annotations. *Journal of biomedical informatics*, 43(4):510–519, 2010.

[14] J. Baldridge. The opennlp project. *URL: http://opennlp. apache. org/index. html,(accessed 2 February 2012)*, page 1, 2005.

[15] J. Bandy, D. Milward, and S. McQuay. Mining protein–protein interactions from published literature using linguamatics i2e. In *Protein Networks and Pathway Analysis*, pages 3–13. Springer, 2009.

[16] R. Barresi and K. P. Campbell. Dystroglycan: from biosynthesis to pathogenesis of human disease. *Journal of cell science*, 119(2):199–207, 2006.

[17] S. Batra and C. Tyagi. Comparative analysis of relational and graph databases. *International Journal of Soft Computing and Engineering (IJSCE)*, 2(2):509–512, 2012.

[18] J. Baumard, F. Osiurak, M. Lesourd, and D. Le Gall. Tool use disorders after left brain damage. *Frontiers in psychology*, 5:473, 2014.

[19] S. Beis, S. Papadopoulos, and Y. Kompatsiaris. Benchmarking graph databases on the problem of community detection. In *New Trends in Database and Information Systems II*, pages 3–14. Springer, 2015.

[20] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic acids research*, page gks1195, 2012.

[21] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[22] J. Blanchard, F. Guillet, and P. Kuntz. Semantics-based classification of rule interestingness measures, 2009.

[23] D. G. Brown, S. Rao, T. L. Weir, J. OMalia, M. Bazan, R. J. Brown, and E. P. Ryan. Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. *Cancer & metabolism*, 4(1):11, 2016.

[24] D. Cameron, O. Bodenreider, H. Yalamanchili, T. Danh, S. Vallabhaneni, K. Thirunarayan, A. P. Sheth, and T. C. Rindflesch. A graph-based recovery and decomposition of swansons hypothesis using semantic predications. *Journal of Biomedical Informatics*, 46(2):238–251, 2013.

[25] D. Campos, S. Matos, and J. L. Oliveira. A modular framework for biomedical concept recognition. *BMC bioinformatics*, 14(1):281, 2013.

[26] C. Cavin, D. Holzhaeuser, G. Scharf, A. Constable, W. Huber, and B. Schilter. Cafestol and kahweol, two coffee specific diterpenes with anticarcinogenic activity. *Food and chemical toxicology*, 40(8):1155–1163, 2002.

[27] B. Chen, Y. Ding, and D. J. Wild. Assessing drug target association using semantic linked data. *PLoS computational biology*, 8(7):e1002574, 2012.

[28] H. Chen and B. M. Sharp. Content-rich biological network constructed by mining pubmed abstracts. *BMC bioinformatics*, 5(1):147, 2004.

[29] X. Chen, L. J. Cheskin, L. Shi, and Y. Wang. Americans with diet-related chronic diseases report higher diet quality than those without these diseases. *The Journal of nutrition*, pages jn–111, 2011.

[30] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart. Polysearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic acids research*, 36(suppl 2):W399–W405, 2008.

[31] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo. How to train good word embeddings for biomedical nlp. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, 2016.

[32] I. Cho and M. J. Blaser. The human microbiome: at the interface of health and disease. *Nature Reviews Genetics*, 13(4):260, 2012.

[33] A. Choromanska, J. Kulbacka, J. Harasym, R. Oledzki, A. Szewczyk, and J. Saczko. High-and low-molecular weight oat beta-glucan reveals antitumor activity in human epithelial lung cancer. *Pathology & Oncology Research*, pages 1–10, 2017.

[34] K. B. Cohen and L. Hunter. Getting started in text mining. *PLoS computational biology*, 4(1):e20, 2008.

[35] F. S. Collins, E. D. Green, A. E. Guttmacher, and M. S. Guyer. A vision for the future of genomics research. *Nature*, 422(6934):835–847, 2003.

[36] F. S. Collins, M. Morgan, and A. Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, 2003.

[37] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to algorithms*. MIT press, 2009.

[38] G. Courtois and T. Gilmore. Mutations in the nf-κb signaling pathway: implications for human disease. *Oncogene*, 25(51):6831, 2006.

[39] S. C. Cunha, L. Senra, R. Cruz, S. Casal, and J. O. Fernandes. 4-methylimidazole in soluble coffee and coffee substitutes. *Food Control*, 63:15–20, 2016.

[40] B. V. da Silva, J. C. Barreira, and M. B. P. Oliveira. Natural phytochemicals and probiotics as bioactive ingredients for functional foods: Extraction, biochemistry and protected-delivery technologies. *Trends in Food Science & Technology*, 50:144–158, 2016.

[41] L. A. David, C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button, B. E. Wolfe, A. V. Ling, A. S. Devlin, Y. Varma, M. A. Fischbach, et al. Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, 505(7484):559, 2014.

[42] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly. The comparative toxicogenomics database's 10th year anniversary: update 2015. *Nucleic acids research*, 43(D1):D914–D920, 2015.

[43] R. De Virgilio, A. Maccioni, and R. Torlone. Converting relational to graph databases. In *First International Workshop on Graph Data Management Experiences and Systems*, page 1. ACM, 2013.

[44] J.-P. Derouette, C. Wong, L. Burnier, S. Morel, E. Sutter, K. Galan, A. C. Brisset, I. Roth, C. E. Chadjichristos, and B. R. Kwak. Molecular role of cx37 in advanced atherosclerosis: a micro-array study. *Atherosclerosis*, 206(1):69–76, 2009.

[45] M. Ding, S. N. Bhupathiraju, M. Chen, R. M. van Dam, and F. B. Hu. Caffeinated and decaffeinated coffee consumption and risk of type 2 diabetes: a systematic review and a dose-response meta-analysis. *Diabetes care*, 37(2):569–586, 2014.

[46] R. A. Dixon, D. R. Gang, A. J. Charlton, O. Fiehn, H. A. Kuiper, T. L. Reynolds, R. S. Tjeerdema, E. H. Jeffery, J. B. German, W. P. Ridley, et al. Applications of metabolomics in agriculture. *Journal of Agricultural and Food Chemistry*, 54(24):8984–8994, 2006.

[47] D. Echeverri, F. R. Montes, M. Cabrera, A. Galán, and A. Prieto. Caffeine's vascular mechanisms of action. *International journal of vascular medicine*, 2010, 2010.

[48] G. B. Ehret, P. B. Munroe, K. M. Rice, M. Bochud, A. D. Johnson, D. I. Chasman, A. V. Smith, M. D. Tobin, G. C. Verwoert, S.-J. Hwang, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103, 2011.

[49] D. El Khoury, C. Cuda, B. Luhovyy, and G. Anderson. Beta glucan: health benefits in obesity and metabolic syndrome. *Journal of nutrition and metabolism*, 2012, 2011.

[50] F. Endel and H. Piringer. Data wrangling: Making data useful again. *IFAC-PapersOnLine*, 48(1):111–112, 2015.

[51] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, R. Haw, B. Jassal, F. Korninger, B. May, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 46(D1):D649–D655, 2017.

[52] S. Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2011.

[53] R. Feldman and J. Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge University Press, 2007.

[54] M. Fenech, A. El-Sohemy, L. Cahill, L. R. Ferguson, T.-A. French, E. S. Tai, J. Milner, W.-P. Koh, L. Xie, M. Zucker, et al. Nutrigenetics and nutrigenomics: viewpoints on the current status and applications in nutrition research and practice. *Journal of nutrigenetics and nutrigenomics*, 4(2):69–89, 2011.

[55] J. F. Ferguson, H. Allayee, R. E. Gerszten, F. Ideraabdullah, P. M. Kris-Etherton, J. M. Ordovás, E. B. Rimm, T. J. Wang, and B. J. Bennett. Nutrigenomics, the microbiome, and gene-environment interactions: new directions in cardiovascular disease research, prevention, and treatment: a scientific statement from the american heart association. *Circulation: Genomic and Precision Medicine*, 9(3):291–313, 2016.

[56] W. W. Fleuren and W. Alkema. Application of text mining in the biomedical domain. *Methods*, 74:97–106, 2015.

[57] I. Flight and P. Clifton. Cereal grains and legumes in the prevention of coronary heart disease and stroke: a review of the literature. *European journal of clinical nutrition*, 60(10):1145, 2006.

[58] N. C. for Biotechnology Information. Database of single nucleotide polymorphisms (dbsnp). 2015.

[59] N. C. for Biotechnology Information. Medline, 2015.

[60] N. C. for Biotechnology Information. Pubmed, 2015.

[61] K.-i. Fukuda, T. Tsunoda, A. Tamura, T. Takagi, et al. Toward information extraction: identifying protein names from biological papers. In *Pac symp biocomput*, volume 707, pages 707–718, 1998.

[62] L. Getoor and C. P. Diehl. Link mining: a survey. *Acm Sigkdd Explorations Newsletter*, 7(2):3–12, 2005.

[63] G. Goldenberg, K. Hartmann, and I. Schlott. Defective pantomime of object use in left brain damage: apraxia or asymbolia? *Neuropsychologia*, 41(12):1565–1573, 2003.

[64] R. Gramatica, T. Di Matteo, S. Giorgetti, M. Barbiani, D. Bevec, and T. Aste. Graph theory enables drug repurposing–how a mathematical model can drive the discovery of hidden mechanisms of action. *PloS one*, 9(1):e84912, 2014.

[65] M. Graves, E. R. Bergeman, and C. B. Lawrence. Querying a genome database using graphs. In *Proceedings of the 3th International Conference on Bioinformatics and Genome Research*, 1994.

[66] G. Grosso, A. Micek, J. Godos, S. Sciacca, A. Pajak, M. A. Martínez-González, E. L. Giovannucci, and F. Galvano. Coffee consumption and risk of all-cause, cardiovascular, and cancer mortality in smokers and non-smokers: A dose-response meta-analysis, 2016.

[67] K. A. Guertin, S. C. Moore, J. N. Sampson, W.-Y. Huang, Q. Xiao, R. Z. Stolzenberg-Solomon, R. Sinha, and A. J. Cross. Metabolomics in nutritional epidemiology: identifying metabolites associated with diet and quantifying their potential to uncover diet-disease relations in populations–. *The American journal of clinical nutrition*, 100(1):208–217, 2014.

[68] M. S. Habib and J. Kalita. Scalable biomedical named entity recognition: investigation of a database-supported svm approach. *International journal of bioinformatics research and applications*, 6(2):191–208, 2010.

[69] U. Hahn, K. B. Cohen, Y. Garten, and N. H. Shah. Mining the pharmacogenomics literaturea survey of the state of the art. *Briefings in bioinformatics*, 13(4):460–494, 2012.

[70] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, 33(suppl 1):D514–D517, 2005.

[71] M. Y. Hardy, J. A. Tye-Din, J. A. Stewart, F. Schmitz, N. L. Dudek, I. Hanchapola, A. W. Purcell, and R. P. Anderson. Ingestion of oats and barley in patients with celiac disease mobilizes cross-reactive t cells activated by avenin peptides and immuno-dominant hordein peptides. *Journal of autoimmunity*, 56:56–65, 2015.

[72] J. Hastings, P. de Matos, A. Dekker, M. Ennis, B. Harsha, N. Kale, V. Muthukrishnan, G. Owen, S. Turner, M. Williams, et al. The chebi reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic acids research*, 41(D1):D456–D463, 2013.

[73] S. Haustein, I. Peters, C. R. Sugimoto, M. Thelwall, and V. Larivière. Tweeting biomedicine: An analysis of tweets and citations in the biomedical literature. *Journal of the Association for Information Science and Technology*, 65(4):656–669, 2014.

[74] J. D. Hayes, M. O. Kelleher, and I. M. Eggleston. The cancer chemopreventive actions of phytochemicals derived from glucosinolates. *European journal of nutrition*, 47(2):73–88, 2008.

[75] D. S. Himmelstein and S. E. Baranzini. Heterogeneous network edge prediction: a data integration approach to prioritize disease-associated genes. *PLoS computational biology*, 11(7):e1004259, 2015.

[76] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, A. Green, P. Khankhanian, and S. E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, 6, 2017.

[77] A. Holzinger, J. Schantl, M. Schroettner, C. Seifert, and K. Verspoor. Biomedical text mining: State-of-the-art, open problems and future challenges. In A. Holzinger and I. Jurisica, editors, *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, chapter 16, pages 271–300. 2014.

[78] D. Hristovski, C. Friedman, T. C. Rindflesch, and B. Peterlin. Exploiting semantic relations for literature-based discovery. In *AMIA annual symposium proceedings*, volume 2006, page 349. American Medical Informatics Association, 2006.

[79] Z. Huang, Y. Zheng, R. Cheng, Y. Sun, N. Mamoulis, and X. Li. Meta structure: Computing relevance in large heterogeneous information networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1595–1604. ACM, 2016.

[80] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[81] K. Jensen, G. Panagiotou, and I. Kouskoumvekaki. Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS computational biology*, 10(1):e1003432, 2014.

[82] 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*, page gku724, 2014.

[83] J. J. Jiang and D. W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*, 1997.

[84] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

[85] A. Kao and S. R. Poteet. *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[86] Y. H. Kim, S. H. Beak, A. Charidimou, and M. Song. Discovering new genes in the pathways of common sporadic neurodegenerative diseases: a bioinformatics approach. *Journal of Alzheimer's Disease*, 51(1):293–312, 2016.

[87] B. L. King, A. P. Davis, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly. Ranking transitive chemical-disease inferences using local network topology in the comparative toxicogenomics database. *PloS one*, 7(11):e46524, 2012.

[88] E. M. B. Laboratory. The chembl home page, 2015.

[89] E. M. B. Laboratory. The european bioinformatics institute home page, 2015.

[90] N. Lao and W. W. Cohen. Fast query execution for retrieval models based on path-constrained random walks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 881–888. ACM, 2010.

[91] N. Lao and W. W. Cohen. Relational retrieval using a combination of path-constrained random walks. *Machine learning*, 81(1):53–67, 2010.

[92] R. R. Larson. Introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 61(4):852–853, 2010.

[93] A. Lasry, A. Zinger, and Y. Ben-Neriah. Inflammatory networks underlying colorectal cancer. *Nature immunology*, 17(3):230, 2016.

[94] R. E. Lee, M. A. Qasaimeh, X. Xia, D. Juncker, and S. Gaudet. Nf-κb signalling and cell fate decisions in response to a short pulse of tumour necrosis factor. *Scientific reports*, 6:39519, 2016.

[95] M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.

[96] U. Leser and J. Hakenberg. What makes a gene name? named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4):357–369, 2005.

[97] Y. Li, C. Shi, P. S. Yu, and Q. Chen. Hrank: A path based ranking framework in heterogeneous information network. *arXiv preprint arXiv:1403.7315*, 2014.

[98] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[99] D. Lin et al. An information-theoretic definition of similarity. In *Icml*, volume 98, pages 296–304. Citeseer, 1998.

[100] C. E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[101] H. Liu and M. Rastegar-Mojarad. Literature-based knowledge discovery. In S. Q. Ye, editor, *Big Data Analysis for Bioinformatics and Biomedical Discoveries*, chapter 14, pages 233–248. CRC Press, 2016.

[102] L. Liu, X. Dai, H. Wang, W. Song, and J. Lu. A weighted multipath measurement based on gene ontology for estimating gene products similarity. *Journal of Computational Biology*, 21(12):964–974, 2014.

[103] V. Liu, M. P. Clark, M. Mendoza, R. Saket, M. N. Gardner, B. J. Turk, and G. J. Escobar. Automated identification of pneumonia in chest radiograph reports in critically ill patients. *BMC medical informatics and decision making*, 13(1):90, 2013.

[104] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[105] P. Louis, G. L. Hold, and H. J. Flint. The gut microbiota, bacterial metabolites and colorectal cancer. *Nature Reviews Microbiology*, 12(10):661, 2014.

[106] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic acids research*, 39(suppl 1):D52–D57, 2011.

[107] J. M. McCracken and L.-A. H. Allen. Regulation of human neutrophil apoptosis and lifespan in health and disease. *Journal of cell death*, 7:JCD–S11038, 2014.

[108] J. R. Mein, D. R. James, S. Lakkanna, et al. Induction of phase 2 antioxidant enzymes by broccoli sulforaphane: perspectives in maintaining the antioxidant activity of vitamins a, c, and e. *Frontiers in genetics*, 3:7, 2012.

[109] J. Menche, A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601, 2015.

[110] X. Meng, C. Shi, Y. Li, L. Zhang, and B. Wu. Relevance measure in large-scale heterogeneous networks. In *Asia-Pacific Web Conference*, pages 636–643. Springer, 2014.

[111] D. Meyer, K. Hornik, and I. Feinerer. Text mining infrastructure in r. *Journal of statistical software*, 25(5):1–54, 2008.

[112] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas. Large-scale gene function analysis with the panther classification system. *Nature protocols*, 8(8):1551, 2013.

[113] C. M. Miller, T. C. Rindflesch, M. Fiszman, D. Hristovski, D. Shin, G. Rosemblat, H. Zhang, and K. P. Strohl. A closed literature-based discovery technique finds a mechanistic link between hypogonadism and diminished sleep quality in aging men. *Sleep*, 35(2):279, 2012.

[114] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.

[115] D. Milward, M. Bjäreland, W. Hayes, M. Maxwell, L. Öberg, N. Tilford, J. Thomas, R. Hale, S. Knight, and J. Barnes. Ontology-based interactive information extraction from scientific abstracts. *Comparative and functional genomics*, 6(1-2):67–71, 2005.

[116] B. H. Monien, K. Herrmann, S. Florian, and H. Glatt. Metabolic activation of furfuryl alcohol: formation of 2-methylfuranyl dna adducts in salmonella typhimurium strains expressing human sulfotransferase 1a1 and in fvb/n mice. *Carcinogenesis*, 32(10):1533–1539, 2011.

[117] M. Muller and S. Kersten. Nutrigenomics: goals and strategies. *Nature Reviews Genetics*, 4(4):315, 2003.

[118] V. Neeha and P. Kinth. Nutrigenomics research: a review. *Journal of food science and technology*, 50(3):415–428, 2013.

[119] Neo4j. Neo4j home page, 2015.

[120] I. Neo4j. Awesome procedures on cypher, 2018.

[121] I. Neo4j. opencypher, 2018.

[122] T. Nguyen, P. Nioi, and C. B. Pickett. The nrf2-antioxidant response element signaling pathway and its activation by oxidative stress. *Journal of Biological Chemistry*, 284(20):13291–13295, 2009.

[123] U. S. D. of Agriculture. Agricultural online access, 2015.

[124] U. S. D. of Agriculture. National agricultural library, 2015.

[125] U. S. D. of Agriculture. National agricultural library thesaurus and glossary, 2015.

[126] E. R. Omiecinski. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(1):57–69, 2003.

[127] J. Ovesná, O. Slabỳ, O. Toussaint, M. Kodíček, P. Maršík, V. Pouchová, and T. Vaněk. High throughput omics approaches to assess the effects of phytochemicals in human health studies. *British Journal of Nutrition*, 99(E-S1):ES127–ES134, 2008.

[128] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[129] R. Paul, T. Groza, J. Hunter, and A. Zankl. Semantic interestingness measures for discovering association rules in the skeletal dysplasia domain. *Journal of biomedical semantics*, 5(1):8, 2014.

[130] G. A. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, P. G. Bagos, et al. Using graph theory to analyze biological networks. *BioData mining*, 4(1):10, 2011.

[131] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto. Metrics for go based protein semantic similarity: a systematic evaluation. In *BMC bioinformatics*, volume 9, page S4. BioMed Central, 2008.

[132] C. Pesquita, D. Faria, A. O. Falcao, P. Lord, and F. M. Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[133] L. A. Peterson. Electrophilic intermediates produced by bioactivation of furan. *Drug metabolism reviews*, 38(4):615–626, 2006.

[134] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.

[135] S. Pletscher-Frankild, A. Palleja, K. Tsafou, J. X. Binder, and L. J. Jensen. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89, 2015.

[136] R. Poole, O. J. Kennedy, P. Roderick, J. A. Fallowfield, P. C. Hayes, and J. Parkes. Coffee consumption and health: umbrella review of meta-analyses of multiple health outcomes. *bmj*, 359:j5024, 2017.

[137] S. Ramaswamy, S. Mahajan, and A. Silberschatz. On the discovery of interesting patterns in association rules. In *VLDB*, volume 98, pages 368–379. Citeseer, 1998.

[138] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch, and A. Jimeno. Text processing through web services: calling whatizit. *Bioinformatics*, 24(2):296–298, 2007.

[139] D. Rebholz-Schuhmann, A. Oellrich, and R. Hoehndorf. Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, 13(12):829, 2012.

[140] D. Rebholz-Schuhmann, A. J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, et al. Assessment of ner solutions against the first and second calbc silver standard corpus. *Journal of biomedical semantics*, 2(5):S11, 2011.

[141] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[142] F. Riaz and K. M. Ali. Applications of graph theory in computer science. In *Computational Intelligence, Communication Systems and Networks (CICSyN), 2011 Third International Conference on*, pages 142–145. IEEE, 2011.

[143] T. C. Rindflesch and M. Fiszman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of biomedical informatics*, 36(6):462–477, 2003.

[144] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12(1):77, 2011.

[145] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615, 2008.

[146] J. A. Rothwell, J. Perez-Jimenez, V. Neveu, A. Medina-Remon, N. M'Hiri, P. García-Lobato, C. Manach, C. Knox, R. Eisner, D. S. Wishart, et al. 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*, 2013, 2013.

[147] R. Sætre, A. Tveit, T. S. Steigedal, and A. Lægreid. Semantic annotation of biomedical literature using google. In *International Conference on Computational Science and Its Applications*, pages 327–337. Springer, 2005.

[148] N. Sales, P. Pelegrini, and M. Goersch. Nutrigenomics: definitions and advances of this new science. *Journal of nutrition and metabolism*, 2014, 2014.

[149] O. Salman, I. H. Elhajj, A. Kayssi, and A. Chehab. Sdn controllers: A comparative study. In *Electrotechnical Conference (MELECON), 2016 18th Mediterranean*, pages 1–6. IEEE, 2016.

[150] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[151] Y. J. Sayers EW, Barrett T. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 37(D):5–15, 2009.

[152] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[153] Y. Sebastian and P. H. Then. Domain-driven kdd for mining functionally novel rules and linking disjoint medical hypotheses. *Knowledge-Based Systems*, 24(5):609–620, 2011.

[154] P. Sharma, R. Senthilkumar, V. Brahmachari, E. Sundaramoorthy, A. Mahajan, A. Sharma, and S. Sengupta. Mining literature for a comprehensive pathway analysis: a case study for retrieval of homocysteine related genes for genetic and epigenetic studies. *Lipids Health Dis*, 5(1):1–19, 2006.

[155] T. Shen, J. Lee, E. Lee, S. H. Kim, T. W. Kim, and J. Y. Cho. Cafestol, a coffee-specific diterpene, is a novel extracellular signal-regulated kinase inhibitor with ap-1-targeted inhibition of prostaglandin e2 production in lipopolysaccharide-activated macrophages. *Biological and Pharmaceutical Bulletin*, 33(1):128–132, 2010.

[156] C. Shi, X. Kong, Y. Huang, S. Y. Philip, and B. Wu. Hetesim: A general framework for relevance measure in heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, 26(10):2479–2492, 2014.

[157] C. Shi, Y. Li, J. Zhang, Y. Sun, and S. Y. Philip. A survey of heterogeneous information network analysis. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):17–37, 2015.

[158] R. Singh, S. De, and A. Belkheir. Avena sativa (oat), a potential neutraceutical and therapeutic agent: an overview. *Critical reviews in food science and nutrition*, 53(2):126–144, 2013.

[159] T. Sivapalan, A. Melchini, M. Traka, S. Saha, and R. Mithen. Investigating the bioavailability of phytochemicals and minerals from broccoli soups. *Proceedings of the Nutrition Society*, 74(OCE3), 2015.

[160] R. Sladek, G. Rocheleau, J. Rung, C. Dina, L. Shen, D. Serre, P. Boutin, D. Vincent, A. Belisle, S. Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881, 2007.

[161] N. R. Smalheiser and D. R. Swanson. Indomethacin and alzheimer's disease. *Neurology*, 46(2):583–583, 1996.

[162] N. R. Smalheiser and D. R. Swanson. Linking estrogen to alzheimer's disease an informatics approach. *Neurology*, 47(3):809–810, 1996.

[163] N. R. Smalheiser and D. R. Swanson. Calcium-independent phospholipase a2 and schizophrenia. *Archives of General Psychiatry*, 55(8):752–753, 1998.

[164] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

[165] C. L. Smith, C.-A. W. Goldsmith, and J. T. Eppig. The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome biology*, 6(1):R7, 2004.

[166] N. T. Strande, E. R. Riggs, A. H. Buchanan, O. Ceyhan-Birsoy, M. DiStefano, S. S. Dwight, J. Goldstein, R. Ghosh, B. A. Seifert, T. P. Sneddon, et al. Evaluating the clinical validity of gene-disease associations: an evidence-based framework developed by the clinical genome resource. *The American Journal of Human Genetics*, 100(6):895–906, 2017.

[167] Y. Sun, R. Barber, M. Gupta, C. C. Aggarwal, and J. Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 121–128. IEEE, 2011.

[168] Y. Sun and J. Han. Mining heterogeneous information networks: principles and methodologies. *Synthesis Lectures on Data Mining and Knowledge Discovery*, 3(2):1–159, 2012.

[169] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003, 2011.

[170] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu. Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 565–576. ACM, 2009.

[171] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):11, 2013.

[172] R. Sur, A. Nigam, D. Grote, F. Liebel, and M. D. Southall. Avenanthramides, polyphenols from oats, exhibit anti-inflammatory and anti-itch activity. *Archives of Dermatological Research*, 300(10):569, 2008.

[173] N. Swainston, R. Batista-Navarro, P. Carbonell, P. D. Dobson, M. Dunstan, A. J. Jervis, M. Vinaixa, A. R. Williams, S. Ananiadou, J.-L. Faulon, et al. biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PLoS One*, 12(7):e0179130, 2017.

[174] D. R. Swanson. Fish oil, raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

[175] D. R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in biology and medicine*, 31(4):526–557, 1988.

[176] D. R. Swanson. Somatomedin c and arginine: implicit connections between mutually isolated literatures. *Perspectives in biology and medicine*, 33(2):157–186, 1990.

[177] P.-N. Tan et al. *Introduction to data mining*. Pearson Education India, 2006.

[178] P.-N. Tan, V. Kumar, and J. Srivastava. Selecting the right interestingness measure for association patterns. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41. ACM, 2002.

[179] C. Tew, C. Giraud-Carrier, K. Tanner, and S. Burton. Behavior-based clustering and analysis of interestingness measures for association rule mining. *Data Mining and Knowledge Discovery*, 28(4):1004–1045, 2014.

[180] C.-W. Tung. Chemdis: a chemical–disease inference system based on chemical–protein interactions. *Journal of Cheminformatics*, 7(1):25, 2015.

[181] J. E. Tym, C. Mitsopoulos, E. A. Coker, P. Razaz, A. C. Schierz, A. A. Antolin, and B. Al-Lazikani. cansar: an updated cancer research and drug discovery knowledgebase. *Nucleic acids research*, 44(D1):D938–D943, 2015.

[182] A. R. S. U.S. Department of Agriculture. Usda national nutrient database for standard reference, release 27, 2014.

[183] C. Vicknair, M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins. A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference*, page 42. ACM, 2010.

[184] B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, C. Dina, R. P. Welch, E. Zeggini, C. Huth, Y. S. Aulchenko, G. Thorleifsson, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nature genetics*, 42(7):579, 2010.

[185] A. E. Wagner, A. M. Terschluesen, and G. Rimbach. Health promoting effects of brassica-derived phytochemicals: from chemopreventive and anti-inflammatory activities to epigenetic regulation. *Oxidative medicine and cellular longevity*, 2013, 2013.

[186] P. Wang, B. Xu, Y. Wu, and X. Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[187] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, et al. The genemania prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic acids research*, 38(suppl_2):W214–W220, 2010.

[188] J. Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity*, pages 217–218. ACM, 2012.

[189] M. Weeber, J. A. Kors, and B. Mons. Online tools to support literature-based discovery in the life sciences. *Briefings in bioinformatics*, 6(3):277–286, 2005.

[190] F. Wild, D. Rstem, and M. F. Wild. The lsa package, 2009.

[191] B. Wilkowski, M. Fiszman, C. M. Miller, D. Hristovski, S. Arabandi, G. Rosemblat, and T. C. Rindflesch. Graph-based methods for discovery browsing with semantic predications. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1514. American Medical Informatics Association, 2011.

[192] D. Wishart. Foodb. *The Metabolomics Innovation Centre*, 469:470, 2016.

[193] L. Wu, M. H. N. Ashraf, M. Facci, R. Wang, P. G. Paterson, A. Ferrie, and B. H. Juurlink. Dietary approach to attenuate oxidative stress, hypertension, and inflammation in the cardiovascular system. *Proceedings of the National Academy of Sciences*, 101(18):7094–7099, 2004.

[194] T. Wu, Y. Chen, and J. Han. Association mining in large databases: A re-examination of its measures. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 621–628. Springer, 2007.

[195] T. Wu, Y. Chen, and J. Han. Re-examination of interestingness measures in pattern mining: a unified framework. *Data Mining and Knowledge Discovery*, 21(3):371–397, 2010.

[196] Y. Wu, Z. Yang, J. How, H. Xu, L. Chen, and K. Li. Overexpression of a peroxidase gene (atprx64) of arabidopsis thaliana in tobacco improves plants tolerance to aluminum stress. *Plant molecular biology*, 95(1-2):157–168, 2017.

[197] Y. Xiao, J. Zhang, and L. Deng. Prediction of lncrna-protein interactions using hetesim scores based on heterogeneous networks. *Scientific reports*, 7(1):3664, 2017.

[198] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, and M. Ishizuka. Unsupervised relation extraction by mining wikipedia texts using information from the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1021–1029. Association for Computational Linguistics, 2009.

[199] H. Zhang and R. Tsao. Dietary polyphenols, oxidative stress and antioxidant and anti-inflammatory effects. *Current Opinion in Food Science*, 8:33–42, 2016.

[200] N. Zhong, Y. Li, and S.-T. Wu. Effective pattern discovery for text mining. *Knowledge and Data Engineering, IEEE Transactions on*, 24(1):30–44, 2012.

[201] C. Zielke, O. Kosik, M.-L. Ainalem, A. Lovegrove, A. Stradner, and L. Nilsson. Characterization of cereal β-glucan extracts from oat and barley and quantification of proteinaceous matter. *PloS one*, 12(2):e0172034, 2017.

[202] M. Žitnik, E. A. Nam, C. Dinh, A. Kuspa, G. Shaulsky, and B. Zupan. Gene prioritization by compressive data fusion and chaining. *PLoS computational biology*, 11(10):e1004552, 2015.

[203] A. Zwyghuizen-Doorenbos, T. A. Roehrs, L. Lipschutz, V. Timms, and T. Roth. Effects of caffeine on alertness. *Psychopharmacology*, 100(1):36–39, 1990.

APPENDIX A: CODE FOR ETL AND RANKING

The code for data extraction, transformation, and loading into the Neo4j graph database consists of over 2,000 lines of code in Python, bash, and Cypher. The entirety of code for ranking was written in Python and Cypher, with over 3,000 lines of code. Full source code, documentation, and usage instructions is available at `https://github.com/rlinchangco/Dissertation_PhD`.