A Context-Driven Subgraph Model for Literature-Based Discovery

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

by

Delroy H. L. Cameron
M.S., University of Georgia, 2007
B.S., Savannah State University, 2005

2014
Wright State University
I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Delroy H. L. Cameron ENTITLED A Context-Driven Subgraph Model for Literature-Based Discovery BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Amit P. Sheth, Ph.D.
Dissertation Director

Arthur A. Goshtasby, Ph.D.
Director, Department of Computer Science and Engineering

Robert E. W. Fyffe, Ph.D.
Vice President for Research and Dean of the Graduate School

Committee on Final Examination

Amit P. Sheth, Ph.D.

Krishnaprasad Thirunarayan, Ph.D.

Michael Raymer, Ph.D.

Ramakanth Kavuluru, Ph.D.

Thomas C. Rindflesch, Ph.D.

Varun Bhagwan, M.S.
ABSTRACT

Cameron, Delroy. Ph.D., Department of Computer Science and Engineering, Wright State University, 2014. A Context-Driven Subgraph Model for Literature-Based Discovery.

Literature-Based Discovery (LBD) refers to the process of uncovering hidden connections that are implicit in scientific literature. Numerous hypotheses have been generated from scientific literature using the LBD paradigm, which influenced innovations in diagnosis, treatment, preventions, and overall public health. However, much of the existing research on discovering hidden connections among concepts have used distributional statistics and graph-theoretic measures to capture implicit associations. Such metrics do not explicitly capture the semantics of hidden connections. Rather, they only allude to the existence of meaningful underlying associations. To gain in-depth insights into the meaning of hidden (and other) connections, complementary methods have often been employed. Some of these methods include: 1) the use of domain expertise for concept filtering and knowledge exploration, 2) leveraging structured background knowledge for context and to supplement concept filtering, and 3) developing heuristics \textit{a priori} to help eliminate spurious connections.

While effective in some situations, the practice of relying on domain expertise, structured background knowledge and heuristics to complement distributional and graph-theoretic approaches, has serious limitations. The main issue is that the intricate context of complex associations is not always known \textit{a priori} and cannot easily be computed without understanding the underlying semantics of associations. Complex associations should not be overlooked, since they are often needed to elucidate the mechanisms of interaction and causality relationships among concepts. Moreover, they can capture the broader aspects of a biomedical sub-domain by segregating associations along different thematic dimensions, such as \textit{Metabolic Function}, \textit{Pharmaceutical Treatment}, and \textit{Neurological Activity}.

This dissertation proposes an innovative context-driven, automatic subgraph creation method for finding hidden and complex associations among concepts, along multiple the-
matic dimensions. It outlines definitions for context and shared context, based on implicit and explicit (or formal) semantics, which compensate for deficiencies in statistical and graph-based metrics. It virtually eliminates the need for heuristics a priori. An evidence-based evaluation of the proposed framework showed that 8 out of 9 existing scientific discoveries could be recovered using this approach. Additionally, insights into the meaning of associations could be substantiated using provenance provided by the system. In a statistical evaluation to determine the interestingness of the generated subgraphs, it was observed that an arbitrary association is mentioned in only approximately 4 articles in MEDLINE, on average. These results suggest that leveraging implicit and explicit context, as defined in this dissertation, is a significant advancement of the state-of-the-art in LBD research.
## Contents

1 Introduction ................................................................. 1
   1.1 Motivation ............................................................. 1
      1.1.1 Mendelian Laws of Inheritance ............................. 3
      1.1.2 Chromosome Theory of Inheritance ....................... 4
      1.1.3 Dyerberg–Bang Hypothesis ................................. 5
   1.2 The Science of Making Discoveries ............................... 5

2 Overview ........................................................................ 9
   2.1 Biomedical Text Mining .............................................. 10
      2.1.1 Information Retrieval ........................................ 10
      2.1.2 Question Answering .......................................... 11
      2.1.3 Document Summarization .................................... 12
      2.1.4 Literature-Based Discovery ................................. 13
   2.2 Biomedical Knowledge Sources .................................... 15
      2.2.1 Assertional Knowledge ....................................... 15
      2.2.2 Definitional Knowledge ...................................... 18
   2.3 Literature-Based Discovery Research ............................. 20
      2.3.1 Paradigms and Modes ........................................ 22
      2.3.2 Methodologies and Limitations ............................ 25
   2.4 Contributions ............................................................ 39

3 Context-Driven Subgraph Model ...................................... 43
   3.1 Query Specification and Document Selection .................. 44
   3.2 Predications Graph Definition .................................... 45
   3.3 Candidate Graph Generation ..................................... 46
   3.4 Path Clustering ........................................................ 48
   3.5 Application to Rediscovery and Decomposition ................ 50
      3.5.1 Literature Selection and Preprocessing .................. 50
      3.5.2 Predications and Candidate Graph Creation ............. 51
      3.5.3 Subgraph Creation ............................................ 52
      3.5.4 Experimental Results ........................................ 54
   3.6 Summary ................................................................. 64
4 Automatic Subgraph Creation
4.1 Path Relatedness .......................................................... 68
4.2 Predication Context ...................................................... 70
4.3 Shared Context ............................................................... 74
  4.3.1 Semantics-enhanced shared context ...................... 80
4.4 Hierarchical Agglomerative Clustering ............................. 82
  4.4.1 Threshold Selection .............................................. 82
  4.4.2 Combination Similarity .................................... 85
  4.4.3 Subgraph Ranking .............................................. 87
  4.4.4 Clustering Algorithm .................................... 89
4.5 Summary ................................................................. 91

5 Evaluation ................................................................. 93
5.1 Evidence-based .......................................................... 94
  5.1.1 Raynaud Syndrome – Dietary Fish Oils .............. 95
  5.1.2 Magnesium – Migraine .................................. 101
  5.1.3 Somatomedin C – Arginine .......................... 106
  5.1.4 Indomethacin – Alzheimer’s Disease ................ 109
  5.1.5 Estrogen – Alzheimer’s Disease ...................... 113
  5.1.6 Calcium-Independent Phospholipase A2 – Schizophrenia 115
  5.1.7 Chlorpromazine – Cardiac Hypertrophy ............ 117
  5.1.8 Testosterone – Sleep .................................... 119
  5.1.9 Diethylethyl Phthalate (DEHP) – Sepsis ............ 120
5.2 Statistical .............................................................. 121
5.3 Summary .............................................................. 124

6 Biomedical Knowledge Exploration ..................................... 126
6.1 The Search-and-Sift Paradigm .................................. 126
6.2 Ontology-Driven Text Exploration ................................. 127
6.3 Semantic Predications-Based Text Exploration ............... 129
6.4 Discovery Browsing .................................................. 131
6.5 Summary .............................................................. 132

7 Future Directions ......................................................... 133

8 Discussion ................................................................. 137

Bibliography ............................................................... 139

A Supplementary Materials .................................................. 153
  A.1 Rediscovery and Decomposition – Experiment II .......... 153
    A.1.1 Platelet Aggregation ................................ 154
    A.1.2 Blood Viscosity ....................................... 155
    A.1.3 Vascular Reactivity .................................. 156

B Tools and Applications .................................................. 157
List of Figures

1.1 Aspects of the Science of Making Discoveries ........................................ 6
2.1 Venn Diagram showing the overall idea of the ABC model .................. 23
2.2 Open Discovery Mode. Solid arrows indicate potentially interesting associations which may lead to discoveries. Dashed arrows indicate uninteresting connections ................................................................. 23
2.3 Closed Discovery Mode. Solid arrows indicate potentially interesting associations which may lead to discoveries. Dashed arrows indicate uninteresting connections ................................................................. 24
2.4 Complex association between Raynaud Syndrome – Dietary Fish Oils . . . 32
2.5 Thematic dimensions of association for Raynaud Syndrome – Dietary Fish Oils ........................................................................................................................................ 35
2.6 Evolution of Literature-Based Discovery Systems and Methodologies . . . 38
3.1 Importance of directionality in Semantic Associations ............................ 46
3.2 Supplementary Association for Platelet Aggregation .............................. 58
3.3 Secondary Association: Platelet Aggregation (path 1.1) ......................... 58
3.4 Secondary Association: Platelet Aggregation ............................................. 59
3.5 Primary Association Subgraph: Blood Viscosity ...................................... 60
3.6 Supplementary Association Subgraph: Blood Viscosity .......................... 61
3.7 Primary Association Subgraph: Vascular Reactivity ............................... 62
3.8 Supplementary Association Subgraph: Vascular Reactivity ..................... 63
4.1 Properties of MEDLINE article [PMID6098049] ................................. 71
4.2 Subset of the G-Tree of the MeSH Hierarchy ........................................... 79
4.3 Raw Distribution of Path Relatedness scores for Raynaud Syndrome – Dietary Fish Oils ................................................................. 84
4.4 Gaussian Distribution of Path Relatedness scores for three initial experiments 85
4.5 Interaction Network for Combination Similarity Scenarios .................... 86
4.6 Intra-cluster Network ............................................................................. 88
5.1 Subgraph1 (k = 3, 3σ) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome ................................................................. 96
5.2 Subgraph1 \((k = 3, 2\sigma)\) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome: (Cellular Activity) ............................ 100
5.3 Subgraph2 \((k = 3, 2\sigma)\) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome (Pharmaceuticals) ....................... 100
5.4 Subgraph3 \((k = 3, 2\sigma)\) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome (Lipids) ............................... 101
5.5 Subgraph4 \((k = 3, 2\sigma)\) on Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome (Lipids) ............................... 101
5.6 Subgraph1 \((k = 2, 2\sigma)\) Magnesium - Migraine ................................. 102
5.7 Subgraph4 \((k = 2, 2\sigma)\) Magnesium - Migraine ................................. 103
5.8 Subgraph9 \((k = 2, 2\sigma)\) Magnesium - Migraine ................................. 104
5.9 Subgraph22 \((k = 2, 2\sigma)\) Magnesium - Migraine ................................. 105
5.10 Subgraph7 \((k = 2, 2\sigma)\) Magnesium - Migraine ................................. 106
5.11 Subgraph5 \((k = 2, 3\sigma)\) Somatomedin C – Arginine .............................. 107
5.12 Subgraph7 \((k = 2, 3\sigma)\) on Somatomedin C – Arginine ...................... 109
5.13 Subgraph2 \((k = 2, 2\sigma)\) on Indomethacin, Lipid Peroxidation and Alzheimer’s Disease ................................................................. 110
5.14 Subgraph4 \((k = 2, 2\sigma)\) on Indomethacin, Acetylcholine and Alzheimer’s Disease ................................................................. 111
5.15 Subgraph3 \((k = 2, 2\sigma)\) on Indomethacin, Hydrocortisone and Alzheimer’s Disease ................................................................. 111
5.16 Subgraph14 \((k = 2, 2\sigma)\) on Indomethacin – Alzheimer’s Disease ............ 112
5.17 Subgraph3 \((k = 2, 2\sigma)\) on Estrogens, Genes and Alzheimer’s Disease .... 113
5.18 Singleton2 \((k = 2, 2\sigma)\) on Calcium-Independent PLA2, Selenium and Schizophrenia ................................................................. 116
5.19 Subgraph12 \((k = 2, 2\sigma)\) on Chlorpromazine, Isoproterenol and Cardiac Hypertrophy ................................................................. 118
5.20 Subgraph5 \((k = 2, 2\sigma)\) on Chlorpromazine, Calcineurin and Cardiac Hypertrophy ................................................................. 119
5.21 Subgraph7 \((k = 2, 3\sigma)\) on Testosterone, Cortisol and Sleep .................. 120

6.1 Document Reachability Framework .................................................. 130

A.1 Primary Association subgraph: Platelet Aggregation (Exp2) ................. 154
A.2 Supplementary Association subgraph: Platelet Aggregation (Exp2) ......... 155
A.3 Primary and Supplementary Association subgraph: Blood Viscosity (ID:2, ID:2.1) ................................................................. 155

C.1 Screenshot of the Obvio Web Application ........................................... 167
List of Tables

2.1 Comparison of capabilities and accomplishments of LBD techniques .... 40

3.1 Primary and Supplementary Associations ................................. 53
3.2 Subgraphs for Primary and Supplementary Associations ............... 55
3.3 Secondary Associations .................................................. 57
3.4 Raynaud Syndrome-Dietary Fish Oils Intermediate Rediscovery Comparison 64

4.1 Terminology ................................................................. 72
4.2 Distribution of MeSH descriptors for Semantic Predications from Raynaud Syndrome – Dietary Fish Oils ........................................ 75
4.3 Distribution of MeSH descriptors for Paths from Raynaud Syndrome – Dietary Fish Oils ....................................................... 77
4.4 Threshold Comparisons .................................................... 83

5.1 Comparison of rediscoveries with other approaches for Raynaud Syndrome - Dietary Fish Oils .................................................... 99
5.2 Comparison of rediscoveries with other approaches for Magnesium - Migraine ................................................................. 105
5.3 Comparison of rediscoveries with other approaches for Somatomedin C - Arginine .............................................................. 108
5.4 Comparison of rediscoveries with other approaches in the literature for Indomethacin - Alzheimer’s Disease ........................................ 112
5.5 Comparison of rediscoveries with other approaches in the literature for Estrogen - Alzheimer’s Disease .............................................. 115
5.6 Comparison of rediscoveries with other approaches in the literature for Calcium-Independent Phospholipase A2 – Schizophrenia ............... 117
5.7 Comparison of rediscoveries with other approaches in the literature for Chlorpromazine - Cardiac Hypertrophy ...................................... 119
5.8 Rarity and Interestingness of the subgraphs in the rediscoveries .......... 123

B.1 Literature-Based Discovery Systems and Tools .......................... 157
Acknowledgment

I would like to thank my advisor Dr. Amit P. Sheth, for his guidance and direction during my sojourn as a member of the Kno.e.sis Center. His vision and leadership always ensured that I pursue challenging real-world problems, with the goal of advancing science as well as delivering practical applications. I would also like to thank other Kno.e.sis faculty, including Dr. Krishnaprasad Thirunarayan for his in-depth questions and discussions on various aspects of my research. Similarly, Dr. Michael Raymer for bringing an outside perspective, which brought into focus many interesting unexplored ideas, many of which have now been addressed. Many thanks to Dr. Ramakanth Kavuluru for his brilliant ideas, which enhanced the quality of many aspects of this dissertation. Special thanks to Dr. Thomas C. Rindflesch, whose role as a leading researcher in the field of Literature-Based Discovery (LBD) continues to be recognized in the broader research community. Interactions with him during both my internships at the National Library of Medicine (NLM) were always intellectually charged, highly scientific and very progressive. I also thank Dr. Olivier Bodenreider for his stimulating discussions and insights. Finally, thanks to Varun Bhagwan, whose mentorship, while I interned at the IBM Almaden Research Center, helped refine my competencies as a researcher. I believe that my ability to appreciate the significance of ‘dialogue with data’ can be attributed to his stewardship.

I will always be indebted to my Kno.e.sis colleagues, who worked tirelessly on various projects, including Gary Alan Smith, Gaurish Anand, Nishita Jaykumar, Lu Chen and Revathy Krishnamurthy. I appreciate your sense of shared responsibility and holistic view of research, which were crucial in making our projects successful. Special thanks to Wenbo Wang, Pavan Kapanipathi, Shreyansh Bhatt and Swapnil Soni, for the many insightful discussions towards the end of my studies. Thanks also to many Kno.e.sis alumni whom I respect, and from whom I learned a great deal. Finally, thanks to my partner and dear friend Catherine Wilkins, for your unwavering support throughout this process.
Dedicated to

my mother Henrietta, my sister Sithendisi and my brother Clayton . . .
Introduction

“All truths are easy to understand once they are discovered; the point is to discover them.”

(Galileo Galilei, 1564 – 1642)

1.1 Motivation

Italian astronomer Galileo Galilei (1564–1642) once noted in his writings that “all truths are easy to understand once they are discovered” – the fundamental challenge is to actually discover them. Galileo’s observation is consistent with an existential human desire; the incessant pursuit of knowledge. Knowledge is considered an important aspect of understanding complex truths; to know is to have developed the capacity for understanding truth. English playwright William Shakespeare (1564–1616) once aptly noted that “Ignorance is the curse of God; [while] knowledge is the wing wherewith we fly to heaven.”

Human history has entailed an unrelenting quest for knowledge, both about ourselves and our surroundings. Some believe that acquired knowledge could promote self-understanding and help give perspective on life’s four fundamental questions: 1) Who am I? 2) Where do I come from? 3) What is my purpose?, and 4) What happens when I die? It has also been argued that the more we understand ourselves, the more it will engender purpose and give meaning to life itself. Ultimately, such self-awareness may be instrumental in enhancing overall quality of life and help optimize the human experience.
Understanding our environs, whether local or universal, is also important to the human experience. Renown physicist, cosmologist, and (self-described) “dreamer” Stephen Hawking (1942–present) once aptly noted that “[his] goal is simple. It is a complete understanding of the universe, why it is as it is and why it exists at all.” This innate curiosity about knowledge of the universe has materialized tangibly – from the subatomic discovery of the Higgs Boson\(^1\) to the discovery of an extragalactic expanding universe.\(^2\)

While the pursuit of knowledge to understand complex truths is ingrained in the human psyche, both the amount and diversity of knowledge dictate that only a subset of knowledge is known to humans at all times, whether holistically, group-wise or individually. Invariably, much of the knowledge required to facilitate human understanding must be discovered. However, to make such discoveries, one must first understand the various constructs that could lead to knowledge discovery. Knowledge itself is one of several components in a hierarchy of elements that inform human awareness. This hierarchy is the Data, Information, Knowledge, and Wisdom (DIKW) continuum [73, 1].

Data are discrete, unorganized, unprocessed observations, such as signs, signals, symbols, and stimuli, which convey no particular meaning when taken without context. Information is derived when data are meaningfully connected and result in a higher awareness, based on the context of the data. Knowledge emerges when seemingly unrelated pieces of information are assimilated, each potentially with disparate contexts. When logically assembled into clear and conceivable awareness, such knowledge may result in discoveries otherwise impossible to derive had the information fragments been taken in isolation. Knowledge discovery is therefore a process. It is characterized, at the most fundamental level, by finding unknown, yet informative connections among pieces of information.

In early human history, the process of knowledge discovery was mainly serendipitous. Chance encounters, trial and error, and \textit{ad hoc} methods were used to make discov-

\(^1\)Higgs Boson – http://en.wikipedia.org/wiki/Higgs_boson
\(^2\)In 1924, American astronomer Edwin Powell Hubble (1889–1953) discovered, not only the existence of galaxies beyond the Milky Way but that they were accelerating apart (see “The Day We Found the Universe” – http://bit.ly/UnivHubble).
eries. However, with the evolution of science, humans have developed more systematic approaches for making discoveries. Emergent discovery support systems are built on robust principles, involving activities such as experimentation, observation, explanation, and theory formulation. To illustrate the role of these activities from the historical perspective surrounding the science of making discoveries, consider three discovery scenarios, chosen arbitrarily among numerous others. These scenarios are the: 1) Mendelian Laws of Inheritance, 2) Chromosome Theory of Inheritance, and 3) Dyerberg–Bang Hypothesis, each of which is examined in the following subsections.

1.1.1 Mendelian Laws of Inheritance

In 1866, Austrian monk and scientist Gregor Johan Mendel (1822–1884) published an article entitled “Experiments on Plant Hybridization” [60], in which he explored the research question regarding the inheritance of traits in peas. Between 1856–1863, Mendel conducted over 10,000 experiments in which he cross-fertilized up to six generations of peas to observe the evolution of specific traits (i.e., pod size, pea shape, flower color, etc.) in the lineage. He made the observation that inheritance of certain traits extended beyond the traits of immediate parents in the lineage. Consequently, he put forth the explanation that inheritance of traits in peas was influenced by the existence of, what he called, dominant and recessive factors, which are encoded in all plants. Pairs of dominant and recessive factors from the parent plants undergo a split (Law of Segregation) and then independently recombine during fertilization (Law of Independent Assortment). When two recessive factors randomly recombine, features that were not recently observed in the lineage, suddenly reappear. Through this process, traits are passed from generation to generation. Mendel therefore presented the theory, which states that inheritance of traits in plants is based on random selection of dominant and recessive factors. This theory was crucial in debunking the erroneous Theory of Blending Fertilization, which assumed that inheritance of traits was governed by a blending of factors contributed by the parents. Mendel's seminal dis-
coveries on the segregation and recombination of factors for inheritance is now known as the *Mendelian Laws of inheritance*, and Mendel is credited posthumously as the founder of genetics.

### 1.1.2 Chromosome Theory of Inheritance

In 1903, American cytologist *Walter Sutton* (1877–1916) conducted a series of *experiments* in which he explored the research question regarding the mechanism of chromosome behavior in the *Brachystola magna* species of grasshopper [88, 87]. By microscopically studying the embryos of these grasshoppers, Sutton made the specific *observation* that chromosomes split during meiosis (cell division). This phenomenon was strikingly identical to the phenomenon observed years earlier by Mendel, whose findings had remained largely obscured in the scientific community. Thus, by linking plant pollination and cell division, Sutton put forth the specific *explanation* that chromosomes occur in pairs, segregate, and then randomly combine. German biologist *Theodor Boveri* (1862–1915) simultaneously produced similar findings while conducting microscopic analysis of the roundworm species *Ascaris megalcephala* [14]. In fact, Boveri provided further evidence for the split of chromosomes after experimentation with the embryos of Sea Urchins [15].

Sutton and Boveri were therefore able to establish the applicability of *Mendel’s Law of Segregation* in plants, to cell division in living organisms. Moreover, they were able to provide insights into the mechanism behind Mendel’s theory, thereby linking cytology and genetics (now cytogenetics) in a most intrinsic way. Together they are credited with the *Sutton-Boveri Chromosome theory*, which states that chromosomes are the basis for genetic inheritance [16]. Dutch botanist *Hugo de Vries* (1848–1935), also in 1903, provided evidence for the recombination aspect of dominant factors in terms of chromosomes.
1.1.3 Dyerberg–Bang Hypothesis

In the 1970s, Danish physicians Jorn Dyerberg (1937–present) and Hans Olaf Bang (1913–1994) conducted a series of epidemiological experiments in which they explored the research question regarding low incidence of Thrombosis and Atherosclerosis in Eskimos from the Umanak district of Greenland [11, 9, 10]. Together they made the observation that the plasma lipid levels of the Greenland Eskimos were especially low (i.e., the so-called “bad cholesterols” LDL and VLDL), compared with Eskimo expatriates living in Denmark. They also found that levels of HDL (i.e., “good cholesterol”) were significantly higher in the Greenland Eskimos. Additionally, they observed that incidence of Acute Myocardial Infarction (AMI) among the Greenland Eskimos was almost “10-fold lower than westernized countries” [10, 52]. Bang and Dyerberg put forth the explanation that this phenomenon was due to the high concentration of marine oils in the diets of the Greenland Eskimos. These diets were notably rich in Eicosapentaenoic Acid (EPA), compared with the average Danish who consumed diets rich in Arachidonic Acid. Additional experiments [8] showed that the Greenland Eskimos exhibited bleeding tendencies in the order of almost 200% more than the average bleeding time of the Danes. These extended bleeding times were later attributed to lower levels of Platelet Aggregation, also arising from elevated levels of EPA in their diets. Bang and Dyerberg therefore introduced their theory, which states that diets rich in marine oils can positively affect cardiovascular diseases, such as acute myocardial infarction, thrombosis, and atherosclerosis. This seminal work is now known as the Dyerberg–Bang Hypothesis [25].

1.2 The Science of Making Discoveries

The three aforementioned scenarios share six aspects commonly involved in the science of making discoveries (shown in Figure 1.1). These aspects are: 1) a human, 2) experimental...
Given a research question, humans first engage in experimentation, which yields observations. The human then reasons on these observations to derive an explanation. If the explanation is consistent across many cases, it may validate or invalidate an existing theory. It may also lead to the derivation of an entirely new theory, if none exists. Ultimately, it is the human who makes the discovery.

Humans are adept at making discoveries under such conditions because they can interpret context more effectively than machines. Scientific discoveries often require context, which manifests in humans through intellect, reasoning, ingenuity, and creativity [94]. For example, recognition of the complementarity between the results on plant hybridization by Mendel and the results on living organisms by Sutton and Boveri, required the human acumen of Sutton, to make the logical connection between the two fields. In spite of recent
advances in computation, these are still difficult tasks for machines to perform.

In today’s technological world, with vast amounts of digital data and the availability of information processing systems that can process big data *in silico*, to aid knowledge discovery, it is important to understand the role of such systems in terms of the science of making discoveries. I espouse the view posited by American information scientist *Don R. Swanson* (1924–2012), who noted that the role of an information processing system should not be to automatically take a series of observations, analyze them, and produce an explanation, then validate, invalidate, or generate a new theory for humans [94]. Rather, the role of an information processing system should be to analyze a series of observations and uncover meaningful *promising links*, which will enhance the human’s ability to make discoveries. This position is plausible for the previously stated point that humans are more effective than machines at higher-order cognitive tasks, such as memory, reasoning, and perception. Furthermore, at this juncture of computing, information processing systems are limited by architecture, technology, language, rules, algorithms, and background knowledge that can be devised or utilized to effectively facilitate these higher-order tasks.

The fundamental question that arises in this complex process of knowledge discovery support is therefore *what are these promising links?* This dissertation outlines a method for creating an information processing system that finds promising links to assist humans with the task of making discoveries from scientific literature. The central thesis is that an information processing system that leverages rich representations of textual content, based on implicit and explicit semantics, can provide an effective means for making discoveries from scientific literature. This has been convincingly demonstrated in this research, through the rediscovery of several well-known associations between biomedical concepts, with their substantiation in the scientific literature.
Dissertation Organization

This dissertation is organized as follows: Chapter 2 gives an overview of LBD research, including paradigms, modes, resources, and techniques. Chapter 3 discusses the context-driven subgraph model, together with a specific application for knowledge rediscovery and decomposition. Chapter 4 presents enhancements to the original context-driven model, which enable automatic subgraph creation. These enhancements include specification of context, shared context, and semantic relatedness to facilitate path clustering and automatic subgraph generation. Chapter 5 presents the evaluation of the enhanced framework on nine rediscovery scenarios. Chapter 6 then examines utilities for knowledge exploration and discovery browsing, which provide humans with additional tools and background knowledge, to understand the meaning of the associations in the subgraphs. Limitations and future directions are explored in Chapter 7. The Appendices provide: 1) additional experimental results, 2) an in-depth review of other LBD systems and approaches, and 3) a description of the Obvio web application, developed in this research.
Overview

Literature-Based Discovery (LBD) is characterized by uncovering hidden but novel information, implicit in non-interacting literatures. The notion of LBD was first proposed by Don R. Swanson (1924–2012) almost three decades ago, when he suggested that discoveries may arise from scientific literature, if logical connections \((B)\) can be found between seemingly disjoint literatures \((A, C)\). Swanson referred to this paradigm as the \(ABC\) model for LBD [89], which has since become an integral part of LBD research, facilitating the generation of several hypotheses [89, 91, 92, 79, 80, 81, 101, 86, 2, 108, 40, 38, 41, 42, 28].

However, the idea of making discoveries from seemingly disjoint scientific resources far precedes Swanson. American biologist James Arthur Peters (1922–1972), in reference to the discovery made by American cytologist Walter Sutton (1877–1916) in 1903 [88, 87], noted the potential for entirely new fields of research based on logically connecting seemingly unrelated fields. Peters stated that “When an author takes a series of apparently unrelated facts and ideas from two areas of investigation, combines them so that they make new sense, and develops a new hypothesis from the combination, he not only aids in the advance of both fields but also is quite likely to open up a new one” [88] (p. 27).

This Chapter presents a detailed examination of various aspects of LBD, cognizant of hidden connections in non-interacting literatures. It first discusses the broader field of biomedical text mining in Section 2.1, which includes Information Retrieval (IR), Question Answering (QA), Automatic (or Document) Summarization, and LBD. Various characteristics of biomedical knowledge sources, which are currently used to support LBD, are
discussed in Section 2.2, while Section 2.3 provides a detailed review of LBD methodologies, limitations, paradigms, and modes. The Chapter concludes with a discussion on the contributions of this dissertation to the field of LBD in general, in Section 2.4.

2.1 Biomedical Text Mining

Harnessing information from biomedical texts for knowledge discovery is generally considered a subtask of Biomedical Text Mining. In Computer Science, Text Mining “refers to the process of deriving high quality information from text” [26]. Biomedical Text Mining therefore encompasses approaches developed specifically to obtain high quality information from biomedical texts. Generally, biomedical text mining [76] can be perceived as being comprised of four different tasks: 1) Information Retrieval (IR), 2) Question Answering (QA), 3) Document Summarization, and 4) Literature-Based Discovery (LBD).

2.1.1 Information Retrieval

Information Retrieval (IR) is a process of finding information, within large collections, to satisfy an information need [58]. It is mainly a retrieval task, in which the information sought by the user is known to be present in the collection. To retrieve the desired information, biomedical IR systems commonly extract potentially meaningful text fragments, and then index the corpus using these fragments, and along other dimensions that likely match the user interests (commonly keyword queries).

To begin the search process, users typically possess a mental conceptualization of their information need – some argue this can be framed using a mental model [97]. The search system then provides an environment, or a user query language, for users to adequately express their information needs in terms of language constructs and primitives that can be understood by the system. The system must then provide a specification for translating the
user query into a system query, based on the interpretation of the user query by the system. Such interpretation is commonly driven by the existence of a knowledge model, which may formally represent the primitive and complex constructs (and their associations), in the search space. System queries must then be matched with the indexed corpus, and highly ranked artifacts presented as potentially relevant results to the user.

One drawback of IR systems is that users must search-and-sift \cite{69} through potentially large volumes of text to find relevant information. While thriving search engines like Google and Yahoo! perform reasonably well at retrieval tasks, it is still incumbent on users to peruse the result set to find the specific information that will satisfy their information need. This could be problematic in situations where complex information needs exist. The number of documents that must be explored to find the information could be overwhelming for information seekers. The second issue is that users must perform query reformulation when the search system fails to retrieve relevant documents. This is often necessary since direct links between atomic facts within documents are typically not provided. To alleviate these and other limitations, techniques designed to retrieve concise information fragments from documents, have been developed. This area of research is known as Question Answering (QA).

### 2.1.2 Question Answering

Question Answering (QA) in biomedical text mining can supplement IR, by focussing on the retrieval of very specific information fragments from text. These fragments often give precise answers to specific information needs \cite{7}. To achieve this, a typical biomedical QA system consists of the following three phases: 1) query processing, 2) document processing, and 3) answer processing.

In the question processing phase, the user provides a question to the system, which is interpreted using the underlying query interpretation framework. The interpreted question is then translated into a system query, which is used to retrieve an initial set of relevant
documents. In the second phase (document processing), the system query is used to select specific candidate passages (or sentences) from within documents, which may be more relevant to the query. In the final phase (answer processing), candidate passages are matched more strictly with the system query. The matching text fragments are delivered to the user, as potential answers.

QA is therefore considered an effective, albeit more challenging, retrieval task than IR. It is challenging because elements in the question must be dynamically typed and matched with elements in passages. The passages must therefore be typed and indexed \textit{a priori}. Type matching is problematic because it often requires comprehensive knowledge models to cover diverse corpora. Such knowledge bases are commonly unavailable. Type matching also requires complex algorithms to resolve ambiguities during query processing. These are difficult to design across diverse texts. QA systems are therefore highly tuned and domain specific, supporting only a restricted set of queries. To supplement IR and QA systems, and allow users to iteratively expand and refine their interests/question, broad themes derived from summarizing search results, have often been used. This area of research is known as Automatic (or Document) Summarization.

2.1.3 Document Summarization

Document Summarization [57] is designed to summarize search results and provide a concise overview of the salient information conveyed across the entire search result set. It mainly supports exploratory search situations in which no specific information need exists, but can be used to help users refine their initial user interests and questions, based on what is contained in the corpus. Summarization is commonly a direct result of information overload. A typical summarization task involves two crucial steps: 1) extraction and 2) abstraction.

Extraction is focused on obtaining text from the corpus in its unaltered form as document summaries – i.e., ‘\textit{what you see is what you get}.’ On the contrary, the abstraction step
is designed to identify topics or themes through automatic matching or manual bootstrapping. These themes serve as broad labels that summarize the content. Document summarization is therefore mainly a matching task, between high level themes (or abstractions) and the content itself – i.e., ‘what you know is what you get’ [46].

A fundamental limitation of IR, QA, and Document Summarization systems is the degree of manual effort required to find meaningful information after the search system returns the result set. Furthermore, these techniques are not very effective at finding implicit associations across the textual content. LBD is focused on finding such implicit connections, which may lead to the discovery of new knowledge.

2.1.4 Literature-Based Discovery

Literature-Based Discovery (LBD) is a challenging aspect of biomedical text mining. It involves uncovering hidden connections that are implicit in scientific literature. In reference to the challenging nature of knowledge discovery, in general, Professor James Caruthers, in a Purdue University News Service Report (October 19, 2004)\(^1\) noted that “knowledge discovery is more like sifting through a warehouse filled with small gears, levers, etc., none of which is particularly valuable by itself. After appropriate assembly, however, a Rolex watch emerges from the disparate parts.” LBD, as a specialized aspect of knowledge discovery, is therefore a complex process, which involves not only identification and extraction of important information from text (as for IR, QA, and Summarization), but also linking them to uncover hidden, complex and meaningful associations.

**Brief History**

Information processing systems developed for LBD must provide functionality to enable finding *promising links* among diverse information fragments. Assuming the existence of

these atomic information fragments to begin with, LBD systems must then model and exploit the context of information, to help connect related fragments, based on shared context. Early representations of context for LBD were mainly keyword-based [36, 53, 34, 95, 82], using distributional statistics from term (co)occurrence to capture implicit context.

Subsequent enhancements were mainly concept-based [85, 100, 12, 66, 108, 106, 107, 29, 45, 98], which mapped keywords in text to standard biomedical concepts using structured background knowledge. While concept-based approaches were more precise, they still only provided an indirect way of identifying the meaning of associations.

Relations-based techniques for LBD were developed [2, 38, 42, 23, 105, 19, 18] to utilize the explicit relationships (or predicates) among concepts and help provide insights into the nature of associations. Predicates were commonly obtained from background knowledge or known a priori by domain experts. While an advancement over keyword-based and concept-based approaches, relations-based approaches remain mainly applicable in scenarios where both the predicates and semantic types of concepts are already known. In particular, it is deemed unsuitable for elucidating complex associations, in which context is often initially unknown.

Contemporary approaches to LBD focus on graph-based approaches [105, 112, 28, 33, 98] that leverage structural properties of graphs to create complex associations (or subgraphs). These subgraphs typically utilize abstractions of assertions extracted from scientific literature, called semantic predications. Semantic predications are binary relations between biomedical concepts, which take the form (subject, predicate, object). For example, the semantic predication which states that [Leptin STIMULATES Serotonin] can be obtained from the MEDLINE article [PMID20660061], whose text notes that “…CNS serotonin activated by leptin modulates sympathetic outflow to the skeleton.”

In spite of the successes of the graph-based approaches to LBD, more effective methods for LBD are desired. One possible reason is because complex associations among biomedical concepts may exist along multiple thematic dimensions, such as Metabolic
Function, Pharmaceutical Treatment, and Neurological Activity (as we will show). A method that can automatically generate complex associations among concepts, along multiple thematic dimensions, that has been proven effective, has never been developed. In this dissertation, such a method has been devised, and its efficacy has been clearly demonstrated through the rediscovery of 8 out of 9 existing discoveries. In the following Section, various biomedical knowledge sources that enable new definitions for context, which enable multi-faceted automatic subgraph generation are covered.

2.2 Biomedical Knowledge Sources

For LBD, biomedical knowledge sources can be classified into two broad categories, based on the type of knowledge they provide. These categories are: 1) assertional knowledge and 2) definitional knowledge.

2.2.1 Assertional Knowledge

Assertional knowledge refers to statements asserted in scientific literature. For example, the following text fragment, which states that “...CNS serotonin activated by leptin modulates sympathetic outflow to the skeleton” is an assertion from the following article [PMID20660061] in MEDLINE. Semantic predications are extracted from MEDLINE using a tool called SemRep², developed at the National Library of Medicine (NLM). The two primary sources of assertional knowledge are: 1) MEDLINE, which is a repository of more than 23 million bibliographic citations maintained by NLM and 2) the Semantic MEDLINE Database (SemMedDB) [51, 72], which consists of 65 million semantic predications extracted from MEDLINE. Assertional knowledge is considered appealing for making discoveries because it is dynamic and rapidly adds new information to an already rich body of accumulated knowledge. Such knowledge has the following characteristics: 1) comple-


15
mentarity, 2) disjointness, 3) dynamism, 4) heterogeneity, and 5) conflictedness.

1. **Complementarity:** Biomedical literature often contain information that is complementary. For example, research on *Testosterone* (A) has shown that high levels of this steroid hormone inhibits the stress hormone *Cortisol* (B). Concurrently, research on *Sleep* (C) has shown that high levels of *Cortisol* disrupts *Sleep*. Together, these fragments from the *Testosterone*-literature and the *Sleep*-literature, help explain the phenomenon of diminished sleep quality in aging men compared to women [61]. As the largest, and largest growing, bibliographic database for cataloging biomedical literature, it is not surprising that numerous assertions exist in MEDLINE, which are complementary but undiscovered, as noted by Swanson [90, 94].

2. **Disjointness:** Complementary literatures are sometimes disjoint. That is, none or very few of the articles in the A-literature, reference the C-literature, despite the existence of logical connections across their content. For example, in the *Raynaud Syndrome – Dietary Fish Oils* discovery [89] only 4 articles cross-referenced the two sets of literature, which contained more than 1000 articles on *Raynaud Syndrome* and 3000 on *Dietary Fish Oils*. Similarly, the link between the observations that led to the *Mendelian Theory of Inheritance* (see Section 1.1.1) based on plants and that on the *Chromosome Theory of Inheritance* by Sutton and Boveri (see Section 1.1.2) on living organisms, was also inconspicuous. This disjointness property of scientific literature is a serious challenge for LBD, as noted by Swanson [90].

3. **Dynamism:** The body of scientific literature from which discoveries can arise continues to grow at an increasing rate. In MEDLINE, more than 850,000 articles were published in 2009, compared with 485,942 in 1999 and 398,120 in 1989. This rapid evolution of asser-

\[3\text{MEDLINE literature growth chart – http://jasonpriem.org/2010/10/medline-literature-growth-chart/}\]
tional knowledge reported in the literature suggests that new solutions for current medical problems may have already been developed. Numerous research articles on various topics are published constantly, detailing new experiments, procedures, and observations from clinical trials, wet labs, reports, and other in vivo studies. Assertional knowledge is therefore replete with information from which discoveries may arise.

4. **Heterogeneity:** Scientific literature is also fraught with abbreviations, and many syntactic and lexical variations of standard concepts. Such heterogeneity can make it challenging to detect concepts and assertions, which convey meaningful associations. Lexical background knowledge sources such as the UMLS SPECIALIST Lexicon\(^4\), together with NLP tools such as MetaMap\(^5\) and the Medical Term Indexer (MTI)\(^6\), provide support for concept identification and ambiguity resolution. In addition tools like MetaMap and SemRep address the linguistic heterogeneity that arise in scientific writing, manifested as negation, infix, prefix, and postfix expressions. Decoding such expressions require complex extraction algorithms, supported by these tools, which ultimately facilitate downstream LBD research.

5. **Conflictedness:** Assertional knowledge may also contain contradictory statements. Conflicting or inconclusive results from different research may lead to statements, which imply opposing semantics. Conflict resolution is a serious issue in biomedical science. Some contradictions may not be resolved for many years, until new findings from new experiments have been published. Relying on a computation system to automatically arrive at explanations for observations is generally considered unwieldy. Hypothesis generation will invariably require adjudication by humans, whenever appropriate. As previously noted (see Section 1.2), the task of the LBD system should be to provide promising links, which will facilitate the human in making discoveries. A second issue is that seemingly contra-

---


ditory statements may not be contradictory at all. Rather they may require context for correct interpretation. For example, an article may report that “…several years ago, it was believed that concept $c_1$ inhibited the production of concept $c_2$. However, recent research now indicate that concept $c_1$ stimulates the production of concept $c_2$.” Distinguishing the two contexts and identifying whether the first assertion $c_1 \rightarrow \text{INHIBITS} \rightarrow c_2$ or the second assertion $c_1 \rightarrow \text{STIMULATES} \rightarrow c_2$ is correct, depends on the robustness of the LBD system. Indeed both assertions are correct, with some temporal window of context.

In this dissertation, an approach for LBD is presented that uses assertional knowledge to create complex associations among concepts. The approach is able to bridge disjoint and complementary literatures by clustering semantic predications based on their shared context. The provenance of the semantic predications are provided to enable users to manually adjudicate and resolve conflicts, where appropriate. Heterogeneity is accounted for by relying on NLM’s tools for information extraction, while dynamism and scalability remain outstanding issues. In the next Section, the characteristics of definitional knowledge and its role for LBD are discussed.

2.2.2 Definitional Knowledge

Definitional knowledge refers to statements about biomedical concepts that are generally known and regarded as common knowledge within the community. Several biomedical knowledge sources, which contain definitional knowledge are available for LBD research. These include: 1) the Unified Medical Language System (UMLS)\textsuperscript{7}, a knowledgebase of more than 13 million statements and 2) the Biomedical Knowledge repository (BKR), an RDF database consisting of close to 650 million statements aggregated from SemMedDB and the UMLS.

The UMLS is a compendium of more than 130 biomedical vocabularies. Some of these include SNOMED-CT, ICD-x, NCBI taxonomy, LOINC, Gene Ontology, MeSH, and

\textsuperscript{7}Unified Medical Language System (UMLS) – \url{http://www.nlm.nih.gov/research/umls/}
the OMIM database. The UMLS consists of three main components: 1) the Metathesaurus, 2) Semantic Network, and 3) SPECIALIST Lexicon. The semantic network consists of 134 high-level categories for classifying biomedical concepts according to broad semantic types. This network also contains 15 higher level semantic groups. The semantic network also consists of 54 unique predicates, which are organized in a hierarchy. Note that while the UMLS is a terminology, in strict terms, it is not an ontology. The UMLS categories in the Semantic Network hierarchy are organized only to a limited degree. Numerous instances in the Metathesaurus refract from the semantic network schema. Such inconsistencies are a natural consequence of attempting to integrate many disparate vocabularies, with different (and sometimes unspecified) semantics in their schema. At the instance level, the UMLS Metathesaurus integrates statements from the member vocabularies, and consists of close to 3 million unique concepts and more than 13 million statements. Synonyms for UMLS concept labels are maintained in the UMLS SPECIALIST Lexicon. This lexicon provides lexical and syntactic services that facilitate various information extraction tasks.

The UMLS subsumes the Medical Subject Headings (MeSH), which is also a controlled vocabulary (and thesaurus) of biomedical terms, organized in a hierarchical structure. MeSH has been used separately for LBD research: 1) to support mapping textual context to structured background knowledge and 2) to provide implicit context among co-occurring concepts. It is organized into a category hierarchy of 16 trees (as of the time of this writing), with a maximum depth of 15 for more than 27,000 distinct descriptors. The focus of MeSH is on categorization and organization of biomedical knowledge. As such, MeSH does not contain any explicit predicates among descriptors. In practice, it is used for indexing, cataloging, and searching MEDLINE. To achieve this, MeSH descriptors are manually assigned to scientific articles, by individuals called MeSH Indexers. The Medical Term Indexer (MTI) is used to first generate a list of candidate descriptors, which are subsequently finalized by the indexers. The quality of these assignments is considered high and relatively good indicators of the semantics of the content of the article to which
they are assigned. The MeSH descriptors therefore serve a dual role: 1) as definitional knowledge in biomedicine and 2) as a layer of abstraction for the semantics of the content of biomedical articles.

The availability of assertional and definitional knowledge for LBD offers opportunities for making discoveries by capturing important elements in text, which are grounded in background knowledge. They also enables filtering and abstraction of biomedical content, while also providing a means capturing context. In the next Section paradigms, modes, and methodologies for LBD that leverage assertional and definitional knowledge are discussed.

2.3 Literature-Based Discovery Research

The field of Literature-Based Discovery was pioneered by American information scientist Don R. Swanson (1924–2012) through the well-known Raynaud Syndrome–Dietary Fish Oils Hypothesis (RS-DFO) [89] in 1986. Raynaud Syndrome (RS) is a circulatory disorder, discovered by French physician Auguste Gabriel Maurice Raynaud (1834–1881), which causes periods of severely restricted blood flow to the fingers and toes [110], resulting in blanching, blackening, and discoloration. By reading the titles of MEDLINE articles, Swanson serendipitously discovered that Dietary Fish Oils (DFO) lower Blood Viscosity, reduce Platelet Aggregation, and inhibit Vascular Reactivity (specifically Vasoconstriction), mainly due to the presence of omega–3 Fatty Acids. Concomitantly, he observed that a reduction in both Blood Viscosity and Platelet Aggregation, as well as the inhibition of Vasoconstriction, appeared to prevent Raynaud Syndrome. Swanson therefore postulated that “dietary fish oil might ameliorate or prevent Raynaud’s syndrome.” DiGiacomo et al. [68] clinically confirmed this hypothesis in 1989.

The RS-DFO discovery encapsulates the ideas of complementarity and disjointness between literatures. This is because explicit associations between DFO and these intermediate concepts (i.e., Blood Viscosity, Platelet Aggregation, and Vascular Reactivity) had
long existed in the literature. In particular, *Dietary Fish Oils* (or *Marine Oils*) were the subject of study by Danish Physicians *Jorn Dyerberg* (1937–present) and *Hans Olaf Bang* (1913–1994) for more than a decade (1971–1982). Dyerberg and Bang conducted a series of epidemiological studies on Eskimos in Greenland [25, 10, 11, 9, 8]. Their research focused on the possible prophylactic effect of marine oils in thrombo-embolic disorders. They initially observed a correlation between diet, plasma lipid levels and acute myocardial infarction (AMI) in the Greenland Eskimo [11]. Furthermore, they found that levels of HDL or the so-called “good cholesterol” were also significantly higher in the Greenland Eskimos. Additionally, they observed that incidence of AMI among the Greenland Eskimos was almost “10-fold lower than westernized countries” [10, 52]. These observations were supported by Kromann et al. [52], who showed that *Marine Oils* increased levels of good cholesterol or high-density lipoproteins (HDL), while decreasing levels of bad cholesterol or low-density lipoproteins (LDL, VLDL) in the Greenland Eskimos. The elevated levels of HDL in Greenlanders was therefore linked to the significant reduction in cardiovascular complications, such as *Thrombus, Atherosclerosis*, and *Hemostasis*.

Concomitantly, explicit associations between the intermediate concepts (i.e., *Blood Viscosity, Platelet Aggregation*, and *Vascular Reactivity*) and RS had been well documented. In particular, Pringle et al. [67] in 1965, showed that the mean *Blood Viscosity* measurements in a sample of 22 patients with *Raynaud Disease* was 5.2 centipoises; a value of 2.7 centipoises above the mean of the control group, which measured at 2.5 centipoises. The first study on *Platelet Aggregation* and *Raynaud Syndrome* was reported in 1980. Zahavi et al. [110] showed that *Platelet Aggregation* in 37 patients were significantly increased compared to the patients in the control group. After 9 of the 37 patients who suffered *Chronic Raynaud Disease* (i.e., longer than 5 years) received plasma exchange treatments, significant reductions in *Platelet Aggregation* suggested that platelets were involved in the pathogenesis of *Raynaud phenomenon*.

The serendipity in Swanson’s Hypothesis lies in the fact that no explicit associations
linking \textit{DFO} and \textit{RS} had been previously articulated in a single document. To arrive at this discovery, Swanson performed a Dialog\textsuperscript{®} Scisearch using Raynaud and Fish Oil terms, on titles and abstracts of MEDLINE and Embase (Excepta Medica) citations, in November 1985. There were approximately 1000 articles in the Raynaud set and 3000 in the Fish Oil set. Swanson found that only four articles among a reduced set of 489 articles (after filtering), contained cross-references spanning both sets. Among these four articles, only two articles \cite{62, 63}, written by Honduran–British pharmacologist \textit{Sir Salvador Moncada} (1944–present) (also noted for his work on nitric oxide)\textsuperscript{8} discussed relevant aspects of \textit{RS} with \textit{DFO}; although not in the context of Swanson’s discovery. Swanson speculated that this phenomenon of logically related but noninteracting literatures alludes to the existence of \textit{undiscovered public knowledge} \cite{89}. Logically related information fragments may exist in the literature, but may have never been assembled, or fully elucidated in any single article. Swanson explained that “\textit{Until those fragments, like scattered pieces of a puzzle, are brought together, the relationships among them may remain undiscovered–even though the isolated pieces might have long been public knowledge}” \cite{90}. He therefore exploited his awareness of the existence of such undiscovered semantic associations \cite{4} across noninteracting literatures, to investigate several other scenarios (three with Smalheiser \cite{79, 80, 81}) that later led to new scientific discoveries \cite{91, 92}. Swanson grounded his observations in a paradigm now commonly known as the \textit{ABC model} \cite{89} for LBD. In the next Section the paradigms and modes for LBD research are discussed.

2.3.1 Paradigms and Modes

Various paradigms have been developed for LBD, among which the \textit{ABC model} is perhaps the most widespread. The \textit{ABC model} is based on the idea that new knowledge can be discovered between two concepts \((A, C)\) from noninteracting literatures, if hidden associ-

\textsuperscript{8}http://www.nature.com/nature/journal/v395/n6703/full/395625a0.html
Figure 2.1: Venn Diagram showing the overall idea of the ABC model

ations involving some intermediate concept (B) can be uncovered (as depicted in Figure 2.1). More specifically, if logical AB relations, linked to BC relations can be uncovered, then the intermediate B-concepts may be candidates for discoveries. This seminal model has been widely adopted for LBD and has been used both to recover many of Swanson’s original hypotheses [53, 36, 100, 85, 66, 38, 42] as well as to propose new hypotheses [2, 38, 105, 86]. Frequently and rarely co-occurring B-terms have generally been considered good indicators of hidden connections.

Figure 2.2: Open Discovery Mode. Solid arrows indicate potentially interesting associations which may lead to discoveries. Dashed arrows indicate uninteresting connections

The ABC model has been utilized for LBD research in two modes: open and closed discovery. Open discovery consists of generation of a hypothesis [86, 100], where none previously existed. Various approaches have been implemented to facilitate open discov-
In open discovery, there is typically a single scientific problem and only a limited idea about which concepts are involved. The hypothesis generation process commonly begins with one concept from the research question and propagates outwards, seeking relevant nearby concepts, which participate in some unknown yet interesting association (see Figure 2.2) – i.e., only A is known, not necessarily B or C.

Closed discovery on the contrary, is characterized by hypothesis testing [86, 100] or elucidating a candidate solution to the research question. It may also lead to the generation of new, more granular hypotheses. Various approaches have been implemented to facilitate closed discovery [84, 53, 35, 85, 45, 2, 38, 61, 30, 19, 21]. In closed discovery, typically the source A and target C are known, but not necessarily the intermediate(s) B. The discovery process begins with A and C, and traverses (or explores) from both ends to discover interesting intermediates, which may expound the association between the source and target (as depicted in Figure 2.3).

![Figure 2.3: Closed Discovery Mode. Solid arrows indicate potentially interesting associations which may lead to discoveries. Dashed arrows indicate uninteresting connections](image)

While the ABC model is a well established part of LBD research, it is not foolproof. In [105], Wilkowski et al. proposed an extension of the ABC model, which was used to elucidate the association among Norepinephrine, Depression, and Sleep. Wilksowski suggested that the ABC model can be decomposed into a more granular model in which several intermediate concepts may be required to expound associations. This logical extension of
Swanson’s *ABC model* is referred to as the *AnC model* (*pronounced ants*) in [19], in which
\[n = (B_1, B_2, \ldots, B_m)\].

Still, both the *ABC model* and the *AnC model* for LBD have limitations. Consequently, numerous approaches [112, 111, 28, 98] have been developed that rely on complex *graph-based models* to address these limitations, which commonly surface when dealing with complex scenarios. For example, van der Eijk et al. [98] in 2004, first utilized a subgraph-driven paradigm for LBD, on the premise that co-occurring MeSH descriptors can be important in providing a holistic understanding of complex associations. Zhang et al. [112] concretely supported this idea and developed a technique for subgraph creation based on degree centrality and clustering of cliques. In this dissertation, a context-driven *subgraph model* for LBD [19, 21] has been developed, which is consistent with this class of graph-based models. In the next Section, methodologies that apply the *ABC, AnC, and subgraphs* paradigms for LBD, using both open and closed discovery modes, are discussed in detail.

### 2.3.2 Methodologies and Limitations

Methodologies for finding unknown connections and elucidating poorly understood associations in scientific literature, whether based on the *ABC, AnC* or *subgraph* model, can be classified according to five general categories based on technique. These are: 1) keyword-based, 2) concept-based, 3) relations-based, 4) graph-based, and 5) hybrid approaches. Approaches that utilize each of these techniques are discussed in the follow subsections.

**Keyword-based**

Much of the early LBD research used keyword-based methods to find intermediates, mostly based on the *ABC paradigm* [36, 53, 34, 95, 82]. The conventional wisdom was that discoveries are likely to arise from logical connections between keywords that frequently, or rarely, (co)occur in a corpus. Metrics such as relative frequency, token frequency, term
frequency-inverse group frequency (TF-IGF) [36, 53] and term frequency-inverse document frequency (TF-IDF) [53] were used to rank potentially informative intermediates and targets; the latter applicable to open discovery.

In ARROWSMITH [95], to improve coverage across intermediates, Swanson iteratively utilized MEDLINE query results to support LBD. For both open and closed discovery, the titles of MEDLINE articles were first queried to obtain an initial set of intermediates. MEDLINE was then queried again, using this initial set of intermediates, to obtain more documents, from which potentially interesting B-terms were obtained based on frequency of (co)occurrence. In [35], Gordon and Dumais successfully applied the popular technique of Latent Semantic Indexing (LSI) for both open and closed discovery. LSI was used to first retrieve relevant documents given a set of search terms, then the vector space model (cosine similarity) was used to find intermediates based on co-occurrence of intermediates, with the source and target. The authors reported that LSI was only slightly more effective than traditional frequency-based metrics, such as token frequency, record frequency, and term frequency-inverse global frequency (tf-igf) [36] for finding intermediates.

In IRIDESCENT [106], Wren et al. used a probabilistic approach to rank keywords for open discovery. The probability (or veracity) of co-occurring intermediates (as n-grams) was computed using maximum likelihood estimates (MLE), derived from sentences and abstracts. The strength of co-occurrence between AB and BC pairs was then computed and normalized, based on expected connectedness (or degree centrality) between terms, given the MLE values. In this way, non-informative terms that were frequently co-occurring and highly connected in the corpus, could be eliminated. In the final step, intermediates were ranked by the ratio of observed/expected connectedness between co-occurrence of AB and BC pairs. In a refinement of this approach [107], mutual information was used to estimate the strength co-occurrence and Wren successfully applied IRIDESCENT to discover new knowledge on Chlorpromazine and Cardiac Hypertrophy [106].
**Limitations:** While many keyword-based techniques have been used to successfully discover [91, 92, 106] and recover [36, 35, 53, 34] numerous discoveries, several problems existed. The first issue was the absence of effective methods for synonym detection. Biomedical concepts can be expressed in a variety of forms. The concept *Interleukin-6* for example, can be expressed as “Interleukin-6,” “Interleukin 6,” “IL6”, “IL 6” and “IL-6.” Early keyword-based approaches were not well suited to detect these syntactic variations. The second issue was the absence of reliable identification of simple, modified, and compound entities. For example, the concept “acute myocardial infarction” implies higher severity than simply “myocardial infarction.” Several approaches relied on *ad hoc* text processing techniques for *n-gram* detection. These included dictionary-based methods, stemming, and stopword filtering. Additional *ad hoc* methods were also used for keyword filtering and removal of non-novel terms, such as “patient,” “clinical,” and “treatment” from corpora.

**Concept-based**

To overcome some of the limitations of keyword-based approaches, several approaches [85, 100, 12, 66, 106, 107, 29, 45, 98] used concept-based techniques. These techniques normalize syntactic expressions to known concepts, using structured background knowledge. For example, in [100], Weeber et al. adopted a concept-based approach, in a system called *DAD* [99], which used MetaMap [6] for entity identification. Metamap is advantageous because it performs entity identification in text and links entities to the UMLS. Mapping expressions in text to UMLS concepts enables use of semantic types from the UMLS semantic network for filtering. Additionally, it provides a basis for gleaning context from the hierarchical and associative relations in the UMLS semantic network. In [101], Weeber et al. successfully applied this approach to filter intermediates and successfully recover the *Raynaud Syndrome–Dietary Fish Oils* discovery. The approach was also used to suggest new treatments for *Acute Pancreatitis, Chronic Hepatitis C, Gastritis,* and
Myasthenia Gravis using the drug Thalidomide.

In BITOLA [44], Hristovski et al. also implemented a concept-based approach that leveraged UMLS semantic types to filter intermediates. This approach used MeSH descriptors assigned to articles, as A-, B-, and C-terms, instead of UMLS concepts. The approach also utilized association rules, together with confidence and support to rank AB and BC MeSH descriptor pairs. Hristovksi expanded the approach by using gene symbols [41, 42] and supplementary concepts records [47, 48] to demonstrate its applicability beyond MeSH descriptors. In [30] Gabetta et al. used a similar approach to reproduce the discovery of genes associated with Dilated Cardiomyopathies, using UMLS concepts instead of MeSH descriptors.

Blake and Pratt [12] initially implemented a concept-based approach for closed discovery that also utilized UMLS semantic types to filter intermediates. Then, in LitLinker, Pratt et al. [66] extended this approach, by using a combination of concept-based, text mining, and data mining approaches for open discovery. Association rules (specifically, the Apriori algorithm) and level of support were used to rank AB, BC concept pairs. Intermediates (or linking concepts) as well as candidate C-terms were filtered based on their UMLS semantic types. Additionally, a ranking scheme was developed to group similar concepts based on syntactic similarity. Both approaches were successfully applied to recover several intermediates from the Magnesium–Migraine discovery [91].

Yestigen-Yildiz et al. [108] further enhanced the concept-based approach for open discovery in LitLinker. The enhancement used hierarchical relationships from both the MeSH hierarchy and UMLS (semantic types and groups) for filtering. The probabilistic z-score metric was used to rank potential intermediates (i.e., MeSH descriptors). In their work, z-score was used to compute the probability of occurrence of a given MeSH descriptor in the selected literature compared with the mean probability of its occurrence in the entire corpus. This was used to rank AB, BC pairs. Target C-terms were also ranked in the final step, based on relative frequency with AB associations. Yestigen-Yildiz used this approach
to rediscover intermediates from the *Magnesium–Migraine* discovery as well as to suggest new insights into the association between: 1) *Alzheimer’s Disease* and *Endocannabinoids*, 2) *Migraine* and *AMPA receptors*, and 3) *Schizophrenia* and *Secretin*.

In *Manjal*, Srinivasan [85, 86] also implemented a concept-based approach for both open and closed discovery. This approach leveraged co-occurrence relationships between UMLS semantic types and MeSH descriptors. Instead of z-score, the approach used the popular TF-IDF metric, to weigh MeSH descriptors. The ranking scheme then measured the co-occurrence between MeSH descriptors and their respective UMLS semantic types. In this way, weighted topic profiles were created, as vectors of scored MeSH descriptor–UMLS semantic type pairs. Relevant intermediates were obtained by restricting UMLS semantic types, then applying the TF-IDF ranking, on the topic profiles. The approach was used in [85] to recover intermediates from 5 out of the 6 discoveries made by Swanson [89, 91, 92, 79, 81]. It was also used to provide new insights into the association among *Tumeric/Curcumin*, *Crohn’s Disease*, retinal diseases, and disorders related to the *Spinal Cord* in [86].

**Limitations:** While concept-based approaches offer improvements over keyword-based approaches, there are two fundamental issues. The first is that the use of frequency of co-occurrence only provides an indirect way of capturing context. These metrics do not provide insights into the meaning of associations among concepts. For instance, consider the scenario in which *Dietary Fish Oils* inhibit *Platelet Aggregation*, and the aggregation of *Blood Platelets* causes *Raynaud Disease*. While *Dietary Fish Oils*, *Platelet Aggregation*, and *Raynaud Disease* may frequently co-occur in a corpus, their precise association is not explicitly captured by their co-occurrence. This limitation of concept-based approaches has far reaching implications in the biomedical domain. Ahlers et al. [2] makes the specific observation that in treatment of diseases, “*Drug therapies are often used effectively, even though the exact cause of action may be either poorly understood or unknown.*”
LBD systems should therefore be designed not merely to uncover co-occurrence relationships among concepts, but also offer additional insights into the underlying mechanisms of interaction and causality relationships among them. Gordon and Dumais made this crucial observation in [35] after successfully applying LSI for LBD. The authors speculated that richer representations of textual content are needed to capture “evidence suggestive of ‘causal’ relationships in the literature (which may be revealed independently of their statistical prominence).” Moreover, they stressed the need for “semantic and category knowledge to improve the step of identifying [intermediate and] terminal concepts.”

The second issue is that most concept-based approaches are based on the ABC paradigm, which assumes that new knowledge depends on one intermediate between A and C. However, the idea of one level of intermediates between concepts (A, C) is somewhat restrictive for LBD. Novel associations between concepts may exist in longer sequences beyond ABC. Smalheiser [78] recently noted this issue and suggested that next generation LBD systems will require more expressive representations beyond the ABC model to uncover hidden associations.

**Relations-based**

To address the limitations of concept-based approaches [2, 38, 42, 23, 105, 19, 18], several relations-based techniques have been developed. Unlike keyword-based and concept-based approaches, relations-based approaches utilize explicit relationships (or predicates) between concepts, to more effectively capture the meaning of associations. To achieve this, some approaches rely on semantic predications extracted from the biomedical literature.

In the development of Semantic BITOLA *(SemBT)* [38], Hristovski used semantic predications for LBD. He developed an approach for closed discovery that used ordered sequences of predicates and classes called **discovery patterns**. These discovery patterns were derived from the UMLS semantic network and also using domain expertise. Matching instance-level information was obtained from the semantic predications extracted from
MEDLINE, using SemRep [71]. Hristovski also used gene-related assertions on genotypic and phenotypic relations, extracted using BioMedLEE, to specify discovery patterns [56]. The discovery patterns were used as filters to complement the earlier approach that used confidence and level of support to rank intermediates. Discoveries could arise from intuitive patterns. For example, Hristovski argues in [38] that if a Disease causes a change in a Body Function/Substance, and a Drug inhibits this change, then the Drug MAYBE_TREATS the Disease. The CAUSES-INHIBITS sequence is used as a discovery pattern to discover new Drug treatments for Diseases. This approach was used to rediscover the Raynaud Syndrome–Dietary Fish Oils discovery. It was also used to suggest new insights into the association among Insulin, Diabetes Mellitus, and Huntington Disease. Furthermore, in [40] discovery patterns were used together with DNA Microarray Data to generate novel hypotheses on Parkinson’s Disease.

Ahlers et al. [2] also implemented a relations-based technique for closed discovery that leveraged the idea of discovery patterns. Ahlers suggested that an antipsychotic drug (A) MAYBE_DISRUPTS cancer (C), if the drug INHIBITS some intermediate bioactive, pathological and/or pharmacologic concept (B) that CAUSES, PREDISPOSES or is ASSOCIATED_WITH cancer. The INHIBITS–CAUSES pattern was used as a discovery pattern to infer MAYBE_DISRUPTS and the approach was used to suggest “five biomolecules that may provide a link between the antipsychotic agents and cancer.”

In BioSbKDS, Hu et al. [45] implemented a relations-based approach for closed discovery that automates the selection of discovery patterns, formulated as UMLS relations. The approach used the semantic type of the starting A-term (MeSH descriptors) to find and then expand the candidate set of semantic types for B-terms based on hierarchical relationships in the UMLS. The semantic types for possible C-terms were obtained in the same way, using the types for B-terms. Instances that conform to the automatically selected UMLS relations were obtained from MEDLINE based on articles whose MeSH descriptors matched the semantic types of the MeSH descriptors from the automatically generated
pattern. A special term filter called Bi-Decision Maker, which uses distributional statistics to eliminate generic MeSH descriptors, was then applied to filter non-informative concepts. This approach was used to rediscover several intermediates from the Raynaud Syndrome–Dietary Fish Oils and Magnesium–Migraine discoveries.

Figure 2.4: Complex association between Raynaud Syndrome – Dietary Fish Oils

**Limitations:** In spite of the successes of relations-based approaches, they are mainly suitable for scenarios where both the predicates and semantic types are already known. While Hu’s approach achieves some degree of automation, it is unsuitable for complex situations, in which many associations exist. For example, consider the complex scenario depicted in Figure 2.4. Dietary Fish Oils produce several Prostaglandins including Prostaglandin I3 (PGI₃) and Prostacyclin (PGI₂) – called Epoprostenol, in the synthetic form. One of these Prostaglandins (PGI₂) treats Raynaud Syndrome and also disrupts Platelet Aggregation. Since Platelet Aggregation causes Raynaud Syndrome, one can reasonably conclude that a plausible mechanism by which Dietary Fish Oils treat Raynaud Syndrome is through the production of Prostaglandins, which actively disrupt Platelet Aggregation, deemed a cause of Raynaud Syndrome. This is a complex situation involving several predicates and concepts of different semantic types, which may not be known initially.
The approach by Hu will require additional enhancements to meaningfully construct discovery patterns from UMLS relations to accommodate such complexity. Moreover, it will require new definitions of context to filter out irrelevant instances from the knowledgebase. The latter of these tasks is particularly challenging, since the semantic types of *Prostaglandins* and *Platelet Aggregation* share no common ancestors in their lineage in the UMLS. It is therefore unclear how one might associate them based on hierarchical associations from the UMLS schema, as specified by Hu. Alternatively, it can be argued that distributional approaches could also be used to create such complex subgraphs. However, an approach that can meaningfully capture this level of complexity based on statistical frequency has not been forthcoming in the literature.

**Graph-based**

Contemporary approaches to LBD focus on creating subgraphs [105, 70, 33, 98], using semantic predications. Wilkowski et al. [105] developed a graph-theoretic approach that uses a greedy algorithm to create the ‘best’ subgraph given a graph of predications. Wilkowski’s approach was the first concrete implementation that went beyond the canonical *ABC* model for LBD, and focused on creating paths of semantic predications, in accordance with the *AnC paradigm*.

To generate subgraphs, a starting concept or *A-term* was first selected. This term was then used to generate a graph of predications (or *predications graph*), based on the set of all MEDLINE articles that contain it. Concepts with high degree centrality were then selected from the graph and used to expand the graph, by performing another MEDLINE search. The graph is iteratively expanded into a larger predications graph, then reduced into to a smaller, more well connected subgraph. To achieve this, only high centrality nodes were manually selected for membership. Wilkowski used this method to elucidate the association among *Norepinephrine*, *Depression*, and *Sleep* [105]. In principle, this method could produce subgraphs consisting of highly connected concepts related to the start term. How-
ever, it requires a considerable degree of manual input to discard uninteresting concepts.

By implementing a greedy approach that leverages edge weights for subgraph creation, Wilkowski’s approach resembles to the approach by Ramakrishnan et al. [70], which also relies on weighted edges to create subgraphs. Ramakrishnan’s approach uses an ensemble of structural and semantic features, including class and property specificity, instance level rarity, and refraction to compute edge weights and subsequently generate subgraphs. This approach was applied to a synthetically generated dataset, not comprised of semantic predications, but ontological statements instead. While Ramakrishnan notes that this approach was used to recover the connections from the Raynaud Syndrome–Dietary Fish Oils discovery in exploratory research, its applicability for LBD in general has not been fully demonstrated. It is unclear how this approach might be adapted to create subgraphs on multiple thematic dimensions, given that the algorithm produces a single subgraph using the greedy strategy. Furthermore, since the class specificity property and the span heuristic would need to rely on the UMLS schema, it is unclear how concepts with vastly different semantic types will be clustered into the same subgraph. Reliance on hierarchical relationships in the UMLS Semantic Network is subject to inconsistencies since the UMLS is a terminology and not a formal ontology. Also, by design, the trees in the UMLS Semantic are fairly disjoint, as observed for Prostaglandins and Platelet Aggregation.

Goodwin et al. [33] developed a hybrid approach, which uses spreading activation, relative frequencies (of concepts and semantic predications) and degree centrality for LBD. This approach successfully rediscovered the connection in the Testosterone–Sleep discovery [61], and also elucidated the Norepinephrine, Depression, and Sleep scenario from [105]. Goodwin’s approach generates a single subgraph by capturing the strength of association between concepts. Association strengths are initialized using concept-based and predications-based degree centrality. The spreading activation algorithm is then used to select relevant concepts. Ultimately, Goodwin generates a list of intermediates instead of a graph. This approach does not leverage predicates or provenance, and therefore does not
provide deeper insights into the meaning of the associations. It is unclear how the spreading activation algorithm might be adapted to capture the context of associations on multiple thematic dimensions.

In [98], van der Eijk et al. implemented an approach for LBD that generates subgraphs based on co-occurrence of MeSH descriptors and Hebbian learning. In this approach, each document (or fingerprint) is represented as a vector of MeSH descriptors, and in turn, each MeSH descriptor is in an n-dimensional location vector in an Associative Concept Space. Each descriptor in a document (fingerprint) is randomly located in the space. MeSH descriptors are clustered based on their euclidean distance from the centroid of the fingerprint. This centroid is computed as the average of the location vectors of the descriptors in a document. The A-star algorithm is then used to find paths between nodes in the graph. van der Eijk’s approach was effectively used to provide new insights into the association between Deafness and Macular Dystrophy, and Insulin and Ferritin. Similar to the approach by Ramakrishnan, it is unclear how this approach might capture subgraphs on multiple thematic dimensions.

Figure 2.5: Thematic dimensions of association for Raynaud Syndrome – Dietary Fish Oils

**Limitations:** In spite of the successes of graph-based approaches for LBD, more effective methods for capturing the context of associations are desired. Complex associations between biomedical concepts may exist along multiple thematic dimensions. Figure 2.5 shows that Dietary Fish Oils and Raynaud Syndrome are associated in at least the three
following ways: 1) through Cellular Activity involving Blood platelets/Prostaglandins (Figure 2.5a), 2) through Pharmaceuticals such as Nifedipine and Verapamil (Figure 2.5b), and 3) through Lipids from Efamol and Evening Primrose Oil (Figure 2.5c).

Zhang et al. [111, 112] developed an hybrid approach for document summarization that uses degree centrality to generate clusters of cliques, which has the potential to capture multiple thematic dimensions of associations among concepts. However, the approach is likely to create subgraphs consisting of only highly connected concepts. For LBD, it is understood that discoveries arise from hidden associations among concepts based on context, not necessarily based on frequency of (co)occurrence or connectivity; as noted by Gordon and Dumais in [35]. This dissertation proposes a context-driven subgraph model to automatically create subgraphs on multiple thematic dimensions, using context, not frequency of co-occurrence, graph-theoretic metrics or specificity.

**Hybrids**

A subset of approaches for LBD utilize a hybrid of techniques to define context and find hidden connections. Torvik et al. [96] implemented an approach uses machine learning and several statistical and concept-based features in ARROWSMITH v2. Features are learned from user feedback and by applying binary logistic regression to tune the weights of the parameters. The following seven features were used to predict intermediates: 1) frequency of B-term in MEDLINE, 2) number of common MeSH descriptors in A and C literature for B, 3) frequency of occurrence of B in the A and C literatures, 4) frequency of B-term in A, C relative to its frequency in MEDLINE, 5) year of first occurrence in MEDLINE, 6) UMLS concept (concepthood), and 7) degree of generality/specificity. The logistic regression function learns the parameters that estimate the weight of each feature empirically, based on training data. The output of the approach is a ranked list of intermediates. Torvik used the approach to predict interesting biomedical concepts based on a gold standard of intermediates created from 6 different scenarios.
In the development of Semantic MEDLINE, Kilicoglu et al. [50, 72] implemented an hybrid approach, based on the notion of discovery browsing. Discovery browsing is enabled when a system guides the user through their exploration of the literature in a process of co-operative reciprocity. According to Wilkowski [105], the “user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems.” Semantic MEDLINE uses a hybrid approach, which combines techniques developed for LBD and document summarization. Given a search term, a MEDLINE search is first executed to obtain the relevant articles. The set of semantic predications in the result set is then fed into an abstraction summarizer [27]. The summarizer uses various features, including relevance, connectivity, novelty, and saliency to filter predications, according to high-level summarization perspectives. These perspectives are specified a priori and include Treatment of Disease, Substance Interactions, Diagnosis, and Pharmacogenomics. The reduced set of semantic predications from this ‘semantic condensate’ step, form a predications graph, which can then be explored using UMLS predicates, semantic types, and groups, as additional filters.

**Limitations:** The main issue with the hybrid approach by Torvik is whether the identified features will be effective predictors of not just novel intermediates, but also the related concepts, which help to elucidate complex associations. Additionally, it is unclear whether the parameter weights will prove effective if this technique is applied on a broad scale. Furthermore, it is unclear how this approach might be used to create complex subgraphs on multiple thematic dimensions. Similarly, the more reliable user-centric approach developed in Semantic MEDLINE, may still need some refinements to better enable focused retrieval of interesting associations.
Summary

This Chapter presented an overview of LBD research. It included discussions on general aspects of biomedical text mining, including Information Retrieval (IR), Question Answering (QA), Document Summarization and LBD. It also presented discussions on the capabilities of various biomedical knowledge sources for LBD research. Specifically, these knowledge sources include assertional knowledge extracted from MEDLINE and definitional knowledge provided by the UMLS and MeSH. The Chapter also presented a detailed review of specific approaches in LBD research, including the paradigms, modes and techniques.

The examination of related work brings the evolution of LBD paradigms, over the years, into focus. LBD approaches continue to shift towards richer representations, aimed at capturing longer sequences, and more complex associations. State-of-the-art approaches focus on the construction of subgraphs, which have the potential to provide insights into the mechanism of interaction and causality relationships among concepts. Figure 2.6 depicts this evolution in terms of the five broad categories of LBD techniques previously discussed.

Figure 2.6: Evolution of Literature-Based Discovery Systems and Methodologies

Section 2.3.2 discussed the benefits of concept-based approaches [100, 101, 85, 86, 66, 108], which improve on early keyword-based approaches [36, 53, 34, 95, 82]. Concept-based approaches are able to leverage background knowledge to resolve syntactic variations for concepts from the biomedical literature. Schema knowledge associated with biomedical concepts also enable filtering and grouping of concepts based on shared semantic types. Relations-based approaches for LBD [44, 2, 45] utilize explicit relationships between con-
cepts to provide insights into the meaning of associations. This is an improvement over concept-based approaches, which typically do not explicitly leverage such relationships. Relations-based approaches therefore highlight the inadequacies of prior approaches in modeling and exploiting context. Graph-based techniques for LBD [112, 105, 19] have the capacity to generate complex associations. However, the viability of such techniques to generate multidimensional subgraphs has not been established. Hybrid approaches to LBD [50, 72] may require considerable domain expertise.

2.4 Contributions

In this dissertation, a novel context-driven, automatic subgraph generation method for closed discovery has been developed. The method uses implicit and explicit semantics to generate complex associations along multiple thematic dimensions. The dissertation is based on ideas from [20], in which semantic predications were used to meaningfully connect biomedical text for QA. Our experiments showed that the content of biomedical literature could indeed be connected using semantic predications. The predications were again applied in [19] to rediscover and decompose the Raynaud Syndrome – Dietary Fish Oils discovery. Domain expertise was used to obtain context and subgraphs were manually created. In [21], a method to automatically create the subgraphs was developed, and implemented in a system called Obvio – http://bit.ly/obviodemo.

To automatically create subgraphs, the context of a semantic predication is first specified, and then used to develop the notion of the context of a path, based on the predications that comprise the path. Related paths are then clustered into coherent subgraphs along multiple thematic dimensions, by exploiting the shared context among such paths. To facilitate understanding the meaning of associations in the subgraphs, the predicates of the semantic predications and their provenance in MEDLINE are provided (see Chapter 5).

The method developed in this dissertation advances the state-of-the-art in several spe-
Table 2.1: Comparison of capabilities and accomplishments of LBD techniques

<table>
<thead>
<tr>
<th>System/Technique</th>
<th>Automatic</th>
<th>Relational</th>
<th>Evidence-Based</th>
<th>Thematic</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRIDESCENT [106]</td>
<td>K</td>
<td></td>
<td></td>
<td></td>
<td>1 0</td>
</tr>
<tr>
<td>ARROWSMITH [82]</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>5 0</td>
</tr>
<tr>
<td>DAD [99, 100]</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>0 2</td>
</tr>
<tr>
<td>BITOLA [44]</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>0 1</td>
</tr>
<tr>
<td>LitLinker [108]</td>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td>0 2</td>
</tr>
<tr>
<td>Manjal [85, 86]</td>
<td>C</td>
<td>×</td>
<td></td>
<td></td>
<td>0 5</td>
</tr>
<tr>
<td>SemBT [38, 39, 40]</td>
<td>R</td>
<td>×</td>
<td>×</td>
<td></td>
<td>0 1</td>
</tr>
<tr>
<td>BioSbKDS [45]</td>
<td>R x</td>
<td>×</td>
<td></td>
<td></td>
<td>0 1</td>
</tr>
<tr>
<td>Wilkowski [105]</td>
<td>G x</td>
<td>×</td>
<td></td>
<td></td>
<td>0 0</td>
</tr>
<tr>
<td>Ramakrishnan [70]</td>
<td>G x</td>
<td>×</td>
<td></td>
<td></td>
<td>0 1*</td>
</tr>
<tr>
<td>Zhang [112]</td>
<td>G x</td>
<td>×</td>
<td></td>
<td></td>
<td>0 0</td>
</tr>
<tr>
<td>Obvio [19, 21]</td>
<td>G x</td>
<td>×</td>
<td>×</td>
<td></td>
<td>0 8</td>
</tr>
<tr>
<td>ARROWSMITH v2 [96, 84]</td>
<td>H x</td>
<td></td>
<td></td>
<td></td>
<td>0 6*</td>
</tr>
<tr>
<td>Semantic MEDLINE [61, 18]</td>
<td>H x</td>
<td></td>
<td></td>
<td></td>
<td>2 0</td>
</tr>
</tbody>
</table>

Specific ways. The approach is: 1) automatic, 2) predicate-driven, 3) evidence-based, and 4) thematic. Table 2.4 compares Obvio on these facets, with several other LBD systems. In column 2, the abbreviations \((K, C, R, G, H)\) represent techniques used for LBD, where: keyword-based \((K)\), concept-based \((C)\), relations-based \((R)\), graph-based \((G)\), and hybrid \((H)\). Columns 7 and 8 indicate the number of original discoveries \((DK)\) and rediscoveries \((RK)\) achieved by each system. Each facet is compared in more detail in the following subsections.

1. **Automation:** Table 2.4, column 3 shows that LBD approaches can be grouped according to whether they provide automation. Some LBD approaches [33, 85] implement a ranking mechanism that generates a list of several intermediates. To meaningfully distinguish these intermediates, the user must then restrict the list according to specific semantic types *a posteriori*. In the case of the relations-based approaches applied in *SemBT* [38] and by Ahlers in [2], these restrictions must be specified *a priori*. While this is not a disadvantage in many cases, the approach developed in this dissertation automatically generates a
list of subgraphs and does not require any filtering \textit{a priori}. Hu [45] and Ramakrishnan [70] also outlined approaches that are automatic. However, since both approaches are not based on semantic predications, they offer no supporting evidence from the literature, to help give further insights into the meaning of associations.

2. \textit{Relational}: Table 2.4, column 4 shows that some LBD approaches utilize predicates to help users better understand associations among concepts. The relations-based approaches in \textit{SemBT} [38] and by Ahlers in [2], as well as the approaches by Zhang [112], Wilkowski [105], Miller [61], Goodwin [33], Cameron [19], and Cairelli [18] are all based on semantic predications. Consequently, these approaches utilize predicates to gain insights into the meaning of associations (so does the approach by Ramakrishnan). The method outlined in this dissertation (which includes [19, 20, 21]) also utilizes semantic predications and therefore predicates are provided to users to help expound associations.

3. \textit{Evidence-based}: Table 2.4, column 5 shows LBD approaches can also be distinguished based on the provision of evidence from scientific literature to support the results of the LBD system. Approaches that do not rely on semantic predications mainly utilize the search-and-sift paradigm [74] to obtain details from MEDLINE. Srinivasan, in [85], provided the PMIDs that supported various claims from the search results. However, given that provenance management was not fully integrated into the LBD system, it is likely that provenance was obtained using a pre-processed corpus or using the method of search-and-sift. The method outlined in this dissertation uses the provenance of semantic predications in MEDLINE. In [21], the evaluation shows that the search-and-sift method can also be used to complement provenance. This is especially true for associations not explicitly contained in the subgraphs, but implied. Cairelli [18] and Miller [61] showed the complementarity between the provenance and the search-and-sift method in their approach as part of the discovery browsing paradigm, in Semantic MEDLINE. Both approaches however,
require considerable manual intervention.

4. **Thematic:** Table 2.4, column 6 shows that the approach outlined in this dissertation generates subgraphs on multiple thematic dimensions (see Figure 2.5). Although the approach by Zhang [112] may be adapted to generate such thematic subgraphs, in principle, this has not been clearly demonstrated. The specific contributions of this dissertation are therefore as follows:

1. Proposes a novel context-driven automatic subgraph creation model for closed discovery, based on rich knowledge representations.

2. Presents a novel definition for the context and shared context among associations, which leverages distributional semantics for implicit context and structured background knowledge for explicit context, to compute the semantic relatedness among associations.

3. Implements an unsupervised hierarchical clustering algorithm to automatically create complex subgraphs on multiple thematic dimensions, without the need for complex heuristics or filtering a priori.

4. Illustrates the role of discovery browsing, through the use of predicates and provenance, to supplement the subgraphs with additional insights from the scientific literature.

5. Confirms the effectiveness of this approach, by applying it to facilitate the rediscovery of 8 out of 9 existing scientific discoveries.
Context-Driven Subgraph Model

This dissertation proposes a novel context-driven subgraph model for LBD. The approach is capable of automatically creating complex subgraphs along multiple thematic dimensions. It belongs to the general class of graph-based methods that evolved from the ABC and AnC paradigms. The approach also uses provenance to facilitate the knowledge exploration process, since it is based on assertional knowledge in the form of semantic predications.

Subgraph creation methods that utilize semantic predications for LBD mainly differ based on the definition of context. Otherwise, in general, the first step towards subgraph creation is query specification and document selection. Given a query and the associated documents, semantic predications present in each document can then be used to build a graph, called a predications graph. This initial predications graph is considered inclusive of the broad contexts covering the user interest. To obtain more specific contexts, relevant semantic predications must then be extracted from the predications graph as paths, and used for candidate graph generation. The final subgraph(s) can then be produced by further refining the candidate graph(s) by grouping paths that share some context. Users interact with the generated subgraphs, to gain insights into the meaning of the generated associations. In accordance with this general procedure, the context-driven subgraph model developed in this dissertation consists of four aspects: 1) query specification and document selection, 2) predications graph generation, 3) candidate graph generation, and 4) path clustering.

This Chapter describes the overall context-driven subgraph model. It provides a formal definition of the predications graph in Section 3.2 and discusses the method for candi-
date graph (or path) generation in Section 3.3. Broad aspects of the path clustering process, which is used refine the candidate graph, is discussed in Section 3.4. And in Section 3.5 our first application of this model to rediscover and decompose the Raynaud Syndrome – Dietary Fish Oils discovery is discussed.

3.1 Query Specification and Document Selection

To select a set of relevant documents $D$ for an LBD exercise, the system requires a query, denoted $q$. The query can be specified by first providing the labels of two concepts of interest $(A, C)$. These terms are then mapped to concept unique identifiers (or CUIs) from the UMLS. Such mappings are obtained by using the UMLS Semantic Navigator\(^1\) for initial suggestions, then manually deciding on the appropriate CUI-label pairs. These initial $A$- and $C$- terms are then augmented with closely related concepts. For example, the concepts Fish oil - dietary (C0016157) and Eicosapentaenoic Acid (C0000545) are sufficiently related to Fish Oils (C0016157) to be included in the expanded query. The query may also include a cut-off date $dt$ for the literature to be included. If none is given, the entire corpus (MEDLINE) will be used. The maximum path length $k$ of paths to be generated between $A$ and $C$, may also be specified. If none is given, then maximum path length defaults to $k = 2$. In this way, the system minimally supports $ABC$ scenarios. The complete query format is therefore as follows: $q = (A, C, dt, k)$. An example query for Raynaud Syndrome – Dietary Fish Oils is therefore $q = \{\{\text{Fish Oils}, \text{Fish oil – dietary}, \text{Eicosapentaenoic Acid}\}, \{\text{Raynaud Phenomenon}, \text{Raynaud Disease}\}, 11/01/1985, 3\}$. Given a query, the system then retrieves the matching set of documents $D$ from MEDLINE. The $E$-utilities\(^2\) service provided by NCBI can be used for this task. The semantic predications in the set $D$ can then be used to build the predications graph.

3.2 Predications Graph Definition

A predications graph is a collection of semantic predications, represented as nodes and labeled edges in a relational graph. The notion of the predications graph was first discussed for the biomedical text mining task of QA in [20]. In a feasibility study, the effectiveness of semantic predications for connecting answer passages was conducted. The idea of the predications graph was then applied to LBD in [19], to reproduce and decompose the Raynaud Syndrome – Dietary Fish Oils discovery. And we again used the predications graph for LBD in [21], to facilitate automatic generation of multifaceted subgraphs. In this section, the graph-theoretic underpinnings of the predications graph is formally described.

Using set notation, let \( S(d_i) \) be the set of semantic predications associated with article \( d_i \). If \( t \) denotes a semantic predication in \( d_i \) and \( D = \{d_1, d_2 \ldots, d_n\} \) is the set of articles then,

\[
\text{For any } t = (s_t, p_t, o_t), \text{ let } D(t) = \{d \mid t \in S(d)\}
\]  

be the set of corresponding articles that contain the semantic predication \( t \) and \( S(D) \) be the set of all semantic predications associated with articles in \( D \). That is,

\[
S(D) = \bigcup_{i=0}^{|D|} S(d_i).
\]  

The semantic predications in \( S(D) \), naturally form a directed labeled graph denoted \( G_{S(D)} \), in which the subject \( s_t \) and object \( o_t \) of each semantic predication \( t \) is a vertex and the predicate \( p_t \) is a labeled edge from subject to object. This is called the predications graph.

A typical predications graph can be very large and very difficult to explore. Zhang et al. [112] and Wilkowski et al. [105] used degree centrality for filtering, while Cairelli et al. [18], Miller et al. [61] relied on the underlying summarization mechanism in Semantic MEDLINE to prune the predications graph. Hristovski [38, 39, 40] used confidence and support as statistical metrics, together with discovery patterns for filtering. And in [19]
and [21], we showed that the predications graph can be reduced into a more concise and relevant candidate graph using background knowledge as context. The reduction of the predications graph into a candidate graph is the third aspect of the subgraph model, which is discussed in the next section.

### 3.3 Candidate Graph Generation

“In graph theory, a path in a graph is a finite or infinite sequence of edges which connect a sequence of vertices” [104]. In a predications graph, a path is a finite or infinite sequence of semantic predications. Paths have the capacity to provide insights into the semantics of the association between source and target concepts at the path extremities. Anyanwu and Sheth [4, 5] formally defined a semantic association between two vertices \((v_i, v_j)\) in a graph, in terms of property sequences and joined property sequences. In these semantic associations, a vertex \(v_i\) may be the origin of a property sequence, while another vertex \(v_j\) may be the origin or terminus of another property sequence. The importance of edge direction is not explicit in Anyanwu’s definition. However, for LBD, the direction of labeled edges could be important in understanding the semantics of associations.

![Figure 3.1: Importance of directionality in Semantic Associations](image)

To illustrate this, consider the association shown in Figure 3.1, which shows that **Nifedipine** inhibits **Vasoconstriction** and also treats **Raynaud Syndrome**. Since **Eicosapentaenoic Acid**...
taenoic Acid also inhibits Vasoconstriction, a plausible conclusion is that it might also treat Raynaud Syndrome, and Vasoconstriction may be a cause of Raynaud Syndrome. Note that this conclusion could also be derived using Hristovski’s discovery pattern (discussed in Section 2.3.2), based on the notion of opposing predicates. The INHIBITS-TREATS pattern could be used to infer CAUSES.

The path from Eicosapentaenoic Acid to Vasoconstriction to Nifedipine to Raynaud Syndrome includes edges in opposing directions. Still, the association is meaningful. For LBD therefore, the direction of edges in a path could be significant. Cognizant of this, a semantic association is regarded as a path, consisting of a sequence of labeled edges between a sequence of vertices, oriented in any direction. This is an adaptation of the $\rho$-pathAssociation by Anyanwu in [3]. In this dissertation, a semantic association is formally defined as follows:

Given a directed graph $G=(V, E)$, where $V = \{v_0, v_1, \ldots, v_n\}$ is the vertex set and $E = \{e_0, e_1, \ldots, e_m\}$ is the set of labeled edges, a semantic association exists between the vertex $s=v_i$ and the vertex $t=v_j$ if, for some arbitrary set of vertices $V_p = \{v_i, v_{i+1}, \ldots, v_{i+d}\}$, $V_p \subset V$, each semantic predication is such that $t = \{v_i, e_t, v_j\}$, $(i=0, j=1)$, $(i=1, j=0)$ and $j=i+1$ or $j=i-1$, for all $1 \leq i \leq d$, and $d = |V_p|

This notion of a semantic association creates a basis for extracting the subset of all potentially meaningful paths in the predications graph, between any pair of source $v_i$ and target $v_j$. In both [19] and [21] it was shown that such paths could be used concretely to create the candidate graph, based on the idea of reachability.

Reachability is the notion of being able to get from one vertex in a directed graph, to some other vertex [20, 103]. The adapted definition of a semantic association, therefore assumes that a vertex $v_j$ is reachable from another vertex $v_i$, if the two vertices are semantically associated. That is, if there exists a set of labeled edges between the two vertices,
oriented in any direction. More generally, the set of all semantic associations between a pair of vertices \((v_i, v_j)\) in the vertex set \(V\) of the graph \(G\), is the transitive closure (or reachability relation \(R\)) between \((v_i, v_j)\). The reachability relation between the source and target concepts \(A = v_i\) and \(C = v_j\), in the query \(q = \{v_i, v_j, dt, k\}\) reduces the predications graph \(G\) to the candidate graph (also \(R\)). The path length \(k\) imposes a limit on the maximum length of paths that can be included in the candidate graph. Given this candidate graph, the next aspect of the subgraph model is to cluster paths into subgraphs, based on a specification of path context.

### 3.4 Path Clustering

While generally more focused than the predications graph, the candidate graph may still be overwhelming to explore. To capture more meaningful associations, the candidate graph must appropriately be partitioned into coherent subgraphs, discriminated based on context. We showed in [19] that domain expertise could be used for graph partitioning. A knowledge of related concepts in a domain could be used as context. In [112], Zhang et al. creates clusters of cliques, using degree centrality as the measure of context. In [21], a method that can automatically create subgraphs by using the explicit and implicit semantics of MeSH descriptors was implemented. The approach utilizes the hierarchical agglomerative clustering algorithm to cluster paths that are semantically related.

The overall objective of the clustering activity should be to produce subgraphs that are coherent and thematic. For example, in the Raynaud Syndrome – Dietary Fish Oil scenario, it is a reasonable expectation that a subgraph on Platelet Aggregation should contain associations on Cell Function, while the subgraph on Blood Viscosity might involve various aspects of Lipid Activity. Figure 2.5 shows that associations related to Pharmaceutical Preparations might also be sufficient to generate a separate subgraph.

The specification of context and the method to generate diverse subgraphs, should
depend on the measure of relatedness among semantic predications, not similarity. While some research situations rely on measures of semantic similarity, the scenarios in Sections 1.1.1 – 1.1.3 and Figure 2.4 suggest that finding hidden connections may extend far beyond similarity. With regards to Raynaud Syndrome and Dietary Fish Oils, the lipid compounds Prostaglandins and the process of Platelet Aggregation are in no way similar. Yet, the action of lipids on blood platelets is crucial in understanding a mechanism by which Dietary Fish Oils treat Raynaud Syndrome. Although methods to compute both semantic similarity and relatedness of UMLS concepts have been developed [54, 59], extending them for semantic predications is still a challenge. The context-driven subgraph model for LBD developed in this dissertation relies on a method for capturing relatedness among paths, to generate multifaceted subgraphs. It can be expressed as follows:

Given a predications graph $G=(V,E)$ and a candidate graph $R_{s,t}$ for a source ($s = v_i$) and target concept ($t = v_j$), the set of subgraphs $S_{s,t} = \{R_{s,t}(c_1), R_{s,t}(c_2), \ldots, R_{s,t}(c_n)\}$ derived from $R_{s,t}$ is desired, where $c_i$ is a context from $C = (c_1, c_2, \ldots, c_k)$, such that $c_i \neq c_j, \forall i, \forall j$, and $c_i$ consists of the set of discriminating features $F$.

The definition of the feature set $F$ is problem specific. In this dissertation, one implementation that utilizes domain expertise was developed in [19]. A more robust approach that enables automatic subgraph creation based on MeSH, was developed in [21]. In the next Section the manually-driven approach for subgraph creation, from [19], based on this model is presented. The techniques for automatic subgraph creation, from [21], are discussed in Chapter 4.
3.5 Application to Rediscovery and Decomposition

The context-driven subgraph model for LBD was first applied to reproduce the Raynaud Syndrome – Dietary Fish Oils discovery in [19]. The overall approach included the following steps: 1) literature (document) selection and preprocessing, 2) predications graph generation, 3) candidate graph (or semantic association) generation, and 4) subgraph generation (or path clustering). The subgraphs were manually created and no provenance of semantic predications in MEDLINE was used to substantiate the associations. Instead, the rediscovery activity relied solely on the predicates that were present in the subgraphs for insights into the meaning of associations. Using this approach, all 3 of the primary associations, generally attributed to Swanson, were retrieved. However, an additional 14 out of the 19 associations, which help elucidate the primary associations were discernible from the subgraphs. A method to decompose Swanson’s RS-DFO hypothesis, to this degree, was not previously reported in the literature. In the following subsections, each task is described in detail.

3.5.1 Literature Selection and Preprocessing

The subgraph generation task first required the following query $q = (A, C, dt, k)$ as input. The concepts Fish Oil - dietary (FOD), Fish Oils (FO), Eicosapentaenoic Acid (EPA), were selected as sources (A), while Raynaud Disease (RD) and Raynaud Phenomenon (RP) were selected as targets (C). The maximum path length was set to $k = 3$. The date range for the literature was restricted to $dt = 1985$. This was achieved by including only the 61 articles cited in Swanson’s original Raynaud Syndrome-Dietary Fish Oils paper [89], as being relevant to the discovery.

Some full text articles were obtained manually by searching several online digital libraries, including PubMed, PubMed Central Google Scholar, Mendeley, Wiley Online Library and Science Direct. Altogether, one article (citation #60) was not found, while
another article was written in French (citation #64). An English version of the latter, consisting of the title and abstract was retrieved from PubMed. Twelve articles from [89] contained titles and abstracts only, but no full text (citations 4, 5, 22, 23, 25, 27, 37, 40, 43, 51, 57 and 63). Five articles (citations 1, 8, 46, 56, 59) contained titles and full text only, but no abstract.

This minimal dataset was selected mainly because we were interested in assessing the feasibility of the context-driven subgraph model. We speculated that if our techniques could successfully rediscover existing knowledge, then moving from rediscovery to actual knowledge discovery would require recalibrating our approaches to deal with scalability and performance issues.

This baseline was split into two sets: 1) baseline 1 (B1) consisted of the full text for articles, where available, together with titles and abstracts and 2) baseline 2 (B2) consisted of titles and abstracts only. The choice of the two datasets was to determine the extent to which titles and abstracts might be adequate for LBD. The articles were formatted to remove special characters and to unify character encodings, and prepare the text for semantic predication extraction. To extract predications, the articles were processed in batch mode using SemRep.

3.5.2 Predications and Candidate Graph Creation

SemRep returned 4434 semantic predications and 1077 unique concepts in baseline B1, while baseline B2 contained 388 semantic predication and 192 concepts. In the output it was observed that SemRep erroneously parsed the phrase “Vascular Reactivity” (C1660757) into two tokens: 1) “Vascular” (Blood Vessels-C0005847) and 2) “Reactiv-ity” (Reactive-C0205332). This problem was also detected by Hristovski et al. in [38] in his attempt at reproducing Swanson’s Raynaud Syndrome-Dietary Fish Oils hypothesis using semantic predications. Hristovski did not solve this problem, and consequently only reported results on Platelet Aggregation and Blood Viscosity. After inspecting the
SemRep output, we observed that only 26 unique mentions of *Vascular Reactivity* (and its lexical variants) occurred in this corpus. We therefore manually augmented the output by correcting these misses, using the postfix “*MAN*” to distinguish manually included semantic predications. For example, the predicate *INHIBITS* *MAN* was used to represent the manually included version of the predicate *INHIBITS*.

This action enabled the flexibility of further distinguishing between *Vasoconstriction* and *Vasodilation* for *Vascular Reactivity* in the SemRep output. Consequently, since *Vasoconstriction* and *Vasodilation* convey opposite semantics, two semantic predications were added to the result set; one for each concept. For example, two semantic predications were added for the semantic predication, which states that [*Epoprostenol INHIBITS Vascular Reactivity*]. These were [*Epoprostenol INHIBITS MAN Vasoconstriction*] and [*Epoprostenol CAUSES MAN Vasodilation*]. These manual additions resulted in an additional 76 semantic predications added to *B1* and one to *B2*.

Given this predication graph *G* constructed from the 61 articles, the next task was creation of the candidate graph. To accomplish this, paths up to length \( k = 3 \) were extracted using the notion of reachability. The depth first (DFS) algorithm was used to implement reachability. Each source-target pair from *A* and *C* were used as the vertices in *G*, from which the transitive closure, up to length *k*, was extracted. The aggregate of all generated paths was used as the candidate graph. There were 189 paths in the candidate graph for *B1* and 34 for *B2*.

### 3.5.3 Subgraph Creation

The goal of the rediscovery task was to determine whether the semantic predications could be used to recover the original associations noted by Swanson in [89]. Coauthors from [19] thoroughly read this original article and manually identified detailed associations, conveyed in natural language by Swanson. These associations were formulated as sequences of semantic predications. For example, in [89] (p. 4), in the Section entitled: “The Effects of
Dietary Fish Oil on Blood Viscosity, Platelet Function and Vascular Reactivity,” Swanson asserted that: “It is known that EPA can suppress platelet aggregation by several different mechanisms, though which among them are of greatest importance is not known.” From this sentence the semantic predication [Eicosapentaenoic Acid INHIBITS Platelet Aggregation] was created, since EPA is a synonym for Eicosapentaenoic Acid, and the term “suppress” corresponds to the UMLS predicate “INHIBITS.”

By reading the article, the three primary associations noted by Swanson (involving Platelet Aggregation, Blood Viscosity and Vascular Reactivity) were expressed as semantic associations. These are shown with a light purple background in Table 3.1 – path IDs: #1, #2, #3. Through this process, eight supplementary associations, which expound the three primary associations were also identified. These are shown in Table 3.1, with path ID:1a on Platelet Aggregation, path IDs:2a-c on Blood Viscosity and path IDs:3a-d on Vascular Reactivity. Eight additional associations called secondary associations were also found, which directly involve Raynaud Syndrome and other concepts related to Dietary Fish Oils. These include Ketanserin, Nifedipine and Alprostadil, which are shown in Table
3.3 (path IDs: 1.1-2, 2.1-3, & 3.1-3). The number of associations that could be recovered by manually creating the subgraphs, using semantic predications extracted from these 61 articles is reported in the experimental results.

3.5.4 Experimental Results

The goal of the experiments was to create subgraphs that would match as many of the three primary, eight supplementary and eight secondary associations as possible. For the first experiment (based on full text articles from $B_1$) there was a total of 17848 associations in the candidate graph for which Eicosapentaenoic Acid (EPA) was the source, but only 172 of these terminated with Raynaud Syndrome (i.e., Raynaud Disease (RD)=48, Raynaud Phenomenon (RP)=124). There were 2124 associations in the candidate graph for which Fish Oils (FO) was the source, but only 14 terminated with Raynaud Syndrome (i.e., RD=3, RP=11). Then there were 382 associations for which Fish Oils - dietary (FOD) was the source, but only three associations terminated with Raynaud Syndrome (i.e., (RD)=1, (RP)=2). Altogether, there was a total of 189 associations for which Dietary Fish Oils was the source and Raynaud Syndrome was the terminal. The 189 paths were manually grouped into subgraphs based on their context, using human judgements. This Section begins by comparing the manually created subgraphs for $B_1$ on Platelet Aggregation, with the associations created by reading Swanson’s article [89] (i.e., our gold standard dataset).

Experiment 1: Platelet Aggregation

Swanson’s first primary association shown in Table 3.1 (ID: #1, repeated in Table 3.2) states that [Dietary Fish Oils INHIBITS Platelet Aggregation] and high levels of [Platelet aggregation CAUSES Raynaud Syndrome]. He further claimed that [Dietary Fish Oils
<table>
<thead>
<tr>
<th>Assoc.</th>
<th>#1</th>
<th>Dietary Fish Oil → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph 77</td>
<td>EPA DISRUPTS Platelet aggregation</td>
<td>Epoprostenol DISRUPTS Platelet aggregation</td>
</tr>
<tr>
<td>78</td>
<td>Epoprostenol INHIBITS_MAN Platelet aggregation</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Epoprostenol PREVENTS_MAN Platelet aggregation</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Epoprostenol TREATS RD</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>Epoprostenol TREATS RP</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>EPA STIMULATES Epoprostenol</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>1a</th>
<th>Dietary Fish Oil → PRODUCES → Prostaglandin (PGI₃) → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph 37</td>
<td>EPA CONVERTS_TO prostaglandin I₃</td>
<td>prostaglandin I₃ ISA Epoprostenol</td>
</tr>
<tr>
<td>38</td>
<td>Epoprostenol ISA Prostaglandins</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>EPA CONVERTS_TO Prostaglandins</td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>Epoprostenol DISRUPTS Platelet aggregation</td>
<td></td>
</tr>
<tr>
<td>77</td>
<td>Epoprostenol INHIBITS_MAN Platelet aggregation</td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>Epoprostenol PREVENTS_MAN Platelet aggregation</td>
<td></td>
</tr>
<tr>
<td>80</td>
<td>Epoprostenol TREATS RD</td>
<td></td>
</tr>
<tr>
<td>81</td>
<td>Epoprostenol TREATS RP</td>
<td></td>
</tr>
<tr>
<td>83</td>
<td>EPA STIMULATES Epoprostenol</td>
<td></td>
</tr>
<tr>
<td>84</td>
<td>Epoprostenol TREATS RD</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>#2</th>
<th>Dietary Fish Oil → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph 71</td>
<td>EPA DISRUPTS blood viscosity</td>
<td>Ketanserin DISRUPTS blood viscosity</td>
</tr>
<tr>
<td>175</td>
<td>Ketanserin TREATS RP</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>EPA STIMULATES Epoprostenol</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>Epoprostenol TREATS RD</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>2a</th>
<th>Dietary Fish Oil → INHIBITS → triglyceride → ISA → Blood Lipid → AFFECTS → blood viscosity → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph 71</td>
<td>EPA DISRUPTS blood viscosity</td>
<td>Ketanserin DISRUPTS blood viscosity</td>
</tr>
<tr>
<td>175</td>
<td>Ketanserin TREATS RP</td>
<td></td>
</tr>
<tr>
<td>135</td>
<td>EPA STIMULATES Epoprostenol</td>
<td></td>
</tr>
<tr>
<td>150</td>
<td>Epoprostenol TREATS RD</td>
<td></td>
</tr>
<tr>
<td>111</td>
<td>EPA ISA Fatty Acids</td>
<td></td>
</tr>
<tr>
<td>112</td>
<td>Epoprostenol STIMULATES Fatty Acids</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>2b</th>
<th>Dietary Fish Oil → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>2c</th>
<th>Dietary Fish Oil → INHIBITS → Serotonin → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assoc.</td>
<td>#3</td>
<td>Dietary Fish Oil → INHIBITS → vascular reactivity → CAUSES → Raynaud Syndrome</td>
</tr>
<tr>
<td>--------</td>
<td>----</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Subgraph 25</td>
<td>EPA AFFECTS_MAN Vascular constriction</td>
<td>Epoprostenol INHIBITS_MAN Vascular constriction</td>
</tr>
<tr>
<td>26</td>
<td>Prostaglandin (PGI₃) INHIBITS Platelet aggregation → CAUSES → vasoconstriction → CAUSES → Raynaud Syndrome</td>
<td></td>
</tr>
</tbody>
</table>

**PRODUCES Prostaglandin (PGI₃)** and **[Prostaglandin (PGI₃) INHIBITS Platelet Aggregation]** as the method by which **[Platelet Aggregation CAUSES Raynaud Syndrome]**. Among the 189 associations, eight associations selected by our domain experts (shown in Table 3.2, IDs: 77, 78, 80, 81, 83, 84, 135 and 150), were deemed relevant to this claim. The subgraph in Figure 3.2 was then constructed from these eight associations (inferred predicates are shown in double green lines).

Figure 3.2 shows that **[Eicosapentaenoic Acid STIMULATES Epoprostenol]**, which treats **Raynaud Syndrome**. It also shows that **[Epoprostenol DISRUPTS Platelet Aggregation]**. It is therefore plausible that one mechanism by which **[Epoprostenol TREATS Raynaud Syndrome]** is by disrupting **Platelet Aggregation**, deemed a cause of the disease. It follows by abduction that since **[Prostaglandin (PGI₃) ISA Epoprostenol]**, it also treats **Raynaud Syndrome** by disrupting **Platelet Aggregation**. Furthermore, since **[Epoprostenol ISA Prostaglandin]**, we can also infer that **[Dietary Fish Oils TREATS Raynaud Syndrome]** by producing a series of **Prostaglandin**, which disrupt **Platelet Aggregation**. The latter conclusion is supported by the first secondary association noted by Swanson, in which he claims that another **Prostaglandin (PGE₁, or Alprostadil)** also treats **Raynaud Syndrome** by inhibiting **Platelet Aggregation**. Figure 3.3 constructed from paths with IDs:1.1 & 73 in Table 3.2, shows the subgraph in which **[Alprostadil DISRUPTS Platelet Aggregation]** and
Table 3.3: Secondary Associations

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>1.1</th>
<th>Prostaglandin (PGE$_1$) → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>73</td>
<td>EPA DISRUPTS Platelet aggregation</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>1.2</th>
<th>Prostacyclin (PGI$_2$) → INHIBITS → platelet aggregation → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>77</td>
<td>EPA DISRUPTS Platelet aggregation</td>
</tr>
<tr>
<td></td>
<td>78</td>
<td>Epoprostenol INHIBITS_MAN Platelet aggregation</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>Epoprostenol PREVENTS_MAN Platelet aggregation</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>Epoprostenol TREATS RD</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>Epoprostenol TREATS RP</td>
</tr>
<tr>
<td></td>
<td>84</td>
<td>EPA STIMULATES Epoprostenol</td>
</tr>
<tr>
<td></td>
<td>135</td>
<td>Epoprostenol TREATS RP</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>Epoprostenol TREATS RP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>1.3</th>
<th>Nifedipine → ISA → CCB → INHIBITS → platelet activation → ASSOCIATED_WITH → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>91</td>
<td>EPA DISRUPTS Platelet function</td>
</tr>
<tr>
<td></td>
<td>140</td>
<td>EPA STIMULATES Epoprostenol</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>2.1</th>
<th>Prostacyclin (PGI$_2$) → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome (in vitro)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>71</td>
<td>EPA DISRUPTS blood viscosity</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>Fish Oils AFFECTS blood viscosity</td>
</tr>
<tr>
<td></td>
<td>135</td>
<td>EPA STIMULATES Epoprostenol</td>
</tr>
<tr>
<td></td>
<td>150</td>
<td>Epoprostenol TREATS RD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epoprostenol TREATS RP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>2.2</th>
<th>Ketanserin → INHIBITS → serotonin → AUGMENTS → Erythrocyte Deformability → INHIBITS → blood viscosity → CAUSES → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>71</td>
<td>EPA DISRUPTS blood viscosity</td>
</tr>
<tr>
<td></td>
<td>175</td>
<td>Fish Oils AFFECTS blood viscosity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ketanserin TREATS RP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>3.1</th>
<th>Prostaglandin (PGE$_1$) → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>3.2</th>
<th>Prostacyclin (PGI$_2$) → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>25</td>
<td>EPA AFFECTS_MAN Vasoconstriction</td>
</tr>
<tr>
<td></td>
<td>26</td>
<td>Epoprostenol TREATS RD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epoprostenol TREATS RP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Assoc.</th>
<th>3.3</th>
<th>Nifedipine → ISA → Calcium Channel Blocker → CAUSES → vasodilation → INHIBITS → Raynaud Syndrome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subgraph</td>
<td>35</td>
<td>EPA AFFECTS_MAN Vasoconstriction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nifedipine TREATS RP</td>
</tr>
</tbody>
</table>
interacts with (and likely treats) Raynaud Syndrome.

Swanson also presented another secondary association (Table 3.2, ID:1.3) in which the drug Nifedipine (a Calcium Channel Blocker), inhibits Platelet Activation and so TREATS Raynaud Syndrome. Among the 189 paths, we found one association (Table 3.2, ID: 91) which states that [Nifedipine DISRUPTS Platelet function]. We also found a second association (Table 3.2, ID:140) which states that [Nifedipine TREATS Raynaud Phenomenon].
Since [Prostaglandin TREATS Raynaud Syndrome] by inhibiting Platelet Aggregation we can also surmise by abduction that [Nifedipine also TREATS Raynaud Phenomenon] as a result of disrupting Platelet Function. The knowledge that Nifedipine ISA Calcium Channel Blocker, which is not explicit in the subgraph, could be determined based on knowledge of the domain.

![Figure 3.4: Secondary Association: Platelet Aggregation](image)

**Experiment 1: Blood Viscosity**

The second primary association between Dietary Fish Oils and Raynaud Syndrome, claims that [Dietary Fish Oils INHIBITS Blood Viscosity] and [Blood Viscosity CAUSES Raynaud Syndrome] (shown in Table 3.1, ID: 2). We found four associations (Table 3.2, ID: 71, 135, 150 and 175), which were deemed relevant to this claim, and from them constructed the subgraph shown in Figure 3.5. By application of Hristovski's discovery pattern from [38] to this subgraph, we can surmise by abduction that if [Epoprostenol TREATS Raynaud Syn-
and [Ketanserin TREATS Raynaud Syndrome], given that [Ketanserin DISRUPTS Blood Viscosity], then perhaps [Epoprostenol TREATS Raynaud Syndrome] by also disrupting Blood Viscosity. If this is true, it follows that [Dietary Fish Oils DISRUPTS Blood Viscosity] by producing Epoprostenol. We can therefore conjecture that Blood Viscosity may also be another cause of Raynaud Syndrome.

Figure 3.5: Primary Association Subgraph: Blood Viscosity

Swanson’s first supplementary association involving Blood Viscosity (Table 3.2, ID: 2a) claims that Dietary Fish Oils inhibits various Blood Lipids (specifically Triglycerides) which directly (or indirectly) increase Blood Viscosity. We found six associations (Table 3.2, ID: 71, 111, 112, 135, 150, 175) deemed relevant to this observation. From them, we constructed the subgraph in Figure 3.6. In this subgraph, notice that [Dietary Fish Oils ISA Fatty Acid] and Dietary Fish Oils also disrupts Blood Viscosity. However, from background knowledge, it is known that Dietary Fish Oils is an essential fatty acid which is known to exhibit several health benefits. One major benefit is the inhibition of Triglycerides (also a Lipid). This inhibition AFFECTS Blood Viscosity. It follows by abduction that [Dietary Fish Oils TREATS Raynaud Syndrome] by inhibiting Lipids, thereby lowering Blood Viscosity.
Swanson discussed two additional supplementary associations, which provide further insights into the role of Dietary Fish Oils in inhibiting Blood Viscosity. In the first (Table 3.2, ID: 2b), he claims that Fish Oils influence the ability of red blood cells to alter their shape under fluid pressure. That is, [Dietary Fish Oils AUGMENTS Erythrocyte Deformability], [Erythrocyte Deformability INHIBITS Blood Viscosity], and high [Blood Viscosity CAUSES Raynaud Syndrome]. In yet another supplementary association (Table 3.2, ID: 2c) Swanson claims that Dietary Fish Oils INHIBITS Serotonin and it is this inhibition that augments Erythrocyte Deformability. Unfortunately, we did not recover any of these two associations.

Finally, Swanson also discussed another secondary association involving blood viscosity (Table 3.3, ID: 2.2) in which [Ketanserin INHIBITS Serotonin] and hence augments Erythrocyte Deformability. While this association was not recovered at this level of granularity, two associations (Table 3.3, ID: 71, 175) provided evidence that [Ketanserin TREATS Raynaud Syndrome] by inhibiting Blood Viscosity.
**Experiment 1: Vascular Reactivity**

Swanson’s third primary association (Table 3.1, Table 3.2, ID: #3) claims that *Dietary Fish Oils TREATS Raynaud Syndrome* by inhibiting Vascular Reactivity (i.e., vasoconstriction, *Vascular constriction (function)*) which causes *Raynaud Syndrome*. Among the 189 paths, we found two relevant associations (Table 3.2, ID: 25, 26) from which the subgraph shown in Figure 3.7 was constructed. This subgraph shows that *Eicosapentaenoic Acid AFFECTS MAN Vascular constriction (function)*, *Epoprostenol INHIBITS MAN Vascular constriction (function)* and *Epoprostenol TREATS Raynaud Syndrome*. Since, we know that *Dietary Fish Oils STIMULATES Epoprostenol* and *Epoprostenol also inhibits Vasoconstriction*, while also treating *Raynaud Syndrome*, we again conjectured that perhaps *Vasoconstriction CAUSES Raynaud Syndrome*.

![Figure 3.7: Primary Association Subgraph: Vascular Reactivity](image)

The first supplementary association involving *Vasoconstriction* (Table 3.2, ID: 3a) states that *Dietary Fish Oils PRODUCES PGI₃*, *PGI₃ CAUSES Vasodilation*, and *Vasodilation INHIBITS Raynaud Syndrome*. The second supplementary association (Table 3.2, ID: 3b) is similar, except it states that *PGI₃ also INHIBITS vasoconstriction* and thereby inhibiting *Raynaud Syndrome*. We found six associations among the 189 paths in the candidate graph (Table 3.2, ID: 25, 26, 37, 38, 66, 67) deemed relevant to this association. From Figure 3.8 we can surmized that perhaps *Epoprostenol TREATS Raynaud Syndrome* by inhibiting Vasoconstriction. If *Dietary Fish Oils PRODUCES Prostaglandin*
and \( \text{Prostaglandin (PGI}_3 \) ISA Epoprostenol}, given that \( \text{Epoprostenol INHIBITS Vasoconstriction} \), \( \text{Prostaglandin (PGI3)} \) possibly also inhibits \( \text{Vasoconstriction} \). Hence, it may be true that \( \text{Epoprostenol TREATS Raynaud Syndrome} \) by inhibiting \( \text{Vasoconstriction} \).

We did not recover the third or fourth supplementary associations (Table 3.2, ID: 3c, 3d) and hence found no evidence that \( \text{Prostaglandin (PGE}_1 \) CAUSES Vasodilation} \) by inhibiting \( \text{Platelet Aggregation} \), although this could be inferred from Figures 3.2 and 3.8 using Hristovski’s discovery patterns.

![Figure 3.8: Supplementary Association Subgraph: Vascular Reactivity](image)

Finally, Swanson also discussed several secondary associations involving \( \text{Vascular Reactivity} \) and \( \text{Raynaud Syndrome} \). Again, we could not recover the secondary association in Table 3.3, ID: 3.1, due the absence of predication which states that \( \text{Prostaglandin (PGE}_1 \) CAUSES Vasodilation} \).

For the secondary association in Table 3.3, ID: 3.2, we found two associations (Table 3.3, ID: 25, 26) from which we conjectured that since \( \text{Epoprostenol INHIBITS Vasoconstriction} \) then \( \text{Epoprostenol CAUSES Vasodilation} \) which inhibits \( \text{Raynaud Syndrome} \). For the last secondary association (Table 3.3, ID: 3.3), we found one association (Table 3.3, ID: 35) from which we conjectured that since \( \text{Nifedipine INHIBITS Vasoconstriction} \)
it causes Vasodilation, which inhibits Raynaud Syndrome. It is known from background knowledge that [Nifedipine ISA Calcium Channel Blocker].

Altogether, were 14 out of the 19 associations were recovered using baseline 1, which was comprised of semantic predications extracted from the full text of the relevant articles. These includes: 1) the three top-level associations (Table 3.4, row 1, column 3) depicted Figures 3.2, 3.5 and 3.7 respectively, 2) 4 out of 8 supplementary associations (row 2, column 3), and 3) 7 out of 8 secondary associations (row 3, column 3). The results from Experiment II that use baseline 2, consisting of titles and abstracts only, show that the full text was much more informative. All datasets and experimental results are also available online³ (detailed in Appendix A).

Table 3.4: Raynaud Syndrome–Dietary Fish Oils Intermediate Rediscovery Comparison

<table>
<thead>
<tr>
<th>Association Type</th>
<th>Swanson [89] (original article)</th>
<th>Cameron et al. [19]</th>
<th>State-of-the-art [38]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Supplementary</td>
<td>8</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Secondary</td>
<td>8</td>
<td>7</td>
<td>0</td>
</tr>
</tbody>
</table>

3.6 Summary

This Chapter presented a broad outline of the context-driven subgraph model for LBD, together with a specific application of its use, to rediscover and decompose the Raynaud Syndrome–Dietary Fish Oils discovery. The application utilized SemRep to extract semantic predications from the minimal set of 61 articles noted as critically relevant by Swanson in [89]. The predications were used to create predications graphs for two baseline datasets: 1) from titles, abstracts and full text of scientific articles and 2) consisting of titles and abstract only (see experimental results in A). The depth first algorithm was then used to

³Datasets - wiki.knoesis.org/index.php/Obvio#RaynaudSyndrome-DietaryFishOils_Hypothesis
reduce the predications graph to a candidate graph, exploiting an adapted definition of a semantic association and the notion of reachability as outlined in [19]. Experiments showed that on baseline 1, all three primary associations, 4 out of 8 supplementary associations and 7 out of 8 secondary associations could be retrieved using this approach. Altogether, 14 out of 19 associations could be recovered using this method.

The supplementary associations IDs:2b, 3a, 3c and 3d were not found. Neither was the secondary association with ID:3.1. The secondary association 2b was not found because the concept Erythrocyte Deformability was not extracted by SemRep and was therefore not present in the candidate graph. The secondary association 3a was not found because the assertion which states that \([\text{Prostaglandin (PGI}_3] \text{ CAUSES Vasodilation}]\) was also not present in the candidate graph. The supplementary association 3c was not found because the association which states that \([\text{Dietary Fish Oils PRODUCES Prostaglandin (PGE}_1])\) was not present in the candidate graph. For secondary association 3.1, the assertion which states that \([\text{Prostaglandin (PGE}_1] \text{ CAUSES Vasodilation}]\) was not present in the candidate graph. The supplementary association 3d was not found because the assertion which states that \([\text{Dietary Fish Oils INHIBITS Serotonin}]\) was not present in the candidate graph.

These results suggest that the use of semantic predications for subgraph creation was effective, not only for recovery, but also to decompose the hypothesis to a degree that had never been shown in the scientific literature. The overall approach is significant for three specific reasons: 1) This was the first experiment to conduct a detailed decomposition of Swanson’s Raynaud Syndrome-Dietary Fish Oils Hypothesis into 19 associations, compared with recovery approaches limited to only the 3 well known associations. 2) Similar to Wilkowski, this approach also showed that the classical ABC model for LBD, can be extended into an more expressive AnC model, which could be more effective for LBD. 3) Further, it showed that meaningful, complex associations could be created along different thematic dimensions, which give broader insights into the nature of associations between concepts.
The main limitation of this approach is the reliance on domain expertise for context, and the absence of a methodology for automatically creating the subgraphs. Overcoming this obstacle requires a definition for context and shared context that would allow automatic partitioning of the candidate graph into meaningful subgraphs, without manual input. This task of automatic subgraph creation is addressed in Chapter 4, and is a key contribution of this dissertation.
Automatic Subgraph Creation

This chapter presents an approach to automatically create subgraphs, using the context-driven subgraph model outlined in Chapter 3. The approach draws from ideas on “semantics for the semantic web” [75]. In discussing the semantic web, Sheth et al. suggested that information processing systems, designed to harness the knowledge expressed in text, will benefit from the use of *implicit*, *explicit* (or formal), and *powerful* semantics when analyzing data. In this dissertation, both implicit and explicit semantics are used to capture the context required for clustering paths and elucidating hidden complex associations. **Implicit semantics** are used first to define the context of a semantic predication – as the distribution of MeSH descriptors associated with the predication in MEDLINE. Predication context is then used to infer the context of a path – as the aggregation of MeSH descriptors from the contexts of each constituent semantic predication in the path. Hierarchical relationships in MeSH are then used to glean **explicit semantics**, which provide additional context across inexact pairs of MeSH descriptors. Ultimately, subgraphs are created by clustering paths that exceed some threshold of semantic relatedness based on the implicit and explicit context derived from MeSH descriptors. This Chapter presents details regarding the use of implicit and explicit semantics for automatic subgraph creation. Sections 4.1-4.3 discuss aspects of path relatedness computation, including the specification of context and shared context among paths. Section 4.4 describes the clustering algorithm, including automatic threshold selection and subgraph ranking.
4.1 Path Relatedness

The fundamental idea behind the approach for automatic subgraph creation is that assertions in scientific literature, which are related, typically occur in similar contexts. To illustrate this point, consider the interplay between relatedness and context for a specific complex scenario.

The Calcium Channel Blockers (CCBs), Nifidepine and Verapamil are related to Raynaud Syndrome (RS) – so are Eicosapentaenoic Acid (EPA), Prostaglandins, and the process of Platelet Activation. In particular, it was reported in the article [PMID3157318] by Malamet et al. that patients with Raynaud Syndrome had high levels of Platelet Activation. Likewise, Sauza et al., in [PMID6376801], Kahan et al., in [PMID6352267], and Rodeheffer et al., in [PMID6339921], confirmed that EPA, Nifidepine, and Verapamil were implicated in treating Raynaud Syndrome. Malamet suggested specifically that the CCBs treated Raynaud Syndrome by disrupting Platelet Activation; an assertion also supported by Goodnight et al., in [PMID7295999]. On the other hand, various Prostaglandins, which are produced by EPA, also treat Raynaud Syndrome. Saynor et al. reported in [PMID6320840] that EPA stimulates Prostaglandin (PGI₃), which in turn disrupts Platelet Activation. CCBs and EPA are therefore related in the context of treatment of Raynaud Syndrome. They both disrupt Platelet Function as the mechanism for treating Raynaud. CCBs provide pharmaceutical treatment, while EPA provides dietary treatment.

The crucial challenge for automatic subgraph creation is capturing the shared context among such associations, and then using shared context to create complex and diverse subgraphs. To achieve this, consider each association expressed more succinctly as paths. Let $p_1 = \text{CCB} \rightarrow \text{DISRUPTS} \rightarrow \text{Platelet Activation} \rightarrow \text{CAUSES} \rightarrow \text{RS}$, and $p_2 = \text{EPA} \rightarrow \text{STIMULATES} \rightarrow \text{Prostaglandins} \rightarrow \text{INHIBITS} \rightarrow \text{Platelet Activation} \rightarrow \text{CAUSES} \rightarrow \text{RS}$.

Since a path consists of a sequence of semantic predications, the problem of path relatedness can be decomposed into one of first expressing the context of a semantic predication.
Path context can then be inferred from semantic predication context. Ultimately, subgraphs can be created by clustering paths whose semantic relatedness exceed a certain threshold of relatedness, computed based on the combination of implicit and explicit context.

Distributional statistics, such as those implemented in Singular Value Decomposition (SVD), Latent Semantic Indexing (LSI), and Mutual Information (MI), may be used alternatively, for subgraph creation. The main issue is that the frequency of co-occurrence of semantic predications across MEDLINE, is sparse. Furthermore, while it may be plausible that semantic predications that frequently (or rarely) co-occur, are highly related, it is difficult to extend this idea to moderately co-occurring semantic predications in MEDLINE. As suggested by Gordon and Dumais in [35], distributional frequency alone is unlikely to distinguish context sufficiently among moderately co-occurring semantic predications, to reliably facilitate creation of meaningful, complex, and diverse subgraphs.

It may also be argued that schema relations from background knowledge can be used effectively for subgraph creation, as illustrated by Ramakrishnan et al. [70] (albeit on a synthetically generated knowledgebase). The issue with Ramakrishnan’s approach is that effectively defining predication context using relations from the UMLS is subject to inconsistencies. The UMLS semantic network is not a formal ontology. It contains various inconsistencies among semantic types. Moreover, the trees in the semantic network are fairly disjoint, by design. The semantic type of Platelet Aggregation is Cell Function, while the semantic types for Prostaglandins are Eicosanoid, Pharmacologic Substance, and Biologically Active Substance. These semantic types do not converge in the UMLS Hierarchy\(^1\). Further, the semantic groups in the semantic network, which segregate the instances in the UMLS metathesaurus, are at a high level of abstraction. Consequently, for a given semantic group, only highly related concepts share highly related semantic types [13], that belong to the given group. For these reasons, use of the hierarchical relations in the UMLS, to glean context for automatic multi-faceted subgraph creation, may be cumbersome.

The core of the approach developed in this dissertation is based on the representation of the context of a semantic predication in terms of MeSH descriptors and leveraging implicit and explicit context, to cluster paths, based on the aggregation of the various predication contexts. Specification of predication context is covered in the next Section.

4.2 Predication Context

Analysis of the properties of MEDLINE articles reveals several important observations, which are crucial to the definition of predication context. These observations involve: 1) attributes of MEDLINE articles and 2) levels of abstraction of the content.

The 65 attributes\(^2\) used to annotate MEDLINE articles can be divided into two broad categories. The first category contains provenance metadata, such as title, author, affiliation, date of publication, and journal title (see Figure 4.1). The second category contains semantic attributes, which give insights into the meaning of the content of the article, such as the MeSH descriptors, supplementary concept records (SCRs), and gene symbols.

MEDLINE articles (i.e., abstracts of full text) provide a level of abstraction of the full text. Consequently, they provide a textual semantic summary of the salient points of the publication. For example, Figure 4.1 shows the 1984 article by Kristensen et al. [PMID6098049], whose title suggests that the production of a Prostacyclin (PGI\(_2\))-like material in the Human Umbilical Veins, increases significantly when an EPA perfusate is used together with an Arachidonic Acid perfusate, as opposed to use of Arachidonic Acid alone. EPA is therefore said to have potentially “beneficial antithrombotic properties in human.” The abstract (or textual semantic summary) of the full text confirms in a sentence that “perfusion with HBA containing 10 mumol/l AA plus 10 mumol/l EPA resulted in a significant greater production of PGI2-like material than perfusion with HBA-AA alone” (highlighted in light purple).

MeSH descriptors (and SCRs – shown as Substances in Figure 4.1) are also intended to reflect the salient aspects of the content of an entire article – albeit at a higher level of abstraction. MeSH descriptors can be thought of as a concept-level semantic summary instead of a textual summary. In particular, for the given article, the descriptors Arachidonic Acids, Eicosapentaenoic Acid, Epoprostenol (Prostacyclin), Humans and Umbilical Veins (Figure 4.1, bottom left), suggest some implicit association among them.

Similar to the MeSH descriptors, the semantic predications extracted from the article can also be perceived as a semantic summary of the content. These predications provide a relational semantic summary of the content, by virtue of expressing the explicit predicates between the concepts. In particular, the semantic predication [Eicosapentaenoic Acid STIMULATES Epoprostenol] (Figure 4.1, bottom right), captures the main point of the article, asserted in the abstract. That is, the finding that EPA potentiates the production of a Prostacyclin-like material in Humans.

The context of a scientific article in MEDLINE can therefore be abstracted in terms of its textual, concept-level, and relational semantic summary. Each of these equivalent
Table 4.1: Terminology

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d$</td>
<td>MEDLINE article</td>
</tr>
<tr>
<td>$m$</td>
<td>MeSH descriptor</td>
</tr>
<tr>
<td>$t$</td>
<td>Semantic Predication</td>
</tr>
<tr>
<td>$p$</td>
<td>Path</td>
</tr>
<tr>
<td>$Z$</td>
<td>Number of paths for an experiment</td>
</tr>
<tr>
<td>$N$</td>
<td>Number of Unique MeSH descriptors $\approx 27000$</td>
</tr>
<tr>
<td>$M(d)$</td>
<td>MeSH descriptors of a MEDLINE article</td>
</tr>
<tr>
<td>$c(t)$</td>
<td>Context set of a semantic predication</td>
</tr>
<tr>
<td>$c(p)$</td>
<td>Context vector of a semantic predication</td>
</tr>
<tr>
<td>$C(p)$</td>
<td>Context set of a path</td>
</tr>
<tr>
<td>$C'(p)$</td>
<td>Context vector of a path</td>
</tr>
<tr>
<td>$f(t, m_i)$</td>
<td>Co-occurrence frequency of a predication and a MeSH descriptor in MEDLINE</td>
</tr>
<tr>
<td>$P2M$</td>
<td>Path to MeSH matrix</td>
</tr>
</tbody>
</table>

(although not equal) perspectives, provide varying levels of detail about the content. We speculate that a MEDLINE article may be represented in terms of its concept-level semantic summary or relational semantic summary, where appropriate. Given this equivalence, we assume that the MeSH descriptors of an article capture some degree of implicit context, which is shared by its semantic predications. A semantic predication may therefore be represented in terms of the MeSH descriptors of the article in which the predication occurs. This is the basis for our interchangeability assumption:

The interchangeability assumption for subgraph creation states that the concept-level semantic summary and relational semantic summary of a MEDLINE article, are interchangeable.

Specifically, given a semantic predication $t$ and a MEDLINE article $d$ such that $t$ is extracted from $d$, the context of $t$ denoted $c(t) = M(d)$, where $M(d)$ is the set of MeSH descriptors assigned to $d$ (see Table 4.1 for various terminology that will be used throughout this section). With regards to the article in Figure 4.1, the context $c(t)$ of the semantic predication $t = \{\text{Eicosapentaenoic Acid STIMULATES Epoprostenol}\}$ is the set
of assigned MeSH descriptors, \( c(t) = \{\text{Arachidonic Acids, Eicosapentaenoic Acid, Endothelium, Epoprostenol (Prostacyclin), Fatty Acids Unsaturated, Humans, Indomethacin, Platelet Aggregation, Umbilical Veins}\} \).

A critical observation is that not all MeSH descriptors are relevant to each semantic predication in the article. We therefore speculate that the distribution of MeSH descriptors for a given semantic predication across the entire corpus will provide a better estimation of its context. That is, the context of an individual predication \( t \) across the entire corpus could be represented more comprehensively, as the aggregation of the MeSH descriptors assigned to each document in which \( t \) occurs. This representation of the context of a semantic predication as a vector of MeSH descriptors is the basis for our second assumption; the context distribution assumption.

The context distribution assumption for subgraph creation states that the context of a semantic predication can be expressed as the distribution of all MeSH descriptors associated with all articles that contain the predication.

Specifically, the context of the semantic predication \( t \) across all documents in MEDLINE is the vector \( \overrightarrow{c(t)} = [f(t, m_1), f(t, m_2), \ldots, f(t, m_{27000})] \), where \( f(t, m_i) \) is the frequency of co-occurrence between the predication and the \( i^{th} \) MeSH descriptor in the vector. Note that the number of dimensions in this vector is the total number of unique descriptors in MeSH \( (N \approx 27000) \).

The idea of expressing the context of an element in terms of the distribution of other associated elements is not arbitrary. The idea is rooted in distributional semantics, which was pioneered by English Linguist John Rupert Firth (1890–1960). Firth is famous for noting the context-dependent nature of meaning. His idea of contextualism is that an expression may not be truly understood unless viewed relative a given context. That is, expressions are context-sensitive. He is famous for the quotation, which states that: “You
shall know a word by the company it keeps” – Firth, J. R. 1957:11.³ In this dissertation, an underlying assumption is that a semantic predication may be known by the distribution of its MeSH descriptors – as the company it keeps.

With regards to the 65 articles from the Raynaud Syndrome – Dietary Fish Oils discovery by Swanson [89], only one article ([PMID6098049]) contained the semantic predication that states [Eicosapentaenoic Acid STIMULATES Epoprostenol]. There were nine MeSH descriptors assigned to this article (see Table 4.2, row 3, columns 1 and 2). Four articles ([PMID6301111], [PMID6321621], [PMID6320840], and [PMID6314583]) contain the semantic predication, which states that [Eicosapentaenoic Acid DISRUPTS Platelet Aggregation]. Among these four articles, there were 52 distinct MeSH descriptors (see Table 4.2, row 3, columns 2 and 3). Seven MeSH descriptors appear in both sets; that is, 7 out of 9 for the first predication ([Eicosapentaenoic Acid STIMULATES Epoprostenol]) and 7 out of 52 for the second predication ([Eicosapentaenoic Acid DISRUPTS Platelet Aggregation]). These statistics suggests that in spite of the out-of-context descriptors, the two predications share some context, indicated by the intersection of MeSH descriptors. In the next Section, the use of this representation to compute the shared context between paths is presented.

### 4.3 Shared Context

The representation of the context of a predication $t$ provides the basis for expressing the context of a path. This path context denoted $C(p)$ is the vector

$$C(p) = \sum_{t \in p} c(t) \quad (4.1)$$

Table 4.2: Distribution of MeSH descriptors for Semantic Predications from Raynaud Syndrome – Dietary Fish Oils

<table>
<thead>
<tr>
<th>EPA STIMULATES Epoprostenol</th>
<th>Intersection (7)</th>
<th>EPA DISRUPTS Platelet Aggregation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection (7)</td>
<td>Intersection (7)</td>
<td>Intersection (7)</td>
</tr>
<tr>
<td>[PMID 6098049]</td>
<td>[PMID 6321623]</td>
<td>[PMID 6314583]</td>
</tr>
</tbody>
</table>

resulting from the vector addition of the component semantic predication vectors (denoted $ \vec{c}(t) $), in the path. This representation that enables the creation of the path-to-MeSH matrix

$$ P2M = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,M} \\ a_{2,1} & a_{2,2} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ a_{Z,1} & a_{Z,2} & \ldots & a_{Z,M} \end{bmatrix} , $$

classified from the context vectors of all paths (in a given experiment $ E $) to be clustered. This $ P2M $ matrix is a $ Z \times N $ matrix, where $ Z $ is the total number of paths, $ N $ is the total number of unique descriptors in MeSH, and the numbers $ a_{i,j} $ are the weights of the elements in $ P2M $. The sequence of numbers

$$ P2M_{(i)} = (a_{i,1}, a_{i,2}, \ldots, a_{i,M}) $$
is the $i^{th}$ row of $P2M$, which represents the frequencies of each MeSH descriptor associated with the path $p_i$. The sequence of numbers

$$P2M^{(i)} = (a_{1,i}, a_{2,i}, \ldots a_{Z,i})$$

is the $j^{th}$ column of $P2M$, which represents the frequencies of a given MeSH descriptor across all paths to be clustered. Given a definition and representation for the path context, subgraphs can then be automatically created by clustering paths whose context vectors exceed a certain threshold of relatedness. To achieve this, the relatedness between the path context vectors must be computed.

Instinctively, the use a traditional metric like cosine similarity, to compute path relatedness between path context vectors, seems appropriate. However, since the clustering approach is based on relatedness, and not similarity, this metric may not be ideal. The main issue is that the number of MeSH descriptors present only in one vector and not the other, will lower the overall relatedness score between them. While this is accurate for similarity, it is problematic for relatedness. The second issue is that the disparity in the weights of inexact matches between MeSH descriptors in different pairs of paths, will produce unreliable estimates of relatedness. To illustrate these shortcomings, consider the two paths $p_i$ and $p_j$, for which the aggregation of MeSH descriptors (from each semantic predication) is shown in Table 4.3$^4$.

$p_i = \{EPA \text{ STIMULATES Epoprostenol}, [Epoprostenol \text{ TREATS RS}]\}$

$p_j = \{EPA \text{ DISRUPTS Platelet Aggregation}, [Epoprostenol \text{ DISRUPTS Platelet Aggregation}], [Epoprostenol \text{ TREATS RS}]\}$

$^4$Note that since both paths contains the semantic predication $[Epoprostenol \text{ TREATS RS}]$, we exclude its MeSH descriptors for simplicity.
Table 4.3: Distribution of MeSH descriptors for Paths from Raynaud Syndrome – Dietary Fish Oils

<table>
<thead>
<tr>
<th>( p_i = \text{EPA STIMULATES Epoprostenol, Epoprostenol TREATS RS} )</th>
<th>( p_j = \text{EPA DISRUPTS Platelet Aggregation, Epoprostenol DISRUPTS Platelet Aggregation, Epoprostenol TREATS RS} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>[PMID6098049]</td>
<td>[PMID6270927, PMID6520840, PMID6374855, PMID2408588, PMID6301111, PMID7295999]</td>
</tr>
<tr>
<td>Arachidonic Acids(1), Eicosapentaenoic Acid(1), Endothelium(1), Epoprostenol(1), Fatty Acids, Unsaturated(1), Humans(1), Platelet Aggregation(1)</td>
<td>Arachidonic Acids(2), Eicosapentaenoic Acid(5), Endothelium(1), Epoprostenol(5), Fatty Acids, Unsaturated(5), Humans(7), Platelet Aggregation(4)</td>
</tr>
<tr>
<td>Indomethacin(1), Umbilical Veins(1)</td>
<td>Animals(3), Aged(2), Angina Pectoris(1), Aspirin(2), Arachidonic Acid(1), Arachidonate Lipoxygenases(1), Arteriosclerosis(2), Adult(3), Aorta Thoracic(1), Blood Platelets(3), Bleeding Time(2), Carbon Radioisotopes(1), Cattle(1), Cyclooxygenase Inhibitors(1), Cholesterol(1), Cholesterol HDL(1), Cerebrovascular Circulation(2), Cerebrovascular Disorders(2), Cyclic AMP(2), Cells, Cultured(2), Fish Oils(1), Female(4), Linoleic Acid(1), Linoleic Acids(1), Lipoproteins HDL(1), Lipids(1), Lipoxygenase Inhibitors(1), Male(3), Muscle, Smooth, Vascular(2), Middle Aged(2), Pregnancy(1), Prostaglandin(3), Platelet Function Tests(1), Rabbits(1), Rats(1), Swine(1), Species Specificity(1), Thrombin(1), Thromboxane A2(2), Thromboxane-A Synthase(2), Triglycerides(1), Umbilical Veins(1), 6-Ketoprostaglandin F1 alpha(2), Adenosine Diphosphate(1), Arteriosclerosis(1), Blood Proteins(1), Dietary Fats(1), Fatty Acids(1), Iloprost(1), Malondialdehyde(1), Mixed Connective Tissue Disease(1), Platelet Count(1), Prospective Studies(1), Scleroderma, Systemic(1), Serotonin(1)</td>
</tr>
</tbody>
</table>

Let the context set of a path \( p \), denoted \( C(p) \), be the unique set of MeSH descriptors in its context vector \( \overrightarrow{C(p)} \), which have a non-zero frequency. Concretely, the context set for path \( p_i \) is \( C(p_i) = \{ \text{Arachidonic Acids, Eicosapentaenoic Acid, Endothelium, Epoprostenol, Fatty Acids, Unsaturated, Humans, Platelet Aggregation, Indomethacin, Umbilical Veins} \} \), while the context set for \( p_j \) is \( C(p_j) = \{ \text{Arachidonic Acids, Eicosapentaenoic Acid, Endothelium, Epoprostenol, Fatty Acids, Unsaturated, Humans, Platelet Aggregation, Animals, Aged, Angina Pectoris, Aspirin, ..., Platelet Count, Prospective Studies, Scleroderma, Systemic, Serotonin} \} \).

In path \( p_j \), the semantic predication [Epoprostenol DISRUPTS Platelet Aggregation] was extracted from three articles ([PMID2408588], [PMID6301111], [PMID7295999]) in MEDLINE. Across these articles there are 33 MeSH descriptors, 21 of which are already present in the context set for the predication [EPA DISRUPTS Platelet Aggregation], and 12 which are not. Five of the 21 descriptors are also among the seven descriptors common to both paths. The two sets therefore share descriptors seven descriptors in common (Arachidonic Acids, Eicosapentaenoic Acid, Endothelium, Epoprostenol, Fatty Acids, Unsaturated, Humans).
saturated, Humans, Platelet Aggregation), with 59 disjoint descriptors. This includes 2 disjoint descriptors from \( p_i \) (Indomethacin, Umbilical Veins) (see Table 4.3, row 4, column 1) and 57 disjoint descriptors from \( p_j \) (see Table 4.3, row 4, column 2).

Using the raw frequencies of the overlapping descriptors (shown in parenthesis in Table 4.3, row 3, columns 1 and 2) to compute to cosine similarity

\[
\cos(\theta) = \frac{X \cdot Y}{||X||||Y||} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \sqrt{\sum_{i=1}^{n} Y_i^2}},
\]

(4.2)

then the dot product of the frequencies produces a value of \( s'(p_i, p_j) = (1 \times 2) + (2 \times 5) + (1 \times 1) + (1 \times 5) + (1 \times 5) + (1 \times 7) + (1 \times 4) = 29 \). The issue is that the overall relatedness between the two paths due to the weights of these exact descriptors is diminished by the weights of the out-of-context descriptors, from the denominator of Equation 4.2. While this is appropriate for similarity, it will adversely affect the relatedness score. The two vectors are as related as their shared MeSH descriptors. The more descriptors they have in common, the more related they are. Normalizing the vectors by the weights of the out-of-context descriptors in the denominator of Equation 4.2 is therefore undesirable. A more practical measure of shared context (or semantic relatedness) between the two paths is simply the numerator (or dot product) of the two vectors. That is, \( s'_{cos} = X \cdot Y \), which is sum of the product of the frequencies. This measure will reduce the relatedness measure to include only the weights of the exact matches between MeSH descriptors.

Still however, while the use of the dot product of the frequencies for the intersecting descriptors is plausible in computing shared context, it presents two problems. First, the variability in the frequencies of intersecting descriptors themselves could result in different relatedness between pairs of paths with the same MeSH descriptors. This will be manifested in pairs of context vectors comprised of the same set of MeSH descriptors in the intersection, but with different frequencies. Our vector representation is therefore downgraded to the Boolean-valued set representation, in which a MeSH descriptor is either
present or absent in the distribution. This frequency variability is also a likely reason why distributional measures such as SVD, LSI and MI may fail to adequately capture the shared context between paths, as alluded to by Gordon and Dumais [35]. Aside from this, a more critical issue is that the intersection alone, does not account for inexact matches between MeSH descriptors. In particular, Table 4.3 shows that in addition to the MeSH descriptor \textit{Fatty Acids, Unsaturated} in context set of $p_j$, the path also contains the descriptors \textit{Fatty Acids, Dietary Fats, Cholesterol, Cholesterol HDL, Lipids and Lipoproteins HDL} in the context set.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{G-Tree.png}
\caption{Subset of the G-Tree of the MeSH Hierarchy}
\end{figure}

Figure 4.2 shows a subset of the MeSH hierarchy, in which the explicit (or formal) hierarchical relationship among some of these concepts is modeled. Indeed, each of these descriptors is closely related to \textit{Fatty Acids, Unsaturated} in MeSH. The solution to this problem of path relatedness given inexact matches between MeSH descriptors therefore requires a measure of \textit{semantics-enhanced shared context} between context vectors. The method for computing semantic relatedness between path context vectors is discussed in the next Section.
4.3.1 Semantics-enhanced shared context

To compute semantic relatedness between path context vectors, given inexact matches between MeSH descriptors, then explicit semantics from the hierarchical relationships in MeSH are used. In particular, the metric of *dice similarity* is used to compute the similarity between pairs descriptors (e.g., *Fatty Acids, Cholesterol*). Dice similarity

\[
dice(m_i, m_j) = 2 \times \frac{|\text{ancestors}(m_i)_{MH} \cap \text{ancestors}(m_j)_{MH}|}{|\text{ancestors}(m_i)_{MH}| + |\text{ancestors}(m_j)_{MH}|}
\]  

(4.3)

is computed as the proportion of common ancestors between descriptors in the MeSH hierarchy (MH), given two MeSH descriptors \(m_i\) and \(m_j\). Notice that the maximum similarity between two descriptors computed using dice similarity is 1. This occurs when the descriptors are equal. (i.e., \(m_i = m_j\)). The range of similarity values is therefore \([0,1]\).

In this computation, pairs of MeSH descriptors, whose dice similarity exceed some manually selected threshold of semantic similarity \(\tau_{\text{sim}} = 0.75\) are normalized to a value of 1. This normalized dice similarity score

\[
dice_N(m_1, m_2) = \begin{cases} 
1 & \text{if } \dice(m_1, m_2) > \tau_{\text{sim}} \\
0 & \text{otherwise} 
\end{cases}
\]  

(4.4)

is computed based on applying the threshold to dice for a pair of MeSH descriptors \(\dice(m_1, m_2)\). The key idea behind this semantics-enhanced shared context metric for computing semantic relatedness, is to *maximize the weights of the in-context descriptors* and *minimize the weights of the out-of-context descriptors*. The overall semantic relatedness

\[
sr''(p_i, p_j) = \sum_{(a,b) \in C(p_i) \times C(p_j)} \dice_N(a, b)
\]  

(4.5)

between the two paths \(p_i\) and \(p_j\) is then computed as the sum of the pairwise dice similarity
scores that exceed the manually defined threshold of dice semantic similarity, across the context sets $C(p_i)$ and $C(p_j)$. In Equation 4.5, the variable $a$ represents an arbitrary descriptor in the path context set $C(p_i)$, while $b$ represents an arbitrary descriptor in the path context set $C(p_j)$.

A consequence of this semantics-enhanced shared context metric is that a broad range of overall scores may exist. Paths that are very similar, which have many exact (and also inexact) MeSH descriptors in common, will have very scores, while others may have significantly lower scores. To dampen the major differences in similarity scores of different path pairs, we apply a log reduction on the normalized dice similarity scores. This is achieved by first computing the relatedness score between a given MeSH descriptor $a$ in context set $C(p_i)$ against the entire set of descriptors in the context set $C(p_j)$. This computation yields the similarity score

$$sim'(a, C(p_j)) = \sum_{b \in C(p_j)} \text{dice}_N(a, b). \quad (4.6)$$

The log reduction

$$sr''_L(p_i, p_j) = \sum_{a \in C(p_i)} \log (1 + sim'(a, C(p_j))) \quad (4.7)$$

is then applied to $sim'(a, C(p_j))$, and the overall semantic relatedness $sr''(p_i, p_j)$ between the two paths is the aggregate of the log-reduced scores for each descriptor $a$ in $C(p_i)$ and the entire set in $C(p_j)$. This metric is the basis for finding and elucidating complex associations among concepts, along multiple thematic dimensions, based on implicit and explicit semantics, alluded to by Gordon and Dumais in [35]. In the next Section, the hierarchical agglomerative clustering algorithm, which is used for path clustering and subgraph generation, based on the measure of semantic relatedness is discussed.

---

5Note that experimentation using the sum of the maximum, instead of the aggregate may also yield useful results.
4.4 Hierarchical Agglomerative Clustering

The representation of the context of a path, as set of MeSH descriptors, is ideal for automatic subgraph creation on multiple thematic dimensions. Paths can be segregated into clusters, based on their underlying shared context, provided by the MeSH descriptors. The key task in generating such clusters, is determining a threshold for path relatedness to distinguish which path pairs belong to the same cluster, and which do not.

Assuming that such a threshold ($\tau_{rel}$) for path relatedness can be automatically computed, then given a source concept $A$, target concept $C$ and a candidate graph $R$, the hierarchical agglomerative clustering algorithm (HAC), outlined by Manning in [58] (Chapter 17), can be effectively used for subgraph generation. The algorithm first requires initializing $|R|$ buckets – one for each path in the candidate graph. For a given path, the relatedness score is computed for each of the remaining $|R| - 1$ paths. If two paths are sufficiently related (i.e., they exceed the threshold of semantic relatedness ($\tau_{rel}$) they are placed in the same cluster. Buckets are then merged, based on their overall similarity – and this process repeats iteratively, until the algorithm terminates. The algorithm yields two set of clusters as output: 1) multi-path clusters; containing more than one path each and/or 2) singleton clusters; containing individual paths, which were not sufficiently related to any other paths to be clustered into the same bucket. The method for automatically computing the threshold for path relatedness is presented in the next Section.

4.4.1 Threshold Selection

The threshold for path clustering was initially determined manually. Table 4.4 column 3, shows the threshold values arbitrarily selected for three experiments: 1) Raynaud Syndrome–Dietary Fish Oils (3.0), 2) Testosterone–Sleep (3.5), and 3) DEHP–Sepsis (4.0). These thresholds were used to generate subgraphs, which led to the rediscovery of 2 out of these 3 discoveries (DEHP–Sepsis was not recovered).
Table 4.4: Threshold Comparisons

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Path Relatedness Score</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 Std. Dev</td>
<td>Manual</td>
</tr>
<tr>
<td>Raynaud Syndrome-Dietary Fish Oils</td>
<td>2.68</td>
<td><strong>3.0</strong></td>
</tr>
<tr>
<td>Testosterone-Sleep</td>
<td>3.35</td>
<td><strong>3.5</strong></td>
</tr>
<tr>
<td>DEHP-Sepsis</td>
<td>3.94</td>
<td><strong>4.0</strong></td>
</tr>
</tbody>
</table>

To automatically determine the threshold for each experiment, the distribution of path relatedness scores between all pairs of paths in the candidate graph was first computed. For example, given the 189 paths in the Raynaud Syndrome – Dietary Fish Oils experiment (from Experiment I, Section 3.5.4), a total of $35 \times 344 = (189 - 1)^2$ scores were computed for all path pairs. These raw scores and their frequencies were then plotted on a graph, and the distribution of the scores was observed. Figure 4.3 shows the raw data for the Raynaud Syndrome-Dietary Fish Oils experiment appear to follow a Normal (or Gaussian) distribution. The mean ($\mu$), variance ($\sigma^2$) and the standard deviation ($\sigma$) of the distribution were therefore computed, and used to compute the expected value for each data point, using the standard Gaussian Function

$$f(x) = \frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sigma\sqrt{2\pi}}.$$ \hspace{1cm} (4.8)

Figure 4.4 shows the Gaussian distributions for the three scenarios in Table 4.4. These distributions show that the data only approximate to a Gaussian. However, this is not surprising. Each distribution is particularly sparse, for low relatedness scores. A possible reason for this discrepancy is the minimum commonality in path distributions. Naturally, each path in the candidate graph is bounded by the same source $A$ and target $C$. Consequently, each pair of paths inherently share some minimum relatedness slightly above zero. Still, given that the data approximates to a normal distribution, certain properties of Gaussian functions can be exploited to help automatically compute a threshold for path relatedness. For instance,
the mean ($\mu$) of the distribution corresponds to the average relatedness score among all the paths. Additionally, gaussian distributions contain two points of inflection; $\pm \sigma$ the standard deviation from the mean, which mathematically are points where the second derivative of the curve is zero, i.e., $(x - \mu)^2 = \sigma$.

Table 4.4, row 1 shows that when the manually selected thresholds were compared with the points of inflection, they were consistently between the 2nd and 3rd standard deviation from the mean. In particular, the manually selected threshold for Raynaud Syndrome–Dietary Fish Oils (3.0) was between the 2nd standard deviation (2.68) and the 3rd standard deviation (3.04). The maximum path relatedness score in the said distribution was 3.38. The second standard deviation ($\tau_{rel} = 2\sigma$) was therefore empirically selected as the threshold for path relatedness, since it approximates to the manual values for thresholds from the 3 experiments. Experimental results in Section 5 show that both the 2nd and 3rd standard
deviation are adequate for rediscovering knowledge.

Given this method for automatically selecting a threshold of path relatedness, the clustering algorithm is then applied. In the first iteration, or bucket population step, paths above $2\sigma$ are automatically added to their respective buckets. In the subsequent bucket merging step, buckets that themselves are related above some threshold of bucket relatedness, are merged. To achieve this, a metric for computing bucket similarity is required. This metric is called the combination similarity, and is discussed in the next Section.

### 4.4.2 Combination Similarity

Four methods for computing combination similarity are commonly used in the literature [58]. Figure 4.5 shows that these four methods are: 1) single-link; maximum similarity, 2) complete-link: minimum similarity, 3) centroid or inter-cluster similarity and 4) group
average (or GAAC): average of all similarities.

In single-link clustering the combination similarity between two buckets is the maximum similarity across all its members. The two buckets are therefore merged if this maximum similarity meets the threshold for clustering. Single-link clustering is desirable for clustering, when any link between the two buckets is sufficient to merge them. Buckets merged using this method are likely to show low cohesion, often leading to straggly clusters.

In complete-link clustering, the combination similarity between buckets is the minimum similarity across all members. Two buckets are therefore merged only if their minimum similarity exceeds the threshold. Buckets merged using complete-link similarity tend to be very compact, as all pairs of items across them must exceed the threshold.

In centroid or inter-cluster similarity, the combination similarity between buckets is the average similarity of all members across the two buckets. Clusters arising from inter-cluster similarity are likely to be more diverse, with some items, above the average and some items below.

In group average similarity (GAAC) the combination similarity is based on the average of all similarities, both within and across two buckets. That is, the average of the intra-
cluster and inter-cluster similarities. GAAC could also lead to clusters that are fairly well connected, but with larger distances between the members than complete-link.

Inter-cluster similarity is ideal for bucket merging, in this case the goal of the clustering process is to produce buckets consisting of paths, which are related but not necessarily similar. Single-link clustering is not appropriate because will merge clusters with significantly different contexts. This may lead to meaningless clusters. Complete-link clustering is not ideal because it may only merge buckets that are similar. Likewise, GAAC is not ideal because the intra-cluster similarity may bias the combination similarity leading to diverse clusters, that are broadly related. The inter-cluster similarity

\[
sim_{inter}(B_a, B_b) = \frac{\sum_{(p_i, p_j) \in B_a \times B_b} sr''_L(p_i, p_j)}{|B_a| \cdot |B_b|}
\]

is computed as the sum of the similarities across pairs of paths \((p_i, p_j)\) from each set, divided by the the product of the sizes of the two buckets \(B_a\) and \(B_b\). Given this method for computing the combination similarity for bucket merging, the final task in the clustering process is ranking of the clusters. The method for ranking is discussed in the next Section.

### 4.4.3 Subgraph Ranking

Clusters containing more than one path are ranked based on the compactness of the cluster. Subgraphs with high compactness are likely to be smaller and highly related, while those with low compactness may be larger and more diverse. The compactness of a cluster is computed using the measure of intra-cluster similarity. Figure 4.6 shows that the intra-cluster network of a cluster is a fully connected network, in which all nodes (paths) are connected to all other nodes. The intra-cluster rank

\[
sim_{intra}(B) = \frac{2 \cdot \sum_{p_i, p_j \in B, p_i \neq p_j} sr''_L(p_i, p_j)}{|B| \cdot (|B| - 1)}
\]
of a cluster \((B)\) is the average similarity of each pair \((p_i, p_j)\) of paths in the cluster, where \(N\) is the size of the cluster.

![Intra-cluster Network](image)

Figure 4.6: Intra-cluster Network

While suitable for multi-path clusters, the intra-cluster similarity is unsuitable for ranking singleton clusters, which contain only one path. Singleton clusters consisting of only one path were ranked in ascending order using the measure of association rarity. That is, given a path \(p_i\) an association \(A(p_i)\), derived of the path, is the set of unique concepts in it. Association rarity is the number of MEDLINE articles \(f(A(p_i))\) that contain only the concepts in the path. For singleton paths, bucket rarity

\[
r(B) = \frac{\sum_{p_i \in B} f(A(p_i))}{|B|}
\]  

(4.11)

is the same as association rarity, since \(B = \{p_i\}\).

The ranked list of clusters is rendered to the user for inspection in the **Discovery Browsing Interface**. This interface is available online: [http://knoesis-hpc.cs.wright.edu/obvio/](http://knoesis-hpc.cs.wright.edu/obvio/) (see video demo – [http://bit.ly/obviodemo](http://bit.ly/obviodemo)). Concepts are color-coded based on semantic groups from the BKR, while predicates are color-coded based on a locally developed coding scheme since, none exists in the BKR. Given methods to: 1) automatically compute the threshold for path relatedness, 2) compute the combination similarity for merging buckets and 3) rank buckets after clustering. The algorithm for automatic subgraph creation is outlined in the next Section.
4.4.4 Clustering Algorithm

The hierarchical agglomerative clustering algorithm for automatic subgraph creation requires a candidate graph $R_{s,t} = \{p_1, p_2, \ldots, p_k\}$ of paths, between a source $s$ and target $t$ as input. It produces a set of subgraphs $S_{s,t} = \{R_{s,t}(c_1), R_{s,t}(c_2), \ldots, R_{s,t}(c_n)\}$ as output, where $c_i$ is a distinct context in $R_{s,t}$, according to the context-driven model outlined in Section 3.4. The specific steps executed to achieve this output are given in Algorithm 1.

Algorithm 1 $HAC(R_{s,t}, s, t)$

1: $buckets := \{p_1, p_2, \ldots, p_k\}$, $tempClusters := empty$, $mergedClusters := empty$
2: $iteration := 0$, $currNumClusters := 0$, $prevNumClusters := k$
3: $\tau_{rel} := computeRelThreshold(buckets)$
4: while stoppingCondition($prevNumClusters$, $numClusters$) := false do
5: \hspace{1em} $prevNumClusters := size(buckets)$
6: \hspace{1em} $tempClusters := populateAndMarkForMerge(buckets)$
7: \hspace{1em} $mergedClusters := merge(tempClusters)$
8: \hspace{1em} $buckets := mergedClusters$
9: \hspace{1em} $currNumClusters := size(mergedClusters)$
10: \hspace{1em} $tempClusters := empty$
11: \hspace{1em} $mergedClusters := empty$
12: \hspace{1em} $iteration++$
13: end while
14: return $rank(buckets)$

Step 1 (Initialization): Initialize $k$ empty buckets and place each path from the candidate graph in a different bucket (line 1). Also initialize the number of iterations and the current number of subgraphs $currNumClusters$ to zero. Initialize $prevNumClusters$ to $k$.

Step 2 (Threshold Selection): Automatically compute the threshold ($\tau_{rel}$) for path relatedness (line 3) as the second standard deviation of the Gaussian distribution of the path relatedness scores, discussed in Section 4.4.1.

Step 3 (Bucket Population): For the set of paths, in each pair of buckets $b_1, b_2$ in $buckets$, 

compute the inter-cluster similarity score (from Equation 4.9). For each pair of buckets whose inter-cluster similarity score exceeds the threshold of semantic relatedness ($\tau_{rel}$) computed in Step 1, mark the two buckets for merging (line 6).

**Step 4 (Bucket Merging):** For each pair of buckets in $tempClusters$ that have been marked for merging, create a new bucket consisting of all distinct paths from each of the two clusters, using the subroutine $merge(tempClusters)$. Add each mergedCluster to the emerging set of clusters ($mergedClusters$, line 7). If a bucket has not been marked for merging or deletion, add it to the set of $mergedClusters$.

**Step 5:** Reinitialize the number of buckets to the newly created set of $mergedClusters$ (line 8). Update variable $currNumClusters$ to reflect the reduced number emerging subgraphs $size(mergedClusters)$ in (line 9).

**Step 6:** Reinitialize the temporarily created set of $tempClusters$ and $mergedClusters$, to empty sets (lines 10–11). Increment the iteration counter (line 12).

**Step 7 (Iterate):** Repeat the algorithm until the stopping conditions are met. One stopping condition is until the number of clusters in the previous iteration $prevNumClusters$ is equal to the number of clusters the current iteration $currNumClusters$ (line 4).

**Step 8 (Bucket Ranking):** Rank the reduced set of buckets using the measure of intra-cluster similarity (from Equation 4.10) or association rarity (from Equation 4.11), depending on the number of buckets.
4.5 Summary

This Chapter discussed an approach to automatically generate multifaceted subgraphs using context. The approach first specifies the notion of semantic predication context, which is used to infer path context. The context of a path is expressed as a binarized vector of MeSH descriptors from MEDLINE, that co-occur with the semantic predications of the path. A semantics-enhanced set intersection metric was developed, which exploits hierarchical knowledge from MeSH, to estimate semantic similarity between inexact matches in MeSH descriptors. A log-reduction is then applied to compute the final shared context between pair of paths.

The hierarchical agglomerative clustering algorithm was then used to cluster paths, based on semantic relatedness. The 2nd standard deviation of the distribution of path relatedness scores, is used to automatically compute the threshold for path relatedness. Buckets are merged based on their inter-cluster similarity, and ranked both by intra-cluster similarity and association rarity, where appropriate.

While an initial step towards automation, several limitations with the current approach exists. The first issue is reliability of the automatically selected threshold for path relatedness, at the second standard deviation from the mean. Figure 4.4 clearly suggests that each distribution only approximates to Gaussian. Exploratory experiments, in which the goodness of fit was computed for the various distributions, using the chi-squared $\chi^2$ test, showed that $p$-values for the initial experiments were more than the commonly accepted value of $p < 0.05$. In practical terms, a low $p$-value indicates a high correlation of independence of the distribution. That is, if several experiments are conducted, where a $p < 0.05$ has been observed, then similar results can be expected, and the observed experiments were likely due to independent events.

For LBD, our higher $p$-values mean that there is less randomness across the path relatedness scores. Thus, across a significant number of experiments, results are expected to decline. While in general this could be problematic, given that our initial experiments
show that 8 out of 9 existing discoveries were rediscovered using this approach, the decline in recall may not be too costly. Furthermore, the current system does not leverage any manual input from domain experts. It is expected that with human involvement to provide additional filters, both precision and recall are likely to improve.

Alternatively, the existing approach could be improved by first normalizing the path relatedness scores, based on the expectation that all paths share some minimum, non-zero relatedness score, inherent in the experiment. Path relatedness could be computed as a function of both relatedness and *unrelatedness*, which captures some minimum relatedness from the distribution. Additionally, the existing model, which is based only on MeSH descriptors, to capture implicit and explicit context, could be extended to an ensemble approach. Torvik et al. in [96], for example utilized a mixture of two models, which produced a more perfect Gaussian.

The most critical issue however, is the complexity of the algorithm for subgraph generation. While the HAC algorithm is an ideal choice, its time complexity is quadratic, in the order of $\Theta(N^2 \log N)$. This may be suboptimal when a large number of bins must be clustered, and many iterations must be performed. In this dissertation, no serious attempt to optimize the running time of the algorithm, or overall scalability of the approach, has been attempted. Scalability is the most pressing task for any future application of this approach.

The technical novelty in this dissertation are as follows: 1) the representation of the context of a path as a binarized vector of MeSH descriptors and 2) the use of structured background knowledge, specifically the MeSH hierarchy, to extrapolate and compute the shared context between two paths. As such, the approach developed in this dissertation, is based on context, instead of distributional statistics and graph-theoretic metrics.
Evaluation

Evaluating LBD systems is generally a challenging problem. The absence of a gold standard makes it difficult to compute standard measures such as precision and recall. Instead, closed discovery LBD systems are commonly evaluated based on their ability to rediscover existing knowledge. The measure of correlation [109], which computes the fraction of existing discoveries (or intermediates) recovered by the system, is commonly used. To assess the effectiveness of the context-driven automatic subgraph creation model developed in this dissertation, two types of evaluation were conducted. The first is an evidence-based evaluation, which reports on correlation. It also provides utilities that enable understanding the meaning of associations. The second is a statistical evaluation, designed to estimate the interestingness of the subgraphs in general. Interestingness is a numerical measure that captures the rarity of associations in MEDLINE. The interestingness measure captures the likelihood that a domain expert might be motivated to explore an arbitrary subgraph in the first place. Intuitively, associations that have never been mentioned in MEDLINE can be considered most interesting. Such zero-rarity associations are likely sources of new knowledge, since they have not been mentioned in any single article. The obvious caveat is that rare associations are not necessarily all interesting. Each evaluation is discussed in the following subsections.
5.1 Evidence-based

The evidence-based evaluation consists of two aspects. The first aspect reports on the number of intermediates from an existing discovery retrieved by our system. The second aspect substantiates the meaning of each association using evidence from the literature. Such evidence is derived first using the predicates of the semantic predications in the subgraph. When this is insufficient or contradictory, evidence is obtained using the provenance of the predications in MEDLINE. Additionally, MEDLINE queries are composed and executed in PubMed (as we will show) to explore relationships that are inferred in the subgraphs but not explicitly stated.

For each rediscovery scenario, no concept filters were specified, to exclude concepts based on semantic types or groups. A generic predicate filter, called the STRICT filter was applied uniformly by the system (not the user), across some experiments, to exclude less informative UMLS predicates. Some of these include ASSOCIATED_WITH, INTERACTS_WITH and AFFECTS. This limited degree of manual filtering is the extent of a priori knowledge required for subgraph generation in the system.

In the following tables, the letter $Y$ (for yes) is used to indicate that the status $S$ of an intermediate as found directly in a subgraph at position $P$ in the list of subgraphs. The symbol $Y^*$ indicates that an intermediate was found through discovery browsing. The identifier $ZR$ indicates that an intermediate was found in the list of zero-rarity singletons (i.e., a subgraph consisting of only one path). The next section discusses the application of our approach to the Raynaud Syndrome – Dietary Fish Oils discovery. Supplementary materials that showcase all results are available online\(^1\).

---

5.1.1 Raynaud Syndrome – Dietary Fish Oils

In November 1985, American Information Scientist Don R. Swanson (1924 – 2012) explored the research question of the role of Dietary Fish Oils (from salmon, mackerel, albacore, etc.) in Raynaud Syndrome. Through the methods described in [89], Swanson discovered that “dietary fish oil might ameliorate or prevent Raynaud’s syndrome.” This is because Dietary Fish Oils: 1) inhibit Platelet Aggregation, 2) increase the flow of blood (thereby reducing Blood Viscosity) and 3) also have a regulatory effect on the smooth muscle (thereby preventing Vasoconstriction and stimulating Vasodilation). Each of these processes is causally implicated in Raynaud.

We seeded our algorithm with three concepts as sources: 1) Fish Oils (C0016157), 2) Fish oil - dietary (C0556145) and 3) Eicosapentaenoic Acid (C0000545), and two concepts as targets: 1) Raynaud Disease (C0034734) and 2) Raynaud Phenomenon (C0034735). The corpus consisted of the relevant 61 full text articles discussed by Swanson [89] in the pre-November 1985 period. There were only 4 articles from the Dietary Fish Oil set, which were in the Raynaud set. The path length was set to 3 and no predicate filter was specified. These choices are consistent with the choices in our earlier experiments in [19], in which we rediscovered and decomposed this hypothesis by manually constructing the subgraphs, using domain expertise as context.

The algorithm terminated in less than 5 minutes, producing four subgraphs (and 134 singletons) at $2\sigma$ and one subgraph (and 164 singletons) at $3\sigma$. There were 1035 unique concepts and 4143 unique predications in the predications graph and the candidate graph contained 171 paths of length 3. Figure 5.1 shows that at $3\sigma$, subgraph1 directly contains the intermediate Platelet Aggregation, which many past rediscovery approaches consider sufficient to constitute a rediscovery. However, to better substantiate the association, we utilize the predicates in the subgraph, together with the provenance of the predications in MEDLINE, along with traditional PubMed search, to provide evidence.
The predication, which states that [Eicosapentaenoic Acid CONVERTS TO Prostaglandins] was extracted from the following corroborating sentence, in the full text of the following article [PMID6827988] by Harris et al. The authors state that the “recent discovery that the prostaglandins derived from eicosapentaenoic acid have biological effects different than those derived from arachidonic acid (C20:4w6) has generated further interest in fish oils.” Two of the other 61 articles [PMID6321621, PMID6314583] contained this predication. Harris also refers to the 1979 article [PMID218223] by Needleman et al., which suggests further that [Eicosapentaenoic Acid CONVERTS TO Prostaglandin (PGI₃)] in its metabolic pathway. And the full text of 1985 article [PMID2997286] by von Schaky et al. confirms that Eicosapentaenoic Acid produces Prostaglandin (PGI₃) and Epoprostenol (Prostacyclin (PGI₂)). von Schaky notes that “dietary EPA is transformed in vivo in humans into prostaglandins I₃, which is as active ... as the vasodilatory and antiaggregatory prostaglandin I₂.”

The subgraph also contains the predication, which states that [Eicosapentaenoic Acid DISRUPTS Platelet Aggregation]. This predication was extracted from the full text of
the article [PMID6320840] by Saynor et al., who refers to the “Mechanisms underlying the inhibition of platelet aggregation by eicosapentaenoic acid and its metabolites.” The predication [Alprostadil DISRUPTS Platelet Aggregation] was extracted from the full text of the article [PMID6302714] by Dyerberg et al., who pointed out that another author “was the first to show that [Alprostadil] PGE\(_1\) inhibited platelet aggregation.” The previously article by von Schaky also alludes to this property of fish oils.

Conversely, the predication [Epoprostenol TREATS Raynaud’s Phenomenon] was correctly extracted from two articles; by Dowd et al. [PMID7037038], who discusses “Treatment of Raynaud’s phenomenon by intravenous infusion of prostacyclin (PGI\(_2\))” and by Belch et al. [PMID3883365], who discusses “Increased prostacyclin metabolites and decreased red cell deformability in patients with systemic sclerosis and Raynauds syndrome.” Since both Alprostadil (PGE\(_1\)) and Epoprostenol (PGI\(_2\)) are Prostaglandins, it is plausible that both Alprostadil and Epoprostenol actually treat Raynaud’s Syndrome by disrupting Platelet Aggregation. Indeed, the 1982 article [PMID6890719] by Pardy et al., obtained through a date-restricted MEDLINE query\(^3\), confirms that Alprostadil (PGE\(_1\)) treats Raynaud Phenomenon, instead of the weaker INTERACTS_WITH relationship in the subgraph. The role of Platelet Aggregation in causing Raynaud, which is inferred in the subgraph, is easily confirmed using another MEDLINE query (Platelet Aggregation AND Raynaud AND 1865:1985/11[DP]), which yields the 1985 article [PMID3985417] by Soro et al.

This subgraph together with discovery browsing suggest a richer relationship among Eicosapentaenoic Acid, Platelet Aggregation and Raynaud Syndrome than would be provided by their co-occurrence. Rather, it appears that one mechanism by which [Eicosapentaenoic Acid TREATS Raynaud Syndrome] is by stimulating a series of Prostaglandins (namely, Prostaglandin I\(_3\) (PGI\(_3\)), Alprostadil (PGE\(_1\)) and Prostacyclin (PGI\(_2\))), which actually disrupt Platelet Aggregation.

---


\(^3\)Query: Alprostadil AND Raynaud AND 1865:1985/11[DP]. Confirmed in search result #12
An important observation is that the subgraph contains contradicting semantic predications. For example, the two predications \textit{[Eicosapentaenoic Acid CONVERTS TO Prostaglandins]} and \textit{[Eicosapentaenoic Acid INHIBITS Prostaglandins]} are opposing. The full text of the article [PMID6827988] by Harris et al., from which the predication \textit{[Eicosapentaenoic Acid CONVERTS TO Prostaglandins]} was extracted supports this claim. However, the full text of the lone article [PMID6301111] by Moncada from which the predication \textit{[Eicosapentaenoic Acid INHIBITS Prostaglandins]} was extracted states that “It is clear, therefore, that both prostaglandin dependent and independent pathways of platelet aggregation are inhibited by EPA in vitro.” This is an incorrect extraction from SemRep. The author is noting that \textit{[Eicosapentaenoic Acid INHIBITS Platelet Aggregation]}, not \textit{Prostaglandins} as the predication suggests. It is important to note that resolution of such discrepancies is part of the discovery browsing process, which requires adjudication by domain experts. We provide the infrastructure for achieving this through provenance.

The second intermediate \textit{Blood Viscosity}, was found in the list of singletons with a zero-rarity in MEDLINE (result \#15 in Table 5.1). The actual singleton, which states that \textit{[Eicosapentaenoic Acid DISRUPTS Blood Viscosity]}, \textit{[Ketanserin DISRUPTS Blood Viscosity]}, \textit{[Ketanserin TREATS Raynaud Disease]}, suggests a causal relationship between \textit{Blood Viscosity} and \textit{Raynaud Syndrome}. This inferred relation that \textit{[Blood Viscosity CAUSES Raynaud Syndrome]} is confirmed in the 1984 article [PMID6707529] by Larcan et. al through a MEDLINE search. The provenance of the other predications were available for inspection in the following articles [PMID401574], [PMID6303363], [PMID2412054] and [PMID6432198], [PMID6209510] respectively, and no MEDLINE querying is required.
Table 5.1: Comparison of rediscoveries with other approaches for Raynaud Syndrome - Dietary Fish Oils

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
<th>Weeber</th>
<th>Gordon</th>
<th>Hristovski</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S</td>
<td>P</td>
<td>S</td>
<td>P</td>
<td>S</td>
</tr>
<tr>
<td>Diet 1 - Raynaud Syndrome</td>
<td>Blood Viscosity Y*</td>
<td>ZR-15</td>
<td>Y</td>
<td>2</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Platelet Aggregation Y</td>
<td>1</td>
<td>Y</td>
<td>1</td>
<td>Y</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Vascular Reactivity Y</td>
<td>–</td>
<td>–</td>
<td>Y</td>
<td>–</td>
<td>19</td>
</tr>
<tr>
<td>Diet 2 - Raynaud Syndrome</td>
<td></td>
<td>–</td>
<td>–</td>
<td>Y</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 5.1 shows the number of intermediates rediscovered for this experiment compared with 4 other approaches. The intermediate *Vascular Reactivity* (in reference to *Vasoconstriction*) was not found explicitly by our approach (although can be inferred from the article [PMID2997286] by von Schacky et al.). This is not completely unexpected however, since it is known from our reports in [19] that SemRep interprets “Vascular” and “Reactivity” as separate concepts. Hristovski in [38] was also subject to the same limitation.

Srinivasan [85] found all three intermediates in the top 2 of the top 30. However, note that Srinivasan’s approach relies on *a priori* knowledge of the semantic types of the intermediates for filtering and is manually intensive. Additionally, that approach does not create complex subgraphs, nor does it provide evidence for the meaning of associations using predicates. Hristovski et. al. [38] and Weeber et al. [100] also require considerable domain expertise, particularly for specification of *a priori relations* (including semantic types and discovery patterns). Gordon & Lindsay [53] find intermediates but make no attempt to elucidate their meaning.

To illustrate that our subgraphs capture different thematic dimensions of association, consider the four subgraphs at $2\sigma$. Subgraph 1 in Figure 5.2a is similar to subgraph 1 (at $3\sigma$) except that it includes the three additional intermediates, *TIMP1, TIMP1 protein, human* and *Thromboembolism*, naturally due to a lower threshold for path relatedness. By inspection, this subgraph elucidates the association between *Dietary Fish Oils* and *Raynaud Syndrome* through *Blood Platelets/Prostaglandins*, similar to the previous subgraph.

Subgraph 2 (shown in Figure 5.3) associates *Dietary Fish Oil* and *Raynaud Syndrome*...
from the perspective of pharmaceutical drugs, including *Nifedipine, Pentifylline, Thyrocalcitonin*, and *Trinitrin* detailed in the article [PMID6352267] by Kahan et al., from which the predication [*Nifedipine TREATS Raynaud Phenomenon*] was extracted.

Subgraph3 in Figure 5.4 discusses the role of various Fatty Acids, which associate *TIMP1, Epoprostenol, Efamol* and *Evening Primrose* (see [PMID4082084, PMID6318123, PMID6321621]).

Subgraph4 in Figure 5.5, which focuses on *Blood Platelets*, is subsumed by sub-
5.1.2 Magnesium – Migraine

In August 1987, Swanson explored the research question of the role of Magnesium in Migraine Disorder. Through the methods described in [91] he discovered 11 neglected connections between them. He found that Magnesium deficiency might exacerbate Migraine due to complications involving Stress (Type A personality), Spreading Cortical Depression, Epilepsy, Platelet Aggregation, Serotonin, Substance P, Inflammation, Vasoconstriction, Prostaglandin formation and Hypoxia. As a natural calcium channel blocker, Magnesium
could further be used to prevent migraine attacks.

We seeded our algorithm with Magnesium (C0024467) as the source and Migraine Disorders (C0149931) as the target. The path length was 2 and no predicate filter was used, to be more consistent with the discovery. The corpus consisted of more than 47,000 articles from the pre-August 1987 period (i.e., 41,507 abstracts on Magnesium and 6,171 on Migraine, 7 overlapping). There were 14697 unique concepts and 73,960 predications in the predications graph and 256 distinct paths of length 2 in the candidate graph. The algorithm terminated in less than one hour, producing 25 subgraphs (and 151 singletons) at 2σ and 6 subgraphs (and 231 singletons) at 3σ.

![Figure 5.6: Subgraph1 (k = 2, 2σ) Magnesium - Migraine](image)

It was known from the 1973 article [PMID4725298] by Vosgeru that Magnesium Glutamate was used to treat Migraine. Figure 5.6 shows that the intermediate Serotonin was found in subgraph1 at 2σ. The lone article [PMID3629724] by Pertseva et al. from which the predication [Magnesium INTERACTS_WITH Serotonin] was extracted, is inconclusive. According to Swanson this association should be that [Magnesium INHIBITS Serotonin]. Conversely, the article [PMID3512233] by Houston et al. from which the predication [Serotonin CAUSES Migraine] was extracted (among three others), suggests that elevated levels of Serotonin can induce Vasoconstriction, which causes Migraine. Houston explicitly states that “much evidence has implicated serotonin (5-hydroxytryptamine) in the pathogenesis of migraine.” The article further notes that Serotonin is released from Platelet Aggregation and might be involved in a positive covariance relationship with Mi-
graine, as noted by Swanson. The 1987 article [PMID2440758] by Briel et al. (through a MEDLINE search) confirms that Magnesium inhibits Platelet Aggregation. It follows that elevated Magnesium levels may inhibit both Serotonin and Platelet Aggregation, and so treat Migraine.

![Figure 5.7: Subgraph4 (k = 2, 2σ) Magnesium - Migraine](image)

Figure 5.7 shows subgraph4, which contains the intermediate Prostaglandins between Magnesium and Migraine. The lone article [PMID3871957] by Friedlander et al. from which the predication [Prostaglandins INTERACTS WITH Magnesium] was extracted, suggests a decrease in prostaglandin synthesis is accompanied by lower levels of magnesium (and calcium). This is based on the title: “Decreased calcium and magnesium urinary excretion during prostaglandin synthesis inhibition in the rat” as noted by Swanson. The 1986 article [PMID3016750] by Nigam et al. confirms that [Magnesium STIMULATES Prostaglandins] as suggested by Swanson. The article [PMID89390] by Hakkarainen et al. from which the predication [Prostaglandins ASSOCIATED WITH Migraine Disorders] was extracted (among only three others) states that “Tolfenamic acid (a potent inhibitor of prostaglandin biosynthesis) was effective in treating acute migraine attacks.” The specific role of Prostaglandins in Migraine was unclear however, even after discovery browsing.

Figure 5.8 shows that the intermediate Vascular Disease was found explicitly in subgraph9. The title of the article [PMID4260015] by Wustenberg et al. from which the predi-
cation [Magnesium ASSOCIATED WITH Vascular Disease] was extracted, suggests a role for magnesium in vascular reactivity. The title of the article reads in part, “…Findings in magnesium metabolism in vascular diseases.” Similar to the predication with Serotonin, it is unclear from this title that [Magnesium INHIBITS Vasoconstriction] as noted by Swan-son. On the other hand, the article [PMID1153064] by Domzal, from which the predication [Migraine Disorders ISA Vascular Diseases] was extracted (among three others), suggests that migraine is also a vascular disorder, although primarily a cerebral disorder. The lone article [PMID3945397] by Coppeto et al. from which the predication [Migraine Disorders AFFECTS Vascular Diseases] was extracted provides more compelling evidence by linking migraine and vascular retinopathy as suggested by Swanson. Coppeto reported that “two migraineurs suffered sudden, persisting loss of vision from retinal vascular occlusion.” This effect is consistent with the observation by Houston et al. from the article [PMID3512233] on Serotonin from subgraph1. Salati et al. in [PMID6225285], from which the predication [Migraine Disorders ISA Vascular Diseases] was extracted, noted a dependency among Migraine, Vascular diseases, Epilepsy and Autoscopy (outer-body hallucination).

The two calcium channel blockers, Nifedipine and Verapamil were the only intermediates in subgraph22, shown in Figure 5.9. All three articles [PMID2425960, PMID3673084, PMID6539877] confirmed that these calcium channel blockers treat Migraine as suggested
Table 5.2: Comparison of rediscoveries with other approaches for Magnesium - Migraine

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate (s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
<th>Weeber</th>
<th>Blake</th>
<th>Gordon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium Channel Blockers</td>
<td>Y 22</td>
<td>Y 3</td>
<td>Y</td>
<td>Y</td>
<td>10</td>
<td>Y 1</td>
</tr>
<tr>
<td>Epilepsy</td>
<td>Y* 9</td>
<td>–</td>
<td>Y</td>
<td>Y</td>
<td>8</td>
<td>Y 3</td>
</tr>
<tr>
<td>Hypoxia</td>
<td>–</td>
<td>Y 5</td>
<td>–</td>
<td>Y 6</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>Inflammation</td>
<td>Y* ZR-3</td>
<td>Y 2</td>
<td>Y</td>
<td>Y 170</td>
<td>Y 82</td>
<td></td>
</tr>
<tr>
<td>Platelet Activity</td>
<td>Y* 1</td>
<td>Y 2</td>
<td>Y</td>
<td>Y 2</td>
<td></td>
<td>Y 8</td>
</tr>
<tr>
<td>Prostaglandins</td>
<td>Y 4</td>
<td>Y 1</td>
<td>Y</td>
<td>Y 42</td>
<td>Y 27</td>
<td></td>
</tr>
<tr>
<td>Type A Personality</td>
<td>–</td>
<td>Y 1</td>
<td>Y</td>
<td>Y 23</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Serotonin</td>
<td>Y 1</td>
<td>Y 1</td>
<td>Y</td>
<td>Y 5</td>
<td>Y 1</td>
<td></td>
</tr>
<tr>
<td>Cortical Depression</td>
<td>–</td>
<td>Y 6</td>
<td>–</td>
<td>Y 45</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Substance P</td>
<td>–</td>
<td>Y 18</td>
<td>Y</td>
<td>Y 38</td>
<td>Y 23</td>
<td></td>
</tr>
<tr>
<td>Vascular mechanisms</td>
<td>Y 9</td>
<td>Y 1</td>
<td>Y</td>
<td>Y 46</td>
<td>Y 16</td>
<td></td>
</tr>
</tbody>
</table>

by Swanson. The article [PMID537283] by Khoda et al. from which the predication [Verapamil INTERACTS_WITH Magnesium] was extracted suggested that Magnesium inhibits Verapamil as noted by Swanson.

Figure 5.9: Subgraph22 (k = 2, 2σ) Magnesium - Migraine

The intermediate Hydrocephalus (accumulation of fluid in the brain), which leads to Brain Edema (referred to as or inflammation by Swanson), was found among the zero-rarity associations (see Table 5.2). The remaining intermediates Hypoxia, Spreading Cortical Depression, Stress (Type A Personality) and Substance P were not found among the subgraphs.
Interestingly, only subgraph22 on the calcium channel blockers was a complex subgraph in which existing knowledge was recovered. While several intermediates related to Vascular Reactivity, such as Vasospasm, Vascular Function, Vasoconstriction and Vascular Disease exists, however their shared context did not meet our threshold for path relatedness, hence they were not grouped into the same cluster. The shortcomings of SemRep in extracting Vascular Reactivity may also have been a limiting factor. Still, altogether 10 out of the 25 subgraphs contained complex associations. Subgraph7 (shown in Figure 5.10) for example, links Theophylline and Caffeine, which have different semantic types, but belong to the general group of Stimulants, with Magnesium. Subgraph6 (not shown) associates Epinephrine and Glucose from the perspective of Metabolism. Table 5.2 shows that ultimately, 7 out of the 11 associations found by Swanson could be found using our approach, often with their substantiation.

5.1.3 Somatomedin C – Arginine

In April 1989, Swanson explored the research question of the role of the dietary amino acid Arginine, in Growth along with the protein Somatomedin C (also called Insulin-Like Growth Factor 1 - IGF1). Through the methods discussed in [92], Swanson discovered 4 implicit connections between the two concepts. He found that Arginine intake could: 1) stimulate Growth and protein synthesis, 2) promote Wound Healing and cell regenera-
tion, 3) facilitate nutritional repletion, thereby overcoming malnourishment and 4) improve overall Body Mass (and Weight), especially in the elderly and debilitated.

We seeded our algorithm with Somatomedins (C0037657) and Insulin-Like Growth Factor I (C0021665) as the sources, and Arginine (C0003765) as the target. The corpus consisted of more than 11,000 articles (819 on Somatomedins and 10,698 on Arginine, with 53 overlapping) in the pre-April 1989 period. The path length was set to 2, this time with the STRICT predicate filter. There were 5195 concepts and 17,058 predications in the predications graph and 239 distinct paths in the candidate graph. The algorithm terminated in less than one hour producing 10 subgraphs (and 153 singletons) at 2σ and 7 subgraphs (and 205 singletons) at 3σ.

![Subgraph5](image)

Figure 5.11: Subgraph5 ($k = 2, 3\sigma$) Somatomedin C – Arginine

Figure 5.11 shows the intermediate Growth Hormone in subgraph5 at 3σ. The sequence of predications [Arginine STIMULATES Growth Hormone], and [Growth Hormone STIMULATES Somatomedins] is entirely correct and requires no further proof (in terms of rediscovery). Still, for verification, we confirmed in the article [PMID6394628] by Chew et al. that dietary Arginine stimulates the release of Growth Hormones. These Growth Hormones then stimulate the production of Somatomedin C (IGF1), which leads to cell
growth and increased body size and muscle (i.e., protein synthesis), as noted in article [PMID7194347] by Clemmons et al. The same association is captured in subgraph6 at 2σ (not shown).

Table 5.3: Comparison of rediscoveries with other approaches for Somatomedin C - Arginine

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Somatomedin C - Arginine</td>
<td>Growth Hormone</td>
<td>Y</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Body Weight</td>
<td>Y*</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Malnutrition</td>
<td>Y*</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Wound healing</td>
<td>– –</td>
<td>– –</td>
</tr>
</tbody>
</table>

Several articles from which the seemingly spurious predication [Arginine TREATS Child] was extracted, upon investigation, were shown to actually discuss Glucagon and Insulin. This includes the article [PMID7204541] by Blethen et al. whose title is “Plasma somatomedins in children with hyperinsulinism.” Likewise, the article [PMID6205015] by Binoux et al. from which the predication [Arginine TREATS Rattus norvegicus] was extracted, discusses observations regarding Insulin-like Growth Factor 1 in the serum of rats. The article [PMID7007553] by Ashby et al. from which the same predication was extracted, discusses the effects of Progesterone and Insulin in rats, resulting from Glucose and Arginine stimulation. Based on these observations, it is reasonable conclusion that this subgraph captures the shared context of role of Insulin in Somatomedin C and Arginine.

Subgraph7 at 3σ, shown in Figure 5.12 contains the concept Growth as an intermediate instead of Growth Hormone (similar to subgraph2 at 2σ, not shown). The sequence of predications [IGF1 CAUSES Growth] and [Growth PRODUCES Somatomedins], is interesting because the article [PMID3748655] by van Buul-Offers et al. from which the predication [IGF1 CAUSES Growth] was extracted states that IGF1 “increases body length and weight, as well as the growth of several organs of Snell dwarf mice,” which is consistent with Swansons report. The association involving Malnutrition and Somatomedin produc-
tion was found in the article [PMID7023246] by McCumbee et al. No obvious association to Wound Healing was found using our methods. Table 5.3 shows that 3 out of 4 intermediates could be found using our approach.

5.1.4 Indomethacin – Alzheimer’s Disease

In June-July 1995, Neil R. Smalheiser and Swanson studied the mechanism by which Indomethacin might affect patients with Alzheimer’s Disease. Indomethacin is an anti-inflammatory used to reduce pain, swelling and stiffness. Through the methods discussed in [79] Swanson and Smalheiser found that Indomethacin decreases Plasma Membrane Fluidity, inhibits M2-muscarinic, Acetocylcholine and Lipid Peroxidation, which are elevated in patients with Alzheimer’s Disease. Indomethacin also stimulates killer T-cell activity and Thyrotropin-Releasing Hormone (TRH), whose levels are reduced in patients with Alzheimer’s Disease.

We seeded our algorithm with Indomethacin (C0021246) and Indocin (C0700798) as sources, and Alzheimer’s Disease (C0002395) as the target. The corpus consisted of more than 40,000 articles in the pre-July 1995 corpus (24,715 articles on Indomethacin and 15,736 on Alzheimer’s, with 9 overlapping). After applying the STRICT predicate filter at
path length 2, the algorithm terminated in less than one hour, producing 15 clusters at $2\sigma$.

![Figure 5.13: Subgraph2 (k = 2, $2\sigma$) on Indomethacin, Lipid Peroxidation and Alzheimer’s Disease](image)

Figure 5.13 shows subgraph2 at $2\sigma$, which contains the predication \textit{[Indomethacin INHIBITS Free Radicals]}. Among the three articles from which this predication was extracted, the article [PMID1315618] by Sagar et al., explicitly states that “GLA-induced free radical generation and lipid peroxidation were also inhibited by indomethacin.” The article [PMID7969721] by Friedlich et al., from which the predication \textit{[Free Radicals DISRUPTS Alzheimer’s Disease]} was extracted, confirms that “quenching free radicals slows the clinical progression of Alzheimer’s disease.” This suggests that \textit{Indomethacin} may treat \textit{Alzheimer’s Disease} by inhibiting \textit{Free Radicals} (and thereby \textit{Lipid Peroxidation}), which are implicated in \textit{Alzheimer’s Disease}.

Figure 5.14 shows subgraph4, which contains the intermediate \textit{Acetylcholine} (to which $M2$-muscarinic = \textit{Muscarinic acetylcholine receptor M2} is related). While it is inconclusive whether \textit{[Indomethacin STIMULATES Acetylcholine]} or vice versa from the articles from which those predications were extracted [PMID6408242, PMID8075854 and PMID8315533], it is clear that \textit{[Acetylcholine PREVENTS Alzheimer’s Disease]} from the article [PMID2655861] by Vida et al. The article states that “Several lines of evidence have implicated acetylcholine (ACh) as one of the neurotransmitters found to be decreased in
Alzheimers disease (AD).” Indomethacin is believed to inhibit the release of Acetylcholine.

Figure 5.15: Subgraph3 (k = 2, 2σ) on Indomethacin, Hydrocortisone and Alzheimer’s Disease

Figure 5.15 shows subgraph3, which contains an association among Indomethacin, Alzheimer’s Disease and Hydrocortisone. This subgraph was initially deemed uninteresting and was therefore neglected. Upon investigation however, the article [PMID3140601] by Goldstein et al. from which the predication [Indomethacin INHIBITS Hydrocortisone] was extracted (among 2 others), discusses the “Suppression of natural killer cell activity by hydrocortisone” from the title. However the predicate INHIBITS extracted by SemRep is incorrect. Indomethacin actually STIMULATES Hydrocortisone. Goldstein noted that “Indomethacin was able to partially reverse hydrocortisone suppression in 7 out of 12 experiments.” Indomethacin therefore stimulates Hydrocortisone, which in turn stimulates
Natural Killer T-cell activity, as noted by Smalheiser.

Figure 5.16: Subgraph14 (k = 2, 2σ) on Indomethacin – Alzheimer’s Disease

Subgraph14, shown in Figure 5.16, contained the concept Cyclic AMP, which is used for intracellular Signal Transduction, and also the intermediate Lymphokines (which are Cytokines produced by Lymphocytes). The Thyrotropin-Releasing Hormone were found in zero-rarity singleton20. Ultimately, 5 out of the 15 subgraphs at 2σ contained complex associations, and Table 5.4 shows that 5 out of 7 intermediates could be found using this approach.

Table 5.4: Comparison of rediscoveries with other approaches in the literature for Indomethacin - Alzheimer’s Disease

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indomethacin - Alzheimer’s</td>
<td>Acetylcholine</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Disease</td>
<td>Lipid peroxidation</td>
<td>Y*</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>M2-muscarinic</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Membrane Fluidity</td>
<td>-</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Lymphocytes</td>
<td>Y*</td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td>Thyrotropin</td>
<td>Y</td>
<td>ZR-20</td>
</tr>
<tr>
<td></td>
<td>T-lymphocytes (T-Cells)</td>
<td>Y*</td>
<td>Y</td>
</tr>
</tbody>
</table>

112
5.1.5 Estrogen – Alzheimer’s Disease

Again in June-July 1995, Smalheiser and Swanson studied the mechanism by which Estrogen protects against Alzheimer’s Disease. Through the methods discussed in [80] they discovered that Estrogen exhibits Antioxidant Activity, which are protective against Alzheimer’s Disease. Several proteases including Cathepsin D and Superoxide Dismutase are involved with Estrogen production. Likewise several proteins are implicated in Alzheimer’s Disease, including Calbindin D28k, Apolipoprotein E (ApoE), which are inhibited by Estrogen. Furthermore, Estrogen enhances Neuronal Activity and induces the Cytochrome C Oxidase Subunit 2 protein, which protects against Alzheimer’s Disease.

We seeded our algorithm with Estrogens (C0014939) as the source and Alzheimer’s Disease (C0002395) as the target. The pre-June 1995 corpus contained more than 38,000 articles (23,167 articles on Estrogens and 15,694 on Alzheimer’s, with only 12 overlapping). There were 4,995 concepts, 12,929 predications in the predications graph and 69 paths in the candidate graph at path length 2, using the STRICT predicate filter. The algorithm terminated in less than 1 hour and produced 4 subgraphs at 2σ.

![Figure 5.17: Subgraph3 (k = 2, 2σ) on Estrogens, Genes and Alzheimer’s Disease](image)

Figure 5.17 shows subgraph3, which contains 6 paths. Among them is the seemingly innocuous association [Estrogens INHIBITS Genes], [Estrogens STIMULATES Genes], [Estrogens INHIBITS RNA, Messenger], [Estrogens STIMULATES RNA, Messenger], [Estrogens CAUSES Growth], [Estrogens CAUSES Alzheimer’s Disease].
and [Genes CAUSES Alzheimer’s Disease]. Upon investigation, to determine which genes were involves, we found the article [PMID7816154] by Kurz et al. from which the predication [Genes CAUSES Alzheimer’s Disease] was extracted (among 3 others). This article states that “apolipoprotein E is involved in the pathogenesis of Alzheimers disease.” Kurz suggests might ApoE might negatively affect Alzheimer’s Disease. The article [PMID8502236] by Berkowitz et al. (among 3 others) confirm that [Estrogens INHIBITS ApoE] after experiments were conducted on chickens. Berkowitz also discusses the role of RNA, Messenger (mRNA). The article [PMID8175820] by Majeska et al. explicitly states that “Estrogen increased the levels of messenger RNA.” The article [PMID8362984] by Kalaria et al. provides evidence in support of the predication [RNA, Messenger CAUSES Alzheimer’s Disease]. Furthermore, Kalaria also suggests that increased levels of another protease, Plasma Protein Antithrombin III (ATIII) is involved in the pathology of Alzheimer’s Disease.

The candidate graph produced only 42 distinct associations, from the 69 paths altogether. Among these, only 5 associations were known to more than 1 document in MEDLINE. In subgraph4, the lone article [PMID2591722] by Nys et al. from which the predication [Estrogens STIMULATES Calbindin] was extracted, explicitly refers to Calbindin D28k, as noted by Smalheiser. The lone article [PMID1317496] by Sutherland et al. from which the predication [Calbindin DISRUPTS Alzheimers Disease] was extracted however, suggests that it is [Alzheimer’s Disease DISRUPTS Calbindin], and remains inconclusive.

We also found the concept Free Radicals in subgraph4, and the list of zero-rarity singletons. It has already been established from the Indomethacin – Alzheimers experiment, that [Free Radicals CAUSES Alzheimers Disease] and “quenching free radicals slows the clinical progression of Alzheimers disease,” as reported in the article [PMID7969721] by Friedlich et al. The more interesting predication which states that [Estrogens STIMULATES Free Radicals], extracted from the article [PMID2924301] by Roy et al. appears to be a SemRep error. Roy argues that Estrogen successfully treated Lipid Peroxidation in hamsters by lowering Oxidative Stress; hence, [Estrogens DISRUPTS Free Radicals]. Ul-
timately, only 3 out of 8 intermediates could be found using our approach, shown in Table 5.5. Srinivasan did not attempt this experiment.

Table 5.5: Comparison of rediscoveries with other approaches in the literature for Estrogen - Alzheimer’s Disease

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estrogen - Alzheimer’s Disease</td>
<td>Antioxidant activity</td>
<td>Y*</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Alipoprotein E (ApoE)</td>
<td>Y*</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Calbindin D28k</td>
<td>Y</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Cathepsin D</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Cytochrome C oxidase</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Glutamate</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Receptor Polymorphism</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>Superoxide Dismutase</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

5.1.6 Calcium-Independent Phospholipase A2 – Schizophrenia

Sometime in 1997, Neil R. Smalheiser and Swanson explored the research question of the role of a Calcium-Independent form of Phospholipase A2 in Schizophrenia. Previous experiments suggested that Calcium-Dependent PLA2, and not the Calcium-Independent PLA2 was elevated in the serum of patients with Schizophrenia. Using the approaches discussed in [81] Smalheiser suggested that Chronic Oxidative Stress elevates Calcium-Independent PLA2 levels, and is also a factor in Schizophrenia. They suggested therefore that Antioxidants, which reduce the level of Calcium-Independent PLA2, might be effective in treating Schizophrenia.

We seeded our algorithm with Calcium-Independent Phospholipase A2 (C2830173), Phospholipase A2 (C0031667) and Phospholipase A2-alpha (C1744635) as sources, and Schizophrenia (C0036341) as the target. The corpus consisted of more than 42,000 articles (343 articles on Calcium-Independent Phospholipase A2 and 42,023 on Schizophrenia, 1 overlapping), pre-1997. Notably, this one overlapping article [PMID9152103], as noted by Smalheiser, piqued their interest and was critical to their discovery. There were 3,851
distinct concepts in the predication graph, and 797 paths in the candidate graph at path length 2. The algorithm terminated in less than one hour, producing 10 subgraphs, all 10 singletons, at $2\sigma$.

Figure 5.18: Singleton2 ($k = 2, 2\sigma$) on Calcium-Independent PLA2, Selenium and Schizophrenia

Singleton2, shown in Figure 5.18, suggests several proteins might be involved in causing Schizophrenia. Upon closer investigation to determine the specific proteins involved, we found the article [PMID7739414] by Berry, from which the predication $[\text{Proteins CAUSES Schizophrenia}]$ was extracted (along with another [PMID9399691] by Jones et al.), which states that “selenoprotein P, a hypothesized selenium transport protein, is a likely candidate for a protein involved in the etiology of a form of schizophrenia.” The article by Jones implicated Polyglutamine but not Selenoprotein P. The lone article [PMID7518449] by Andersson et al. from which predication $[\text{Proteins INHIBITS Phospholipase A2}]$ was extracted, does not mention Selenium or Selenoprotein P, neither did any of the remaining singletons. A PubMed search for PLA2 and Selenium prior to 1997 (i.e., Query: phospholipase a2 AND selenium AND 1865:1997[DP]) returned 13 articles. The critical overlapping article [PMID7782894] by Kuo et al., noted by Smalheiser, was fourth in the search result list. This article clearly suggests that oxidative stress, through “deficiency of both vitamin E and selenium activates and/or induces unique calcium-independent forms of phospholipase A2.” Ultimately, all three intermediates could be found using our approach, as shown in Table 5.6.
Table 5.6: Comparison of rediscoveries with other approaches in the literature for Calcium-Independent Phospholipase A2 – Schizophrenia

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
<th>Srinivasan</th>
</tr>
</thead>
<tbody>
<tr>
<td>Calcium-Independent PLA2 - Schizophrenia</td>
<td>Oxidative stress</td>
<td>Y*</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Selenium</td>
<td>Y*</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Vitamin E</td>
<td>Y*</td>
<td>3</td>
</tr>
</tbody>
</table>

5.1.7 Chlorpromazine – Cardiac Hypertrophy

In January 2002 Jonathan D. Wren explored potentially new treatments for Cardiac Hypertrophy (CH), which is characterized by abnormal enlargement of heart muscle tissue. Using open discovery [107], Wren et al. found that the antipsychotic Chlorpromazine (CPZ) could be used as a new treatment for Cardiac Hypertrophy, which they later confirmed through laboratory testing. In their experiments groups of mice received equal amounts of the beta-adrenergic agonist Isoproterenol. Another group that received larger amounts Chlorpromazine, showed less Cardiac Hypertrophy than mice. Wren also suggested a role for Calmodulin-Independent Phosphatase Calcineurin as another mechanism by which [Chlorpromazine TREATS Cardiac Hypertrophy].

We seeded our algorithm with Chlorpromazine (C0008286) as the source, and Cardiac Hypertrophy (C1383860) and Cardiomegaly (C0018800) as targets. The corpus consisted of more than 40,000 articles (12,500 on Chlorpromazine and 27,689 on Cardiac Hypertrophy, 2 overlapping), for the pre-January 2002 corpus, as specified by Wren. Only 2 articles overlapped in this set. There were 7,234 unique concepts, 21,967 predications and 255 paths in the candidate graph at path length 2, using the STRICT predicate filter. The algorithm terminated in less than one hour, and produced 14 subgraphs and 60 singletons at $2\sigma$.

Figure 5.19 shows subgraph12, in which the intermediate Isoproterenol was found. The article [PMID6165961] by Rossi et al. from which the predication [Isoproterenol CAUSES Cardiomegaly] was extracted, confirms that it causes Cardiac Hypertrophy when

\[117\]
Figure 5.19: Subgraph12 (k = 2, 2\(\sigma\)) on Chlorpromazine, Isoproterenol and Cardiac Hypertrophy

... elevated. On the other hand, one of the two articles from which the predication [Isoproterenol STIMULATES Chlorpromazine] was extracted, [PMID2420847] by Van der Ven et al., claimed that Chlorpromazine has no effect on Isoproterenol. The other however [PMID203365] by Tsang et al., clearly states the correct association, which is that “Chlorpromazine completely blocked the increase induced by noradrenaline, isoproterenol, clonidine, serotonin or dopamine;” that is [Chlorpromazine INHIBITS Isoproterenol]. Subgraph14 is also interesting because it contains both Propranolol, which also a beta-adrenergic agonist and Norepinephrine also implicated in Cardiac Hypertrophy (reported in both [PMID9054858] by Yamazaki et al).

Calcineurin was found in subgraph5, shown in Figure 5.20, which contains only two paths. The first predication states that [Chlorpromazine INHIBITS Calcineurin], supported by [PMID9001710] by Gong et al. All articles articles [PMID9568714, PMID10679475, PMID11248077, PMID11773940, PMID10189350] confirm that [Calcineurin CAUSES Cardiac Hypertrophy] instead of the opposing predication [Calcineurin DISRUPTS Car-
Figure 5.20: Subgraph $5$ ($k = 2, 2\sigma$) on Chlorpromazine, Calcineurin and Cardiac Hypertrophy

Table 5.7: Comparison of rediscoveries with other approaches in the literature for Chlorpromazine - Cardiac Hypertrophy

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Intermediate(s)</th>
<th>Cameron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chlorpromazine - Cardiac Hypertrophy</td>
<td>Calcineurin</td>
<td>S Y 5</td>
</tr>
<tr>
<td></td>
<td>Isoproterenol</td>
<td>Y 12</td>
</tr>
</tbody>
</table>

*diaic Hypertrophy*, which is the result of a recurring SemRep error.

### 5.1.8 Testosterone – Sleep

In 2012 Miller et al. conducted research which provided insights into the link between Hypogonadism and diminished sleep quality in men. Hypogonadism is a hormonal disorder that is characterized by “a diminished functional activity of the gonads – the testes and ovaries in males and females, respectively – that may result in diminished sex hormone biosynthesis and impaired gamete production and/or regulation.”

That sleep quality in diminishes in aging men, more than women was well known. However, the mechanism of diminished sleep quality in aging men and hypogonadism was poorly understood. Based on the methods in [61], Miller explained that a consequence of hypogonadism is high levels of Cortisol. Miller postulated that as men age and testosterone levels diminish, (hypogonadism sets in), cortisol level increase. This leads to diminished sleep quality,

more in men than women, since testosterone levels in women are more stable in women throughout aging.

We seeded our algorithm with Testosterone (C0039601) as the source and Sleep (C0037313) as the target. There were 113,058 articles (57,540 articles on Testosterone and 55,725 on Sleep, with 207 overlappins) in the pre-2012 corpus. The algorithm terminated in less than one hour, and produced 11 subgraphs at $2\sigma$ and 10 subgraphs at $3\sigma$. Cortisol (or Hydrocortisone) was found in subgraph7 at $3\sigma$, shown in Figure 5.21, and also in subgraph11 at $2\sigma$. The article [PMID10484567] by Kraemer et al. from which the predication [Testosterone INHIBITS Hydrocortisone] was extracted states that “With training the older group demonstrated a significant increase in total testosterone in response to exercise stress along with significant decreases in resting cortisol.” On the other hand, several articles, including the article [PMID8548511] by Kern et al. from which the predication [Hydrocortisone DISRUPTS Sleep] was extracted, confirmed that Cortisol disrupts Sleep. Kern notes that “Both changes in GH and cortisol secretion may act together to reduce anabolic functions of sleep in the aged.”

Figure 5.21: Subgraph7 ($k = 2, 3\sigma$) on Testosterone, Cortisol and Sleep

5.1.9 Diethylethyl Phthalate (DEHP) – Sepsis

In 2013, Michael J. Cairelli and Thomas C. Rindflesch investigated the underlying mechanism of the obesity paradox articulated in [18]. Abhyankar et al., reported that patients who
are slightly overweight and mildly obese, exhibit better mortality rates than normal weight patients up to 1 year after admission from intense care units (ICU). Using the methods described in [2] Cairelli discovered a link between the toxic organic compound Diethylhexyl Phthlate (DEHP) and Sepsis. Specifically they found that both DEHP and Obesity increase the levels of the nuclear receptor PPAR gamma in humans. PPAR gamma has been shown to inhibit Inflammation. Unfortunately, we did not find this association in any of the generated subgraphs using our approach. Instead, several other interesting zero-rarity associations were found. Novel associations involving Leptin, Lactoferrin and Immunoglobulin 1 (IgG1) were observed, but require further investigation.

The evidence-based evaluation shows that several intermediates across a number of discoveries could be retrieved using the context-driven automatic subgraph creation model, either directly or indirectly in the subgraphs. Moreover, the predicates provided a means for understanding the meaning of the associations. Provenance in MEDLINE was also shown to be very useful. In cases where both predicates and provenance were insufficient, inferred associations in the subgraphs were explored through MEDLINE queries and searching and sifting. Ultimately, the provision of numerous subgraphs, which contained promising links to facilitate making rediscoveries from scientific literature, was shown to be effective. And we believe that this approach plausible for future applications.

5.2 Statistical

In the previous section, we showed that our context-driven subgraph method facilitated the rediscovery of 8 existing discoveries with substantiation in MEDLINE. While these are encouraging results, one might argue that our experiments were biased since we know the intermediates to be found in the first place. Hence, it was easy to find them in the subgraphs. A more important question is how interesting are subgraphs in general, such that an arbitrary domain expert might be motivated to explore them altogether? To address this
question, we conducted a statistical evaluation, which uses association rarity to compute interestingness. If the interestingness score of the subgraphs across an entire experiment is low, then the rediscoveries were fortuitous and the associations that led to the rediscoveries were serendipitous, rather than systematic. While this not a complete loss, it is still less than ideal.

To perform the evaluation, for each path in each subgraph across the 8 rediscoveries (excluding singletons), a PubMed query was executed using the eUtils Web Service\(^5\). This is used to determine the number of documents that contain the association in MEDLINE, with the date restriction enforced. For example, for the path \([\text{Arginine STIMULATES Growth Hormone}], \ [\text{Growth Hormone STIMULATES Somatomedins}],\) the query “Arginine AND Growth Hormone AND Somatomedins AND 1865:1989[DP]” is composed, where Arginine, Growth Hormone, and Somatomedins represent an association. The rarity of the associations

\[
r(E) = \frac{\sum_{p_i \in E} f(A(p_i))}{|E|} \tag{5.1}
\]

across all subgraphs in an experiment \(E\) denoted \(r(E)\), is the average of the association rarity, where \(f(A(p_i))\) is the frequency of a unique association in MEDLINE. The interestingness of an experiment

\[
I(E) = \frac{1}{r(E) + 1} \tag{5.2}
\]

can be computed as the reciprocal of rarity. Table 5.2 shows the rarity and interestingness scores for each of the eight successful rediscoveries.

For the Raynaud Syndrome – Dietary Fish Oils experiment, there were 10 unique intermediates/associations, all of which were zero-rarity in MEDLINE. This is not surprising, since Swanson noted in [89] that only four articles from the Raynaud literature overlapped with the Fish Oil literature by 1986. The rarity of these subgraphs is therefore 0.00, while the interestingness is 1 (meaning absolutely interesting). For Magnesium – Migraine there

Table 5.8: Rarity and Interestingness of the subgraphs in the rediscoveries

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Unique Association</th>
<th>MEDLINE Frequency</th>
<th>r(E)</th>
<th>I(E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raynaud-Fish Oil</td>
<td>10</td>
<td>0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Magnesium-Migraine</td>
<td>48</td>
<td>27</td>
<td>0.56</td>
<td>0.64</td>
</tr>
<tr>
<td>Somatomedin C-Arginine</td>
<td>18</td>
<td>306</td>
<td>17.00</td>
<td>0.06</td>
</tr>
<tr>
<td>Indomethacin-Alzheimer</td>
<td>21</td>
<td>9</td>
<td>0.43</td>
<td>0.70</td>
</tr>
<tr>
<td>Estrogen-Alzheimer</td>
<td>42</td>
<td>36</td>
<td>0.86</td>
<td>0.54</td>
</tr>
<tr>
<td>PLA2-Schizophrenia</td>
<td>10</td>
<td>0</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>CPZ-Cardiac Hypertrophy</td>
<td>21</td>
<td>2</td>
<td>0.10</td>
<td>0.91</td>
</tr>
<tr>
<td>Testosterone-Sleep</td>
<td>61</td>
<td>654</td>
<td>10.72</td>
<td>0.09</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>29</strong></td>
<td><strong>129</strong></td>
<td><strong>3.71</strong></td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>

were 48 unique associations. The most commonly known intermediates were Hypertensive Disease (3), Individual (3) and Vascular Diseases (4), respectively among a total of 27 documents. There were also 35 zero-rarity associations. The overall rarity of the subgraphs was therefore 27/48 = 0.56 and the interestingness is 0.64. For Somatomedin C – Arginine there were 18 unique associations/intermediates among a total of 306 documents and the most commonly known intermediates were Child (16), Somatropin (63) and Growth Hormone (63). There were only two zero-rarity associations, with intermediates Mus (0) and Falls (0). Clearly these are not interesting. Not surprisingly, the overall interestingness score of these subgraphs was 306/18 = 17 and the interestingness is low (0.06). This suggests that this field is better studied than others. It also partially supports the observation by Gordon and Dumais [35] that frequency of intermediates may be sufficient for finding intermediates in some cases, but independent for finding related concepts to elucidate associations. There were 21 unique associations for Indomethacin – Alzheimer and 16 were zero-rarity. Hydrogen Peroxide (2), Interleukin-1 (2) and Free Radicals (3) were the most commonly known among a total of 9 documents. The overall rarity score was 9/21 = 0.43 and the interestingness is 0.70. For Estrogen – Alzheimer there were 42 unique associations, with 36 zero-rarity associations. Metabolism (6), Dementia (10) and Senile dementia (10) were the most commonly known intermediates among a total of 36 documents. The rarity score is 36/42 = 0.86 and the interestingness is 0.54. For Calcium-Independent PLA2
– Schizophrenia there were 10 unique intermediates/associations (singletons described in Section 4.1.6), each of which was zero-rarity. Hence, the rarity of this subgraph is 0.00 and the interestingness is high (1.0). For Chlorpromazine – Cardiac Hypertrophy there were 21 unique associations, and 19 at zero-rarity. The most commonly known were Catecholamines (1) and Hypertensive disease (1) among a total of 2 documents. The rarity is therefore 2/21 = 0.10 and the interestingness is high (0.91). For Testosterone – Sleep, there were 61 unique associations/intermediates with 20 at zero-rarity. The most commonly known were Proteins (63), Symptoms (91) and Hormones (207) among a total of 654 documents. The overall rarity score is therefore 654/61 = 10.72 and the interestingness is low (0.09). This is not surprising, since these two domains (Testosterone and Sleep) are fairly well studied. Again, another scenario suggesting that frequency may not be as effective for elucidating associations.

Across all 8 rediscoveries, the average rarity score is therefore 3.71 and the average interestingness is 0.62. This suggests that an association chosen at random from the rediscoveries is likely to be known only to 4 documents in MEDLINE. Such a low rarity score suggests that the subgraphs themselves might be interesting. This is however not surprising, since most of the discoveries, at the time when made would have been inherently interesting situations and possibly not well studied in the literature. Testosterone – Sleep (2011) and Somatomedin C – Arginine (1990) are exceptional since each field was fairly well studied. The key issue was that the mechanism of lack of Testosterone on Sleep in aging men compared with women was not well understood. Neither was the reason for higher incidence of Alzheimer’s Disease in men then women.

5.3 Summary

This Chapter presented the results of an evidence-based evaluation and a statistical evaluation, to assess the context-driven automatic subgraph creation approach to LBD. Using the
approach, several intermediates from 8 out of 9 existing discoveries, at an average rarity of four in MEDLINE were retrieved. Predicates and provenance were also available to help understand the meaning of each association. These results provide evidence that this approach is plausible for LBD in future scenarios.
Biomedical Knowledge Exploration

Many LBD systems succeed in finding interesting and novel intermediates. However not all provide utilities to facilitate browsing the scientific literature, to enable further exploration and hypothesis generation. A considerable degree of domain expertise is still often required to substantiate the associations detected by the system. This Chapter examines various paradigms for biomedical knowledge exploration. It includes a review of: 1) the search-and-sift paradigm in Section 6.1, 2) ontology-driven text exploration in Section 6.2, 3) semantic predications-based text exploration in Section 6.3, and 4) the notion of discovery browsing in Section 6.4.

6.1 The Search-and-Sift Paradigm

Ramakrishnan et al. [69] discussed a search-and-sift paradigm for accessing information on the web. In this method, users first formulate keyword queries, then manually aggregate relevant information from the search results returned by the system. In the context of closed LBD, the user first issues a query \((A, C)\), then explores a list of intermediates returned by the system, which contains \(B\)-terms. A secondary set of queries may then be formulated, based on interesting \(AB\) and \(BC\) terms to obtain the associated documents from the corpus. Only after such documents have been retrieved, can the search-and-sift process begin. As the process unfolds, users develop curiosities from assertions in the documents and iteratively reformulate their secondary \((AB, BC)\) queries after each search cycle.
This process is often laborious, as users must pore over large volumes of text to find meaningful information. The ponderous nature of this task is not altogether surprising, since the system typically delivers documents, not necessarily information fragments, in response to user queries. This document-centric model of information persists, in spite of the general belief that users are information seekers, who are more interested in the information contained in documents, rather than the documents themselves. White et al., aptly described this phenomenon in [102], noting that information seeking is not just about a destination (i.e., a single or group of documents), instead the bits and pieces of knowledge acquired during the journey, are also important.

The manual effort required for text exploration using the search-and-sift approach can be alleviated in two ways. First, the system may fragment documents into smaller distinct information fragments, as often done in Question Answering (see Section 2.1.2). Second, the search system may provide users with a method for directly linking content from document-to-document based on the semantics of associations from the content. One way to achieve this, is through ontology-driven text exploration, which is discussed in the next Section.

6.2 Ontology-Driven Text Exploration

Sheth et al. introduced the concept of the relationship web in [74]. This concept envisions a web in which entities are connected using implicit, explicit linguistic and formal relationships. A document collection in which entities have been annotated according to formal background knowledge might then be browsed using semantic relationships from background knowledge, instead of searching and sifting. This vision is consistent with the memory extender, or Memex Vision, outlined by Vannevar Bush (1890–1974) in 1945 [17]. Bush noted that the human brain “operates by association. With one item in its grasp, it snaps instantly to the next that is suggested by the association of thoughts, in accordance
with some intricate web of trails carried by the cells of the brain.” This logic is the basis for semantic trailblazing, in a relationship web, where semantic metadata from formal background knowledge is superimposed over text.

Gonen [32] first implemented a prototype web application called the Semantic Browser, which enabled semantic trailblazing. Starting with an entity of interest $A$ in some arbitrary document $d_i$, the user could browse to another entity $B$, in another document $d_j$. To accomplish this, the user could select a relationship $r_1$, between $A$ and $B$. Then the user could browse to another entity $C$, in another document $d_k$, through another relationship $r_2$, between $B$ and $C$, and so on. The semantic browser was extended by Cameron et al. [22] into a discovery support system called Trellis, which provided functionality for bookmarking and search result reorganization. In another enhancement, called Scooner, Kavuluru et al. [49] integrated domain knowledge from Wikipedia with statements from the UMLS as background knowledge for knowledge exploration. The system provided support for collaborative filtering by enabling users to save and share the search results.

Doms et al. [24] implemented GoPubMed, which is a web service that allows users “to explore PubMed search results with the Gene Ontology (GO).” GoPubMed facilitates category-driven document browsing and neighborhood-based exploration of scientific literature. Users first execute a search for MEDLINE articles, which are classified according to the classes in the ontology. Users can explore the search results using this hierarchical category structure. Additionally, the system also provides highly related concepts to the initial search term, which facilitate browsing. Spotted entities in text, which are linked to the ontology can also be used for browsing. A similar tool, called XplorMed, restricts exploration to eight MeSH categories and several molecular biology databases, including OMIM, SMART, SWISS-PROT, and SpTrEMBL.

In [64], Müller et al. implemented Textpresso, which is a text mining and ontology-driven system for exploring biomedical literature. Textpresso implements information extraction methods to identify biological concepts in text, then enables browsing, based on
annotations linked to the ontology. The system was applied to a corpus related to genes and organisms. Hoffmann et al. in [37], developed a similar tool called IHOP, which enables sentence-level browsing for interactions between proteins and genes using background knowledge.

The idea of ontology-driven text exploration, while an advancement over the search-and-sift paradigm suffers several limitations. The first limitation is that the context of the relationships encountered during browsing reside in the background knowledge base, but not necessarily in the text. While the resulting trail of statements may be sufficient to provide interesting knowledge, there is often a disconnect between the context of the statements in the background knowledge and the assertions in the corpus. This subtle distinction is the key difference between ontology-based knowledge discovery and literature-based discovery. To more ably support LBD, systems should enable browsing of assertions directly present in the literature, leveraging background knowledge, where appropriate. In the next Section, semantic predications-based text exploration is discussed.

6.3 Semantic Predications-Based Text Exploration

Hristovski et al. [38] implemented a semantic predications-based approach for text exploration in SemBT. The application leveraged semantic predications extracted from MEDLINE using SemRep, together with gene-related assertions extracted using BioMedLEE [56]. Articles containing AB and BC associations could then be explored based on the sentences from which the associations were extracted. Entities and predicates are color-coded in the search results, but are not hyperlinked to the background knowledgebase.

In [20], we implemented an approach for semantic predications-based text exploration. The approach simulated user activity by automatically connecting semantic predications in answer passages from the 2006 TREC Challenge. By mimicking user activity, the system enabled predication-based text exploration (depicted by relations bounded with
black nodes in Figure 6.1) for Question Answering. While the approach has not been deployed in a live tool, it is based on the fundamental principle of predications-based text exploration. Cohen et al. [23] also developed a semantic predications-based framework for text exploration, called *Epiphanet*. Users express keyword queries, which are translated into a predication-based format. Search results are then visualized as subgraphs of semantic predications, linked to MEDLINE through provenance.

The fundamental limitation of semantic predications-based text exploration approaches is flexibility. None of the implemented approaches provide utilities for trailblazing, or realizing Vannevar Bush’s Memex vision. The application developed by Cohen, is promising, but requires more utilities to facilitate knowledge exploration. Likewise, our simulation in [20] is insufficient because it connects the maximum number of answer passages for a given question, regardless of whether the question has been answered. This leads to long and convoluted paths, in which subsequences make sense, but no particular path as a whole. While the path length limitations in [20], were addressed by: 1) imposing a maximum path length constraint and 2) specifying the source and target concepts, for closed discovery in [19], both approaches do not provide sufficient functionality to allow users to quickly and clearly gain insights into the full scope of associations in the literature.
6.4 Discovery Browsing

To systematically address this subject of biomedical text exploration, Wilkowski et al. introduced the notion of discovery browsing in [105]. Discovery browsing is enabled when a system guides the user through their exploration of the literature in a process of cooperative reciprocity. According to Wilkowski, the “user iteratively focuses system output, thus controlling the large number of relationships often generated in literature-based discovery systems.” This approach addresses the shortcomings of prior approaches to biomedical text exploration by providing utilities for navigating selected aspects of a domain in the corpus. It is a direct application of the idea of information access as a journey, rather than a destination, articulated by White [102].

Discovery browsing integrates tenets of LBD, with principles from automatic summarization (see Section 2.1.3). The discovery browsing system interleaves between LBD and summarization functionality, enabling an interplay between specificity and generalization of user interests. The system first performs general keyword-driven search, then presents various summarization perspectives to appeal to more specific user interests. The refined search results are then provided as a navigable predications graph (i.e., an LBD artifact), with semantic predications whose provenance is provided in the literature. Insights from the predications graph are then used to activate new searches in an iterative process of cooperative reciprocity between the system and user. Both Cairelli et al. in [18] and Miller et al. [61] exploited this idea of discovery browsing, in Semantic MEDLINE, to discover new knowledge.

The Obvio platform developed in this dissertation is geared towards discovery browsing. Although the current application only showcases preprocessed search results for nine concept pairs, it illustrates the combination of the LBD paradigm with the search and sift paradigm. A fully developed application will accept two keywords \((A, C)\), a cut-off date \(dt\), and a maximum path length \(k\) as the search query \(q = (A, C, dt, k)\), for any pair of concepts. Given such a query, the system will produce search results as a list of subgraphs,
in which each semantic predication is linked to MEDLINE through provenance. Subgraphs will consist of labeled directed edges, which provide insights into the meaning of the associations between concepts \((A, C)\). The generated subgraphs capture multiple dimensions of associations between concept pairs on different themes. Conflicting semantic predications present in subgraphs can be resolved using provenance. Additionally, inferred associations from the subgraphs will be explored by constructing MEDLINE queries (see Chapter 5), and searching and sifting through the search results. Enriching the current framework by providing functionality for summarization, is a natural future enhancement. However, the key enhancement will be enabling users to activate new searches in Obvio, which generate new subgraphs, to refine the user interest and guide the knowledge exploration process.

6.5 Summary

This Chapter presented various paradigms for biomedical text exploration. These include methods for: 1) search-and-sift, 2) ontology-based text exploration, 3) semantic predications-based text exploration, and 4) discovery browsing. The search-and-sift approach requires considerable manual effort and lacks the specificity often sought by information seekers. Ontology-based text exploration approaches typically leverage relationships between entities through implicit or explicit labeled edges. However, the context of association derived from the ontology is often not present in the text. Semantic predications-based exploration includes the context of assertions in the literature, but lack the flexibility and utilities for discovery browsing.
Future Directions

This dissertation implemented and evaluated a context-driven, automatic subgraph creation method to facilitate LBD. The approach outlines definitions for context and shared context, which compensate for deficiencies in distributional and graph-theoretic metrics. The approach was used to facilitate the rediscovery of 8 out of 9 existing scientific discoveries. In addition, it also provides support for understanding the meaning of complex associations. To achieve this, the approach first provides labeled directed edges between nodes in the generated subgraphs. Such labeled edges are complemented by the provenance of the semantic predications in MEDLINE. Further, the approach captures complex associations between concepts \((A, C)\), along multiple thematic dimensions (e.g., Cellular Function, Pharmaceutical Treatment, and Neurological Activity). Capturing the diversity in complex associations enables a broader understanding of the nature of associations among concepts in a domain. A beta-version of the Obvio framework has been deployed (http://knoesis-hpco.cs.wright.edu/obvio/), which showcases the rediscoveries and the applicability of the dissertation to LBD, in general.

While the rediscoveries support the validity of this approach, there are several limitations – some inherent in the approach, and some systemic. The first inherent limitation is the impracticability of the manually specified threshold for semantic similarity of MeSH descriptors. While similarity is a very subjective measure, a manually determined threshold is only acceptable to a certain degree. To determine a threshold for semantic similarity of MeSH descriptors, we believe that insights gained from the distribution of path relat-
edness scores (see Section 4.4.1) can be applied to the MeSH descriptors. In particular, a mixture of Gaussian models can be generated for the distribution of each of the normalized similarity scores, which may enable automatic selection of a threshold.

The second limitation surrounds the inexact Gaussian distribution of path relatedness scores discussed in Section 4.4.1. In exploratory experiments, the \( p \)-values for each observed distribution exceeded the limit \( p < 0.05 \), normally regarded as indicative of the null hypothesis. Typically, when this value holds, it suggests that there is no difference between the expected values and the observed values. Alternative techniques such as the central limit theorem may be useful for transforming the observed distribution into a more consistent distribution. Another consideration is that semantic relatedness could be defined in terms of both relatedness and unrelatedness, where unrelatedness captures the extent to which two paths differ.

The third limitation is the assumption that the context of a semantic predication, expressed in terms of the MeSH descriptors, is reliable for generating meaningful subgraphs. Not all MeSH descriptors assigned to an article are relevant to all its semantic predications, and hence the predication context vectors could be very noisy. Ideally, direct mappings between semantic predications and MeSH descriptors would address this discrepancy. However, such mappings are unavailable and our specification of context is subject to limitations of distributional semantics. Obviously, other metrics such as frequency of (co)occurrence, predicate relatedness, degree centrality, and temporal features (such as first known occurrence) could also be incorporated into an ensemble approach for defining context, similar to Ramakrishnan in [70].

Another limitation is the manual specification of predicate filters, which were used to eliminate non-informative predications. Semantic predications that contained a select set of relationships (i.e., ASSOCIATED_WITH, INTERACTS_WITH, AFFECTS, etc.) were excluded from the clustering process. This is a compromise to achieve scalability. Ideally, the system should not require any predicate filters. In fact, the omission of some predicates
may be responsible for low recall in some of our experiments. Still, given that most experiments terminated in less than one hour, higher recall may not be too costly in terms of loss in performance.

Finally, the choice and implementation environment for the HAC algorithm is another serious limitation. HAC was selected because it is an unsupervised, deterministic clustering algorithm, for which the number of clusters does not have to be specified \textit{a priori}. This comes at a cost in performance. The time complexity of HAC is $O(N^2 \log N)$. While other approaches, such as those by Ramakrishnan et al. [70] and Anyanwu et al. [3], may achieve better performance, it is unclear how they might be adapted to generate complex subgraphs along multiple thematic dimensions. Similarly, it is unclear how to create multiple subgraphs using the approach by van der Eijk et al. [98], which uses Hebbian learning. With the emergence of: 1) big data infrastructure (such as Apache Spark) and 2) algorithms for distributed clustering (such as Locality Sensitive Hashing, etc.), the performance limitations of the clustering task may be easily resolved.

Conversely, the first systemic limitation is the extent to which domain expertise is still required for LBD, in general. Although impractical to eliminate humans (i.e., domain experts) from the LBD process, one improvement could be providing additional background knowledge to supplement the subgraphs where appropriate. In this way, assertional knowledge from the literature would have been complemented with definitional knowledge from structured knowledge sources. This is currently done only to a limited degree in the deployed Obvio web application (see Appendix C). Metrics for determining interesting neighboring concepts in background knowledge, will need to be developed to overcome this limitation.

The second systemic limitation is the inability to detect (and perhaps resolve) contradicting semantic predications. While the provenance of predications in MEDLINE allows domain experts to adjudicate, a method for predicting conflicting predications could be beneficial. We believe that temporal analysis of semantic predications could enable con-
Conflict resolution. However, since many unresolved paradoxes, which are reported in the literature, are inherent to science itself, it is unclear whether one might ever be able to reliably detect and resolve such contradictions automatically.

The third systemic limitation surrounds the reliability of the statistical evaluation. Rare associations are generally interesting, but not always. While alternative methods for conducting statistical evaluations for LBD have been discussed [108], it is cumbersome to coordinate cut-off dates for each predication across the rediscoveries. The suggested techniques are therefore impractical to implement. We use the measure of association rarity as an indication of interestingness, similar to existing research in the literature [85].

Finally, the absence of subgraph labels impedes the identification of subgraph context. Subgraph labeling is a crucial task, since our approach is predicated on the idea that each subgraph captures a different thematic dimension of association between two concepts. These and other limitations prescribe the next steps for this line of LBD research.
Discussion

Research on discovering knowledge is complex but potentially very rewarding. Hidden information is often difficult to capture. This dissertation proposes a context-driven approach for LBD that has shown some initial successes. However, many factors (aside from those identified) can be responsible for the obscurity among important connections in scientific literature. Defining a generic approach to LBD is unlikely to succeed in all (perhaps even most) cases. A key issue is the ability to define, capture, and exploit context among related associations.

This dissertation first discussed context in terms of semantic predication context. Semantic predications are said to inherit the MeSH descriptors of articles in which they are present, based on the interchangeability assumption (see Section 4.2). The context of a semantic predication is therefore initially defined as a vector of MeSH descriptors – subsequently downgraded to a set. This core idea was then extended to define path context. Since a path consists of a sequence of semantic predications, the context of a path is expressed as the aggregate of the component semantic predication contexts. The idea of path context is then used to extrapolate shared context. This is achieved by using structured background knowledge (specifically the MeSH hierarchy) to determine inexact semantic similarity between MeSH descriptors. Path context is then used to infer the context of entire subgraphs. The hierarchical agglomerative clustering algorithm is used to group paths, which share some underlying context.

A more intangible notion, which was not addressed in this dissertation, is the idea of
dimension context. Each subgraph represents a context, topic, or theme, which is shared among paths within the subgraph. These individual topics or contexts are not necessarily disjoint, within or across domains. For example, had Walter Sutton (see Section 1.1.2) been unfamiliar with Mendel’s work (however he stumbled upon it), science would have experienced a delay in linking genetics and cytology (now cytogenetics). The impact of this important integration would invariably, have also been delayed. Yet, the two scenarios were inherently linked based on context in different dimensions.

The extent to which information processing systems can be developed to reason across such dimensions, simulating perception and making analogies in vastly distinct situations, is both perplexing and intriguing. This dissertation explores the capabilities of an information processing system and human ingenuity in designing algorithms to harness the multiplicity of resources available for creating dimension context. However, it provides no discussion on reasoning across dimensions.

It is my view that the future of information processing systems for LBD, should focus on the methodologies that can reason, not just on facts (or semantic predications), paths, and complex subgraphs, but rather on dimensions. There is a higher awareness awaiting the human mind, which can liberate human understanding, if we succeed in harnessing such broader knowledge. This prospect could not be more articulate than the quote by Howard Phillips Lovecraft (1890–1937), “…some day the piecing together of dissociated knowledge will open up such terrifying vistas of reality, …that we shall either go mad from the revelation or flee from the deadly light into the peace and safety of a new dark age” – from The Call of Cthulhu, The Horror in Clay [55]. It is my genuine hope that this dissertation proves to be important towards this end.
Bibliography


[18] Michael J. Cairelli, Christopher M. Miller, Marcelo Fiszman, T. Elizabeth Workman, and Thomas C. Rindflesch. Semantic medline for discovery browsing: Us-
ing semantic predications and the literature-based discovery paradigm to elucidate a mechanism for the obesity paradox. In *AMIA*, 2013.


144


[102] Ryen W. White, Steven M. Drucker, Gary Marchionini, Marti Hearst, and m. c. schraefel. Exploratory search and hci: Designing and evaluating interfaces to sup-


Appendix A

This Appendix briefly describes results from *Experiment II* (see Section 3.5.4) on the application of the context-driven subgraph model in [19], to recover and decompose Swanson’s *Raynaud Syndrome–Dietary Fish Oils* discovery from [89]. *Experiment II* involves use of titles and abstracts only, from the 65 articles noted as relevant to the discovery.

### A.1 Rediscovery and Decomposition – Experiment II

In this experiment, we aimed to determine the extent to which titles and abstracts only, may be adequate for supporting LBD – if indeed adequate at all. The predications graph for the experiment contained 388 semantic predications and 192 concepts, in baseline (*B2*) dataset. After applying the DFS algorithm to generate the candidate graph from the predications graph, there were 22 associations for which *Fish Oil – dietary (FOD)* was the source, but no vertices that terminated with *Raynaud Syndrome* (*i.e.*, *RD = 0, RP = 0*). Additionally, there were 67 associations for which *Fish Oils (FO)* was the source, among which two terminated with *Raynaud Syndrome* (*i.e.*, *RD = 0, RP = 2*). And finally, there were 224 associations, for which *Eicosapentaenoic Acid (EPA)* was the source of the candidate graph, among which 34 terminated with *Raynaud Syndrome* (*i.e.*, *RD = 8, RP = 26*). There was therefore a total of 36 associations for which *Dietary Fish Oils* was the source and *Raynaud Syndrome* was the target. From these paths, subgraphs were once again manually constructed and then compared with Swansons original findings.
A.1.1 Platelet Aggregation

For Swanson’s first primary association, which states that Dietary Fish Oils disrupt Platelet Aggregation, which causes Raynaud (see Table 3.1, ID: 1), four out of the 36 associations were deemed relevant to this hypothesis. From them, we constructed the subgraph shown in Figure A.1. The subgraph shows no direct link between Dietary Fish Oils and Platelet Aggregation. The absence of such a direct link to Platelet Aggregation therefore made it difficult to infer the role of Dietary Fish Oils and Platelet Aggregation in treating Raynaud Syndrome. The B2 dataset therefore did not contain sufficient information to support recovery of this specific intermediate.

![Figure A.1: Primary Association subgraph: Platelet Aggregation (Exp2)](image)

For the first supplementary association (Table 3.2 ID: 1a) involving Platelet Aggregation, again no direct association existed between Dietary Fish Oils and Platelet Aggregation were found after constructing the subgraph shown in Figure A.2. While we can conjecture that [Dietary Fish Oils TREATS Raynaud Syndrome] through Prostaglandins from the two associations; [Dietary Fish Oils PRODUCES Prostaglandin PGI3] and [Epoprostenol ISA Prostaglandin], the role of Platelet Aggregation is still not apparent.
Similarly, for the secondary associations (Table 3.3, ID: 1.1, 1.2, 1.3), although there is sufficient information to surmise that \textit{Dietary Fish Oils TREATS Raynaud Syndrome}, the role of Platelet Aggregation is again obscured. No associations involving Platelet Aggregation were recovered using titles and abstracts only.

### A.1.2 Blood Viscosity

For Swanson’s second primary association, which states that \textit{Dietary Fish Oils disrupt Blood Viscosity}, which in turn causes Raynaud (Table 3.1, ID: 2), virtually the same four associations from \textit{Experiment 1} were present in the candidate graph (shown in Figure A.3).
The main difference was that for two of the four associations, the weaker relation, which states that \([\text{Ketanserin AFFECTS Blood Viscosity}]\) was observed, instead of more precise association that \([\text{Ketanserin DISRUPTS Blood Viscosity}]\). We can conclude that \([\text{Dietary Fish Oils TREATS Raynaud Syndrome}]\) by lowering \textit{Blood Viscosity}, on the assumption that \([\text{Ketanserin AFFECTS Blood Viscosity}]\) is synonymous with \([\text{Ketanserin DISRUPTS Blood Viscosity}]\) and by applying Hristovski’s discovery pattern to infer that \([\text{Epoprostenol DISRUPTS Blood Viscosity}]\).

By contrast, for Swanson’s first supplementary association (Table 3.1, ID: 2a), we recovered four associations out of six associations from \textit{Experiment I}. The critical association between \textit{Dietary Fish Oils} and \textit{Fatty Acids}, which states that \([\text{Eicosapentaenoic Acid ISA Fatty Acids}]\) and \([\text{Epoprostenol STIMULATES Fatty Acids}]\) were notably absent. It was difficult to make the association that \([\text{Dietary Fish Oils TREATS Raynaud Syndrome}]\) by inhibiting \textit{Blood Viscosity}.

### A.1.3 Vascular Reactivity

There were two distinct mentions of \textit{Vascular Reactivity} in the \textit{baseline (B2)} dataset. Among these one predication was manually identified. This predication states that \([\text{Vasoconstriction ASSOCIATED WITH Eicosapentaenoic Acid}]\). Unfortunately, this was insufficient for linking \textit{Dietary Fish Oils} and \textit{Raynaud Syndrome} through \textit{Vascular Reactivity}. Hence, no associations, which involved \textit{Vasoconstriction} or \textit{Vasodilation} were recovered from the \textit{B2} dataset, consisting of titles and abstracts only. These results collectively suggest that titles and abstracts alone, might \textit{NOT} be sufficient for LBD. They were certainly not sufficient to rediscover Swanson’s \textit{Raynaud Syndrome–Dietary Fish Oils} hypothesis.
Appendix B

This Appendix reviews several tools developed for LBD, which help put the innovations of Obvio in perspective. Each tool implements techniques derived from the set of approaches discussed in Section 2.3, which include keyword-based, concept-based, relations-based, graph-based and hybrid approaches for LBD. Table B.1 provides online resources for a variety of these tools, including those that are now defunct (depicted by an asterisk). While not an exhaustive review, this Appendix provides considerable coverage across LBD systems, together with the previous reviews by Hristovski et. al. [43] and Ganiz [31].

Table B.1: Literature-Based Discovery Systems and Tools

<table>
<thead>
<tr>
<th>System</th>
<th>Web Application</th>
<th>Video Demo</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARROWSMITH (1997)</td>
<td><a href="http://arrowsmith.psych.uic.edu/arrowsmith_uic">http://arrowsmith.psych.uic.edu/arrowsmith_uic</a></td>
<td></td>
</tr>
</tbody>
</table>

1. **ARROWSMITH**: In 1997, Don R. Swanson and Neil R. Smalheiser launched ARROWSMITH [95, 82, 77, 83, 96, 84], a tool, which is generally regarded as the first tool developed for LBD. The initial version [95, 82] relied on an interactive (and iterative) approach for exploring biomedical literature. It consisted of three components: 1) the bibliographic literature, 2) computer software and 3) the human operator. To generate hy-
hypotheses, the system ingested the titles ($T$) of articles from the $A$- and $C$- literature. These titles were obtained by searching MEDLINE using the PubMed search interface. The system supported both open discovery (Procedure I) and closed discovery mode (Procedure II). Open discovery mode required only a title term ($c$) from the research question in the $C$ literature as input, while closed discovery mode required: 1) two title terms ($a, c$ – one from each literature) and 2) the titles of articles from the $A$- and $C$- literatures as input.

**Procedure I:** In open discovery mode hypothesis generation was performed in six steps. **Step 1 (Generic Word Filtering):** In the first step, stop words and non-novel words (such as ‘clinical,’ ‘drugs,’ ‘treatment,’ etc) were removed from the titles in the $A$ and $C$ literatures.\(^1\) **Step 2 (Corpus-based Intermediate Filtering):** The remaining terms (considered B-term candidates) were then filtered by retaining only those that returned search results in MEDLINE, above some relative frequency w.r.t. the titles in the $C$ literature. Note that the search sequence applied for open discovery in ARROWSMITH follows a $C \rightarrow B \rightarrow A$ pattern, instead of the well-known $A \rightarrow B \rightarrow C$ pattern. **Step 3 (Manual Intermediate Filtering):** The human operator then removes candidate B-terms considered irrelevant based on human judgements. Additionally, the titles of the intermediate $B$-terms were filtered based on category (topic) restrictions configured into the system 	extit{a priori} [93]. It is unclear whether these categories are MeSH descriptors. The content of the corresponding articles for each title, was then processed for additional B-term candidates. The final list of $B$-terms was constructed from the union of relevant terms from the titles and abstracts. **Step 4 (BA Pair Selection):** This union of $B$-terms was then used to execute a new MEDLINE search for titles. After filtering, new terms, which were disjoint from the $C$-terms and the union of $B$-terms, were considered $A$-term candidates. Such terms were used to form $B \rightarrow A$ pairs. **Step 5 (Target Selection):** Given the initial $A$-term candidates, a filtered list was then obtained by executing a new MEDLINE search and restricting to $A$-terms which belong to categories from the restricted set, ranked by their $B \rightarrow A$ co-occurrence frequencies.\(^1\)

\(^1\)The manually stoplist consisted of around 5000 words. A version of ARROWSMITH is available online: [http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html](http://arrowsmith.psych.uic.edu/arrowsmith_uic/index.html)
Step 6 (Target Filtering): The ranked A-terms list was then presented to human experts for analysis.

Procedure II: In closed discovery mode, hypothesis generation was performed in three steps. Category restrictions and the probabilistic measures for filtering B and A terms were removed. The updated system supported processing n-grams up to 6 tokens in length, instead of only unigrams. Additionally, MEDLINE searches were extended to titles and abstracts. Step 1 (Intermediate Candidate Generation): In the first step, candidate B-terms from titles and abstracts of articles from the C- and A- literature, were generated and then merged into a “B-LIST.” Step 2 (Manual Intermediate Filtering): The B-LIST was then manually edited. Step 3 (Intermediate Inspection): The list of titles from A that contain B-terms, which are also in the title of C-terms was then produced for manual inspection.

The more recent version of ARROWSMITH [77, 83, 96, 84] (developed in 2005), is limited to closed discovery (or two-node search). Smalheiser noted that ARROWSMITH now focused on delivering tools for: 1) “identifying likely mechanisms to explain [scientific] findings,” which can 2) “assess whether the existing literature lends indirect support to a hypothesis.” This version utilizes machine learning and labeled training data from a manually created gold standard dataset to learn features and their weights, for predicting B-terms. Step 1 (Query Formulation): In the first step, the research question is formulated as two separate queries, which may consist of multiple phrases from the A- and C-liternatures. Step 2 (Intermediate Candidate Generation): These phrases are then used as search terms in MEDLINE searches. The computer software then extracts unigram and bigram B-terms from the resulting document set. Each n-gram is ranked based on the probability of its importance in the corpus. Torvik et. al. [96] computes this probability by defining 7 features to estimate relevance. These features include: 1) number of common MeSH descriptors in A and C; 2) frequency of occurrence in the A and C literatures; 3) is UMLS concept – concepthood; 4) cohesion; 5) frequency in MEDLINE; 6) year of first
occurrence in MEDLINE and 7) frequency in A, C relative to frequency in MEDLINE. The weight of each parameter in the feature set is determined using logistic regression. **Step 3 (Intermediate Candidate Ranking):** The ranked list of B-terms is presented in the user interface, which links provides links to the titles of AB and BC MEDLINE articles.

Using *ARROWSMITH*, Swanson and Smalheiser reported several hypotheses [92, 79, 80, 81]. This tool is considered one of the first systems to illustrate the role of distributional statistics (or term co-occurrence) and the *ABC model* to facilitate LBD. It also demonstrated the considerable degree of manual input used in developing early LBD systems. Numerous approaches have since adopted the design principles of *ARROWSMITH* for hypothesis generation.

2. **DAD:** In 2000, Weeber et. al. developed a concept-based NLP-based framework to detect Disease-Adverse Drug Reaction-Drug associations, and vice versa, called *DAD* [99]. The system used a “two-step model of discovery in which new scientific hypotheses can be generated and subsequently tested.” It was used to reproduce both the Raynaud Syndrome-Dietary Fish Oils and the Magnesium-Migraine discoveries [100] and also to suggest new treatments for Acute Pancreatitis, Chronic Hepatitis C, Gastritis, and Myasthenia Gravis with Thalidomide [101]. *DAD* supported both open and closed discovery. Its main innovation was the use of MetaMap to maps keywords in biomedical texts to standard concepts from the UMLS.

To generate hypotheses, *DAD* applies the following four steps. **Step 1 (Query Expansion):** In the first step, *DAD* requires a C-term for which a hypothesis is to be generated (or confirmed). The C-term is then expanded under manual supervision, with closely related UMLS concepts and their synonyms, as deemed appropriate. **Step 2 (Intermediate Candidate Selection):** Using the query, a relevant collection of documents is obtained by querying MEDLINE. Intermediate candidates are then selected from each document, by using MetaMap again, to obtain UMLS concepts from the titles and abstracts. Concepts
that belong to a list of UMLS Semantic Types, specified *a priori*, are then retrieved as B-terms. **Step 3 (Target Selection):** The selected B-terms are then used to query MEDLINE for additional documents, which contain targets or C-terms. The specific set of C-terms that are disjoint from the A and B lists are considered most interesting. **Step 4 (Hypothesis Verification):** In this final step, terms common to both the A and C literature of given semantic types are presented to the user for inspection.

3. **BITOLA:** In 2001, Hristovski launched a concept-based discovery support system called **BITOLA** [44]. The tool supported both open and closed discovery, using MetaMap to detect MeSH descriptors in biomedical texts. These descriptors were also linked to the UMLS and therefore enabled filtering of intermediates based on semantic types. **BITOLA** was used initially to reproduce the *Raynaud Syndrome – Dietary Fish Oils* discovery and subsequently to explore novel *Disease–Gene* associations in [41, 42].

**BITOLA** supports LBD using the following three steps. **Step 1 (Document Representation):** MeSH descriptors assigned to MEDLINE articles are used as an abstraction of the content of the article. The tool then identifies gene mentions in the titles of these MEDLINE articles and augments this representation with gene symbols from several gene databases. The representation is further augmented with chromosomal locations derived based on disease mentions in each article. **Step 2 (Intermediate Candidate Generation):** **BITOLA** then uses association rules together with the co-occurrence based measures of confidence and support to rank AB and BC concept pairs. Starting with all MEDLINE articles and a Disease X, **BITOLA** ranks all XY pairs in which the Y intermediates are filtered by semantic type. **Step 3 (Target Generation):** Upon selecting an appropriate Y-term, a MEDLINE query is executed to retrieve all articles that contain Y. Target (Z) concepts of semantic type Gene, for example, above the threshold of support and confidence are then filtered out for inspection.
4. **Telemakus**: In 2004, Fuller et. al. introduced *Telemakus* [29] for both open and closed discovery. *Telemakus* relies on the idea of a structured research report schema, to first outline relevant aspects of scientific reports that are important for LBD. It then creates a graph of concepts, based on a thesaurus of curated information, extracted from the literature, and from MEDLINE.

*Telemakus* provides LBD through the following steps. **Step 2 (Schema Population):** It uses concepts in the biomedical literature to populate a Telemakus thesaurus. The knowledge is represented in a relational database and indexed according to data types. **Step 2 (Search and Indexing):** The result of a search is a list of documents, each with facets from other data types. Additionally, articles can also be browsed using the metadata provided by MEDLINE, such as author, year and journal. **Step 3 (Visualization):** The results set can then be visualized as connections between concepts from the facets. Discoveries may arise through an iterative process of searching and exploration of the concept space.

5. **Litlinker**: In 2003, Pratt and Yetisgen-Yildiz implemented a concept-based LBD system called *LitLinker* [66, 108]. The first version of LitLinker [66] was used to reproduce the Magnesium – Migraine discovery through open discovery in the following four steps. **Step 1 (Corpus Selection):** The user specifies the start term \( A \), which is then used to retrieve titles and abstracts from MEDLINE. MetaMap is then used to extract UMLS concepts from these titles, to serve as \( B \)-term candidates or (i.e., linking concepts). **Step 2 (Concept Filtering):** LitLinker then uses a frequency-imposed threshold to eliminate common terms in MEDLINE, from the \( B \)-term candidate set. UMLS semantic types are also used as filters. An additional method used to filter intermediates eliminates closely related \( AB \) term pairs. To achieve this, a syntactic pattern-based technique is used to cluster concepts. **Step 3 (Intermediate Ranking):** Intermediates are then ranked using the frequency-based metric of *level of support*. **Step 4 (Target Concept Ranking):** Target \( C \) terms are aggregated from all \( B \)-terms, but eliminated if also an \( A \)-term. The same method for ranking linking
concepts was applied to rank targets.

In an enhanced version of LitLinker, Yetisgen-Yildiz et. al. [108] used MeSH descriptors as intermediates instead of UMLS concepts, together with the probabilistic \( z\)-score metric, to rank intermediates.

6. **Manjal**: In 2004, Srinivasan [85, 86] developed *Manjal*, a concept-based tool that supports both open and closed discovery. This approach leveraged co-occurrence of relationships between UMLS Semantic Types and MeSH descriptors. Instead of \( z\)-score, the approach used term frequency-inverse document frequency (TF-IDF) to rank MeSH descriptors based on their co-occurrence frequency with their respective UMLS semantic types in MEDLINE. Srinivasan’s framework enabled the construction of weighted topic profiles, as vectors of scored MeSH descriptor–UMLS semantic type pairs. Relevant intermediates were first retrieved by restricting to specific UMLS semantic types, then subject to the TF-IDF ranking. The approach was used in [85] to recover the intermediates from 5 out of the 6 discoveries made by Swanson and to provide new insights into the association between *Tumeric/Curcumin, Crohn’s Disease*, retinal diseases and disorders related to the spinal cord in [86].

7. **IRIDESCENT**: In 2002, Wren et. al. [106, 107] developed *IRIDESCENT* to support LBD. *IRIDESCENT* was used to discover new knowledge on the possible therapeutic effects of *Chlorpromazine* in *Cardiac Hypertrophy*. To rank intermediates, the probability (or veracity) of co-occurring intermediates (as n-grams) is computed using maximum likelihood estimates (MLE), derived from sentences and abstracts. The strength of co-occurrence between \( AB \) and \( BC \) pairs is then computed and normalized, based on the expected connectedness (or degree centrality) between terms, given the MLE values. In this way, noninformative terms that were frequently co-occurring and highly connected in the corpus, could be eliminated. In the final step, intermediates were ranked by the ratio of
observed/expected connectedness between co-occurrence of $AB$ and $BC$ pairs. In a refinement of this approach [107], mutual information was used to estimate the strength of co-occurrence.

8. **RaJoLink**: In 2009, Petrič et. al. [65] developed RaJoLink to support open discovery for LBD. RaJoLink was used to provide new insights into the association between *Calcinerin* and *Autism*. RaJoLink connects rare, joint and linked terms in MEDLINE. **Step 1 (Ra)**: Rare term candidates $C$, are first identified in the literature, then subject to verification by the domain expert. This includes term filtering using MeSH. **Step 2 (Jo)**: The rare terms $C$, from the $Ra$ step are used to select additional documents from the literature, for ($A$). Terms from the intersecting documents within the new set are identified and subject to human verification. **Step 3 (Link)**: In the final step, the TF-IDF metric is used to rank documents, which are similar to those from the combined $Ra$ and $Jo$ steps, for $A$ and $C$ terms. Documents that are similar to those containing rare $A$− and $C$− terms are used to find linking terms.

9. **SemBT**: In [38] Hristovski et. al. introduced *SemBT* for closed discovery. The application leveraged relations extracted using SemRep [71]) and BioMedLEE [56]. The main innovation was the use *discovery patterns*, specified *a priori*, to find potentially interesting associations. This approach was used to rediscover the *Raynaud Syndrome–Dietary Fish Oils* discovery. It was also used to suggest new insights into the association among *Insulin*, *Diabetes Mellitus* and *Huntington Disease*. Furthermore, in [40] discovery patterns were used together with DNA Microarray Data to generate novel hypotheses on *Parkinson’s Disease*. To support LBD, *SemBT* implements the following steps: **Step 1 (Document Representation)**: Semantic predications extracted from MEDLINE articles, and relations extracted using BioMedLee were used as atomic elements in the corpus. **Step 2 (Intermediate Candidate Generation)**: Given an $A$-term, semantic predications were retrieved
using BITOLA, based on metrics such as support and confidence. Intermediates were then filtered using specific UMLS semantic types. The matching documents containing interesting AB concept pairs, were then submitted to BioMedLee to extract additional relations.

**Step 3 (Target Generation):** The combined set of relations was then subject to discovery patterns, specified *a priori*, to obtain interesting ABC associations.

10. **Semantic MEDLINE:** In 2008, Kilicoglu et. al. launched Semantic MEDLINE, an interactive graph-based tool for LBD [50, 72]. Semantic MEDLINE utilizes the discovery browsing paradigm, in which common LBD tasks are performed iteratively and complemented by document summarization. In [105], Wilkowski used Semantic MEDLINE to elucidate the association among Norepinephrine, Depression and Sleep. It was also used to provide new knowledge on Testosterone–Sleep [61] and DEHP–Sepsis [18]. To facilitate LBD, Semantic MEDLINE performs the following steps: **Step 1 (Search):** The user executes a document search using an appropriate set of keywords. The system queries MEDLINE for an appropriate result set. **Step 2 (Summarization):** The set of semantic predications in the result set is fed into an abstraction summarizer [27]. The summarizer uses various metrics, including relevance, connectivity, novelty and saliency to filter predications, according to high-level summarization perspectives. These perspectives, which are specified *a priori*, include Treatment of Disease, Substance Interactions, Diagnosis, and Pharmacogenomics. **Step 3 (Visualization):** The reduced set of semantic predications from this ‘semantic condensate’ step, form a predications graph, which can then be explored using UMLS predicates, semantic types and groups as additional filters. The provenance of each predication in MEDLINE is also provided.

11. **CoPub Discovery:** In 2010, Frijters et. al. [28] developed CoPub Discovery to support LBD. CoPub uses a keyword-based term co-occurrence method to support open discovery. This approach was used to: 1) elucidate genetic aspects of Grave’s Disease and
Programmed Cell Death 1 (PDCD1), 2) provide news insights into the therapeutic implications of Milnacipran on Obsessive Compulsive Disorder (OCD), 3) hypothesis generation between Pitavastatin and Monocyte Activity and to 4) discover new knowledge relating Dephostatin and Cell Proliferation.

To support LBD, CoPub applies the following steps. **Step 1 (Corpus Selection):** CoPub combines several thesauri, related to human genes, diseases, drugs, etc, thereby using a keyword-driven search framework. It then performs a regular expression search on MEDLINE for articles that contain each search term. **Step 2 (Ranking):** To rank both B-term candidates, and C-terms, CoPub uses an adaptation of the mutual information measure defined by Wren et. al. [106]. Similar to the approach by Yetisgen-Yildiz [108], cut-off dates were imposed on the literature to create a training set of observed associations, and a test set. The ranking mechanism was then used to confirm the presence of predicated associations in the post cut-off literature.

**12. Obvio:** In this dissertation (2014) I introduce Obvio, a graph-based framework for exploring biomedical literature, which supports LBD [21]. Obvio was used to re-discover and elucidate numerous associations from 8 out of 9 existing discoveries (see Chapter 5). Obvio uses implicit and explicit context to generate complex associations among biomedical concepts based on their shared context. The hierarchical agglomerative clustering is applied to generate subgraphs on multiple thematic dimensions. Additionally, it exploits the provenance of semantic predications in MEDLINE to provide insights into the meaning of complex associations. A video demo of Obvio is available online: [http://bit.ly/obviodemo](http://bit.ly/obviodemo), along with the live version of the web application: [http://knoesis-hpco.cs.wright.edu/obvio](http://knoesis-hpco.cs.wright.edu/obvio).
Appendix C

Obvio Web Application

This Appendix describes the Obvio web application (shown in Figure C.1) developed to showcase the rediscoveries. The system consists of 11 components, which can be used to explore subgraphs generated for closed discovery scenarios, using the following steps.

1. The user must first select a start term (A) using component 1. For example, the concept Chlorpromazine can be selected as an A-term.

Figure C.1: Screenshot of the Obvio Web Application
Step 2: The user must then select the target term ($C$) using component 2. For example, the concept *Cardiac Hypertrophy* has been selected as the target term, for the given source.

Step 3: The user must then select the ‘Search’ button to submit the search request. Obvio retrieves the metadata for the search terms, which are then displayed in the ‘metadata panel’ immediately below search terms (component 3).

Step 4: The identifiers of the preprocessed subgraphs are shown in the ‘subgraph panel’ in component 4.

Step 5: The user must then select the identifier of a subgraph from the subgraph panel. The corresponding subgraph will be displayed in the ‘viewer’ (component 5).

Step 6: Interesting semantic predications may then be explored by clicking on the edge between concepts of interest in the viewer.

Step 7: The number of MEDLINE articles that contain the visualized semantic predications is shown the ‘Result Metadata Panel’ (component 6). The identifier for the MEDLINE article is also shown (currently shown, 2000 Feb 18). The title of the article is shown in component 7, while the date of publication is shown in component 8. The selected semantic predication is shown in component 9 (currently shown, *Calcineurin-CAUSES-Cardiac Hypertrophy*). The set of MEDLINE articles that contain the predication are also available for inspection in component 10. More importantly, the sentence from which the semantic predication was extracted will be highlighted.
Step 8: The user may also utilize the functionality from the ‘Filtering panel’ in component 11, to view different perspectives in the subgraphs based on semantic types and groups. Note that the original subgraph can be restored by clicking an arbitrary point in the viewer. Also, when any node in the subgraph has been selected, only the inlinks and outlinks connected to the selected node are displayed.

Technologies: The Obvio web application was developed using the following technologies: Apache Tomcat 7, Apache Lucene, Apache Hadoop, Openlink Virtuoso 6.1 RDF Database, MySQL 6.5, Eclipse Helios, J2EE, JavaServlets, Cytospace.js and jQuery.