School of Computing Science

Systems Approaches to Drug Repositioning

Joseph Mullen

Submitted for the degree of Doctor of Philosophy in the School of Computing Science, Newcastle University

June 2017

Abstract

Drug discovery has overall become less fruitful and more costly, despite vastly increased biomedical knowledge and evolving approaches to Research and Development (R&D). One complementary approach to drug discovery is that of drug repositioning which focusses on identifying novel uses for existing drugs. By focussing on existing drugs that have already reached the market, drug repositioning has the potential to both reduce the timeframe and cost of getting a disease treatment to those that need it. Many marketed examples of repositioned drugs have been found via serendipitous or rational observations, highlighting the need for more systematic methodologies.

Systems approaches have the potential to enable the development of novel methods to understand the action of therapeutic compounds, but require an integrative approach to biological data. Integrated networks can facilitate systems-level analyses by combining multiple sources of evidence to provide a rich description of drugs, their targets and their interactions. Classically, such networks can be mined manually where a skilled person can identify portions of the graph that are indicative of relationships between drugs and highlight possible repositioning opportunities. However, this approach is not scalable. Automated procedures are required to mine integrated networks systematically for these subgraphs and bring them to the attention of the user. The aim of this project was the development of novel computational methods to identify new therapeutic uses for existing drugs (with particular focus on active small molecules) using data integration.

A framework for integrating disparate data relevant to drug repositioning, Drug Repositioning Network Integration Framework (DReNInF) was developed as part of this work. This framework includes a high-level ontology, Drug Repositioning Network Integration Ontology (DReNInO), to aid integration and subsequent mining; a suite of parsers; and a generic semantic graph integration platform. This framework enables the production of integrated networks maintaining strict semantics that are important in, but not exclusive to, drug repositioning. The DReNInF is then used to create Drug

Repositioning Network Integration (DReNIn), a semantically-rich Resource Description Framework (RDF) dataset. A Web-based front end was developed, which includes a SPARQL Protocol and RDF Query Language (SPARQL) endpoint for querying this dataset.

To automate the mining of drug repositioning datasets, a formal framework for the definition of semantic subgraphs was established and a method for Drug Repositioning Semantic Mining (DReSMin) was developed. DReSMin is an algorithm for mining semantically-rich networks for occurrences of a given semantic subgraph. This algorithm allows instances of complex semantic subgraphs that contain data about putative drug repositioning opportunities to be identified in a computationally tractable fashion, scaling close to linearly with network data.

The ability of DReSMin to identify novel Drug-Target (D-T) associations was investigated. 9,643,061 putative D-T interactions were identified and ranked, with a strong correlation between highly scored associations and those supported by literature observed. The 20 top ranked associations were analysed in more detail with 14 found to be novel and six found to be supported by the literature. It was also shown that this approach better prioritises known D-T interactions, than other state-of-the-art methodologies.

The ability of DReSMin to identify novel Drug-Disease (Dr-D) indications was also investigated. As target-based approaches are utilised heavily in the field of drug discovery, it is necessary to have a systematic method to rank Gene-Disease (G-D) associations. Although methods already exist to collect, integrate and score these associations, these scores are often not a reliable reflection of expert knowledge. Therefore, an integrated data-driven approach to drug repositioning was developed using a Bayesian statistics approach and applied to rank 309,885 G-D associations using existing knowledge. Ranked associations were then integrated with other biological data to produce a semantically-rich drug discovery network. Using this network it was shown that diseases of the central nervous system (CNS) provide an area of interest. The network was then systematically mined for semantic subgraphs that capture novel Dr-D relations. 275,934 Dr-D associations were identified and ranked, with those more likely to be side-effects filtered.

Work presented here includes novel tools and algorithms to enable research within the field of drug repositioning. DReNIn, for example, includes data that previous comparable datasets relevant to drug repositioning have neglected, such as clinical trial data and drug indications. Furthermore, the dataset may be easily extended using DReNInF to include future data as and when it becomes available, such as G-D association directionality (i.e. is the mutation a loss-of-function or gain-of-function). Unlike other algorithms and approaches developed for drug repositioning, DReSMin can be used to infer any types of associations captured in the target semantic network. Moreover, the approaches presented here should be more generically applicable to other fields that require algorithms for the integration and mining of semantically rich networks.

Declaration

I declare that this thesis is my own work unless otherwise stated. No part of this thesis has previously been submitted for a degree or any other qualification at Newcastle University or any other institution.

Signed ....

Date ....

Publications

Portions of the work within this thesis have been documented in the following publications:

Mullen J, Cockell SJ, Tipney H, Woollard PM, Wipat A. (2016) Mining integrated semantic networks for drug repositioning opportunities. *PeerJ* 4:e1558
doi:10.7717/peerj.1558

Mullen J, Cockell SJ, Woollard P, Wipat A (2016) An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations. *PLoS ONE* 11(5): e0155811. doi:10.1371/journal.pone.0155811

Acknowledgements

First and foremost I would like to thank my academic supervisors Prof. Anil Wipat and Dr Simon Cockell for guiding me through my research with their invaluable advice and support. I would also like to thank my industrial supervisors from GlaxoSmithKline (GSK), Dr Hannah Tipney and Peter Woollard for their valued input. I am privileged to have had all four as my supervisors. I gratefully acknowledge the funding from the European and Physical Sciences Research Council (EPSRC) and GSK that made my research possible.

During my research, I was lucky enough to complete a 3-month industrial placement with GSK in their Computational Biology department, Stevenage. I am thankful to all there who made me feel welcome and who took the time to partake in insightful discussions regarding my research. I particularly extend my thanks to my industrial supervisor Peter Woollard as well as Prof. Phillipe Sanseau and Dr Pankaj Agarwal, the latter of whom provided invaluable advice.

Thanks to all members of the writing group for providing detailed feedback on my thesis and manuscript preparation. I was fortunate to complete my PhD within a department consisting of many amazing people, both past and present. A sincere thank you is extended to all of these for providing a friendly and enjoyable place of work.

A special thank you is also extended to my amazing family, including parents and siblings, whom I love dearly. My family have been my motivation throughout life and studies, and have been there for support and advice whenever I have stumbled.

Glossary

Active Pharmaceutical Ingredient (API) the ingredient in a drug that is biologically active.

Adverse Drug Reaction (ADR) an appreciably harmful or unpleasant reaction, resulting from an intervention related to the use of a medicinal product [1].

Attrition the process of reducing something's strength or effectiveness through sustained attack or pressure.

Antagonist a substance which interferes with or inhibits the physiological action of another.

Allosteric allosteric regulation (or allosteric control) is the regulation of a protein by binding an effector molecule at a site other than the enzyme's active site.

Affinity how tightly a substance binds to the receptor.

Agonist a substance which initiates a physiological response when combined with a receptor.

Benefit-risk ratio an evaluation of two dimensions. Benefits are measured in terms of therapeutic efficacy, quality of life, cost or pharmacoeconomic aspects. Risks include the safety profile (sum of all ADRs) and potential risk of unobserved ADRs.

Blockbuster drug an extremely popular drug that generates annual sales of at least US$1 billion for the company that creates it.

Disease a condition of the body, or of some part or organ of the body, in which its functions are disturbed or deranged [2].

Disease (Rare) *US*- any disease that affects fewer than 200,000 people in the United States, in *Europe*- a disease is considered to be rare when it affects 1 person per 2000.

Disease (Common) any disease that is not classed as a rare disease.

Drug Activity (IC_{50}) is a measure of how effective a drug is. It indicates how much of a particular drug or other substance is needed to inhibit a given biological process.

Drug Activity (K_d), or dissociation constant, is a specific type of equilibrium constant that measures the propensity of a larger object to separate (dissociate) reversibly into smaller components, as when a complex falls apart into its component molecules, or when a salt splits up into its component ions.

Drug Activity (K_i), or inhibitory constant, is an indication of how potent an inhibitor is; it is the concentration required to produce half maximum inhibition.

Drug Activity (Potency) amount of drug needed to achieve a specific biological response.

Druggability the likelihood of being able to modulate a target with a drug [3].

Efficacy measurement of the maximum biological response of a compound.

Genome the complete set of genes or genetic material present in a cell or organism.

Indication of a drug refers to the use of that drug for treating a particular disease.

International Nonproprietary Name (INN) or the generic drug name, such as sildenafil, as opposed to brand names used for marketing, such as Viagra.

Monogenic designation describes an inherited characteristic that is specified by a single gene.

Multigenic designation describes an inherited characteristic that is specified by a combination of multiple genes.

New Molecular Entity (NME) designation describes a new chemical drug [4].

New Biological Entity (NBE) designation describes a new biological drug [4].

New Therapeutic Entity (NTE) the combined designation refers to either an NME or an NBE [4].

Orphan Drug a pharmaceutical agent that has been developed specifically to treat a rare disease.

Pharmacogenomics is the study of how genes affect a person's response to drugs.

Pharmacokinetics is the study of drug ADME (absorption, distribution, metabolism, and excretion) properties.

Phenome the set of all phenotypes expressed by a cell, tissue, organ, organism, or species.

Phenotype an organism's observable characteristics [5].

Polypharmacology can mean one drug, multiple targets, or several drugs hitting different targets in one or more pathways.

Semantic is the study of meaning.

Side-effect an effect, whether therapeutic or adverse, that is secondary to the one intended.

Side-effect (Off-target) side effects caused by a drug hitting 'other' targets than those it was designed to hit.

Side-effect (ADME) side-effects resulting from the drugs process through the ADME systems.

Symptom is a departure from normal function or feeling which is noticed by a patient, reflecting the presence of an unusual state, or of a disease.

Topology (Network) a schematic description of the arrangement of a network, or graph, including its nodes and connecting edges.

Contents

List of Figures

List of Tables

Acronyms

ADR Adverse Drug Reaction

API Active Pharmaceutical Ingredient

ADMET Absorption, Distribution, Metabolism, Excretion and Toxicity

BBB Blood-Brain Barrier

BioSSIP Bioinformatics Semantic Integration Platform

CTD Comparative Toxicogenomics Database

DBMS Database Management System

DReNIn Drug Repositioning Network Integration

DReNInO Drug Repositioning Network Integration Ontology

DReNInF Drug Repositioning Network Integration Framework

DReSMin Drug Repositioning Semantic Mining

Dr-D Drug-Disease

D-T Drug-Target

EBI European Bioinformatics Institute

EMA European Medicines Agency

FDA Food and Drug Administration

GO Gene Ontology

GOA Gene Ontology Annotation

GoF Gain of Function

GWAS Genome-Wide Association Studies

G-D Gene-Disease

GPCR G protein-coupled receptors

JVM Java Virtual Machine

KEGG Kyoto Encyclopedia of Genes and Genomes

LoF Loss of Function

MeSH Medical Subject Headings

MGD Mouse Genome Database

NDF-RT National Drug File - Reference Terminology

NBE New Biological Entity

NLM U.S. National Library of Medicine

NME New Molecular Entity

NMs Network Motifs

NP Nondeterministic Polynomial Time

NUI Alphanumeric Unique Identifier

NHGRI National Human Genome Research Institute

OMIM Online Mendelian Inheritance in Man

ORDO Orphanet Rare Disease Ontology

OWL Web Ontology Language

PDB Protein Data Bank

R&D Research and Development

RDF Resource Description Framework

RDFS RDF Schema

RGD Rat Genome Database

RTA Rich Therapeutic Area

SDC Semantic Distance Calculator

SE Semantic Simplicity

SS Semantic Score

SIDER Side-Effect Resource

SCR Supplementary Concept Records

SNP Single-Nucleotide Polymorphism

SPARQL SPARQL Protocol and RDF Query Language

ST Semantic Threshold

TAU Therapeutic Area Unmet Score

UI Unique Identifier

UMLS Unified Medical Language System

VHA U.S. Department of Veterans Affairs, Veterans Health Administration

1

Introduction

1.1 Motivation

The *Oxford English Dictionary* defines a disease as:

> “a condition of the body, or of some part or organ of the body, in which its functions are disturbed or deranged [2]”

Diseases can result in a reduced quality of life, increased morbidity, and mortality. Through the practice of medicine, doctors focus on preventing, diagnosing and treating diseases. Recent advances in the development of genome editing technologies have improved the ability to make precise changes in the genomes of eukaryotic cells [6]. It is hoped that in the future, through the application of programmable nucleases, genome editing will enable the correction of genetic mutations associated with current life-limiting diseases, such as Cystic fibrosis [6]. Gene editing approaches are still in their infancy and as such, drugs are the most common form of disease intervention. Many drugs can treat a plethora of disorders in a relatively safe fashion. Unlike biologics (which are made up of sugars, proteins, nucleic acids, or a combination of these) active small molecules are composed of organic and inorganic compounds, and it is these active small molecules that form the basis of the majority of drugs on the market. As long as there are diseases or clinical conditions without suitable medical products available for their treatment, or ideally cure, the motivation for drug discovery projects will remain [7].

Drug discovery is the process through which potential new medicines and drug targets are identified. Strategies for drug discovery rely heavily on Research and Development (R&D) and may either be target-based (or molecular approaches) or rely on phenotypic measures of response (also known as empirical approaches) [5]. Since the pursuit of the human genome project in the 1990s, drug discovery has focussed on target-based approaches. Emphasis on target-driven approaches has come with reasonable successes, including, for example, the tyrosine kinase inhibitor for the treatment of chronic myelogenous leukaemia, imatinib (Glivec, Novartis) [8]. In target-based drug discovery, a research-driven hypothesis states that altering the functionality of a protein or pathway will result in a therapeutic effect on a disease state. After a relevant target has been identified, a candidate drug is proposed during lead discovery. Before the advent of target-based approaches, candidate drugs were determined by evaluating different chemicals against phenotypes. These phenotypic assays identified candidate drugs and were conducted in biological systems, such as animals or cells, often with no *a priori* knowledge of the mechanism of action [5]. After its identification, using either a target-based or phenotypic R&D approach, a candidate drug then progresses into preclinical, and if successful, clinical development (phases I-IV) before becoming a marketed therapeutic [7]. The whole development process can take 12-15 years with average costs, when including the price of failure and a 'time cost', of US$2.6 billion [7, 9].

Despite a resurgence in productivity in the last two years [10], in 2009, only 26 New Molecular Entities (NMEs) were approved by the Food and Drug Administration (FDA), despite R&D costs reaching US$168 billion. The increased expenditure and decreased productivity is not helped by the fact that a large number of candidate drugs fail in, or before, the clinic [11]. The situation has left many questioning the current state of R&D [12, 13]. It is, therefore, crucial to advance strategies that reduce this time frame while also decreasing associated costs without compromising safety. One complementary approach to traditional R&D drug discovery is that of drug repositioning.

Drug repositioning is concerned with the application of existing drugs to a novel therapeutic area and is also known as redirecting, repurposing and reprofiling [14]. Although drug repositioning tends to focus on drugs that have already reached the market, drugs that have failed to do so for reasons other than safety concerns may also be considered for potential repositioning opportunities.

Drugs already approved for the treatment of a particular disease will have already been tested in humans. As a result, detailed information is available on pharmacology, formulation and toxicity [14]. This prior knowledge greatly reduces the time taken for an already approved drug to be reviewed by the FDA and subsequently integrated into health care. Although examples of successfully repositioned drugs exist, these drugs have often been identified through rational observations [15] or serendipitous findings [16]. The search space for potential repositioning opportunities is vast and so relying on chance or human intervention to identify these opportunities is not efficient. Systematic approaches can enable the efficient and exhaustive exploration of this search space. Although numerous systematic methods to aid in drug repositioning have been proposed, many focus on a single entity type (such as targets) or the inference of a single interaction type (such as Drug-Target (D-T) interactions).

There is a plethora of biological and pharmacological data available describing multiple biological entity types and interactions between them. As data availability increases, especially data from proteomics, transcriptomics and pathway inference, the building blocks of our knowledge are enhanced. A holistic view of a drug and its associated entities can be achieved through the integration of these data, potentially providing a comprehensive understanding of a drug's actions. Using these rich data sets, inferences can be made with a higher confidence than approaches utilising a single data type.

A systems approach to the task of drug repositioning is described in this thesis. In this approach, data integration platforms relevant to drug repositioning were developed to allow the implementation of appropriate strategies for semantic data integration using networks. Topological and semantic network structures, termed semantic subgraphs, were defined. The mappings of these structures were shown to be indicative of repositioning opportunities. Finally, algorithms and automated approaches were developed to allow searching for these connected sub-components.

1.2 Contribution of the work presented

There are three key results of this work: a Drug Repositioning Network Integration Framework (DReNInF); an integrated data set, Drug Repositioning Network Integration (DReNIn); and an algorithm for mining integrated data sets, Drug Repositioning Semantic Mining (DReSMin).

The first of these outcomes, DReNInF, provides a framework for integrating disparate data relevant to drug repositioning. DReNInF includes a high-level ontology, Drug Repositioning Network Integration Ontology (DReNInO), to aid integration and subsequent mining; a suite of parsers; and a generic semantic graph integration platform. This framework allows a user to produce integrated networks maintaining strict semantics that are important in, but not exclusive to, drug repositioning.

The DReNInF is then used to create DReNIn, a semantically-rich Resource Description Framework (RDF) data set. A web-based front end is provided, which includes a SPARQL Protocol and RDF Query Language (SPARQL) endpoint for querying this data set. The user is also provided with the ability to download query results.

DReSMin was developed to systematically mine these semantically-rich integrated drug repositioning networks. DReSMin allows for the exact, exhaustive searching of semantic subgraphs; connected sub-components of a target network that enable the inference of edges or associations that are not captured in the network. This algorithm makes use of the VF2 [17] subgraph isomorphism algorithm, which is optimised and extended to consider and score the semantics, as well as topology of any potential mappings.

The application of DReSMin to the inference of novel D-T associations using historical data was also explored. The outcome of this analysis is a set of scored and ranked D-T associations and of a set of specialised semantic subgraphs. An extended version of DReSMin, which utilises values attached to attributes, was also implemented. Associations between Gene-Disease (G-D) were scored using a Bayesian approach and along with the updated version of DReSMin used to identify novel Drug-Disease (Dr-D) associations.

1.3 Aims and objectives

The aim of this project was the:

> “Development of Novel Computational Methods to Identify New Therapeutic Targets for Existing Drugs Based on Data Integration”

The following objectives were defined to help achieve the project aim:

1. To extend existing data integration platforms relevant to drug repositioning.
2. To research and implement appropriate strategies for semantic data integration of network construction including Ondex, RDF and others.
3. To develop algorithms to search for topological and semantic network structures indicative of repositioning opportunities.

1.4 Thesis structure

This thesis is divided into the following chapters:

- Chapter 2 provides background information and a literature review of current computational approaches to drug repositioning.

- Chapter 3 describes research in the area of data integration with a particular focus on its application to drug repositioning. Data integration methods and platforms are introduced and compared. This Chapter also presents data sources and data types relevant to drug repositioning. Finally, DReNIn, an RDF data set with a Web-based SPARQL endpoint, to aid in the discovery of drug repositioning opportunities is described.

- Chapter 4 introduces a framework for mining integrated networks for the inference of novel associations. The Chapter introduces and defines semantic subgraphs, which are connected sub-components of a target network, whose mappings allow us to infer novel associations from a target network. An exact, exhaustive subgraph matching algorithm, DReSMin, is also described. DReSMin identifies all mappings of a predefined semantic subgraph that are captured in a target graph. Mappings maintain exact topological properties of the semantic subgraph, while semantics are matched to a user defined threshold.

- Chapter 5 describes the application of DReSMin to the identification of novel D-T associations. An automated framework is described which makes use of historical data for the development of a set of relevant semantic subgraphs. Furthermore, a method for the scoring and ranking of inferred D-T associations is described.

- Chapter 6 discusses the application of DReSMin to the identification of novel Dr-D indications. First, a Bayesian approach for scoring and ranking G-D associations from multiple data sources is introduced. Scored G-D associations are then integrated into the network before a data-driven approach to drug repositioning that makes use of these G-D associations is taken.

- Chapter 7 reviews the outcomes of this work in the broader context of drug repositioning and describes possible future work.

2

Background: computational approaches to drug repositioning

2.1 Introduction

Over the past decade, Research and Development (R&D) approaches to drug discovery have become less productive while costs continue to rise, in spite of the rapid advances in genomics, life sciences and technology over the same period [14, 18, 19]. Drug discovery failures are spread across the drug development pipeline [20]; in 2004 it was calculated that only 11% of drugs investigated in clinical trials reach the market [21]. Despite some signs of a very recent resurgence over the last two years [10], the drug development process for successfully marketed drugs can still take between 10-17 years [22]. Disregarding the extensive time scale, the cost of getting a drug to market is also extreme. When including the price of failure and opportunity cost, it has been shown that the cost of drug discovery has more than doubled in the past decade [9]. In 2003, it was estimated that the cost of a new drug, including the price of failures and a 'time cost', averaged US$800 million [23]. The cost of a new drug in 2013, when considering only inflation of the 2003 cost, was projected to be just over US$1 billion. However, a similar analysis to that completed in 2003, put the average cost to US$2.6 billion per marketed drug in 2013, a rise of 145% [9]. It was also shown, in 2011, that the minimum cost of developing a drug not containing a previously approved Active Pharmaceutical Ingredient (API), was US$204 million [24]. Problems resulting in a reduction in R&D productivity are complex and include escalating clinical trial costs and an increased aversion to risk [25]. These complex challenges have resulted in only 20 to 30 New Molecular Entities (NMEs) being approved per year by the Food and Drug Administration (FDA) between 2005 and 2012 [10].

For these reasons, improving R&D productivity remains the most important priority within the pharmaceutical industry [13]. Recently, frameworks such as the 'five R' framework' from AstraZeneca [25], have been developed to try and increase the lagging productivity in R&D. These frameworks have contributed to the recent resurgence in R&D successes of late. 41 drugs (30 NMEs and 11 New Biological Entities (NBEs)) were approved by the FDA in 2014 and a 19 year high of 45 in 2015 (33 NMEs and 12 NBEs) [10]. 2015 represents an approval rate of double the average during 2005-2009 when approvals were at their lowest with an average of 22 drugs per year [10]. Of the drugs approved in 2015, 16 are expected to achieve ‘blockbuster’ status including the Cystic Fibrosis treatment ivacaftor plus lumacaftor (Orkambi; Vertex) [10]. Oncology drugs, developed for the prevention, diagnosis and treatment of cancer and associated neoplasms, have accounted for 30% of approved drugs in three of the past four years [10]. Of the 45 drugs approved in 2015, 14 are oncology drugs such as palbociclib (Ibrance; Pfizer), approved for the treatment of breast cancer. 21 of the approved drugs in 2015 are for orphan, or rare, diseases (half of these are for rare cancers) [10]. An increase in treatments for rare diseases has resulted in a slump of around 35% in average forecast sales between 2014 and 2015 (US$1.4 billion per 2014 drug and US$900 million per 2015 drug) [10], with rare diseases less rewarding, financially. Due to economic incentives, such as the FDA Orphan Drug Act (ODA), it is not coincidental that 2015 represents the highest number of drugs approved for rare diseases to date.

A disease is considered to be rare in Europe if it has a prevalence of fewer than 1 in 2,000 people. In the US, a disease that affects fewer than 200,000 individuals is considered rare. Collectively, rare diseases affect 6-7% of the developed world, representing a small market opportunity for the pharmaceutical industry [26]. Incentives, such as the FDA Orphan Drug Act (ODA) of 1983, were introduced to encourage pharmaceutical companies to invest in R&D for the development of rare disease therapies [26]. The Act grants special status, referred to as orphan designation (or sometimes “orphan status”), to a drug or biological product to treat a rare disease upon request of a sponsor. Orphan designation qualifies the sponsor of the drug for various development incentives of the ODA, including tax credits for qualified clinical testing, more rapid pipeline progression and market exclusivity for seven years[1]. Legislation to promote the development of orphan drugs has also been implemented in other countries, such as Japan in 1993, Australia in 1998, and Europe in 2000 [27]. As a result of these incentives many pharmaceutical companies and specialist biotechnology companies now have groups focussing on rare diseases, also incentivised by the fact that rare diseases have highly active focus groups that mean sometimes clinical trials can be easier. More recent initiatives, such as the 100,000 Genomes Project from Genomics England, focus

[1] www.fda.gov Accessed:01-03-2016

on patients with rare diseases as well as their families, and patients with common cancers[2]. Data from the 100,000 Genomes Project will provide a gene mechanistic-based understanding of multiple rare diseases, a potentially massive opportunity for drug development for rare diseases.

Despite the recent peak in productivity, R&D approaches to drug discovery remain incredibly costly. Complementary approaches to the typical R&D paradigm have come to the fore of late, with examples including personalised medicine to find tailored therapies for individual patients [28]. Personalised medicine takes into account the fact that 30% of drugs investigated in clinical trials fail because of a lack of efficacy (the two foremost reasons for clinical drug attrition are inefficacy and toxicity) [21, 28]. Personalised medicine is particularly interesting when considering the fact that even the top 10 selling approved drugs in the US help as little as 1 in 25 of those that take them [29]. Factors such as the bias towards white Western participants in clinical trials, results in some drugs even being harmful to certain ethnic groups [29]. The idea behind precision medicine is that stratifying patients and diseases into molecular subtypes and treating with subtype-specific drugs will improve specificity and efficacy [28]. The interest in personalised medicine has come at a time when efforts such as the UK 100,000 Genomes Project and the US Precision Medicine Initiative seek to scale up population-based genome sequencing and integrate it with clinical data [30]. Although an attractive alternative to R&D, precision medicine is still a young and growing field. Due to its immaturity, precision medicine faces many immediate problems, not just the fact that many of the required technologies are still in their infancy. Another complementary approach to R&D drug discovery is drug repositioning.

2.2 Drug repositioning

A recent analysis of 217 relevant scientific articles, found no standard definition for drug repositioning [31]. For the purpose of this project the original definition, as provided by Ashburn and Thor in their landmark paper, will be used to define drug repositioning as:

[2] www.genomicsengland.co.uk Accessed:01-03-2016

> “the process of finding new uses outside the scope of the original medical indication for existing drugs [14]”

The word ‘repositioning’ may also be used interchangeably with ‘redirecting’, ‘repurposing’, and ‘reprofiling’ [14]. In this setting, a ‘drug’ is defined as a well-understood chemical or biological ingredient that regulatory agencies, such as the US FDA or the European Medicines Agency (EMA), have approved for use in the context of disease. Furthermore, a drug is a particular combination of API, formulated under defined and consistent conditions for human use [32].

Drug repositioning has traditionally been part of the drug development process. During drug development, drug repositioning is used as a strategy to preserve and extend the value of patents through reformulation strategies [33]. However, as R&D approaches become less efficient and more costly, drug repositioning is becoming ever more important, providing an effective alternative to traditional R&D approaches in drug discovery. The savings arise from the fact the drugs have already reached the market. Marketed drugs will already have established formulations and manufacturing methods and will have been tested in humans [14, 28]. As a result, detailed information will be available describing Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties [14, 28] as well Phase IV (post-marketing) safety data, which are expensive and time-consuming to obtain [34]. Although ‘safety’ to treat one disease, does not necessarily mean ‘safety’ to treat another, since a significant proportion of drugs are deemed toxic. In fact, around 30-40% of FDA-approved drugs display such severe side-effects that they are ‘black boxed’ [32] or receive the EMA’s equivalent (a black triangle along with the statement ‘This medicinal product is subject to additional monitoring’). ‘Boxing’ is used by the FDA to highlight drugs deemed tolerable for the situation. Tolerability means the drug is on the favourable side of risk-benefit analysis, while also carrying a significant risk of serious side-effects [32]. Drug repositioning offers many benefits to traditional R&D approaches to drug development, some of which are described below.

As mentioned, approved drugs will have already been tested for safety in humans. The time taken for an already approved drug to be reviewed by the FDA and their subse-

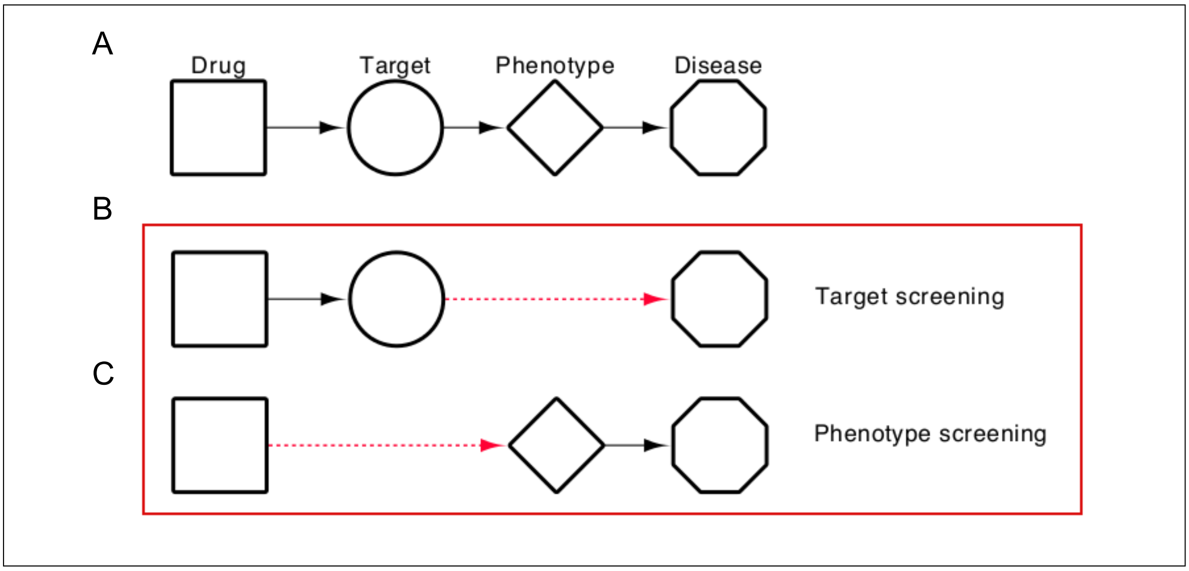


Figure 2.1: **Typical R&D approaches to drug discovery.** (A) Abstract drug-disease connection, known by some as the central dogma of drug action (drug-target-disease) [36]. In the red box different approaches, target screening (B) and phenotype screening (C), to drug discovery are shown in relation to the drug-disease connection. Figure adapted from [22].

quent integration into health care can be significantly reduced. Potentially, repositioning can reduce the traditional timeline of 10-17 to 3-12 years, making drugs available to patients much more quickly than following a traditional R&D route [14, 22]. The financial savings of drug repositioning are also appealing to pharmaceutical companies. Bringing a repositioned drug to market costs approximately 160 million times less than typical R&D approaches to developing an NME [35], with a low barrier to entry [32]. In fact, it takes an average of US$8.4 million dollars to get a repositioned drug to market [35]. In 2003 sildenafil (Viagra; Pfizer) had a global revenue stream of US$1.88 billion [14]. Although this is a 'best case' situation, it does highlight the potential disparity in upfront investment and financial gains of a repositioned drug, supporting the belief that a repositioned drug represents a better return on investment than an NME [35].

Typically, a drug discovery programme initiates as there is a disease, or clinical condition, missing from the portfolio of a pharmaceutical company, or without suitable medical products available [7]. An extensive collection of chemicals is then experimentally tested to see if any can produce a relevant effect, in a process known as *screening*. Screening may be either target-based (see Fig. 2.1B) or phenotypic (see

Fig. 2.1C). Target-based screening has been popular since the advent of the genomic era, and selects chemicals based on their ability to bind to a biological target relevant to the therapeutic area of interest [22]. These targets are, more often than not, single proteins [37]. The more specifically a chemical interacts with the active site of a target, the 'cleaner' the action of the chemical will be. Historically, however, the best drugs are often promiscuous (e.g. some of the anti-kinases). On the other hand, a phenotypic screen does not make any assumptions about the underlying pathological mechanism and proteins involved. Instead, the target is 'skipped' and a cell line or model organism representative of the therapeutic area of interest is directly screened in a direct attempt to find a chemical that exhibits the desired phenotypic behaviour [22]. Both methods help to efficiently discover active chemicals, or *lead drugs*. It takes between 2.5 to 4 years to complete the process of lead discovery; 2 to 3 years for target discovery and validation and 0.5 to 1 year to screen or design chemicals with biological activity [22]. Lead drugs are further optimised for efficacy using medicinal chemistry during *lead optimisation*, a process which takes 1 to 3 years. Lead optimisation aims to avoid side-effects due to either off-target binding or unintended physiologic roles of the target [22]. It then takes 1 to 2 years to ascertain ADMET properties of drugs using animal models, 5 to 6 years to assess the safety and efficacy of drugs in clinical trials, and 1 to 2 years to obtain approval [14]. Drug repositioning opportunities can be derived from three key observations of the traditional workflow as described below.

Clinical trials are expensive and only a small set of diseases are analysed. As a result of inexhaustive clinical trials, once a drug has reached the market, potential indications that were not investigated in the clinic can be identified as a repositioning opportunity. Secondly, the pleiotropic effects of genes and other drug targets mean they may be involved in diseases other than those that the drug was developed to treat (see Fig. 2.2B). Understanding the involvement of a target in the biological system can help to identify new roles for a drug. Thirdly, due to drug promiscuity, a drug may also bind to multiple targets for which it was not primarily designed to affect. Some state that drug promiscuity is a pre-requisite for drug repositioning [38]. Known as off-target molecules, a known drug binding to these secondary targets may provide repositioning opportunities while also highlighting potential side-effects [39] (see Fig. 2.2C).

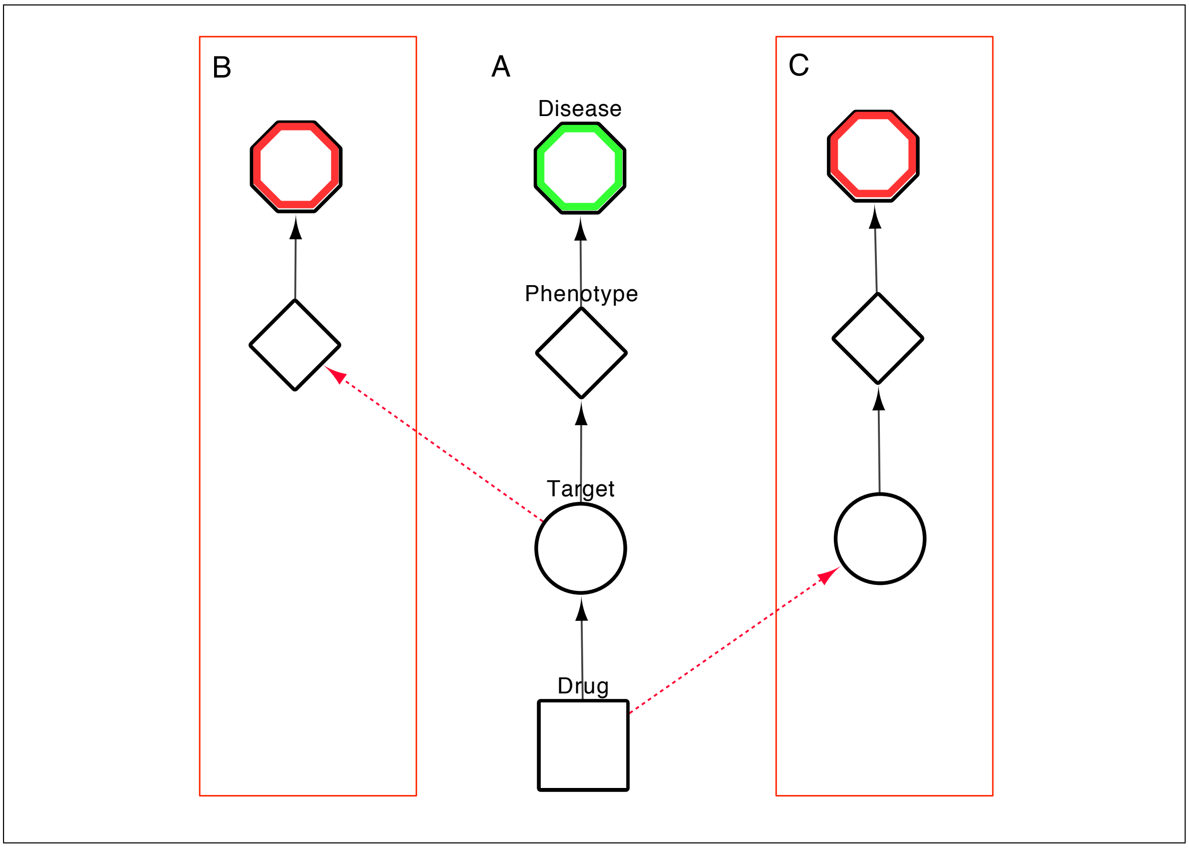


Figure 2.2: **Drug repositioning opportunities arising during typical drug discovery process.** (A) The abstract drug-disease connection, also known as the central dogma of drug action. (B) A drug target may be involved in multiple pathways or diseases and so a secondary indication may be identified. (C) Often small molecules bind to other targets (off-targets), allowing for potential new indications, as well as side-effects, to be highlighted.

In recent years, drug repositioning has become actively supported by governments, non-trading organisations and academic institutions alike [18, 40]. For example, both the UK (Medical Research Council) and the US (National Centre for Advancing Translational Sciences) have launched large-scale funding programs in this area [18]. The goal of the funding is to extend the indication set for molecules that have already undergone significant R&D in the pharmaceutical industry [18]. Many pharmaceutical companies, including GSK and Pfizer, have also implemented intense systematic repositioning efforts. In academia and industry alike there are also substantial economic incentives to reposition marketed drugs for the treatment of rare disorders [41].

Although examples of successfully repositioned drugs exist, they have often been identified through rational observations [15] or serendipitous findings [16], such as the examples provided next.

2.2.1 Marketed examples

Here, examples of successful repositioning involving approved drugs are described. Although some investigational drugs have been successfully repositioned, including thalidomide (Thalidomide Celgene; Celgene) for the treatment of leprosy and multiple myeloma, and zidovudine (Retrovir; ViiV Healthcare) for the treatment of AIDS, only examples including approved drugs are discussed here. Focus is placed on the identification of the repositioning opportunity to highlight the serendipity involved in these cases.

2.2.1.1 Sildenafil: unexpected side-effects

Pfizer was seeking a drug for the treatment of angina when sildenafil (Viagra; Pfizer) was developed in the 1980s [14]. Inhibiting phosphdiesterase-5, sildenafil was to relax coronary arteries, allowing greater coronary blood flow and thus reducing patient symptoms. The desired cardiovascular effects were not observed in the healthy volunteers, with the drug lacking efficacy for angina. Development of sildenafil was thus discontinued, until, that is, many of the patients involved in the clinical trials reported that they had unusually strong and persistent erections [14].

This lead to Pfizer setting up a trial on 3,700 impotent men with a total of 1,631 years of exposure worldwide [42]. In 1998 sildenafil received clearance for the treatment of erectile dysfunction and the drug was repositioned as a first-in-class accordingly. By 2003, sildenafil had cemented its position as a blockbuster drug with annual sales of US$1.88 billion with nearly 8 million men taking sildenafil in the United States alone [14]. In 2005, after successful clinical trials, sildenafil was also approved for the treatment of pulmonary arterial hypertension (Revatio; Pfizer) [43].

The repositioning of sildenafil arose from the secondary functions of its targeted enzyme, PDE5. The pleiotropy of PDE5 and the mechanism which allowed for repositioning is represented schematically by Fig. 2.2B.

2.2.1.2 Duloxetine: pathway involved in multiple diseases

Duloxetine (Cymbalta; Eli Lilly) is a balanced dual serotonin (5-HT)-norepinephrine (NE) reuptake inhibitor [44]. Discovered in the late 1980's, duloxetine was the result of efforts to identify an improved version of the anti-depressant fluoxetine (Prozac; Eli Lilly) [14]. The action of duloxetine is associated with the re-uptake inhibition of 5-HT and NE at the presynaptic neurone in Onuf's nucleus (the pudendal nerve motor nucleus) of the sacral spinal cord [45]. Duloxetine has also been developed for the treatment of pain caused by diabetic neuropathy [44].

Duloxetine was also shown to be beneficial in the treatment of stress urinary incontinence (SUI), the most common form of urinary incontinence. With a prevalence rate of around 50%, SUI is more frequently found in women than in men [44]. The effectiveness of duloxetine in the treatment of SUI was studied in a cat model of acetic acid induced bladder irritation [44]. Here in the UK, duloxetine (Yentreve; Eli Lilly), is approved by the EMA as a first in class add-on medication instead of surgery for those suffering from SUI [44].

The repositioning of duloxetine resulted from a shared mechanism of action between the two diseases, depression and SUI [14]. Like sildenafil, the repositioning of duloxetine can be represented schematically by Fig. 2.2B.

2.2.1.3 Imatinib: drug potently inhibits target in another disease

Chronic myelogenous leukaemia (CML) is characterised by the reciprocal translocation between chromosomes 9 and 22, known as the Philadelphia translocation [8]. A consequence of this inter-chromosomal exchange is the expression of the *BCR-ABL* oncogene [8], a mutation present in 95% of CML patients [46]. Using target-based approaches, specifically to identify a tyrosine kinase inhibitor for BCR-ABL, imatinib (Glivec; Novartis) was identified as a potent and selective inhibitor of BCR-ABL. Imatinib was also found to be an inhibitor of the platelet-derived growth factor (PDGFR) [47] and KIT tyrosine kinases (both the wild and mutant type) [46, 48]. The drug was subsequently approved in 2001 for the treatment of CML by the FDA, after only a two and a half month review time.

Accounting for less than 1% of primary gastrointestinal neoplasms, Gastrointestinal Stromal Tumor (GIST), is a rare, and, therefore, somewhat neglected, disease [49]. Most commonly it originates in the stomach (60%), followed by the small intestine (30%), the colon and rectum (5%), and the oesophagus (5%) [50]. GIST is associated with a Gain of Function (GoF) point mutation in the tyrosine kinase proto-oncogene, *KIT*, responsible for around 85% of GISTs [49]. Platelet-derived growth factor receptor alpha $PDGFR\alpha$ point mutations are also found in 5-10% of GISTs [50]. After the early success of imatinib in chronic phase CML, investigators decided to test its effects on c-KIT receptor tyrosine kinase activity [48]. Testing was based on rational observations regarding the sensitivity of KIT to imatinib [46]. The FDA subsequently approved imatinib for GIST patients in 2002.

The repositioning story of imatinib may be represented schematically, as shown in Fig. 2.2C. In this case, the drug binds to another target involved in another disease.

2.2.2 Computational approaches

Human intervention and luck have been central to the repositioning success stories introduced. The search space for potential repositioning opportunities is enormous; relying on serendipity or rational observations to identify these is, therefore, not efficient. Systematic approaches can enable the exhaustive exploration of this space. Although it is assumed that experimental bioassays are more reliable and predictive than computational assays [32], testing all drugs against all targets experimentally is extremely costly and, at present, technically infeasible [51]. The lower cost and barrier to entry, in comparison to experimental methods, have made computational approaches to drug repositioning of high interest and effort in the research community [32].

Computational repositioning is the process of designing and validating automated workflows that can generate hypotheses for new indications [22]; providing promising and efficient tools for discovering new uses for existing drugs [18]. As such a wide range of computational approaches and strategies have been proposed [32]. These can be divided, roughly, into five categories: drug-based approaches, protein similarity-based approaches, genome-based approaches, phenome-based approaches, and computational strategies, as shown in Fig.2.3.

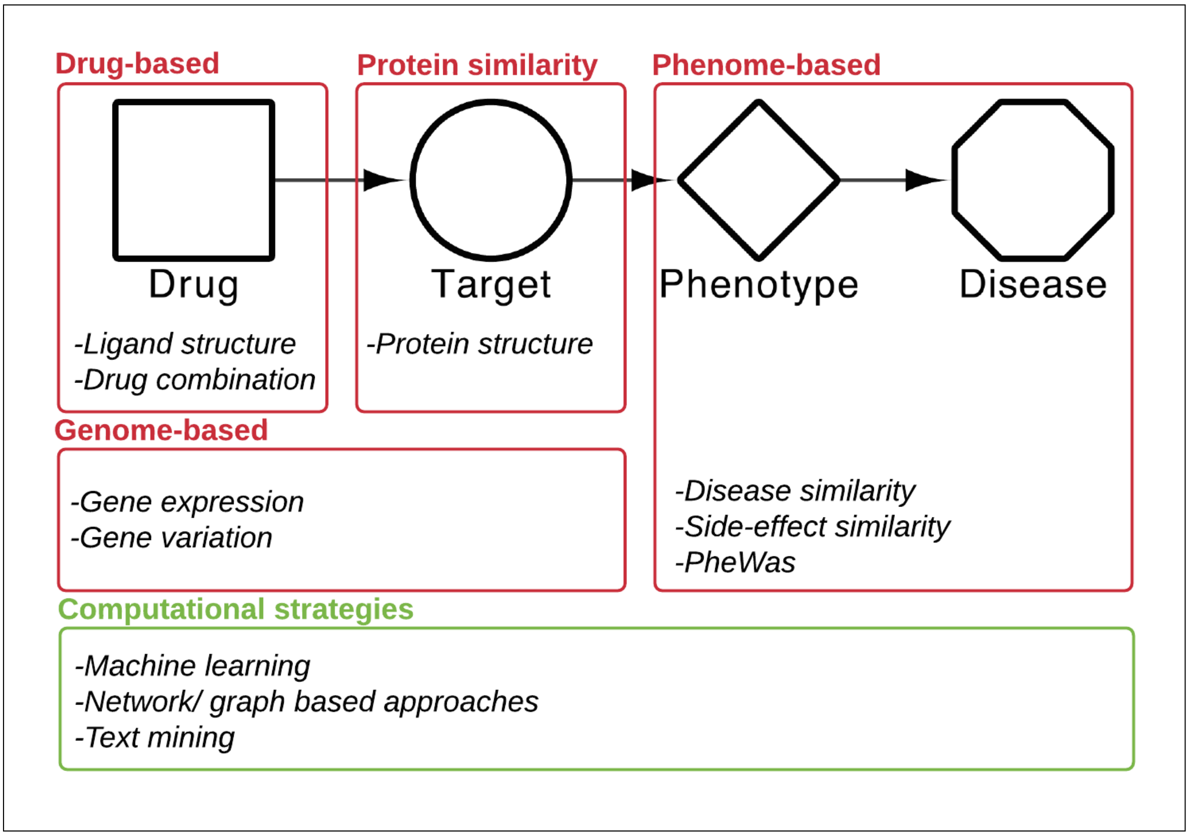


Figure 2.3: **Computational approaches to drug repositioning.** Computational approaches introduced in this Chapter can roughly be divided into five categories: drug-based approaches (Drug-based), protein similarity-based approaches (Protein similarity), genome-based approaches (Genome-based), phenome-based approaches (Phenome-based), and computational strategies. Above, each methodology is categorised and shown in context with the elements involved in the drug-disease connection.

2.2.2.1 Ligand structure-based approaches

Ligand similarity analysis is based on the accepted 'similar property principle' (SPP) [52]. The SPP states that similar molecules should have similar biological activities, affecting proteins and biological systems in similar ways [52]. As such the use of chemical similarities is a common target-based approach for drug repositioning [53]. There are many methods for measuring the potential similarity between molecules, including 1-dimensional $logP$, 2-dimensional topological fingerprints as well as 3-dimensional conformations [54]. It is these similarity measures that allow for ligand-based virtual screenings to be performed. When applying the approach to drug repositioning the aim is to search for, ideally approved, compounds that are structurally similar to other drugs that have known indications. It can then be assumed,

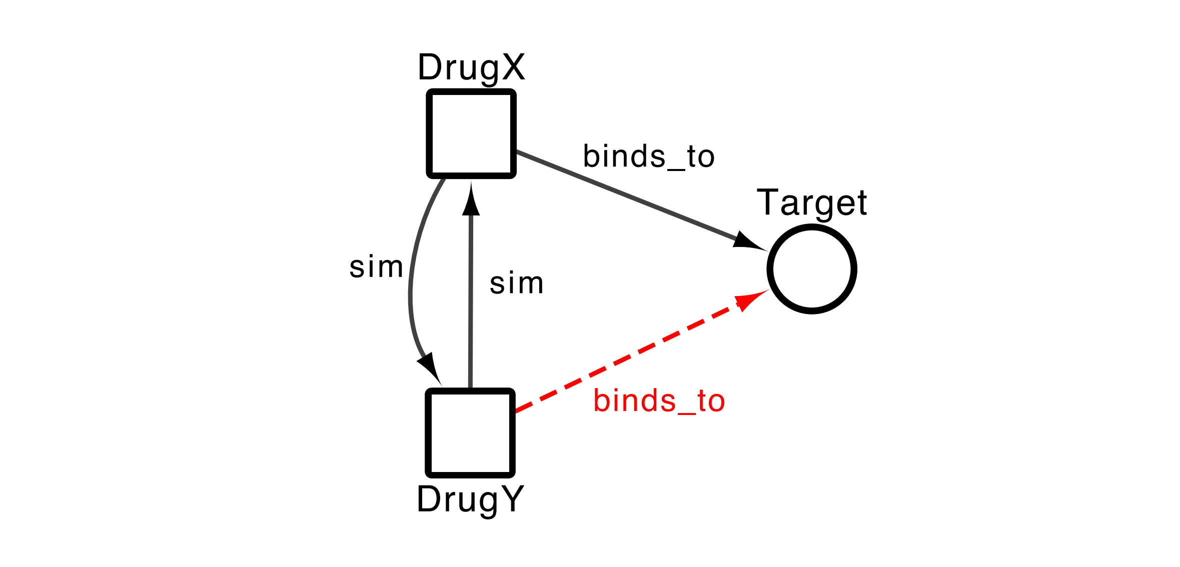


Figure 2.4: **Drug repositioning using compound similarity.** If a drug, `DrugX`, shares a similar structure to another drug, `DrugY`, then, using the similar property principle it can be assumed that `DrugY` will share similar properties with `DrugX`. If `DrugX` is known to bind to a `Target`, it can be inferred that `DrugY` will also bind that `Target`. *Note: sim* = structural similarity between drugs, red relation represents inferred association.

using the SPP, that the similar compound may also be used to treat the same disorder (Fig.2.4 shows an abstracted summary of the logic).

Examples include an approach described by Keiser *et al.* [55], who described a Similarity Ensemble Approach (SEA). Known approved drugs were grouped based on their known target binding partners and chemical features. The method then calculated, using a statistical model, the likelihood a molecule will bind to a target based on the chemical features it shares with those of other known drugs.

There is also a recent trend to integrate chemical structure similarity data with other types of data [18]. For example, Wang *et al.* [56] describe a drug repositioning model that makes use of chemical similarity data integrated with molecular activity and side-effect data. A kernel function is defined and used by a state vector machine classifier, with the approach showing high efficiency when compared to others [56].

Tan *et al.* [57] produced a drug similarity network incorporating 3-dimensional drug chemical structure similarity data, drug-target interactions and gene similarity data. Using this network, 33 modules of drugs with similar modes of action and indications were identified. Using these modules new indications were predicted for 143 drugs.

Focussing purely on chemical similarity is a risky approach to drug repositioning. Many drugs differ in structure from database to database as many structures, as well as other chemical properties of known drug compounds, contain errors, or are held as proprietary information [58]. Furthermore, many physiological effects cannot be predicted by chemical properties alone because drugs undergo complex and largely uncharacterised, metabolic transformations as well as other pharmacokinetic transformations as they are metabolised and physiologically distributed [58]. The structure of a ligand that is captured in a database may differ dramatically to that of the active form— particularly true when considering pro-drugs (a compound that, after administration, is metabolised into the pharmacologically active drug), such as tamoxifen (generic drug).

2.2.2.2 Drug combination approaches

Many diseases are driven by complex molecular and environmental interactions [18]. As such, treating a single component, target or pathway may not suffice to dislodge the mechanisms leading to disease. In cancers, for example, diseases are regulated by interactions of multiple signalling pathways interacting with one another, as such, it is not sufficient to target a single pathway. Furthermore, if a single drug is used to target a disease signalling pathway in cancers, then alternative signalling pathways may be activated to maintain tumour development, also known as acquired resistance [59]. Acquired resistance is a major problem in cancer patients. To increase cancer treatment outcomes, and reduce the drug resistance effect, drug combinations are seen as the optimal option. Drug combinations also have the potential to improve efficacy and reduce side-effects [60]. The majority of strategies in place rely mainly on clinical and empirical evidence, with most picked manually by clinicians depending on their experience and expertise [59]. Considering all possible combinations between drugs the potential search space is huge and impractical [61], as such computational prediction of combinatorial therapies is of high interest.

For example, Zhao *et al.* [60] used integrative approaches focussing on molecular and pharmacological features of drugs. Drugs were represented by sets of their properties, or features, such as targets or indications. A machine-learning method was then used to classify drugs into combinations using these feature patterns trained on pairwise combinations from the FDA Orange Book. The Orange Book contains information regarding drug products approved on the basis of safety and effectiveness by the FDA. 16 possible drug combinations were proposed during their work, with 11 being supported by the literature.

Huang, H. *et al.* [61] used clinical side-effects (from post-marketing surveillance and the drug label) as features for drug-drug combinations. The approach focussed on the safety of potential combinations and is based on the hypothesis that drugs that can be co-prescribed usually do not have or share an Adverse Drug Reaction (ADR). Interestingly they identified three FDA boxed side-effects; ‘Pneumonia’, ‘haemorrhage rectum’, and ‘retinal bleeding’ as top features contributing to model performance and thus developed a ‘Rule of Three’ criterion; a candidate drug combination with any of these side-effects is likely unsafe. The work identified 1,508 ‘safe’ candidate drug combinations.

Sun *et al.* [62] made use of genomic profiles to aid the drug combination prediction task. A model was developed to predict effective drug combinations by integrating gene-expression profiles of multiple drugs. During this approach existing drug combinations were extracted from the Drug Combination Database (DCDB) [63]. Statistical methods were used to identify significant features related to drug combinations in terms of side-effects, genes or disease pathways that would be affected by a known drug combination. These features were then used to construct a machine-learning classifier for predicting potential drug combinations.

One of the main limitations to identifying drug combinations computationally is the potential safety aspects. Safety issues are critical for co-prescribing drugs or developing fix-dose combinations [61]. Although simplifying and abstracting known safety information allows for unsafe combinations to be pruned, there is a significant risk that those identified as being safe computationally may indeed act very differently *in vivo*, due to an incomplete understanding of off-target effects.

2.2.2.3 Protein (ligand site) structure-based approaches

Most drugs are known to bind to more than one target [64], and the majority of these targets are proteins [37]. Drugs, such as the HIV treatment maraviroc (Celsentri; Pfizer) and the calcium mimicking cinacalcet (Mimpara; Amgen), target the allosteric areas of targets. However, the majority of drugs bind ligand sites of the target protein. Furthermore, it has been shown that there is a correlation between drug promiscuity and the structural similarity, as well as binding site similarity, of protein targets [38]. The interaction between a drug and a protein may be analysed computationally using an approach known as molecular docking. Molecular docking is a computational method that predicts how two molecules interact with each other in 3-dimensional space [51]. Docking allows for the optimisation of binding affinities, which can increase the potency of a drug. Many pharmaceutical companies, including Genentech and Melior, use docking to identify new indications for drugs [65]. Docking approaches are popular in drug discovery as they enable researchers to screen drug candidates against a panel of similar proteins to determine their specificity to the intended target [51] within a few days [65]. Due to the popularity of docking during the development process, it has inevitably been applied to the task of drug repositioning.

In the drug repositioning setting, the aim is to identify potential off-targets, based on similarities between the protein structure or the binding site structure of the on-target protein [51]. If an off-target is known to be involved in another disease, then the drug has the potential to treat the second disease (Fig.2.5 shows an abstracted summary of the logic). As such a series of studies have focussed on comparing the similarity between binding sites as a means of identifying these potential off-targets.

For example, Haupt *et al.* [38] used such an approach to identify properties that enable drug promiscuity. A systematic study of drug promiscuity was carried out based on structural data of Protein Data Bank (PDB)[3] target proteins, using a set of 164 promiscuous drugs. The approach identified 71% of the promiscuous drugs having at least two targets with similar binding sites.

Kinnings *et al.* [66] focussed on the repositioning of entacapone (Comtan; Novartis)

[3] www.rcsb.org/pdb

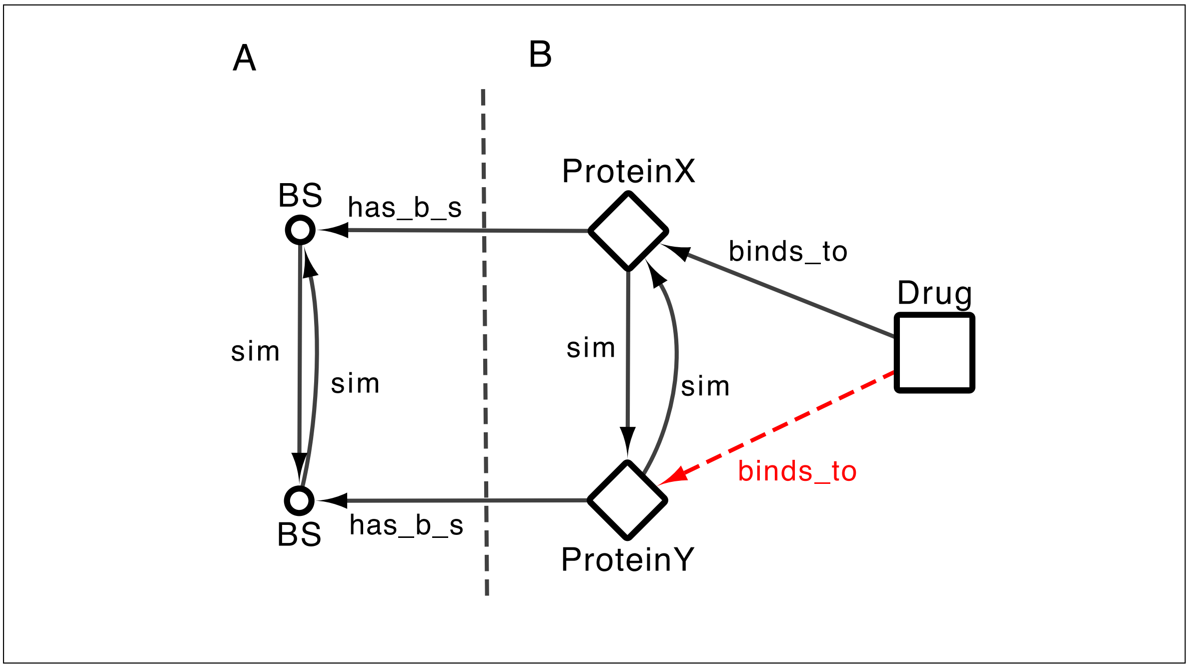


Figure 2.5: **Drug repositioning using protein structure and/or binding site similarity.** If a protein, `ProteinX` shares a similar structure (B), or similar ligand binding site structure (A), to another protein, `ProteinY`, then it can be assumed that a `Drug` that binds to `ProteinX` may also bind to `ProteinY`. *Note: sim* = structural similarity between proteins, or protein ligand sites, *BS* = binding sites, *has_b_s* = has binding site, red relation represents inferred association.

to treat multi-drug and extensively drug resistant tuberculosis. A chemicals systems biology approach to identify off-targets was used to make the prediction. During this approach, the binding site of a commercially available drug was first identified from a 3-dimensional structure of the target protein. Next, off-targets with similar ligand binding sites were identified using an efficient and accurate functional site search algorithm. Atomic interactions between the putative off-targets and the drug were then evaluated using protein-ligand docking before the drug was optimised to enhance potency, selectivity and ADMET properties.

Approaches to drug repositioning utilising only protein structure rely on 3-dimensional structural data being available. One of the most frequently used sources for such information is the PDB. The PDB is still far from containing the whole proteome, limiting inferences to a subset of potential drug targets, and thus cannot be used to predict new mechanisms beyond the known targets [65]. Furthermore, the structures of many physiologically relevant proteins are not fully resolved; including whole families of GPCRs, which are favoured as drug targets for many approved drugs [58]. Protein structure methods are also prone to producing great numbers of false positives due to errors in resolved protein structures and incomplete modelling of atomic and molecular interactions [58].

2.2.2.4 Gene expression-based approaches

Rapid advances in genomics have led to the generation of large volumes of genomic and transcriptomic data for a diverse set of disease samples, normal tissue samples, animal models and cell lines [63]. Much of these data are publicly available. Together with other phenotypic, and clinical databases, these data sets provide a unique opportunity to understand disease mechanism, elucidate drug mechanism of actions and identify new uses for old drugs. Among these, transcriptomic profiles, such as gene expression data are most widely used [63].

Transcriptional drug-treatment databases, such as the Connectivity Map (CMap) [67] project and its extended project, Library of Integrated Network-Based Cellular Signatures (LINCS)[4] [68], provide measurement data. Measurement data is usually produced using a microarray with probes for a set of genes of interest. Provided data includes various experimental factors, including multiple cell types, doses, and time points [69]. CMap aims to construct a detailed map of functional associations among diseases, genetic perturbations and drug actions by observing the behaviour of an organism's gene expression in a particular setting [70]. Differentially expressed genes can serve as a proxy to characterise a molecular effect, the so-called gene expression signature [70]. Comparison of differentially regulated gene expression profiles from cultured human cells treated with bioactive molecules is enabled as well as cross-platform comparisons. The data captured in these databases has been the principal source behind several drug repositioning studies [63], as well as being integrated with other functional genomics databases (such as the NCBI Gene Expression Omnibus (GEO) [71]).

One of the most common approaches utilising CMap data involves identifying new potential interactions based on the mapping of opposite expression profiles of drugs and diseases. This method is known as 'signature reversion' and is depicted in Fig.2.6A. For example, Dudley *et al.* [58] systematically compared gene expression signatures

[4] www.lincsproject.org

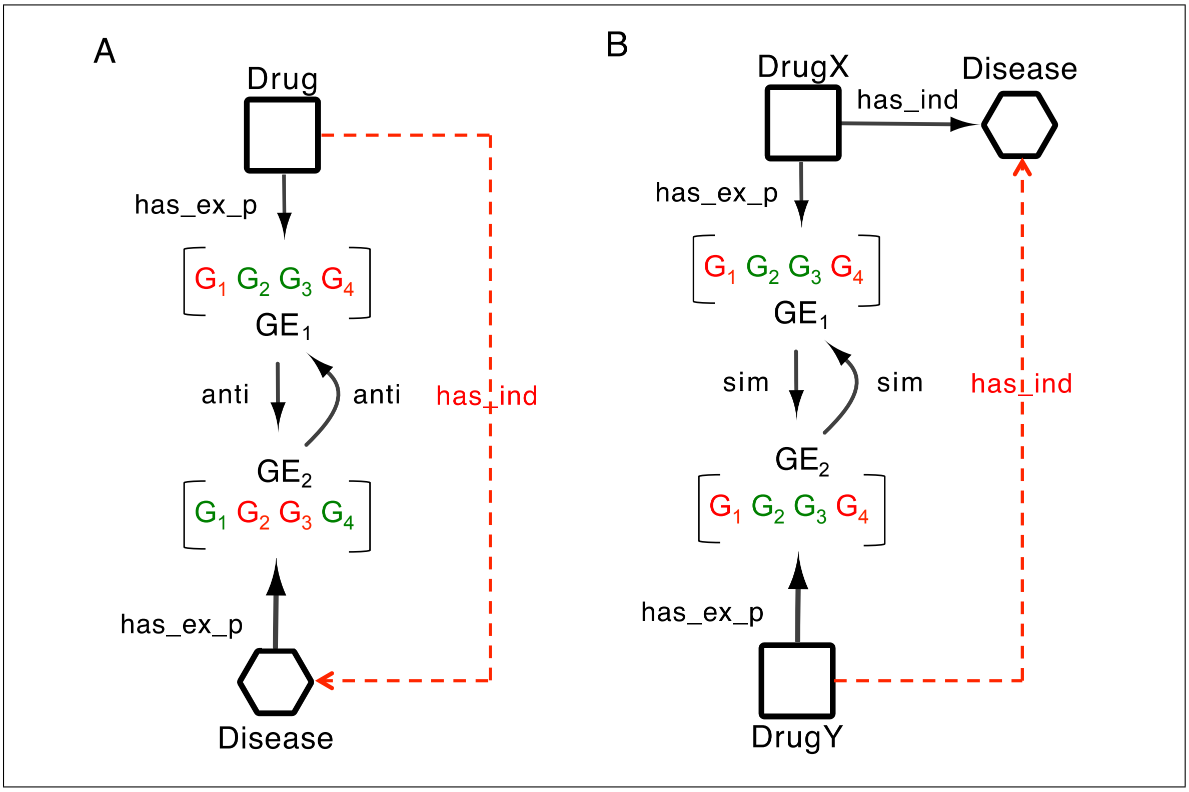


Figure 2.6: **Drug repositioning using CMap data.** (A) An overview of signature reversion approaches to drug repositioning using CMap data. (B) Guilt-by-association approaches using CMap data may allow for novel uses of drugs to be identified. *Note: G* = gene, *GE* = gene expression profile, *has_ex_p* = has expression profile, *has_ind* = has indication, *sim* = similar expression profile, *anti* = anti-correlated expression profile, red relation represents inferred association.

of inflammatory bowel disease (IBD) derived from GEO against a set of drug-gene expression signatures comprising 164 drug compounds from CMap, inferring several new interesting drug-disease pairs. One of these pairs was validated in IBD preclinical models. Jachan *et al.* [72] used a similar strategy to query a large set of gene expression profiles. The aim of the study was to reposition antidepressant drugs to the treatment of small cell lung cancer.

Other approaches apply 'guilt-by-association' methodologies to analyse CMap data. These techniques look at drugs which create similar expression profiles, in the hope that this indicates a similar mode of action [73]. Fig.2.6B shows an abstracted summary of the logic used during such an approach.

Although a potentially interesting source for drug repositioning, transcriptomics data present considerable challenges in terms of statistical analysis [67]. Methods based on expression activity similarity rely heavily on the quality and assumptions of the means used to derive the expression profiles. For example, the CMap date is created by exposing drug compounds to isolated cell lines. Isolated cell lines likely do not accurately reflect a complete physiological system, leaving the accuracy of the observed biological activity of the drug in doubt. Many drugs undergo chemical transformations after they are metabolised *in vivo*, and it is these drug metabolites that often provide the eventual therapeutic effect [58]. Furthermore, the pathology of many disease conditions, such as Type 2 diabetes, spans multiple tissues and organ systems; therefore, it is difficult to represent and compare such diseases on the basis of a single expression signature.

2.2.2.5 Genetic variation-based approaches

Genetic variations can also provide valuable insights regarding drug repositioning opportunities. Due to recent developments in high-throughput DNA sequencing methods and analysis pipelines, it has become increasingly affordable to sequence individuals and study their genotypes. From the information generated, common mutations can be isolated in the DNA that are significantly associated with a phenotypic trait, a method known as Genome-Wide Association Studies (GWAS). It must also be noted that historically, GWAS were array-based studies requiring large cohorts and thus were costly. Sequencing-based approaches are only just beginning to become prevalent, but still face significant limitations compared to array-based assays.

GWAS data has been used to unveil potential new indications for protein targets. For example, Sanseau *et al.* [74] integrated disease associated genes from GWAS with targets of drugs from pharmaceutical projects. In this approach links between genes and disease traits were analysed. Using knowledge that a drug targets the given gene product, the indication of the drug was expected to be the same as the trait studied in the GWAS. If this was not the case, and a mismatch between the trait and known indications was present, then a potential repositioning opportunity was proposed (Fig.2.7 shows an abstracted summary of the logic).

Okada *et al.* [75] performed a three-stage GWAS meta-analysis of rheumatoid arthritis (RA) patients, linking the risk loci to known RA drug targets. In the study, logistic

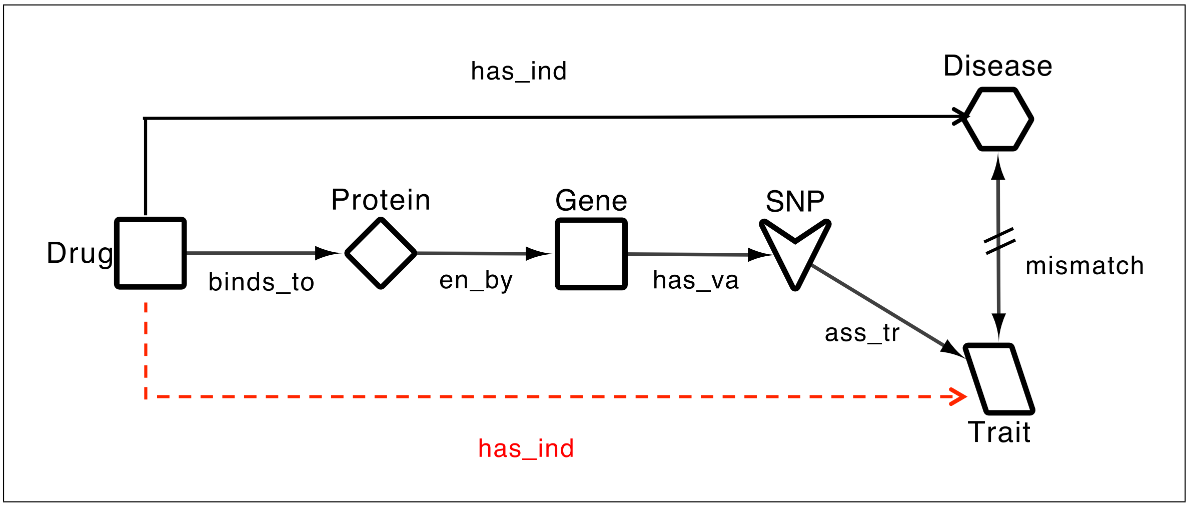


Figure 2.7: **Drug repositioning using GWAS data.** An abstract view of how GWAS data may be used to infer novel drug repositioning opportunities as described in [74]. *Note: en_by* = encoded by, *has_va* = has variation, *ass_tr* = associated trait, *has_ind* = has indication, red relation represents inferred association.

regression models assuming additive effects on the allele dosages were used to assess the relationship of single nucleotide polymorphisms (SNPs) and RA. In total 101 RA risk loci (42 novel) were identified and showed significant overlap with approved RA drug targets. Finally, several approved drugs are connected to RA risk genes, indicating repositioning opportunities.

Any approaches that make use of genetic data are limited to predicting drug repositioning opportunities for disorders that have been investigated by GWAS studies. It is also difficult to assess ascertainment bias when using such data [74].

2.2.2.6 Phenotype-based approaches

Phenotype-based approaches to drug repositioning include those that focus on diseases or side-effects.

Diseases can be linked together based on multiple shared features, such as the cause of the pathology or the biological dysfunction observed. Approaches generating networks of diseases and the similarities between these diseases, aim to develop a 'diseasome' view. For example, Li *et al.* [76] used data relating to disease associated genes and pathway associated diseases, where disease genes were enriched. Diseases were then linked together based on shared pathways, hypothesising that diseases with commonly deregulated pathways were similar. Novel disease relationships were then introduced in the hope that the work would enable pathway-guided therapeutic interventions for diseases.

Hoehndorf *et al.* [77] applied a semantic text-mining approach to identifying the phenotypes associated with over 6,000 diseases. Furthermore, they made use of a phenotypic similarity measure to generate a human disease network where clusters contained diseases that had similar signs and symptoms. Finally, the network was used to identify closely related diseases, based on common etiological, anatomical as well as physiological underpinnings. Although not directly addressing drug repositioning, disease maps can provide insight regarding the usage of a drug.

The phenome-wide association study (PheWAS) has, in recent years, become an increasingly popular approach to identifying relevant genetic associations with human diseases [78]. Denny *et al.* performed a large-scale application of the PheWAS using electronic medical records (EMRs), demonstrating the utility of the PheWAS as a useful tool for detecting novel associations between genetic markers and human diseases.

Side-effects (resulting from off-targets, as opposed to ADMET genes) analysis has enabled the discovery of novel therapeutic uses of drugs. Methods using the pretence that similar side-effect profiles may give rise to similar therapeutic profiles have frequently been applied to drug repositioning [79–81]. Yang *et al.* [80], for example, used side-effects from SIDER [82] to link diseases and extract drug repositioning opportunities. Chemicals were linked to pathologies using information available in the pharmacogenomics knowledge base. The approach valued evidence showing that drugs used to treat similar diseases have similar side-effects. In this respect, side-effects can be indicators of a common underlying mode of action, and two drugs sharing a similar side-effect profile can be used to treat the same pathology.

Finally, phenotypic information may also be integrated with other data. Hoehndorf *et al.* [83] for example, integrated data involving genotype-disease associations with drug-gene associations. During this work, a semantic similarity-based score was derived to measure genotype-disease associations.

The most apparent limitation of the side-effect similarity approaches is the necessity for having well-defined side-effect profiles for a drug [58]. Despite rigorous preclinical

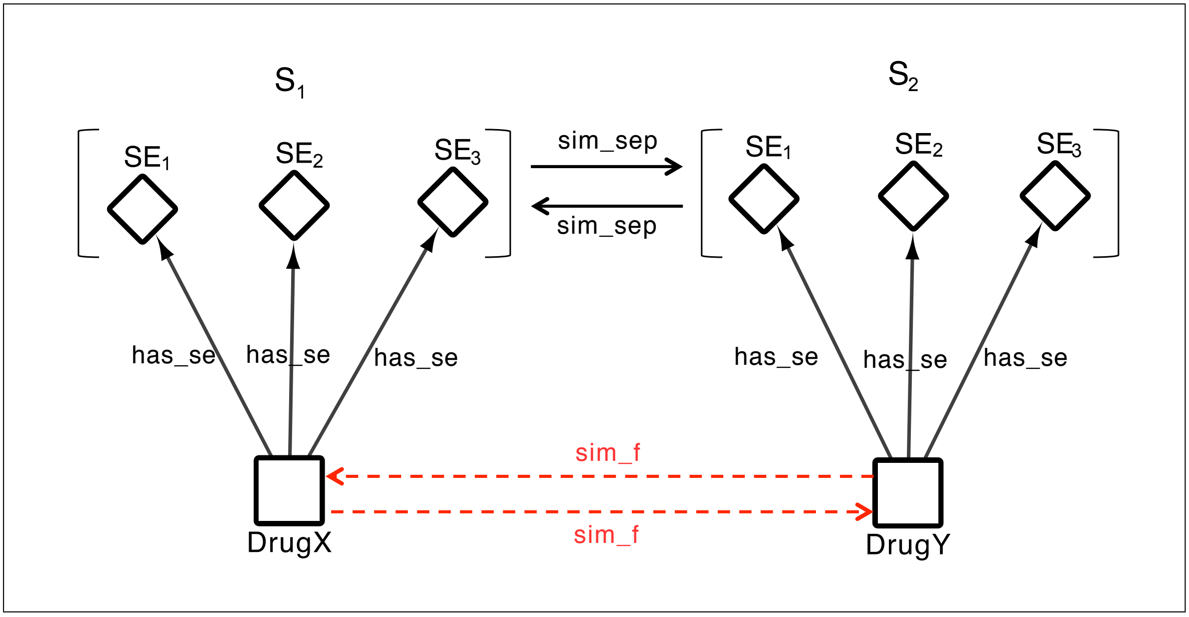


Figure 2.8: **Drug repositioning using side-effect similarity.** An abstract view of how side-effect similarity profiles may be used to infer novel drug repositioning opportunities. *Note:* SE = side-effect, S_1 = set of side-effects associated with `DrugX`, S_2= set of side-effects associated with `DrugY`, *sim_sep* = similar side-effect profiles, *sim_f* = inferred functional similarity.

assessment, the side-effect profile for a newly approved drug may only be fully discerned after years of clinical use and post-market surveillance [58]. In addition, the assumption that similar phenotypic expression of a drug side-effect implies a common pathophysiological basis may not always hold. For example, the side-effect of 'hair loss' can arise when 1) a drug interferes with hormonal systems that regulate hair growth, or 2) a drug causes harm to the cells comprising the hair follicle via disrupting immune function [58].

2.2.2.7 Machine learning approaches

A sub-field of computer science, machine learning evolved from the study of pattern recognition and computational learning theory in artificial intelligence. Machine learning and can be defined as:

> “The set of computer algorithms that automatically learn from experience [84]”

In the field of drug repositioning, machine learning models can leverage various relevant data to study the underlying systems for the prediction of novel uses for drugs [18]. Such approaches provide a way to combine various descriptors into one statistical model, with the aim of increasing the accuracy of the predictions. A great number of machine learning methodologies have been developed.

For example, Gottileb *et al.* [85] introduced PREDICT, a method capable of inferring novel indications for both approved drugs and novel compounds. Based on disease-related features, drug-drug similarities (characterised by the structure and the associated side-effects) and disease-disease similarities (characterised based on their semantic similarities calculated over the Human Phenotype Ontology) measures were computed to construct classification features. A logistic regression classifier was then used to predict novel drug indications. Inferred indications were then validated against known indications. However, negative samples are needed in their model to implement the prediction procedure. Experimentally verified negative drug-disease associations are not available due to lack of research value.

Napolitano *et al.* [86] focussed on a drug-centric approach. Within this methodology, the therapeutic class of FDA-approved compounds were predicted. Making use of data integration drug-related features (drug structure similarity, target structure similarity, drug-gene expression similarity) were merged into a single drug similarity matrix. In turn, this similarity matrix was used as a kernel for Support Vector Machine (SVM) classification. The algorithm was used to predict therapeutic categories of the Anatomical Therapeutic Classification System (ATC), with misclassification, interpreted as a potential repositioning opportunity.

Menden *et al.* [87] developed machine-learning models to predict the response of cancer cell lines to drug treatment, quantified through IC_{50}. In the model cancer genomic features (oncogenes) and chemical properties (structural fingerprints) were used to build a feed-forward perception neural network model and a random forest regression model. Models were used to optimise the experimental design of drug-cell screening by estimating a significant proportion of missing IC_{50} values rather than needing to measure them experimentally. Predicted IC_{50} values were further validated by a cross-validation and an independent blind test. It was proposed that this work could be used as a computational framework for the identification of new drug repositioning opportunities as well as aiding personalised medicine, by linking the genomic traits of patients to drug sensitivity.

Finally, Zhang *et al.* [88] proposed a computational framework for integrating multiple aspects of drug similarity and disease similarity. Genome (e.g. drug target protein, disease gene), phenome (e.g. disease phenotype, drug side-effect) and chemical structure (e.g. drug chemical structure) data were integrated to create multiple drug similarity networks, multiple disease similarity networks, and known drug-disease associations. Using all this information, the authors turned the drug-disease network analysis into an optimisation problem. In addition, the authors showed how three of the top ten drug-disease inferences involving Alzheimer's disease were supported by clinical trials.

Machine learning approaches to drug repositioning come with their caveats. Essentially the statistical model used in machine learning is a black box, hiding the rational pieces of evidence of why a compound, or indeed a disease, may have been chosen, meaning interpreting the repositioning hypothesis is a difficult task.

2.2.2.8 Network theory-based inference methods

Iorio *et al.* [89] developed a network theory-based approach that exploits similarity in gene expression profiles following drug treatment. This technique was also implemented in and released as a tool; MANTRA. Using multiple cell lines and dosages, similarities in drug effect and mode of action (MoA) were first predicted. A 'drug network' was then constructed; made up of 1,302 drugs (nodes) and 41,047 drug similarity scores (edges). Network theory, in particular a clustering algorithm, was then used to partition the network into groups, or communities, of densely connected nodes. Communities were significantly enriched for compounds with similar MoAs. Using this approach, the MoA for nine anti-cancer compounds were correctly identified. It was also proposed that the potent Rho-kinase inhibitor and vasodilator, Fasudil (approved in Japan and China but not in Europe or the US), may be repositioned as an enhancer of cellular autophagy.

2.2.2.9 Text mining

It is unrealistic to expect manual curations to keep up with the amount of data reported in the literature. As such automated approaches for the extraction of data, be it biomedical terms or their inter-relations, are greatly beneficial [90]. This automated analysis of the literature essentially allows for the access of information that would otherwise be 'locked' in a plethora of free-text documents. Central to these automated approaches are the ontologies that are used as controlled vocabularies to provide a framework for mapping relations between concepts [90]. Although the accuracy of such data is not as high as for manually curated sources, the systematic approach to the methodologies means that they are more inclusive of true positives. It is hypothesised that finding relevant knowledge through text mining may allow for the identification of novel indications for existing drugs [91].

The potential for literature mining to be used in drug discovery was initially identified in the 1980s. Swanson [92] used multiple lines of evidence to propose the use of fish oil to treat Raynaud's syndrome, a hypothesis that was later validated in a clinical trial [22]. The model used was referred to as the ABC model. For example, a study finds that disease A was caused by a loss of function mutation of gene B. Furthermore, another study, from another scientific discipline, reports that drug C is an agonist of gene B. Using the ABC model, an implicit link between A and C may be made— drug C may be repositioned to treat disease A.

In this tool gene names as well as other biomedical concepts are extracted from Medline abstracts. Furthermore, these concepts are related to each other using co-occurrence. A mutual information-based metric was used to assess the strength of co-citations. Finally, a series of case studies with novel open and closed model discoveries were presented. These included disease-gene, drug-disease, drug-biological process and biological process-drug relationships. The latter scenario resulted in the novel association between two drugs and cell proliferation: dephostatin, a tyrosine phosphatase inhibitor; and damnacanthal, a tyrosine kinase inhibitor. In vitro cell proliferation experiments validated the influence of these two compounds in the process of cell proliferation at low micromolar concentrations. *In vitro* cell proliferation experiments validated the influence of these two compounds in the process of cell proliferation at low micro molar concentrations.

Although text mining approaches are highly sensitive, their specificity is questionable. Any hypotheses formed using text mining approaches will need to be followed by manual curation [90]. Furthermore, methods making use of the ABC model will have to be wary of inferring potential side-effects.

2.2.2.10 Summary of computational approaches described

Each of the aforementioned computational repositioning approaches described has their methodological advantages and limitations. A combination of these methods is often desired for achieving better results [18]. As such, there is a belief that drug discovery, and by proxy, drug repositioning, may be improved by taking a holistic integrative approach, considering the interaction of existing drugs with target proteins as well as other biological molecules [93]. Holistic approaches are enabled by the ever-increasing volumes of drug-related and -omics data, such as pharmacological, genetic, chemical, or clinical data [18]. As such, bioinformatics finds itself with an increasingly important role in the discovery of new drug indications [94]. By considering a wider range of data, holistic approaches can produce more confident predictions in comparison to methods that utilise a single, or even only a few different data types. In this work the term 'systems approach' is used to describe an integrative approach that enables a holistic view of a drug. Systems approaches provide a method of utilising data increase and can accumulate evidence supporting the discovery of new uses for existing drugs [18].

2.2.3 Systems opportunities

At this point a clear differentiation between 'systems biology' and the term 'systems approach' must be made. Systems biology, a term first coined in the 1960s, was initially used to describe the mathematical models created by theoretical biologists to model biological systems [95]. Nowadays, systems biology is a term used to describe many areas of work, including: (i) network reconstruction through data integration (often involving a single type of entity, such as a protein-protein interaction network)

(ii) network analysis methods (iii) modelling (be it mechanistic or Bayesian) (iv) simulations (such as stochastic and deterministic simulations) [95]. In this thesis the term 'systems approach' will be used to describe the development of integrative, holistic representations of drugs (with regard to pharmaceutical, genotypic and phenotypic entities) as well as methodologies that consider all data captured in such representations. Unlike systems biology approaches, the systems approach defined does not consider modelling or simulation.

An essential part of systems approaches is data integration. The importance of data integration in the drug discovery setting has not gone unnoticed with numerous data integration projects in place. Commercial efforts include DistilBio[5], developed by Metaome[6], which provides an integrated view of over 30 life science focussed data sources. DistilBio provides a graph-based data set using a custom ontology and allows for queries to be created visually, directed by their ontology. Open source efforts include Open PHACTS[7]. Providing an example of a large-scale data integration project relevant to drug discovery, Open PHACTS is funded by the European Innovative Medicines Initiative (IMI). Open PHACTS aims to provide an integrated view of drugs and their targets and pulls together data from multiple pharmacological data sources. Another chemical-based open integration effort is ChEMBL[8], from the European Bioinformatics Institute (EBI). ChEMBL previously included mostly chemistry data that was accessed systematically using tools from vendors only. Although Open PHACTS and ChEMBL collate a large amount of information regarding a compound and its pharmacology, they miss data that would be useful to the drug repositioning setting and, as such, further integration of data is required to enable systems approaches to be taken in the drug repositioning setting. Relevant data are deposited in many distributed, heterogeneous and voluminous data sources, and thus, integration must be achieved to produce homogeneous technical, semantic and syntactic data.

Taking a state-of-the-art snapshot of drugs and their known interactions with other biological entities can allow for more confident predictions of possible novel uses to be

[5] www.distillbio.com

[6] www.metaome.com

[7] www.openphacts.org

[8] www.ebi.ac.uk/chembl

made. For this reason, the work presented in this thesis focusses on systems approaches to drug repositioning. In order to take a systems approach, a clear data representation must be defined. The chosen data format to be used during this work is the graph. Network-theory, a subset of graph-theory, approaches have been introduced in this Chapter. There are different definitions for graphs and networks, with each providing subtle differences between the two structures. For example, a graph is said to be the mathematical definition of nodes and edges whilst a network is said to have discrete labels. As a graph can also, and in the case of this work does have, discrete labels, the terms graph and network will be used interchangeably.

2.2.4 The benefits of graph-based data representation

Graph representations of complex systems are widely used in computer science, social and technological network analysis science, and is especially relevant to many studies in bioinformatics [96]. In *semantic graphs* both edges (or relations) and vertices (or nodes) are *typed*: each vertex and/or edge in the graph is assigned a type from a predefined set, vertices and edges are also *attributed*: vertices and edges are annotated with attributes. The properties of graphs make them ideal for use in data integration applications as they enable data from multiple sources to be stored and interconnected. For example, for each computational approach described in Section 2.2.2, an abstract graph-based summary of the underlying logic was provided. In every image, there is at least one node of type drug, with edges of different types connecting these drug nodes to other nodes of alternative data types. It is, therefore, easy to imagine how all the types of data presented in these images may be collated, or integrated, into one graph-based representation. At the centre of this graph would be a node of type drug, with edges connecting this drug to nodes representing all the other data types present in the images. With a single data set containing all of these data a single query can cover multiple data sources.

To extract any useful information that may be captured by a graph-based integrated data set, querying, or data mining, techniques must be applied. Such techniques fall under the umbrella of graph theory, where graph pattern matching, including subgraph isomorphism, allows for the direct extraction of subgraphs that may be representative of a particular function or process. Since the data in this project is presented as graphs, one of the main computational challenges of the work was the development of novel graph mining techniques. Furthermore, novel graph-based mining approaches that make maximum use of semantically marked up graphs were required to most efficiently mine integrated data.

2.3 Conclusion

In this Chapter the concept of drug repositioning was introduced with the benefits and opportunities in relation to typical R&D approaches to drug discovery discussed. Marketed examples of drug repositioning have been described and the need for systematic methodologies to enable this field of research highlighted. Current systematic computational approaches have been categorised and described, and examples provided. Furthermore, the advantages and limitations of each computational approach have been detailed. Although many approaches have been described it can be seen that many of these are limited to the inference of a particular type of association. Although machine learning methods tend to take a more integrative approach to data than other approaches, the fact that they make use of a 'black box' leaves questions regarding the accuracy, or at least provenance, of assumptions made.

Integrative approaches, such as system approaches, enable a holistic view of pharamceutical and biological entities and have the potential to provide more confident inferences, whereby the evidence used in the prediction can be decomposed and critically analysed. System opportunities have been characterised in this Chapter and identified as a methodology with the potential to enable a holistic approach to drug repositioning. A systems approach can enable the integration of all data types considered in the approaches described in this Chapter, allowing for more confident inferences to be made.

For a systems approach to drug repositioning to be taken, a holistic view of a drug and its interactions is required and is achieved using data integration techniques. Data mining techniques, based on graph theory, can then be used to make novel inferences from an integrated data set captured in a graph-based representation. These novel inferences are often a result of emergent properties of an integrated network. As such, data integration for drug repositioning is discussed in Chapter 3, before an algorithm for the inference of novel associations from the resulting integrated data sets is introduced in Chapter 4.

3

Data integration for drug repositioning

3.1 Introduction

High-throughput sequencing technologies have revolutionised life sciences and lead to a dramatic expansion of data-rich resources for bioinformatics [97]. The shift to large-scale sequencing of individual genomes and the availability of new techniques such as positional cloning and microarray analysis allows probing of thousands of genes that can return tens to hundreds of candidate Gene-Disease (G-D) associations [98]. As a result of this data explosion, there are now more than 1,500 publicly available data sources in the life sciences [99]. Research in the life sciences has a general goal to identify the components that make up a living system and to understand the interactions among them that result in the (dys)functioning of the system [100]. The collection of biological data is, therefore, a method to catalogue the elements of life, but understanding a system requires the integration of these data to describe the relations between their components [100]. Systems-based approaches enable a holistic view of an organism to be realised and is achieved through the task of data integration.

In this Chapter, data integration is introduced along with challenges facing its usage, focusing on its application to the field of drug repositioning. Methods and platforms used for data integration, such as Ondex and Neo4j, are described as well as relevant data formats such as Resource Description Framework (RDF) and ontologies. An integrated data set previously developed in Ondex for *in silico* approaches to drug discovery [101] is first detailed and is then extended as part of this work to include data types relevant to drug repositioning. A comparison between Ondex and Neo4j, regarding performance, is presented— highlighting Neo4j as a more scalable platform for data set development. As Neo4j has no innate means of controlling semantics, Bioinformatics Semantic Integration Platform (BioSSIP) is introduced. BioSSIP is a light-weight module that sits in front of a datastore 'backend', such as Neo4j, and allows for integration projects to be created independently of the users datastore of choice. A Drug Repositioning Network Integration Framework (DReNInF) is then outlined. This framework makes use of BioSSIP and is made up of: Drug Repositioning Network Integration Ontology (DReNInO), a high-level drug repositioning ontology; a suite of more than 20 parsers for drug repositioning relevant data sources; and a data integration strategy. Finally, an RDF exposed data set, Drug Repositioning Network Integration (DReNIn), is characterised. DReNIn is accessible via a dedicated Web site at www.drenin.ncl.ac.uk and can be queried using a SPARQL Protocol and RDF Query Language (SPARQL) endpoint.

3.2 Background

3.2.1 Data integration

The term 'data integration' refers to the situation where, for a given system, multiple sources and types of data are available, and it is beneficial to study them integratively to improve knowledge discovery [100]. Data integration describes the process used to produce a homogeneous syntactic and semantic representation of multiple, often heterogeneous, data sources, that ideally provides a single, unified query interface.

Data integration was born out of necessity and aims to provide a redundancy-free representation of information from a collection of data sources with overlapping content. Used to bring together data from multiple data sources, data integration enables a non-redundant, normalised, holistic view of a system. 1,062 papers mentioned 'data integration' in their abstract or title in 2006. By 2013, this number had more than doubled to 2,365 [100]. Despite the widespread application of data integration to multiple fields including the life sciences, data integration also poses multiple challenges [100].

3.2.2 Challenges involved in data integration

Data integration in the life sciences is particularly burdensome for a number of reasons. As data sets continue to grow the term 'Big Data' becomes ever more widespread, describing data sets infeasible for processing on a single machine. Furthermore, these data are spread across a plethora of databases accessible via a multitude of conventions and available in varying data formats. As a result, multiple challenges must be considered prior to a data integration project. A non-exhaustive overview of these difficulties is provided here.

Identification of quality data sources The first question to ask is 'what data sources are to be included?'. Data is produced by a multitude of research communities and subfields, with varying levels of expertise and aims. It is, therefore, important to understand the quality of the source that is to be used. Does the data source contain *primary* data (stored in operational or working databases) or *derived / secondary* data (refined and presented at a higher level) [102]? One must also be aware of the fact that data sets are constantly updated and may vary dramatically from one release to the next, whether that be in data format or data content.

Data heterogeneity, syntax and semantics At higher levels of integration the heterogeneity of data increasingly becomes the major issue, if only because the more disparate the data types the more likely there are to be mismatches [102]. Mixing data types from disparate sources almost inevitably creates issues related not just to the format, or *syntax* of the data, but also its *semantics*, or underlying meaning.

Inference of equivalence Combining data sources requires recognising data 'types', such as genes or diseases and identifying equivalence, or mappings, between entries of the same type stored in two different data sources. This task can be straight forward if there are unique simple keys and a standardised accession for the data (e.g. HGNC for genes), but this is not often the case for more complex entities (e.g. diseases) [102]. For example, achondroplasia, the most common form of chondrodysplasia, in Orphanet Rare Disease Ontology (ORDO) has the label Orphanet_15, whereas, in the Medical Subject Headings (MeSH) hierarchy, it has the label D000130. Both of these entries refer to the same disease in different vocabularies and so it can be stated that $Orphanet\_15 \equiv D000130$. During an exercise of graph-based data integration, this disease would be captured as a single node containing both labels Orphanet_15 and D000130, respectively.

The size of a data set It is important to understand, prior to any integration project, any limitations that may affect the size of the data set. If data is to be stored on a local machine, for example, one must be aware of the amount of memory available.

The ability of a system to handle a growing amount of work, known as scalability, as well as data set size, should be well managed.

Metadata management To manage variations in sources, an integrated data source needs to provide *meta-data* (data that describes the data) including *data provenance* (describing where the data has come from and when). Provenance becomes even more necessary when one wishes to integrate multiple versions of the same data set and compare updates. The amount of metadata provided can vary between different data sources.

Representation and access Different types of data representation may be required depending on the origin of the data, the method of access and the manner in which the data is to be queried or visualised [102]. There is no single method of access to biological databases with access points including Representational State Transfer (REST) services, SQL databases, flat files via File Transfer Protocol (FTP), screen scraping as well as many others.

What questions are to be asked? Before starting an exercise in data integration, it is important to understand the use of and questions to be asked of, the resulting integrated data set. These questions will provide direction to the task and can help when considering all of the challenges described above. For example, if the research question was 'identify all proteins that interact with proteinX that is encoded by geneY' data describing tissue types would be redundant.

Once challenges facing a data integration project have been identified and considered it is important to select a data integration platform that will allow for all objectives to be achieved.

3.2.3 Data integration methods and platforms

Many integration approaches and platforms have been designed to solve different types of integration problems. For example, federated approaches to data integration involve leaving data on several, distributed servers. An integrated view is then produced by drawing these data sources together within a client application, usually located on a different machine. Alternatively, data warehousing relies on centralised data management and retrieval. In this case, the task of data integration results in a homogeneous data set which may be stored using a number of different database models with a common schema. Both approaches require either consistency in IDs so that data items can be matched across data sets or IDs that map onto a common ontology.

3.2.3.1 Database models

There are three main database models used for data storage and mining tasks in computing: *in-memory*, relational (SQL) and key-value (or NoSQL) stores. There are, however, situations whereby some crossover may exist, for example, *in-memory* SQL databases. *In-memory* formats, such as those employed by Cytoscape [103], Gephi [104] and Ondex [105] are fast when used to their strengths. However, they are designed to enable the complex analyses of 'small' data sets i.e. small enough to fit in the memory of a single machine. Thus, *in-memory* approaches are limited by memory availability. Due to their ability to handle structured and semi-structured data, graphs are often the data representation of choice for integrated data sets in bioinformatics. Often, graphs are made up of too many nodes and edges to represent in RAM and are consequently stored in databases [106].

Traditionally, relational databases such as MySQL[1]and PostgreSQL[2] have been used for this purpose. Developed in the late 1960s, relational databases have decades of research toward their query optimisation [106]. Relational databases maintain tables which are defined by sets of rows and columns. A row can be perceived as an object whose columns are attributes/ properties of that object [107]. The strengths and limitations of relational databases are well known, and a wide talent pool of trained professionals exists. However, an innate inability to explicitly capture required semantics limits the application of the relational database. Schema-based data models put in

[1] www.mysql.com

[2] www.postgresql.org

place limits as to how data will be stored, with a manual process of redesign required to adapt to new data. Relational databases are a better option when complex queries and set operations are needed [106] and are optimised for aggregated data, whereas graph databases are optimised for highly connected data.

The term 'NoSQL' refers to schema-less databases such as: key/value stores (e.g. Apache Cassandra[3]); document stores (e.g. MongoDB[4]); and graph databases (e.g. Neo4j[5] and Virtuoso[6]). These databases, which do not fit within the traditional relational paradigm, are gaining popularity due to their scalability and flexibility in comparison to the relational approach [106]. Graph databases store edges as directed pointers between nodes which, in turn, depending on implementation, can be traversed in constant time. When information about data interconnectivity or topology is important, graph databases become increasingly relevant and are optimised for graph traversals (degrees of separation or shortest path algorithms) [106].

Like graph databases, triplestores, which handle a specific data format, RDF, are designed to store linked data. Data points are called nodes, and the relationship between one data point, and another is an edge. Triplestores are a kind of specialised graph database with some subtle differences. Unlike graph databases, which are node, or property-centric, triplestores store lists of graph edges, many of which are just node 'properties' and not critical to the underlying graph structure. Graph databases are capable of storing a variety of different graph types, including undirected graphs, whereas triplestores specifically handle edge labelled directed graphs. Most graph databases do not possess a declarative query language (Neo4j is an exception, with its query language, Cypher, designed for expressing graph queries, being currently under development), whereas triplestores can be semantically queried using SPARQL and also allow for inferences to be made between nodes. The cost of traversing an edge, using SPARQL, tends to be logarithmic. To use a triplestore directly, one must understand the RDF data format.

[3] www.cassandra.apache.org

[4] www.mongodb.com

[5] www.neo4j.com

[6] www.virtuoso.openlinksw.com

3.2.3.2 Data formats

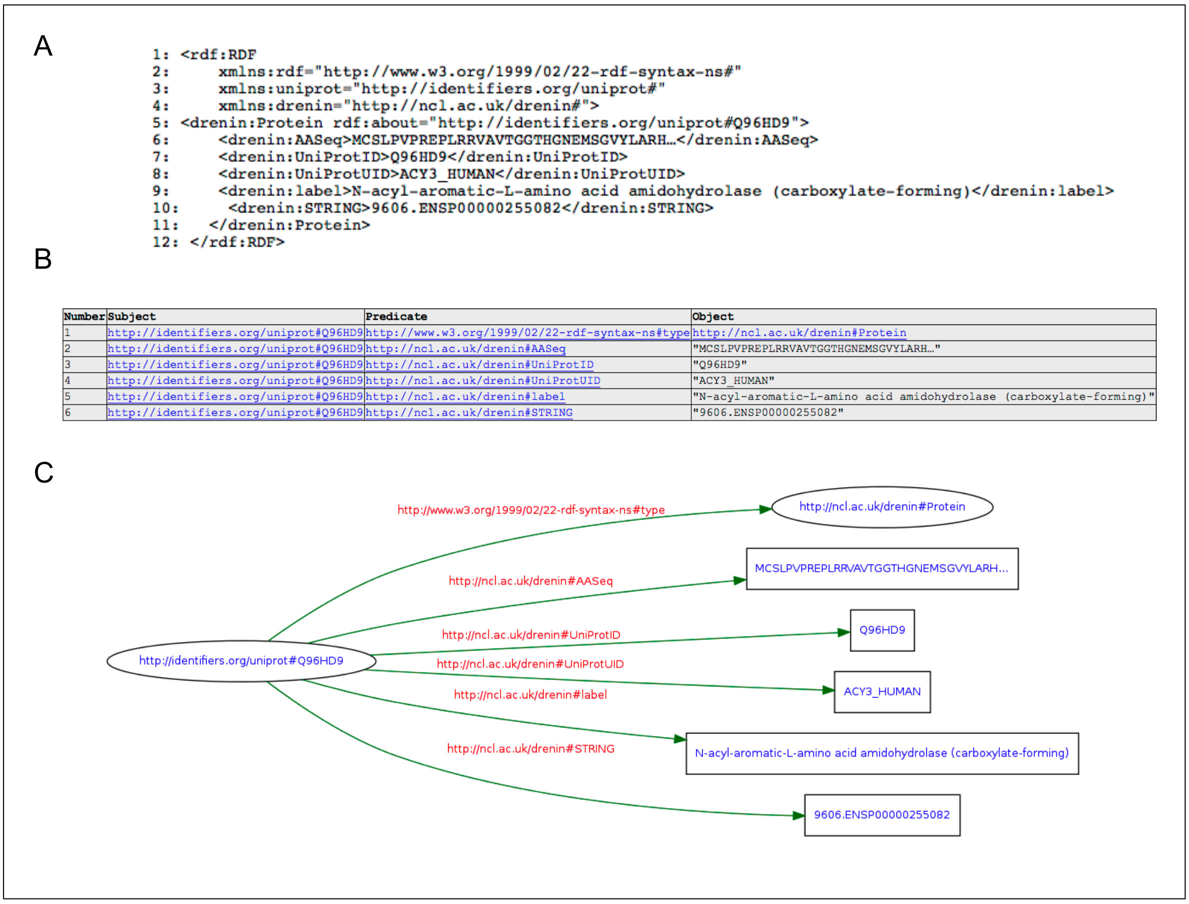
There is no single standardised data format for the storage or exchange of data within the life sciences and a plethora of different data representations are used. RDF, however, is becoming increasingly popular as a common data format for describing, publishing and linking data. For example, several databases and knowledge bases such as UniProt, ChEMBL [108] and Reactome [109] now provide their data in RDF format [110].

RDF A W3C[7] backed data format, RDF, exists as a set of triples, which are statements about entities in a model. Triples take the form of subject-predicate-object. RDF identifies entities using Web identifiers known as Uniform Resource Identifiers (URIs), and describes resources with properties (a resource that has a name, such as "UniProtID") and property values (the value of the property, such as "Q96HD9"). RDF has a graph-based structure that provides a schema-less model. In these directed graphs, nodes represent resources and property values, while the edges represent properties. An RDF document can be serialised for storage and exchange, in several machine-readable formats, such as XML, Turtle and N3.

Fig.3.1A shows how RDF describing the UniProtKB entry 'Q96HD9' is represented in XML. The first line of the RDF document is the XML declaration, followed by the root element `<rdf:RDF>`. The `<xmlns:rdf>`, `<xmlns:uniprot>` and `<xmlns:drenin>` namepsaces specify that elements with the `rdf`, `uniprot` and `drenin` prefixes are from the namespaces `http://www.w3.org/1999/02/22-rdf-syntax-ns#`, `http://identifiers.org/uniprot#` and `http://ncl.ac.uk/drenin#` respectively. The `<drenin:Protein>` element contains the description of the resource identified by the `rdf:about` attribute. The elements `<drenin:AASeq>`, `<drenin:UniProtID>`, `<drenin:UniProtUID>`, `<drenin:label>` and `<drenin:STRING>` are properties of the resource.

As previously mentioned, an RDF document can be serialised in several syntaxes. Many triplestores are available to allow for the efficient storage of RDF using these

[7] www.w3.org



```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:uniprot="http://identifiers.org/uniprot#"
    xmlns:drenin="http://ncl.ac.uk/drenin#">
<drenin:Protein rdf:about="http://identifiers.org/uniprot#Q96HD9">
    <drenin:AASeq>MCSLPVPREPLRRVAVTGGTHGNEMSGVYLARH…</drenin:AASeq>
    <drenin:UniProtID>Q96HD9</drenin:UniProtID>
    <drenin:UniProtUID>ACY3_HUMAN</drenin:UniProtUID>
    <drenin:label>N-acyl-aromatic-L-amino acid amidohydrolase (carboxylate-forming)</drenin:label>
     <drenin:STRING>9606.ENSP00000255082</drenin:STRING>
   </drenin:Protein>
 </rdf:RDF>
```

Number	Subject	Predicate	Object
1	http://identifiers.org/uniprot#Q96HD9	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://ncl.ac.uk/drenin#Protein
2	http://identifiers.org/uniprot#Q96HD9	http://ncl.ac.uk/drenin#AASeq	"MCSLPVPREPLRRVAVTGGTHGNEMSGVYLARH…"
3	http://identifiers.org/uniprot#Q96HD9	http://ncl.ac.uk/drenin#UniProtID	"Q96HD9"
4	http://identifiers.org/uniprot#Q96HD9	http://ncl.ac.uk/drenin#UniProtUID	"ACY3_HUMAN"
5	http://identifiers.org/uniprot#Q96HD9	http://ncl.ac.uk/drenin#label	"N-acyl-aromatic-L-amino acid amidohydrolase (carboxylate-forming)"
6	http://identifiers.org/uniprot#Q96HD9	http://ncl.ac.uk/drenin#STRING	"9606.ENSP00000255082"

C
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://ncl.ac.uk/drenin#Protein
MCSLPVPREPLRRVAVTGGTHGNEMSGVYLARH...
http://ncl.ac.uk/drenin#AASeq
Q96HD9
http://ncl.ac.uk/drenin#UniProtID
http://identifiers.org/uniprot#Q96HD9
http://ncl.ac.uk/drenin#UniProtUID
ACY3_HUMAN
http://ncl.ac.uk/drenin#label
N-acyl-aromatic-L-amino acid amidohydrolase (carboxylate-forming)
http://ncl.ac.uk/drenin#STRING
9606.ENSP00000255082

Figure 3.1: **RDF example.** (A) An example XML serialised RDF document for a single protein, Q96HD9 (B) Triples from A in the form subject-predicate-object (C) Graph representation of RDF in A. Edges show predicates from the protein object to every property value (attribute). *Note:* Images were produced using the W3C RDF validation service[8].

serialised formats. Services are provided for the efficient querying of triplestores, known as endpoints. These endpoints provide a means of querying the RDF using standard query languages, such as SPARQL.

A SPARQL query is a graph pattern used to find matching RDF triples. In these queries the subject, predicate, or object of a triple can be substituted for query variables. Although RDF documents can be effectively queried based on their graph structures, RDF annotation of resources is not expressive enough to facilitate complex queries. Raw RDF annotations are meaningless to a machine. To provide machine-understandable annotations, ontology languages and data formats have been developed.

Ontologies A specification of conceptualization, ontologies provide an abstract or simplified view of a domain of interest [111]. Furthermore, ontologies provide a description of concepts and relationships that can exist for elements within a particular domain [111]. Ontologies enable the organisation of information in a machine interpretable fashion; ontologies are not simply dictionaries prepared for humans. Common components of an ontology include classes (or concepts), individuals, attributes and object properties. Classes represent types of objects within a domain, such as *Disease* where individuals represent instances of classes, for example, *Crohn's Disease* as a *Disease* individual. Attributes describe features of a class or instances. For example, Disease instances, such as Crohn's Disease may have an attribute *definition* containing a string such as *An intestinal disease that involves inflammation located_in intestine.* Object properties, represent relations and describe the way in which classes and individuals can be related to one another. Terms (i.e. classes and instances) within an ontology are structured hierarchically with higher level terms have more general meanings than their lower level counterparts (see Fig.3.2).

Formal languages used to encode ontologies (and describe RDF data) include both the RDF Schema (RDFS) and Web Ontology Language (OWL). The RDFS language defines a vocabulary to describe objects in an RDF data model, extending RDF in order to define special RDF resources, such as `Class` and `subClassOf`. These special resources are used to describe a hierarchy of other RDF resources. It is, however, OWL that is recommended by the W3C Consortium for expressing ontologies. This recommendation is made due to OWL being more expressive and having a far greater vocabulary than RDFS.

3.2.3.3 Ondex

Ondex is an *in-memory* graph tool that combines semantic data integration with graph-based analysis techniques [105]. Designed to both model and integrate (biological) data as networks, Ondex also provides tools to visualise and analyse resulting integrated data sets [105]. More recently, a Web-based implementation of Ondex, Ondex-Web, has been released [112]. In Ondex networks, nodes, termed *concepts* represent the biological entities, such as 'genes' or 'diseases', while edges, termed *rela-*

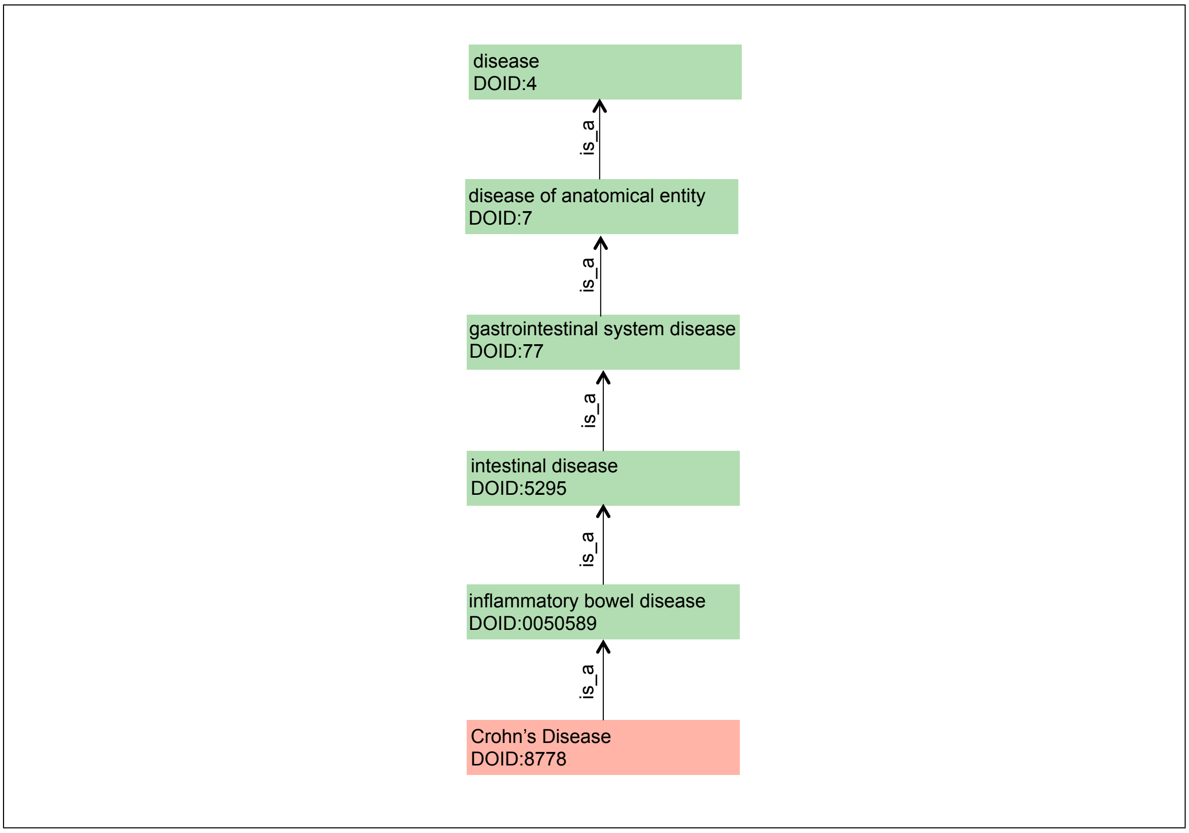


Figure 3.2: **Ontology example.** An example of a hierarchical ontology of terms. The diagram shows the Crohn's disease term and its parents from the Disease Ontology[9]. Classes are shown in green while instances are shown in red.

tions represent the interactions between these entities, such as geneX 'is involved in' diseaseY.

All elements within an Ondex graph (both concepts and relations) are annotated with semantically rich metadata. Each concept is assigned a $c \in C$, where C is a finite set of `concept_classes`. Similarly, each relation is assigned an $r \in R$ where R is a finite set of `relation_types`. The types that nodes are mapped to are captured in a strict fashion, with an XML metadata file describing a hierarchical representation of the biological 'types' that may be included in an Ondex network. This metadata is not an ontology, rather a hierarchical description of terms, and is limited to a handful of types. Although metadata may be extended depending on a particular set of requirements, there is no strict framework for doing so.

Data from external sources are transformed into the Ondex graph-based data model using custom *parsers*. Data is stored in a custom XML format designed for the exchange of Ondex integrated data sets, Ondex Exchange Language (OXL). There is also a framework to enable the development of plugins, allowing for the development of Ondex specific:

1. *Mappers* allow biological concepts to be linked.
2. *Filters* enable the removal of unwanted/ redundant elements from the graph.
3. *Transformers* enable graphs to be either topologically or semantically altered.
4. *Exporters* allow for Ondex graphs to be exported in custom formats, such as comma-separated files, tab-delimited files and the Systems Biology Markup Language.

In Ondex, data integration strategies are explicitly described in custom workflows. These workflows, in turn, make calls to parsers, mappers, transformers, filters and exporters to allow for the production and manipulation of integrated data sets.

Work carried out at Newcastle University led to the creation of an Ondex data set for *in silico* drug discovery [101]. The data set includes `Compound` and `Target concept_classes` from DrugBank[10] [113] and `Protein concept_classes` integrated from UniProtKB[11] [114], as well as information from eleven other databases and analysis methods [101]. The data set contains 150,853 concepts from 17 `concept_classes` which are linked via 787,360 relations from 37 `relation_types`. The data set was successfully used to highlight known examples of repositioned drugs, including chlorpromazine. Chlorpromazine is marketed as a non-sedating tranquillizer but is also known to be effective as an antihistamine. Although there was no relation indicative of a binding between Chlorpromazine and the H1 Histamine receptor in the graph, Chlorpromazine was shown to be similar to Trimeprazine, for which there was a binding to the receptor [101].

Although this data set provides a comprehensive representation of drugs at a systems level, it has not been updated since its production in 2010. The data set was also missing important data types required for it to be used in the identification of novel drug repositioning opportunities. Fig. 2.1 highlights the minimal types of data used in

[10] www.drugbank.ca

[11] www.uniprot.org

drug repositioning prediction tasks and the possible interactions that can be inferred. Using this model, three data types were highlighted as being missing from the original data set:

1. *Disease concepts.* The `Disease concept_class` in the data set was taken from the Online Mendelian Inheritance in Man (OMIM) (see 3.2.4.2), an online catalogue of human genes and genetic disorders.
2. *Target (Gene)-Disease relations.* The `involved_in` associations, which link a `Disease` and `Gene`, were extracted from OMIM. Although OMIM is a reliable source of such data, if the `Disease concept_class` was to be updated it was also necessary to identify a data source capturing associations between genes and the up-to-date disease representation.
3. *Drug-Disease relations.* In order for novel uses of a drug to be identified it is important to capture data detailing the diseases it is currently marketed to treat. In Ondex terms that would require a `has_indication` association between a `Compound` and a `Disease`.

As well as the above data types, the data set included data taken from the Kyoto Encyclopedia of Genes and Genomes (KEGG). KEGG is now a subscription-based system and no longer an open source database. Furthermore, the data set takes all targets from DrugBank, regardless of the species, yet only includes proteins from UniProtKB that are from *Homo sapiens*.

3.2.3.4 Neo4j

Neo4j is an open source NoSQL graph database implemented in Java and is a popular graph Database Management System (DBMS). Unlike Ondex, Neo4j uses a disk-backed storage system, enabling it to store complex and dynamic data (including images, videos, etc.). Like all graph databases, Neo4j stores edges as direct pointers between nodes, which, depending on connectivity, can thus be traversed in constant time. Neo4j makes use of a property graph model meaning that nodes and edges can have properties associated with them. Neo4j provides a varied set of integration possibilities for Java Virtual Machine (JVM)-based languages.

There are two ways of using Neo4j from the JVM. First of all the standalone Neo4j Server can be installed on any machine and then accessed via its HTTP API, using a REST library of choice. Alternatively, Neo4j can be embedded in a JVM process. The second approach is best suited for unit testing, as well as high performance and no-network set-ups.

3.2.4 Relevant data sources

As described in Section 3.2.2, one of the most important aspects to be considered before an exercise in data integration is the identification of relevant and reliable data sources. Many relevant data sets exist that could potentially be used to enable a systems approach to drug repositioning. These data sources can be categorised as being either protein-related, disease-related, drug-related, pathway-related or additional and examples of each are detailed below.

3.2.4.1 Protein-related data sources

UniProtKB UniProtKB[12] [115] is one of the most comprehensive, high quality and freely available resources for proteins and functional information. UniProtKB assigns IDs for each of the `Protein` entries in the database. Each UniProtKB ID consists of 6 or 10 alphanumerical characters (e.g. O95264). As well as primary accessions, a protein entry also contains multiple secondary accessions and maps to many other databases, such as STRING. UniProtKB also details the gene, using the common HGNC_ ID (e.g. 5298) that a given protein is encoded by. The data set is available for download in multiple formats, including XML.

STRING The STRING[13] database is an integrated database that includes known and putative protein-protein interactions, such as binding, gene fusion, co-occurrence

[12] www.uniprot.org

[13] www.string-db.org

and co-expression [116]. Data in STRING are integrated from high-throughput experimental data sets and databases such as KEGG, HPRD[14], Reactome and many others. Available for download in multiple formats, STRING provides confidence scores for each interaction it provides, which is based on the source(s) from which the association is derived.

GO The Gene Ontology (GO)[15] was developed as a means of standardising the annotation of proteins [117]. The GO Project was founded in 1998 to address the challenges of interpreting functional information attached to gene product entries in databases. The GO includes three ontologies that describe: the molecular function that a gene product normally carries out; the biological process that gene products are involved in; and the subcellular locations (or cellular component) that gene products are located in [118]. To make full use of the ontology, it is important to have a means of linking these to the relevant gene products, such as proteins.

Gene Ontology Annotation Database (GOA) The Gene Ontology Annotation (GOA)[16] provides algorithmic and manual annotations of GO terms to UniProtKB protein entries [119]. Annotations are provided from the GOA project (based at the European Bioinformatics Institute (EBI)) and are collated with those from many external databases to provide an extensive publicly available GO annotation resource. From the GOA, GO terms can be linked to a protein.

3.2.4.2 Disease-related data sources

MeSH ® MeSH [120][17] is a controlled vocabulary thesaurus developed by the U.S. National Library of Medicine (NLM). MeSH consists of sets of naming descriptors in a hierarchical structure, thus permitting searching at various levels of specificity. The vocabulary is made up of 'descriptor' records that contain a heading and multiple entry terms. For example, the descriptor 'Ascorbic Acid' has the entry term 'Vitamin C'. In addition to these headings are the Supplementary Concept Records (SCR), supplied

[14] www.hprd.org

[15] www.geneontology.org

[16] www.ebi.ac.uk/GOA

[17] www.nlm.nih.gov/mesh

in a separate file. These SCRs contain specific examples of headings and are updated more regularly than the 'descriptors'. Each SCR is assigned to a related descriptor via the heading. To identify which of the MeSH concepts describe common diseases and which are rare diseases, it is important to determine a good representation of rare disease and to identify the crossover of these with diseases captured in the MeSH hierarchy.

ORDO The ORDO[18], a joint venture between EBI and Orphanet. This data set provides a structured vocabulary for rare diseases and also captures relationships between these diseases, genes and other relevant features. ORDO is derived from the Orphanet database.

Having identified representations of common diseases and rare diseases, it is important to find sources linking these to genes involved in these diseases. Many different data sources that include these associations, derived from differing approaches, such as those:

1. *Manually curated* from the literature.
2. Automatically extracted from the literature via *text mining* methods.
3. Taken directly from *genetic studies.*
4. *Inferred* from mammalian models.

Many sources must be considered to create an integrated view of G-D associations that span all four of the above approaches.

OMIM ® and Orphanet OMIM ®[19] and Orphanet are examples of data sources that contain manually curated G-D associations. OMIM is one of the original sources of G-D associations and provides a catalogue of human genes, genetic disorders and traits. OMIM particularly focusses on the molecular relationship between genetic variation and phenotypic expression [121]. Updates to the OMIM data set are done

[18] www.orphadata.org

[19] www.omim.org

manually by biocurators. The Orphanet database, on the other hand, is a multilingual database dedicated to rare diseases. Populated from the available literature Orphanet is validated by international experts. Orphanet includes a classification of rare diseases, G-D relations and connections/ mappings with other terminologies and also includes a proportion of associations taken from OMIM.

GWAS Catalogue The National Human Genome Research Institute (NHGRI) Catalogue of Genome-Wide Association Studies (GWAS)[20] contains G-D associations taken directly from genetic studies. The catalogue provides a publicly available set of manually curated, published GWAS assays [122]. All Single-Nucleotide Polymorphism (SNP) data and SNP-trait associations with a $p \leq 1 \times 10^{-5}$ are provided [123].

BeFree and SemRep BeFree[21] [124] and SemRep[22] [125] extract G-D associations from the literature using automated text mining approaches. This automated analysis of the literature essentially allows for the access of information that would otherwise be 'locked' in a plethora of free-text documents. BeFree, along with supporting statements and provenance, is available for download and uses the EU-ADT and GAD corpora to extract associations from the text. It is worth noting that any associations documented in the full text or supplementary of articles will be missed by BeFree. SemRep differs from BeFree, as it has been designed to identify a broad range of semantic predictions that take into account the hierarchical relationships between concepts. SemRep allows for relations between the same entity types as BeFree to be extracted from the literature (G-D, drug-disease and drug-target). When using the same corpus as BeFree, SemRep has a higher precision but a lower recall [124].

MGD and RGD The Mouse Genome Database (MGD)[23] [126] and the Rat Genome Database (RGD)[24] contain G-D associations that have been identified in animal models but are statistically inferred to represent human associations.

[20] www.ebi.ac.uk/gwas

[21] www.ibi.imim.es/befree

[22] www.semrep.nlm.nih.gov

[23] www.informatics.jax.org

[24] www.rgd.mcw.edu

ClinicalTrials.gov ClinicalTrials.gov[25] is a database of publicly and privately supported studies of human participants from all around the world. Law (section 801 of the Food and Drug Administration (FDA) Administration Amendments Act in the U.S[26] and the Clinical Trials Directive 2001/20/EC, Article 11[27] in the EU) requires certain clinical trials to submit their trials and results to a publicly accessible database. ClinicalTrials.gov, therefore, provides access to these trials, meaning the novel uses of drugs that are currently being investigated can be determined. From this data source the clinical trials, and the associated drugs and diseases are captured.

3.2.4.3 Drug-related data sources

DrugBank DrugBank[28] is an online database containing biochemical and pharmacological information about drugs, their mechanisms and their targets [127]. The database is maintained and enhanced by extensive literature mining performed by domain-specific experts and biocurators. Most of the data in DrugBank are curated from primary literature sources, and it is thus used as the primary drug data source for many databases (e.g. PharmGKB, PDB, PubChem, KEGG). DrugBank classes drugs as either small molecules or bio tech drugs, while also containing many binds to associations between drugs and target proteins. Although DrugBank offers a rich data set, it does not capture any activity values that may be associated to the binds to associations.

ChEMBL ChEMBL[29] is a large-scale bioactivity database containing data mainly manually curated from the medicinal chemistry literature [128]. ChEMBL contains data regarding compounds, targets, and the biological or physicochemical assays from which the data was extracted. Unlike DrugBank, ChEMBL captures activity values that may be associated to the compound-target binds to associations.

[25] www.clinicaltrials.gov

[26] www.gpo.gov

[27] www.eortc.be

[28] www.drugbank.ca

[29] www.ebi.ac.uk/chembl

SIDER Side-Effect Resource (SIDER)[30] is a resource containing drugs and their adverse drug reactions (ADR)— containing both side-effects and known indications [82] [129].

NDF-RT The National Drug File - Reference Terminology (NDF-RT)[31] is produced by the U.S. Department of Veterans Affairs, Veterans Health Administration (VHA). NDF-RT is an extension of the VHA National Drug File, and organises the drug list into a formal representation and is updated monthly. From here, known indications can be extracted via a RESTful API. The NDF-RT representation contains 18 semantic types (including 'Clinical Drug' and 'Disease or Syndrome'), 20 attributes and 44 relation types. For example, two of the relation types included in NDF-RT are 'may_treat' and 'may_prevent', linking 'Clinical Drug' to 'Disease or Syndrome'. NDF-RT also provides a hierarchical representation of the 'Disease or Syndrome' terms used in the data representation, linked via `has_parent` and `has_child` relations.

3.2.4.4 Pathway-related data sources

Open PHACTS One collaborative project that is dedicated to data integration within the field of pharmacology is Open PHACTS [130]. A linked data platform for integrating multiple pharmacology data sets, Open PHACTS forms the basis for several drug discovery applications. Driven by real business questions gathered from the projects partners, the project provides an Open Pharmacological Space (OPS). The OPS integrates data from a total of 9 distributed open pharmacological and biomedical databases and is accessible via a RESTful API.

The OPS platform builds on previous work delivered by the Semantic Web community in creating Resource Description Framework (RDF)-based data sources. In recent years, several key data sets for drug discovery have been published in Semantic Web formats including those provided by Chem2Bio2RDF and Linking Open Drug Data (LODD). These two data sets, as well as the Bio2RDF, Neurocommons and Linked Life Data, have all made important sets of biology and chemistry data available in RDF [130].

[30] www.sideeffects.embl.de

[31] www.nlm.nih.gov

The Open PHACTS API wraps a number of 'canned' SPARQL queries that applications can call. As a result, application developers do not need to formulate their own SPARQL queries for many commonly-used operations [130].

ConceptWiki, included in Open PHACTS, offers a more thorough mapping of elements to alternative accessions and is far beyond what could have been achieved during this project. However, at the time of this work, there are relatively few mappings for genes in this data source, while some that have been mapped appear to have been done so incorrectly (see Fig. A.1).

3.2.4.5 Additional data sources

CTD The Comparative Toxicogenomics Database (CTD)[32] is a publicly available database whose initial aim was to annotate the response of genes and proteins from diverse species to various toxic agents [131]. The scope of the database has somewhat expanded since its inauguration and at present captures data involving chemical-disease, chemical-gene and G-D interactions, as well as pathway data. Interactions are curated from the scientific literature by professional biocurators who are aided in the task by controlled vocabularies, ontologies and structured notation [131].

3.3 Materials and methods

All Ondex development was done using the compiled source code from v0.5.0[33]. For updating the original data set a bespoke Ondex parser was developed for the integration of disease concept types from NDF-RT. The parser made calls to the NDF-RT RESTful API[34]. First, all DISEASE_KIND concepts were extracted using the '/allconcepts' REST resource which returned the Alphanumeric Unique Identifier (NUI) identifier for each DISEASE in NDF-RT. Next, for every disease NUI, the '/allInfo' REST resource was used to retrieve all information associated to that concept, including the parent and child DISEASES (Table 3.1).

[32] www.ctdbase.org

[33] www.ondex.org

[34] www.rxnav.nlm.nih.gov Accessed: 22-09-2013

Relations of type may_treat and may_prevent were integrated in a two-step process. First, the relations were extracted from NDF-RT using calls to their RESTful API. All DRUG_KIND concepts were extracted using the '/allconcepts' REST resource which returned the NUI identifier for each drug. To ensure that all drugs searched for could be mapped to the drugs captured in the data set, an external mapping between NUI and DrugBank accessions was performed using the drugs mapping file available from PharmGKB[35]. For every NDF-RT drug that mapped to a DrugBank compound, a call to the '/allInfo' REST resource was used to retrieve all concept information, including the may_treat and may_prevent relations. A mapper was then developed to integrate these relations using the relation_types may_treat and may_prevent, using the Disease NUI and the DrugBank accessions (Table 3.1). DisGeNET G-D associations that contained a UniProtKB ID corresponding to a known Target or Protein in the Ondex network were then extracted. A bespoke Ondex mapper was written to integrate the involved_in associations, using the UniProtKB IDs to map to the outgoing Protein or Target nodes, while the incoming Indication nodes were identified using the Unified Medical Language System (UMLS) ID present on diseases in DisGeNET (Table 3.1).

Table 3.1: Data sources used to extend *Dat*.

Type	Name	Number	Source	From-To
concept_class	Indication	4,463	NDF-RT[36]	-
relation_type	has_parent	6,553	NDF-RT[36]	Indication-Indication
relation_type	has_child	2,018	NDF-RT[36]	Indication-Indication
relation_type	may_treat	3,744	NDF-RT[36]	Compound-Indication
relation_type	may_prevent	343	NDF-RT[36]	Compound-Indication
relation_type	involved_in	16,098	DisGeNET[37]	Target-Indication

For Ondex and Neo4j performance testing, the command line accessible ondex-mini (snapshot 0.5.0) was used. For convenience (and to minimise the risk of exposing potentially sensitive data), the embedded approach of Neo4j was used throughout the work described in this Chapter. Neo4j Core-java-API[38] v2.1.2 was used. For Ondex, a single parser was developed that took graph type and graph size arguments and

[35] www.pharmgkb.org Accessed: 22-09-2013

[36] www.rxnav.nlm.nih.gov Accessed: 22-09-2013

[37] www.disgenet.org Accessed: 22-09-2013

[38] www.neo4j.com

ran from the command line. Similarly, for Neo4j a single parser was developed, with transactions limited to 10,000 nodes. All performance tests were single-threaded and ran on an 8GB RAM Mac (1.8 GHz Intel Core i5) with an allocated heap size of 6GB with five repeats.

For DReNInF, the DReNInO was created in OWL using the open-source ontology editor, Protege[39] v4.3. Parsers and mappers were developed as sub-components of a BioSSIP maven archetype. Written in Java, parsers and mappers enabled data sources to be converted from their accessible format to BioSSIP nodes and edges. Finally, the data integration strategy was also implemented in Java.

An RDF exporter for BioSSIP was implemented in Java. The RDF exporter made use of the Apache Jena RDF API[40] v2.11.2. To store and query the DReNIn RDF data set, the triplestore Sesame[41] v2.8.6 was used. An Apache Tomcat server[42] v8.0.26 was used to allow remote access to the stored data and the user interface. Tomcat was installed on a Ubuntu v15.04 virtual machine and the openrdf-sesame.war and openrdf-workbench.war copied to the webapps sub-folder of the Tomcat installation.

A Web-based user interface was developed for the RDF data set. The interface was implemented in JSP (java server pages) and is a Web site made up of five main pages. The project makes use of Maven and the WAR (web application archive) architecture. The compiled WAR file was added to the webapps/ROOT sub-folder of the Tomcat installation.

3.4 Results

An update of the original Ondex data set described by Cockell *et al.* [101] was carried out as part of this work. A comparison between Ondex and Neo4j, regarding performance, is presented— highlighting Neo4j as a more scalable platform for data set development. As Neo4j has no innate means of controlling semantics, BioSSIP was also developed during this project. BioSSIP is a light-weight module that sits in front of

[39] www.protege.stanford.edu
[40] www.jena.apache.org
[41] www.rdf4j.org
[42] www.tomcat.apache.org

a datastore 'backend', such as Neo4j, and allows for integration projects to be created independently of the users datastore of choice. A framework for data integration in the drug repositioning setting (DReNInF) is then outlined. This framework makes use of BioSSIP and is made up of three components, comprising: (i) DReNInO, a high-level drug repositioning ontology (ii) A suite of more than 20 parsers for drug repositioning relevant data sources and (iii) A data integration strategy. Finally, an RDF exposed data set, DReNIn, is characterised.

3.4.1 Updating the original Ondex data set

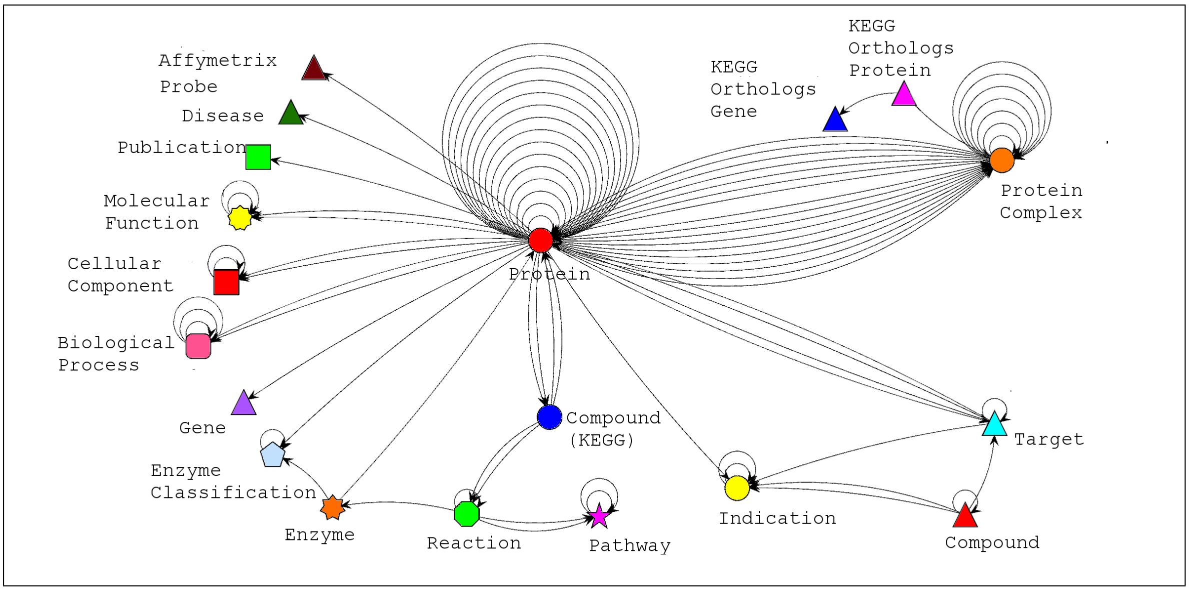


Figure 3.3: **Metagraph of *Dat*, the Ondex drug discovery data set.** Metagraph of *Dat*, the updated data set for *in silico* drug discovery. Nodes represent the `concept classes` and edges the `relation types` included. *Note:* due to space, the `relation types` are not labelled above and are instead listed in Table A.3.

4,463 NDF-RT 'Disease or Syndrome' concepts were integrated into the initial data set using the `concept_class Indication` so as to differentiate between the diseases from NDF-RT and those from OMIM. Therefore, the updated data set, *Dat*, had an additional 4,463 vertices (totalling 155,316) made up of one (and a total of 19) `concept_classes` (see Table A.1) in comparison to the original. Furthermore, *Dat* has an additional 28,736 edges (total of 816,096), representing an additional five (total of 42 `relation_types`) (see Table A.2). Fig 3.4 captures the neighbourhood of chlorpromazine in *Dat* after the update. *Dat* shows a high degree of connectivity with a $d_S(G)$ (average node degree) of 10.42, whereby degrees of vertices range from $\delta(G)$ (minimum degree) of 1 and $\Delta(G)$ (highest degree) of 15,004. Average Connectivity differs between each `concept_class`, with `Protein` displaying the highest $d_S(G)$ of any `concept_class` at $\sim$ 45. Other notable connectivity averages include `Target` $\sim$ 13, `Compound` $\sim$ 7 and `Disease` $\sim$ 4.

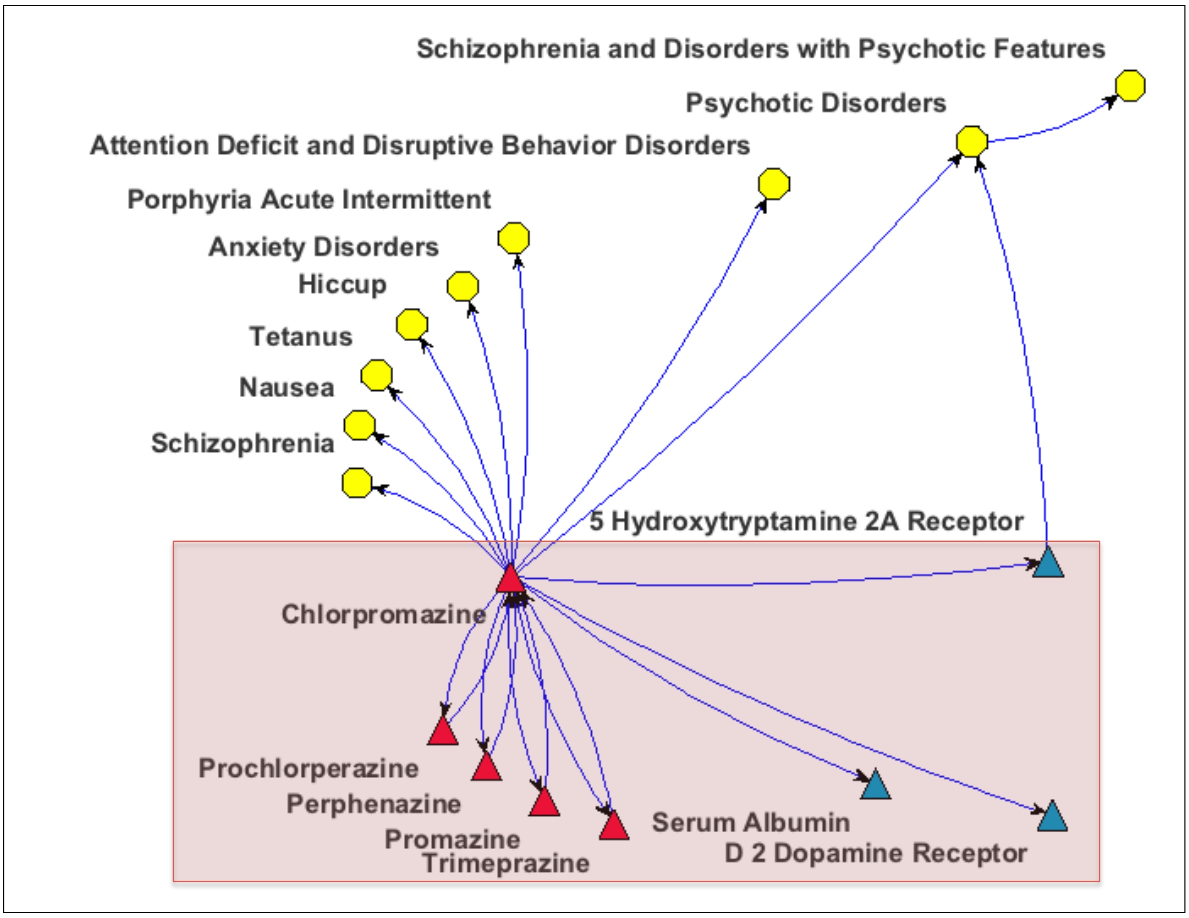


Figure 3.4: **Chlorpromazine neighbourhood as captured in *Dat*.** Nearest neighbours of chlorpromazine from the updated data set, *Dat*. Already included are the symmetrical `similar_to` associations between trimeprazine, prochlorperazine, perphenazine and promazine as well as three `binds_to` relations to Serum Albumin, D-2 Dopamine Receptor and 5HT2A Receptor (captured in the red box). Nine indications are included after the update, linked to chlorpromazine via `may_treat` relations including Nausea, Tetanus, ADHD, Hiccup, Anxiety Disorders, Schizophrenia and Psychotic disorders. Also present is an `involved_in` association between the 5HT2A receptor and Psychotic disorders and the `has_parent` relationship from Psychotic disorders to Schizophrenia and Disorders with Psychotic Features. *Note*: yellow = `Indication`, blue = `Target` and red = `Compound`.

3.4.2 Ondex V Neo4j performance testing

Performance tests were completed on linear as well as more complex (with a higher connectivity) graphs and are shown below in Fig 3.5. In linear graphs, every $v \in G|V|$ is connected by a relation to the previously created v. In the more complex graphs, every $v \in G|V|$ is connected to the previously created $v \in G|V|$ and one other random $v \in G|V|$.

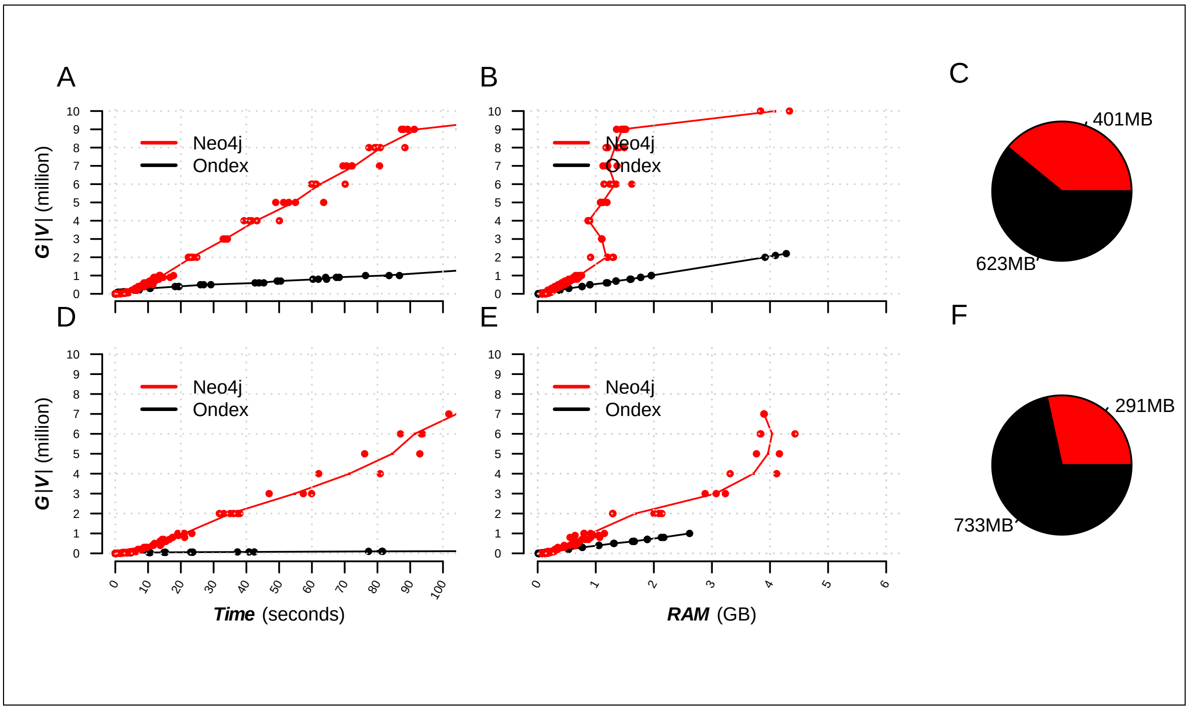


Figure 3.5: **Ondex V Neo4j build performance.** Comparison of Ondex and Neo4J build statistics. (A), (B), (C) Data from linear graph builds ($v \in G|V|$ is connected by a relation to the previously created v). (D), (E), (F) Data from more complex graph builds ($v \in G|V|$ is connected to the previous $v \in G|V|$ and one other random $v \in G|V|$). (A) and (D) Time taken to complete a graph build. (B) and (E) How much of the allocated RAM was used as an average during the build. (C) and (F) Show, as a proportion of a GB what size the graphs were when stored to disk. Graph storage was linear for both platforms, and so a single value is used. *Note*: All performance tests were single-threaded, ran on a 8GB RAM Mac, with an allocated heap of 6GB. At least 5 replicates per $G|V|$ with Neo4j transactions limited to 10,000 nodes.

3.4.3 BioSSIP: a light-weight semantic handling module

BioSSIP is a light-weight module that sits in front of but is completely separate to, a datastore 'backend' of choice. The module provides a framework for parsers, data integration strategies and mining strategies to be developed independently of the storage system. Central to BioSSIP is a user provided application-specific ontology. This ontology has two main purposes: (1) to manage the semantics of the data included in BioSSIP integrated data sets and (2) as a means of expanding queries used to explore the resulting integrated data. BioSSIP can be used for both integration and subsequent querying, even if the datastore of choice does not have an innate means of implementing and defining these semantics. Parsers, exporters, integration strategies and mining strategies can be developed in the BioSSIP module, packaged, and used to run the same analysis pipeline using differing data storage solutions.

BioSSIP was a collaboration with my colleague Matthew Collison (MC) at Newcastle University. Although the majority of the implementation was carried out by MC, this project contributed ideas and discussions and contributed to the implementation of metadata classes and exporters. BioSSIP is an open source project developed in Java using Maven, and the overall architecture of the system is presented in Fig. 3.6.

For data integration using BioSSIP, a `ParserInterface` is defined, which defines three basic methods; `parseFile`, `setHandler` and `setFilePath`. During a process of integration, multiple parsers will be created, this collection of application-specific parsers is referred to as a ‘suite’ of parsers. An ‘integration strategy’ defines the order in which parsers are to be executed in the main class of a BioSSIP project. Parsers are written to convert data of any format to BioSSIP java ‘beans’. In BioSSIP, a `MetaDataInterface` is used to define these java beans, whose instances essentially represent elements (nodes and edges) of the integrated graph, defined in the `SSIPNode` and `SSIPRelation` classes. These BioSSIP Java beans are then converted to the specific datastore format and stored via the data access layer. The `BackendInterface` defines methods to be used when creating a data access layer. The interface includes methods such as `initialiseDatabaseConnection`, `addNode`, `returnAllNodes` and `finaliseDatabaseConnection`.

Once an integrated data set has been produced, ‘step descriptions’ are used as the basic building block to define queries in BioSSIP. These step descriptions contain the information to allow for a simple graph traversal during a query, such as a depth-first traversal of `has_neighbour` relation types at a depth of one. Step descriptions make

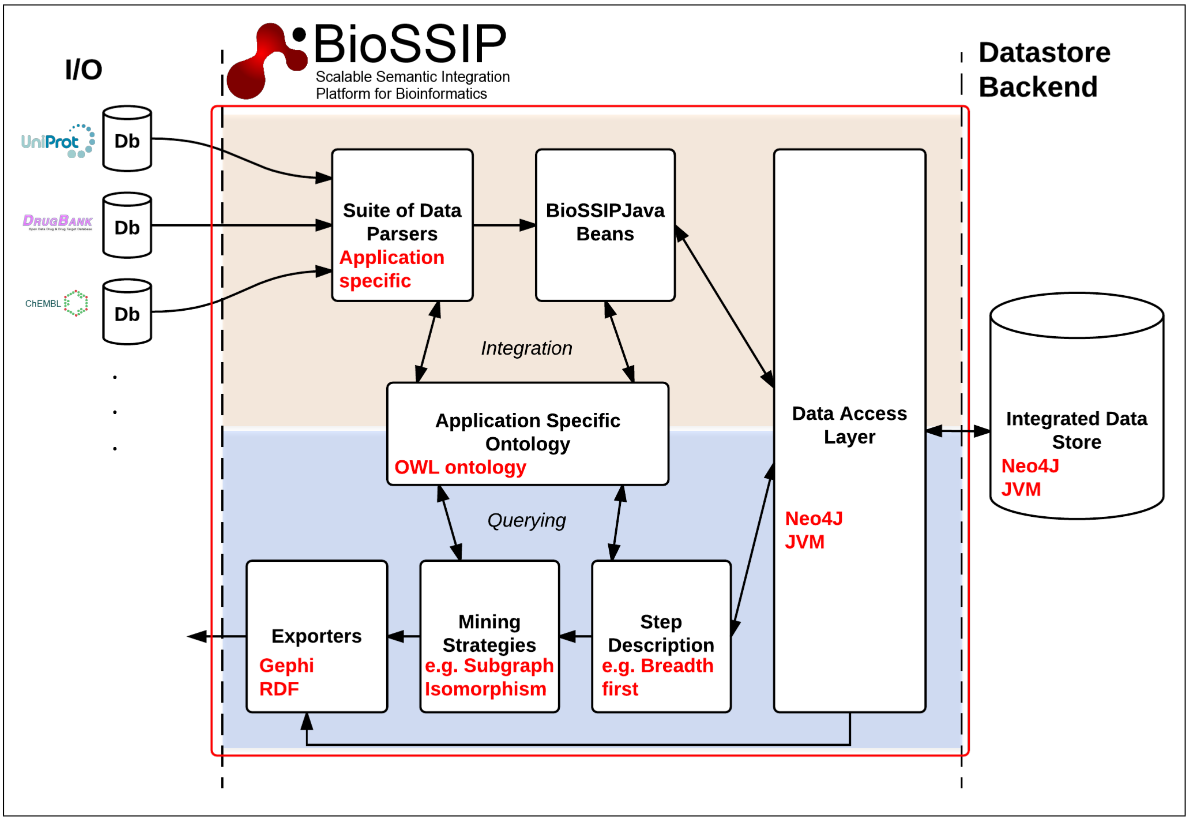


Figure 3.6: **BioSSIP architecture.** BioSSIP architecture is captured inside the central red box. 'I/O' refers to data from external data sources, such as UniProt, DrugBank, ChEMBL, and many more. The 'Datastore Backend' is interchangeable, however, currently, there is a single Neo4j implementation. BioSSIP provides a framework for parser development and allows for the creation of an application-specific suite of parsers. Data, nodes and relations, are represented as Java Beans, where types are dictated by an ontological representation of semantics. A data access layer allows for these Java beans to be converted to the specific datatore format. BioSSIP also provides graph traversal rules, accessing data via the data access layer. These traversal rules can be adjoined to allow for the design of more complex mining strategies, and can also use the high-level ontology for tasks such as 'query' expansion. Finally, multiple exporters are available which allow for part of, or indeed the entirety of the integrated network to be analysed or stored in alternative data formats that can be visualised. The orange box highlights the elements of BioSSIP concerned with data integration and the blue box highlights the components of BioSSIP concerned with querying and exporting.

use of the data access layer to convert the BioSSIP query into one that is recognised by the datastore 'backend'. Furthermore, these step descriptions may be used to create more complex 'mining strategies' or algorithms, such as a subgraph isomorphism algorithm. Finally, exporters have been implemented for integrated data sets to be exported from BioSSIP to Gephi and RDF, with the former allowing for visualisation of the integrated network or indeed the results of a query.

3.4.4 DReNInF: a data integration framework

DReNInF enables the production of integrated data sets relevant to drug repositioning. The project is open source and is made up of three main components:

1. DReNInO: A high-level drug repositioning ontology (3.4.4.1).
2. A suite of parsers to transform data sources relevant to the data types captured in DReNInO into BioSSIP.
3. A data integration strategy (3.4.4.3).

The domain (or more specifically, application) specific ontology DReNInO captures the semantics of both the nodes and the edges to be included in any data sets created using the framework. A suite of parsers convert data from external sources to a format recognisable by the integration platform BioSSIP. With over 20 parsers developed for the inclusion of multiple data types a data integration strategy defines which parsers will be called and when. Ultimately this framework allows for the production of drug repositioning data sets with a strict semantics. Each of the three most important components of DReNInF will now be introduced in more detail.

3.4.4.1 DReNInO: a high-level drug repositioning ontology

First of all, semantic data types that were to be included in DReNInF and the types of interactions that would capture relevant associations between these were identified. These terms were used to create a metagraph depicted in Fig. 3.7. The design of the metagraph was an iterative process and took inspiration from the original Ondex data set [101], the metadata captured in Ondex [105] and considered the research questions that were to be asked of the data set. For the sake of consistency, the same notation used by Ondex to discuss semantic types (3.2.3.3) is used here, as such the metagraph can be said to contain 15 `concept_classes` and 18 `relation_types`.

Data in the metagraph was then formalised in the high-level DReNInO. DReNInO is made of 25 classes (defining `concept_classes`) and 18 object classes (defining `relation_types`), as shown in Fig 3.8. The requirements for the ontology were not

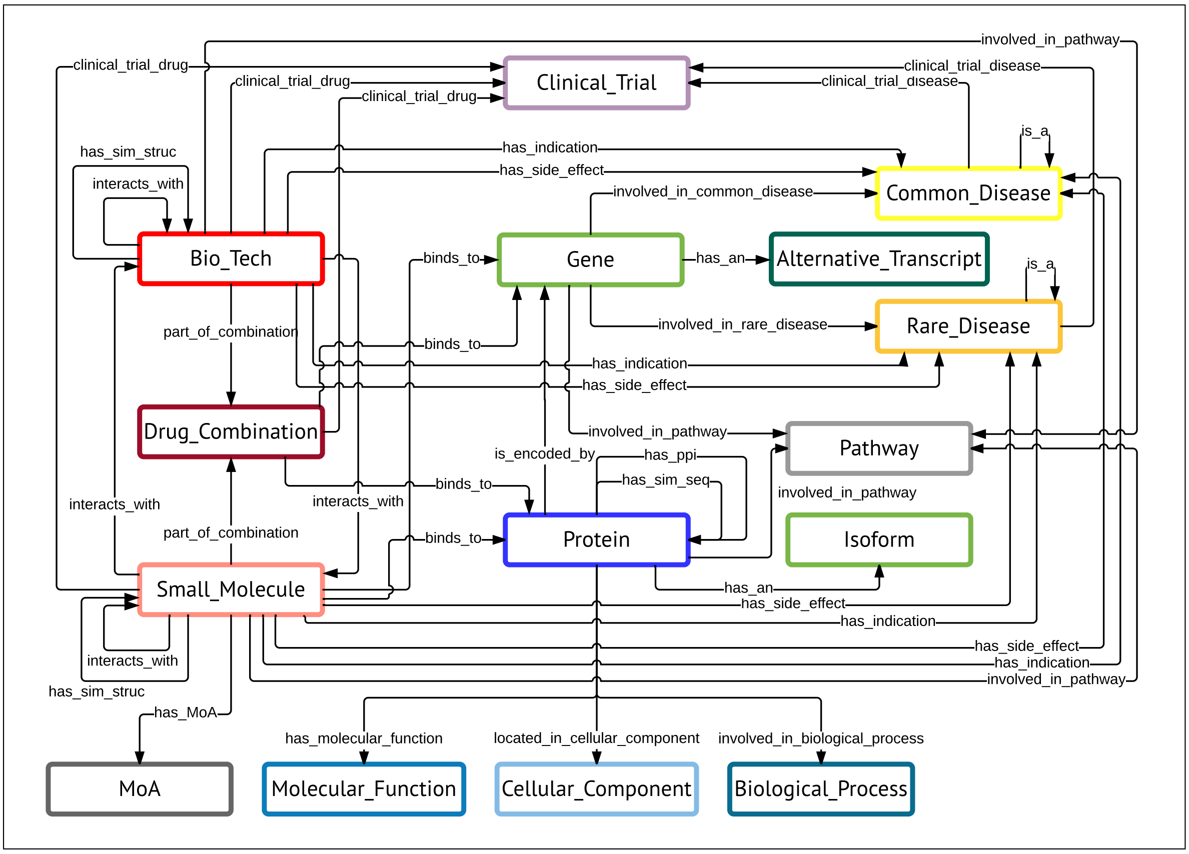
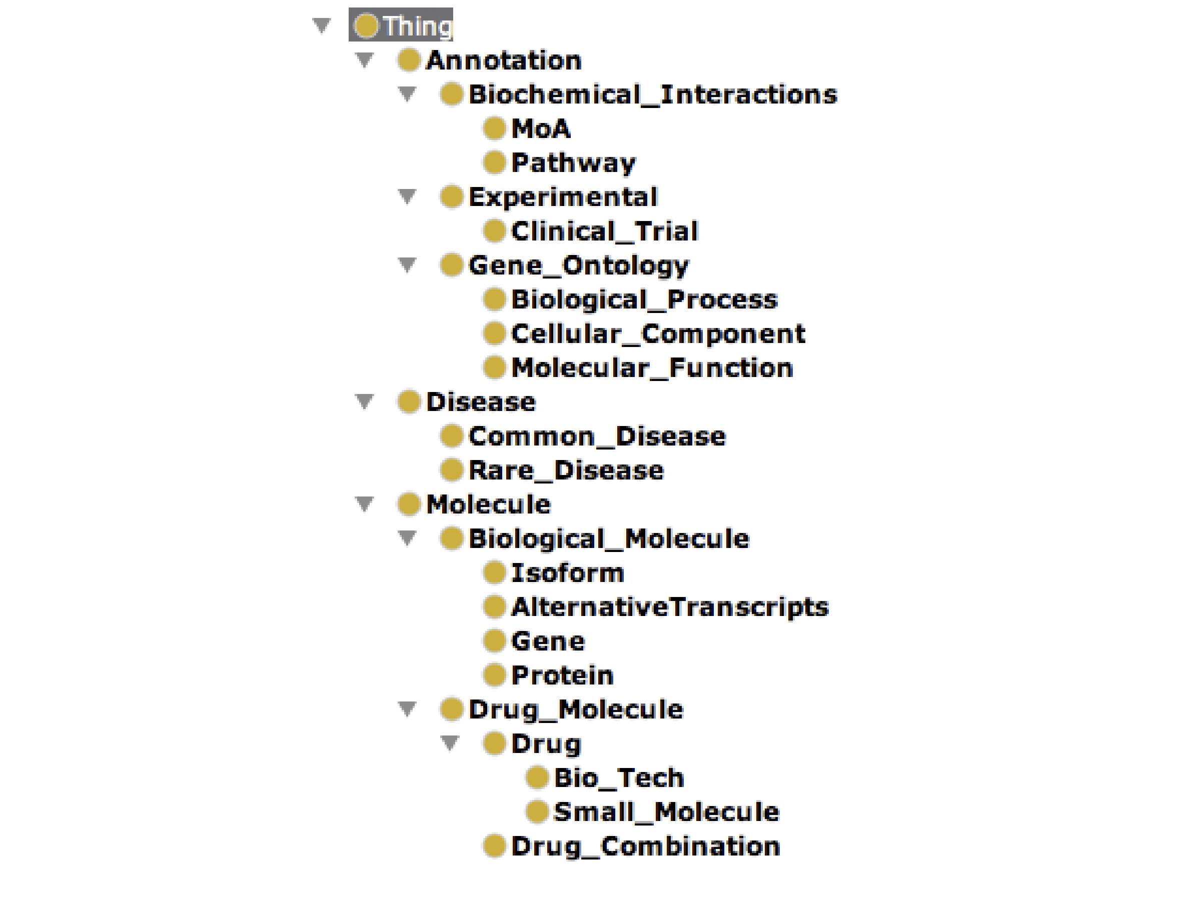


Figure 3.7: **DReNInF metagraph.** Metagraph includes the 15 Concept classes and 18 relation types included in DReNInF. Boxes represent the types of biological concepts, for example, `Rare_Disease` represents the type rare diseases in the data set. Directional edges represent the type of interactions between biological and pharmacological concepts, for example, the `is_encoded_by` association between `Protein` and `Gene` describes a situation whereby the protein (from node) that is encoded by a gene (to node).

necessarily to correctly capture all data types to be included in a strict ontological fashion, but to reflect the data types included in the central dogma of drug action (drug-target-disease), shown in Fig. 2.1. The central dogma of drug action is an abstract representation of the drug-disease connection and drug repositioning approaches aim to 'fill in the blanks' between these data types (e.g. identify drug-target associations or drug-disease associations). It, therefore, makes sense to utilise this representation when aggregating data relevant to the drug repositioning setting and it is for this reason that DReNInO makes clear distinctions between drugs (`Drug_Molecule`), targets (`Biological_Molecule`) and diseases (`Disease`).

DReNInO allows for queries involving a parent type, such as `Disease` to include its child terms, `Common_Disease` and `Rare_Disease` (see Fig. A.2A for illustration).



- Thing
  - Annotation
    - Biochemical_Interactions
      - MoA
      - Pathway
    - Experimental
      - Clinical_Trial
    - Gene_Ontology
      - Biological_Process
      - Cellular_Component
      - Molecular_Function
  - Disease
    - Common_Disease
    - Rare_Disease
  - Molecule
    - Biological_Molecule
      - Isoform
      - AlternativeTranscripts
      - Gene
      - Protein
    - Drug_Molecule
      - Drug
        - Bio_Tech
        - Small_Molecule
      - Drug_Combination

Figure 3.8: **Classes included in DReNInO.** The high-level ontology, DReNInO contains 25 classes, representing the `concept_classes` to be included in DReNInF, with `relation_types` captured as 18 object properties.

Object properties capture `Relation_Types` to be included. For example, there are two associations between `Biological_Molecule` and `Disease` in DReNInO— `involved_in_common_disease` and `involved_in_rare_disease`, and these are captured as object properties (see Fig. A.2B for illustration).

3.4.4.2 Suite of parsers

21 parsers were developed for external data sources described in Section 3.2.4, four mappers for the extraneous mapping sources listed in Table. 3.2 and two mappers for the internal mapping sources[43] (used to map OMIM terms to MeSH terms and GWAS traits to MeSH terms, see Table. 3.2). To produce heterogeneous data sources using the semantics captured in DReNInO and the suite of parsers, a data integration strategy was next defined.

[43] kindly provided by GSK

3.4.4.3 Data integration strategy

Similar in purpose to an Ondex workflow, BioSSIP integration strategies define which parsers are to be used and in what order. The data integration strategy used in DReNInF is shown in Fig 3.9. A characteristic of bioinformatics data sets is the existence of various types of 'standard' identifiers. The data sources that were primarily integrated (as described in previous sections) often use different accession IDs. To effectively integrate these data sets, a set of additional data sources were used as ID-ID mapping resources (Table. 3.2).

Table 3.2: Mapping sources used in DReNIn.

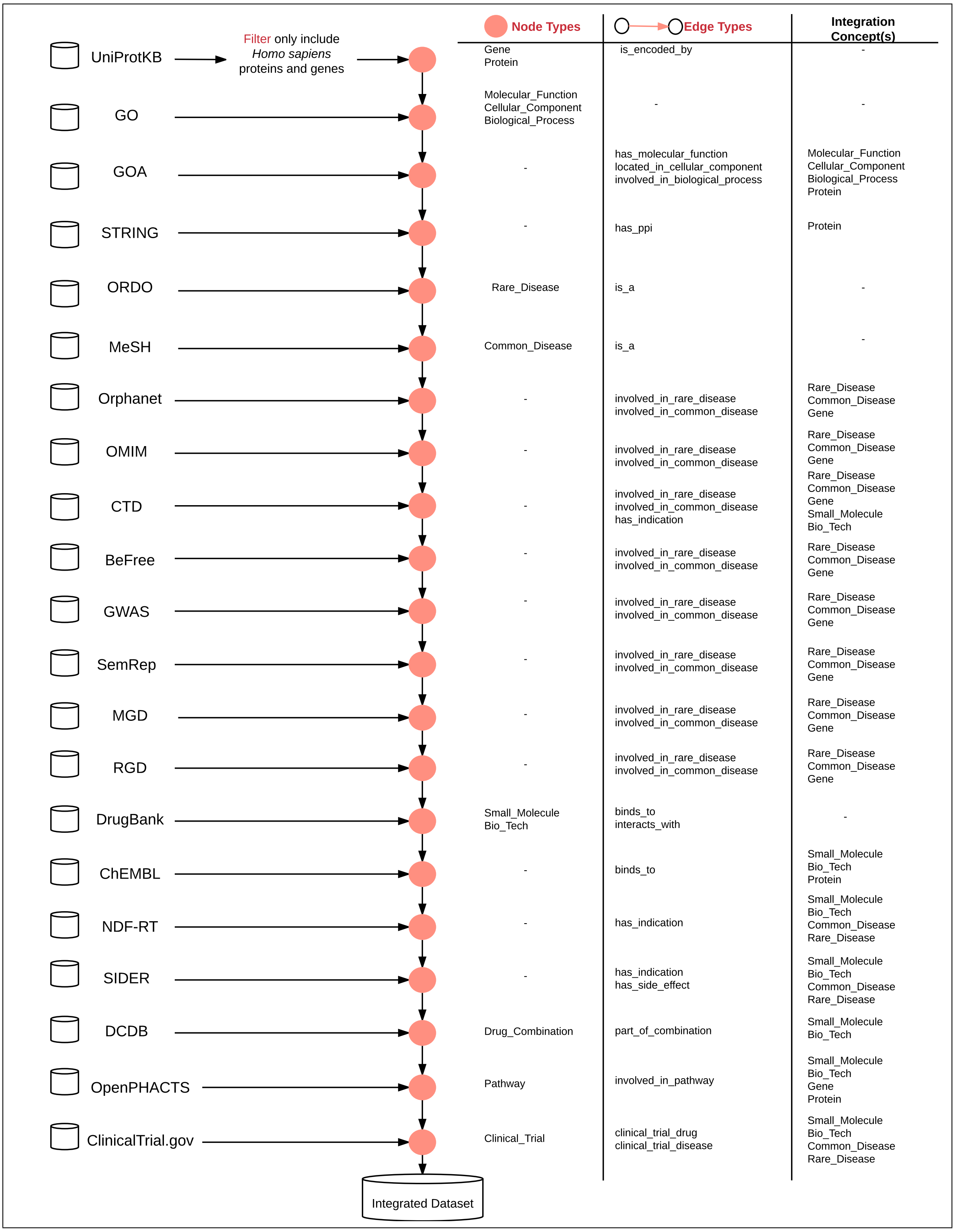
Source	Version/Accessed	Source ID from	Source ID to
Metathesaurus ®[132]		UMLS ®	MeSH ®
UniChem[44] [133]	26-08-15	DBID	ChEMBL
PharmGKB[45] [134]	05-08-2015	NDF-RT (NUI)	DBID
ORDO [135]	31-07-2015	OMIM	MeSH
in-house data	-	OMIM	MeSH
in-house data	-	GWAS trait	MeSH

3.4.5 DReNIn: an RDF exposed data set

An integrated data set was then created using DReNInF, creating a Neo4j graph containing 466,540 nodes, connected via 2,688,436 relations. In Neo4j, the data set uses 3.26 GB of disk space and takes 100 minutes to build. In order to provide a simple access point to the community, an RDF version of the graph was created using the RDF exporter. In RDF, the integrated data set is made up of over 8.5 million triples and is 660 MB in disk size.

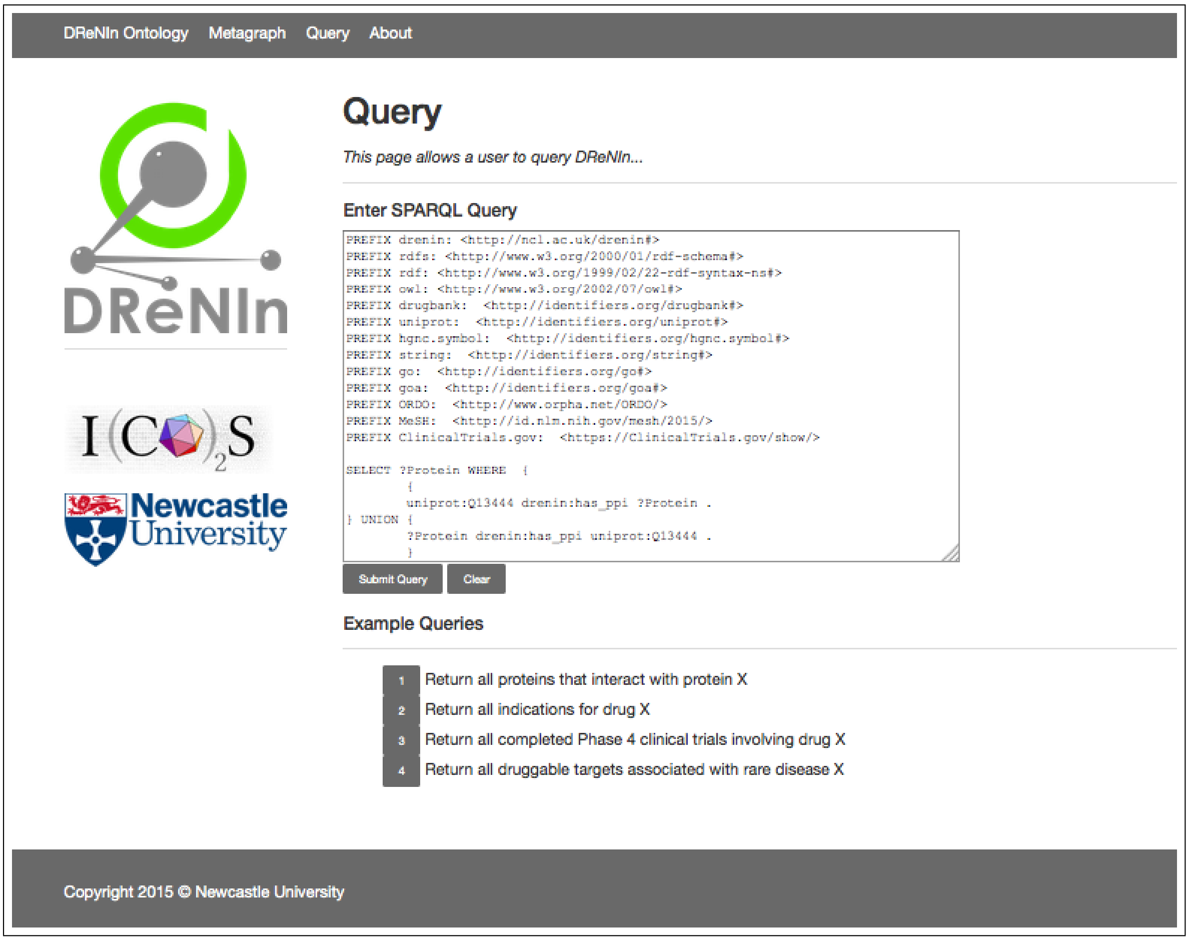
3.4.5.1 User interface

The five pages include home, about (detailing data sets included in DReNIn), metagraph (displays a count of each datatype in the network), ontology (details the classes and relations captured in DReNInO) and a query page. The query page of the Web site allows a user to submit either one of the predefined SPARQL queries, or to compose their own query. Query results are presented in tabular format but may be downloaded in RDF for further analysis or visualisation tasks.



	Node Types	Edge Types	Integration Concept(s)
UniProtKB	Gene Protein	is_encoded_by	-
GO	Molecular_Function Cellular_Component Biological_Process	-	-
GOA	-	has_molecular_function located_in_cellular_component involved_in_biological_process	Molecular_Function Cellular_Component Biological_Process Protein
STRING	-	has_ppi	Protein
ORDO	Rare_Disease	is_a	-
MeSH	Common_Disease	is_a	-
Orphanet	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
OMIM	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
CTD	-	involved_in_rare_disease involved_in_common_disease has_indication	Rare_Disease Common_Disease Gene Small_Molecule Bio_Tech
BeFree	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
GWAS	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
SemRep	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
MGD	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
RGD	-	involved_in_rare_disease involved_in_common_disease	Rare_Disease Common_Disease Gene
DrugBank	Small_Molecule Bio_Tech	binds_to interacts_with	-
ChEMBL	-	binds_to	Small_Molecule Bio_Tech Protein
NDF-RT	-	has_indication	Small_Molecule Bio_Tech Common_Disease Rare_Disease
SIDER	-	has_indication has_side_effect	Small_Molecule Bio_Tech Common_Disease Rare_Disease
DCDB	Drug_Combination	part_of_combination	Small_Molecule Bio_Tech
OpenPHACTS	Pathway	involved_in_pathway	Small_Molecule Bio_Tech Gene Protein
ClinicalTrial.gov	Clinical_Trial	clinical_trial_drug clinical_trial_disease	Small_Molecule Bio_Tech Common_Disease Rare_Disease

Figure 3.9: **Integration strategy utilised in DReNInF**. Strategy makes use of all data sources described in 3.2.4 as well as the mapping sources listed in Table. 3.2. Red circles represent strategy steps, parsers, which involves the integration of the two input sources. Node types and edge types taken from each source as well as the concept types used to integrate the data for each source is also provided.



Query

This page allows a user to query DReNIn...

Enter SPARQL Query

```
PREFIX drenin: <http://ncl.ac.uk/drenin#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX drugbank:  <http://identifiers.org/drugbank#>
PREFIX uniprot:  <http://identifiers.org/uniprot#>
PREFIX hgnc.symbol:  <http://identifiers.org/hgnc.symbol#>
PREFIX string:  <http://identifiers.org/string#>
PREFIX go:  <http://identifiers.org/go#>
PREFIX goa:  <http://identifiers.org/goa#>
PREFIX ORDO:  <http://www.orpha.net/ORDO/>
PREFIX MeSH:  <http://id.nlm.nih.gov/mesh/2015/>
PREFIX ClinicalTrials.gov:  <https://ClinicalTrials.gov/show/>

SELECT ?Protein WHERE  {
        {
        uniprot:Q13444 drenin:has_ppi ?Protein .
} UNION {
        ?Protein drenin:has_ppi uniprot:Q13444 .
        }
```

Example Queries

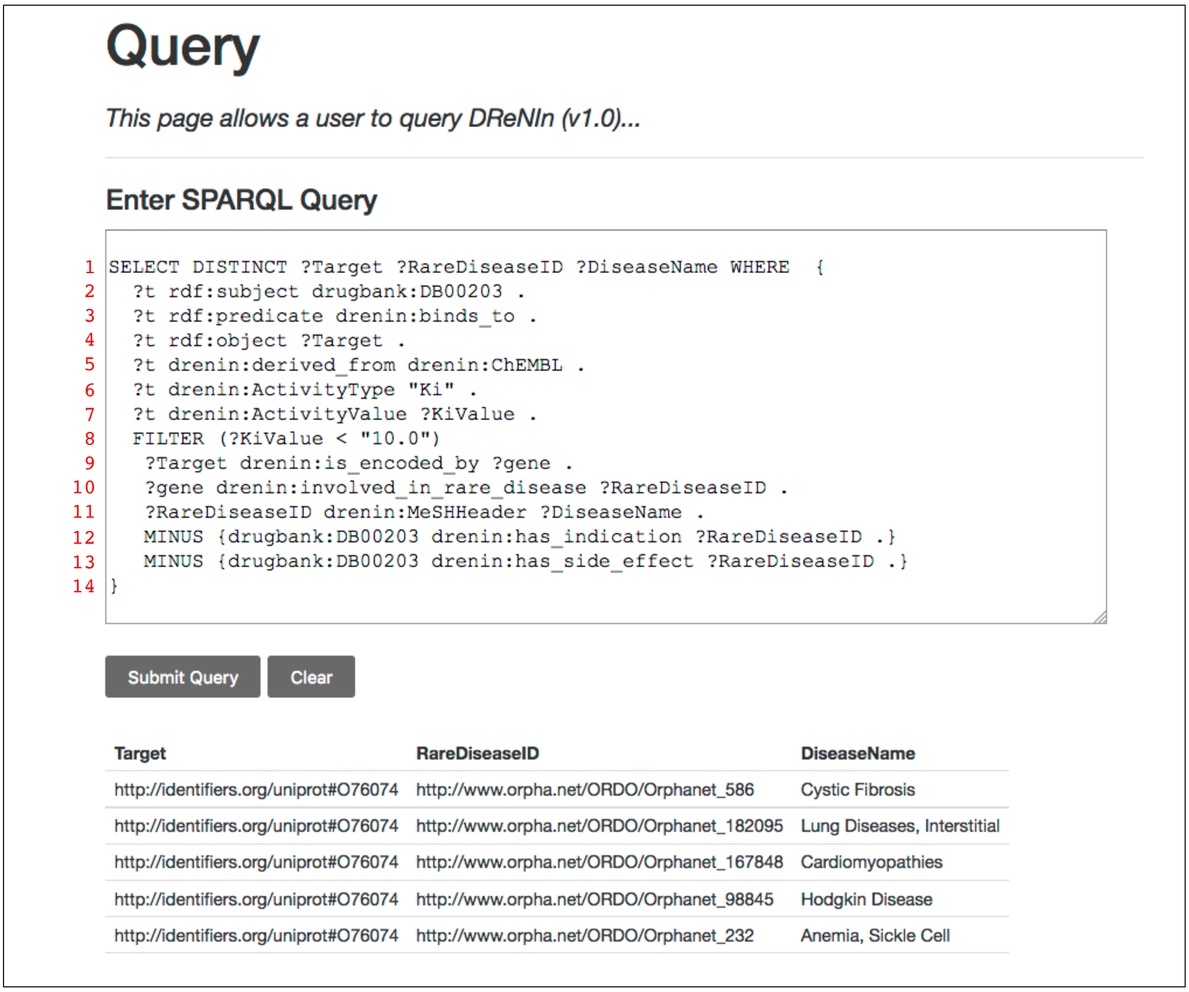
1 Return all proteins that interact with protein X
2 Return all indications for drug X
3 Return all completed Phase 4 clinical trials involving drug X
4 Return all druggable targets associated with rare disease X

Copyright 2015 © Newcastle University

Figure 3.10: **DReNIn query interface**. The DReNIn Web site is made up of five pages: home; about; metagraph; ontology; and the query page which is shown above. The query page provides a user with pre-canned SPARQL queries as well as an open text field for user specified queries SPARQL.

3.4.5.2 Example queries

Some predefined sample queries are presented on the query page (Fig. 3.10), while a text box is also given to allow input of user-specified queries. The data set allows a user to ask questions that would otherwise not be possible to answer systematically, such as those described in Fig. 3.11, Fig. 3.12 and Fig. 3.13.



Query

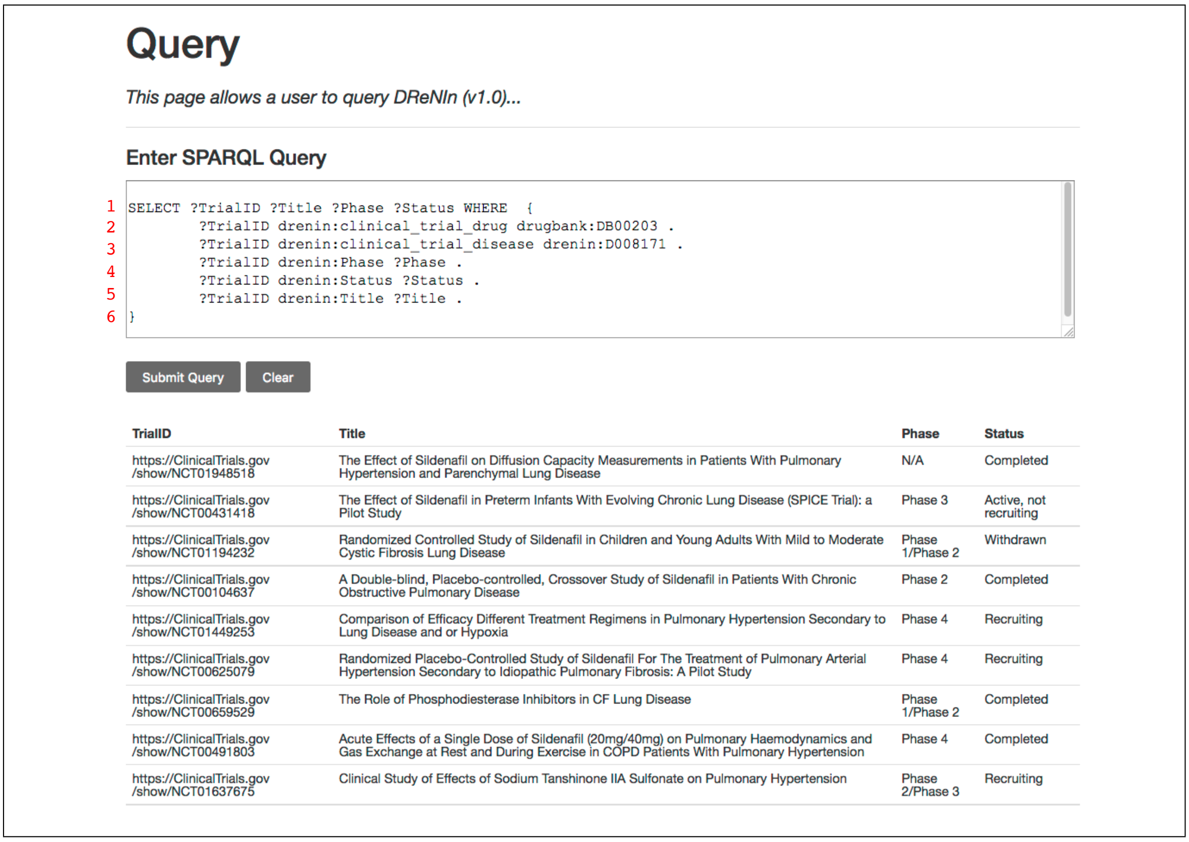
This page allows a user to query DReNIn (v1.0)...

Enter SPARQL Query

```
1  SELECT DISTINCT ?Target ?RareDiseaseID ?DiseaseName WHERE  {
2    ?t rdf:subject drugbank:DB00203 .
3    ?t rdf:predicate drenin:binds_to .
4    ?t rdf:object ?Target .
5    ?t drenin:derived_from drenin:ChEMBL .
6    ?t drenin:ActivityType "Ki" .
7    ?t drenin:ActivityValue ?KiValue .
8    FILTER (?KiValue < "10.0")
9     ?Target drenin:is_encoded_by ?gene .
10    ?gene drenin:involved_in_rare_disease ?RareDiseaseID .
11    ?RareDiseaseID drenin:MeSHHeader ?DiseaseName .
12    MINUS {drugbank:DB00203 drenin:has_indication ?RareDiseaseID .}
13    MINUS {drugbank:DB00203 drenin:has_side_effect ?RareDiseaseID .}
14 }
```

Target	RareDiseaseID	DiseaseName
http://identifiers.org/uniprot#O76074	http://www.orpha.net/ORDO/Orphanet_586	Cystic Fibrosis
http://identifiers.org/uniprot#O76074	http://www.orpha.net/ORDO/Orphanet_182095	Lung Diseases, Interstitial
http://identifiers.org/uniprot#O76074	http://www.orpha.net/ORDO/Orphanet_167848	Cardiomyopathies
http://identifiers.org/uniprot#O76074	http://www.orpha.net/ORDO/Orphanet_98845	Hodgkin Disease
http://identifiers.org/uniprot#O76074	http://www.orpha.net/ORDO/Orphanet_232	Anemia, Sickle Cell

Figure 3.11: **DReNIn SPARQL example 1.** A SPARQL representation of the following research question: *"I am interested in inferring novel uses for drugX. Identify all targets that drugX potently binds to and the rare diseases associated to these targets. Do not include diseases that drugX is marketed to treat or are known off-target effects of the drug".* Query line numbers are shown in red. Line 1 defines the data to be returned by the query: `Target` refers to the id of the target; `RareDiseaseID` refers to the id of the rare disease; and `DiseaseName` refers to the name of the rare disease. Line 2 specifies *drugX* as `drugbank:DB00203` (sildenafil). Lines 3-7 of the query describe the type of association required between the drug and the potential targets. It is stated that a `binds_to` association (line 3) derived from ChEMBL (line 5) is required, and that this association should be annotated with a `Ki ActivityType` (line 6). Line 8 removes any associations that have an `ActivityValue` less than 10.0nM. Next, targets that the drug bind to and that are encoded by a gene (line 9) and associated with a rare disease (line 10) are identified. Finally, line 12 removes diseases that the drug is already known to treat and line 13 removes diseases that are known to be a side-effect of the drug.



Query

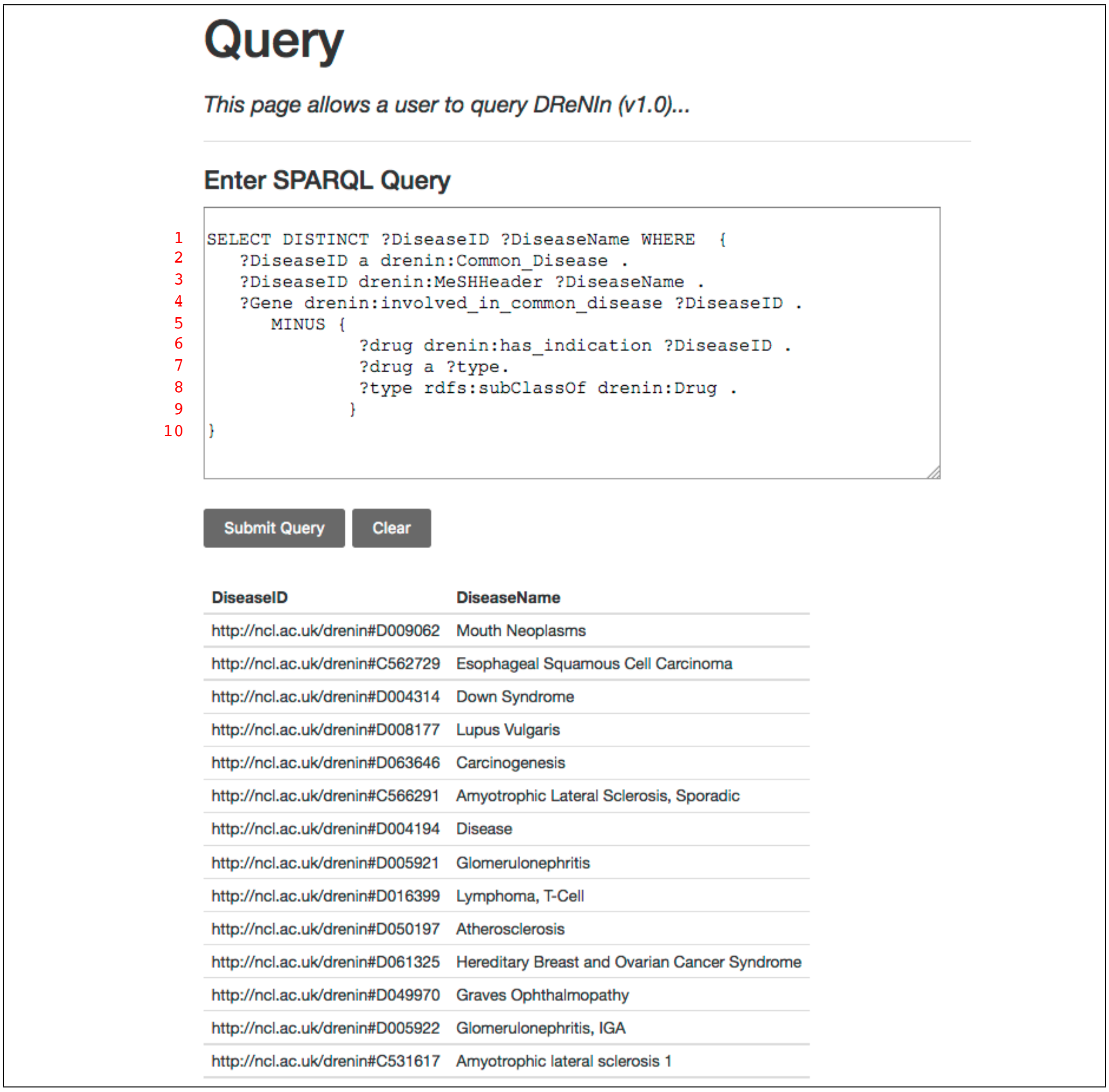
This page allows a user to query DReNIn (v1.0)...

Enter SPARQL Query

```
1 SELECT ?TrialID ?Title ?Phase ?Status WHERE  {
2         ?TrialID drenin:clinical_trial_drug drugbank:DB00203 .
3         ?TrialID drenin:clinical_trial_disease drenin:D008171 .
4         ?TrialID drenin:Phase ?Phase .
5         ?TrialID drenin:Status ?Status .
          ?TrialID drenin:Title ?Title .
6 }
```

TrialID	Title	Phase	Status
https://ClinicalTrials.gov /show/NCT01948518	The Effect of Sildenafil on Diffusion Capacity Measurements in Patients With Pulmonary Hypertension and Parenchymal Lung Disease	N/A	Completed
https://ClinicalTrials.gov /show/NCT00431418	The Effect of Sildenafil in Preterm Infants With Evolving Chronic Lung Disease (SPICE Trial): a Pilot Study	Phase 3	Active, not recruiting
https://ClinicalTrials.gov /show/NCT01194232	Randomized Controlled Study of Sildenafil in Children and Young Adults With Mild to Moderate Cystic Fibrosis Lung Disease	Phase 1/Phase 2	Withdrawn
https://ClinicalTrials.gov /show/NCT00104637	A Double-blind, Placebo-controlled, Crossover Study of Sildenafil in Patients With Chronic Obstructive Pulmonary Disease	Phase 2	Completed
https://ClinicalTrials.gov /show/NCT01449253	Comparison of Efficacy Different Treatment Regimens in Pulmonary Hypertension Secondary to Lung Disease and or Hypoxia	Phase 4	Recruiting
https://ClinicalTrials.gov /show/NCT00625079	Randomized Placebo-Controlled Study of Sildenafil For The Treatment of Pulmonary Arterial Hypertension Secondary to Idiopathic Pulmonary Fibrosis: A Pilot Study	Phase 4	Recruiting
https://ClinicalTrials.gov /show/NCT00659529	The Role of Phosphodiesterase Inhibitors in CF Lung Disease	Phase 1/Phase 2	Completed
https://ClinicalTrials.gov /show/NCT00491803	Acute Effects of a Single Dose of Sildenafil (20mg/40mg) on Pulmonary Haemodynamics and Gas Exchange at Rest and During Exercise in COPD Patients With Pulmonary Hypertension	Phase 4	Completed
https://ClinicalTrials.gov /show/NCT01637675	Clinical Study of Effects of Sodium Tanshinone IIA Sulfonate on Pulmonary Hypertension	Phase 2/Phase 3	Recruiting

Figure 3.12: **DReNIn SPARQL example 2.** A SPARQL representation of the following research question: *"I have a potential indication, diseaseX, for a known drug, drugY, and I would like to check if this indication for drugY has been investigated previously. Identify the id, title, phase and status of clinical trials that involve both drugY and diseaseX"*. Query line numbers are shown in red. Line 1 defines the data to be returned by the query to describe relevant clinical trials: `TrialID` refers to the id of the clinical trial as provided by clinicaltrials.gov; `Title` refers to the title of the trial; `Phase` refers to the phase that the trial has reached; and `Status` refers to the status of the trial. Furthermore, in line 2 *diseaseX* is defined as `drenin:D008171` (Lung Diseases) and in line 3 *drugY* is defined as `drugbank:DB00203` (sildenafil).



Query

This page allows a user to query DReNIn (v1.0)...

Enter SPARQL Query

```
1  SELECT DISTINCT ?DiseaseID ?DiseaseName WHERE  {
2     ?DiseaseID a drenin:Common_Disease .
3     ?DiseaseID drenin:MeSHHeader ?DiseaseName .
4     ?Gene drenin:involved_in_common_disease ?DiseaseID .
5        MINUS {
6                ?drug drenin:has_indication ?DiseaseID .
7                ?drug a ?type.
8                ?type rdfs:subClassOf drenin:Drug .
9               }
10 }
```

DiseaseID	DiseaseName
http://ncl.ac.uk/drenin#D009062	Mouth Neoplasms
http://ncl.ac.uk/drenin#C562729	Esophageal Squamous Cell Carcinoma
http://ncl.ac.uk/drenin#D004314	Down Syndrome
http://ncl.ac.uk/drenin#D008177	Lupus Vulgaris
http://ncl.ac.uk/drenin#D063646	Carcinogenesis
http://ncl.ac.uk/drenin#C566291	Amyotrophic Lateral Sclerosis, Sporadic
http://ncl.ac.uk/drenin#D004194	Disease
http://ncl.ac.uk/drenin#D005921	Glomerulonephritis
http://ncl.ac.uk/drenin#D016399	Lymphoma, T-Cell
http://ncl.ac.uk/drenin#D050197	Atherosclerosis
http://ncl.ac.uk/drenin#D061325	Hereditary Breast and Ovarian Cancer Syndrome
http://ncl.ac.uk/drenin#D049970	Graves Ophthalmopathy
http://ncl.ac.uk/drenin#D005922	Glomerulonephritis, IGA
http://ncl.ac.uk/drenin#C531617	Amyotrophic lateral sclerosis 1

Figure 3.13: **DReNIn SPARQL example 3.** A SPARQL representation of the following research question: *"I would like to identify a therapeutic area of interest for my upcoming study. Identify all common diseases for which there are currently no marketed drugs but that have known genetic associations that can be targeted by a potential treatment"*. Query line numbers are shown in red. Line 1 defines the data to be returned by the query: `DiseaseID` refers to the id of the common disease; and `DiseaseName` refers to the name of the common disease. Line 2 gets the `DiseaseID` of all common diseases in the dataset, while line 3 retrieves the `DiseaseName` of these diseases. Line 4 retrieves all of the genes that are associated with the set of diseases. Finally, the `MINUS` function of the query (lines 5-9) discards all of the diseases that have a drug known to treat them in the dataset.

3.5 Discussion

In this section the work that has been presented in the Chapter will be discussed and summarised. Furthermore, limitations will be presented as well as potential future directions.

3.5.1 Updated Ondex data set, Dat

DISEASE_KIND concepts from NDF-RT were used to represent the `Indication concept_class` in *Dat*. The NDF-RT uses a relatively shallow vocabulary to capture DISEASE_KIND concepts, which explains why there is a limited number of G-D associations included from DisGeNET that map to diseases in *Dat*. It must also be noted that there exists a plethora of data sources storing G-D associations derived using many different approaches. Although DisGENet includes data from multiple G-D sources and provides a score for these associations, it is done so using an arbitrary metric. In future iterations of the data set a broader view of the G-D landscape needs to be taken; with a method to integrate and score these essential.

Dat is used as the test bed for much of the preliminary work described in further Chapters. However, this data set has several limitations. First and foremost is the fact that the data included is outdated. Some data sets, such as KEGG, are now unavailable, making a rebuild of the data set impossible. *Dat* includes DrugBank v2.5 data, which is only available for download as a text file[46]. Newer versions of this data set are now available for download, but it is now distributed in XML format only; leaving the original Ondex parsers incapable of parsing newer versions.

More important than the data release formats are the limitations imposed by the initial design of the semantics. *Dat* uses DrugBank as the focal point for all `Compound` and `Target` concepts and the `binds_to` interactions between these. A `Target` taken from DrugBank is a single protein known to bind a drug, and so should be assigned the type `Protein`. The original semantic definition of a `Target` has repercussions when trying to mine *Dat* for novel interactions, such as Drug-Target (D-T) interactions. The process of inferring novel targets is made more difficult since known targets

[46] www.drugbank.ca/downloads/archived Accessed: 22-12-2015

would be overrepresented when data mining for novel D-T inferences. A more thoroughly designed semantics would need to remove the abstracted `Target` type. Instead, `binds_to` relations should be between the drug and the type of target. In most cases, the target types will be single proteins, or genes, but could also be protein families, protein complexes or more generally tissues or organisms [136].

DrugBank contains a comprehensive set of D-T associations. However, there are no details regarding the activity type and values associated with these associations (e.g. IC_{50}, K_d, K_i and Potency). Targets are also included from multiple species, yet the proteins included in *Dat* are limited to those from *Homo sapiens*. Also missing is knowledge regarding the currently marketed state of the drug, i.e. what it is used to treat, whether there are known side-effects of the drug and if the drug is currently being investigated to treat any disease. These are all questions that need to be addressed in later iterations of the data set.

3.5.2 Ondex V Neo4j performance testing

Ondex was used as the integration platform for both the initial data set and *Dat*. As previously mentioned (see Section 3.2.3.3), Ondex is capable of implementing a strict semantics which can be used to aid data integration tasks via a controlled vocabulary captured in an XML file. This controlled vocabulary is essentially a list of terms and not a full ontological representation. With this limitation in mind, further aspects of Ondex are discussed to evaluate systematically the benefits of using Ondex during the development of further data sets.

Although a compiled jar is available for Ondex, to develop against the project the source code must be downloaded and compiled. Compiling Ondex on a local machine is not an easy task and takes, at best, numerous attempts. Although well documented, one must have a well-rounded understanding of the Ondex framework to build any type of plugin (parser, mapper, etc.) and this task can take a day of development time for more complex implementations. Building integrated networks that include a significant number of individual data sets, each with subtle differences in their data format, can result in months of development time. Even after the production of a data set, rather counter-intuitively (Ondex holds the graph in memory), exploring the graph using filters is a time-consuming task. It can take up to an hour to correctly filter and visualise even the nearest neighbours of a node of interest (for example, Fig.3.4 took an hour to produce). The poor performance is likely due to a combination of poorly implemented code along with the size of the graph that was held in memory (around a million graph entities, that is nodes plus edges).

Ondex shows linear scalability with linear graphs (see Fig.3.5A), with a time complexity of $O(n)$. For more complex graphs, this time complexity shows properties closer to polynomial growth $O(n^2)$ (see Fig.3.5D). Memory usage, taken as a measure of the average RAM used during a graph build, in the Ondex system remains linear for both linear and more complex networks. Available RAM is quickly saturated, however, with linear graphs able to reach a $|V|$ of 2.1M (see Fig.3.5B) and more complex networks a $|V|$ of 1.2M (see Fig.3.5E) before the 6GB is saturated. To save an Ondex graph to disk an extra plugin must be called from the XML workflow, the OXL exporter. The amount of memory used to save graphs, with a $|V|$ of 500,000, in OXL format to disk differs quite dramatically between the two graph types. For linear graphs this is 236MB, reaching 330MB for more complex graphs of the same size.

Due to the fact that the development of a data set for drug repositioning would likely take numerous iterations and be highly connected (no doubt displaying similar topological attributes to *Dat*), it was decided that an alternative integration platform would be investigated.

The development times and build performance of the highly popular Neo4j were, therefore, investigated. The aim of this comparison is to directly contrast the two platforms to identify which offers a more viable platform for the development of integrated data sets. Unlike Ondex, there is very little pre-requisite in terms of knowledge required. As long as a user has a basic understanding of Java and Maven development then the learning curve is minimal. A simple parser for Neo4j may be developed in a couple of hours, however, it must be pointed out that this is mainly due to the fact that there is no mechanism for strictly controlling semantics and no framework to guide a user during parser development, and so there is no need to consider a formal model.

Neo4j displays linear scalability for both linear (see Fig.3.5A), and more complex graphs (see Fig.3.5D). Memory usage, taken as a measure of the average RAM used during a graph build, in Neo4j does not display the same linear pattern as observed in Ondex. Available RAM allows for linear graphs with a $|V|$ of 10M (see Fig.3.5B) and more complex networks a $|V|$ of 7M (see Fig.3.5E) to be built on a local machine. The amount of memory used to save graphs, with a $|V|$ of 500,000, in Neo4j is a lot lower than in Ondex, 80MB for a linear graph and 131MB for more complex graphs of the same size.

Although Neo4j, when compared to Ondex, allows for a more scalable approach to data set development, unlike Ondex, semantics are not controlled in a strict fashion in Neo4j. To make use of the inherent scalability of Neo4j, a lightweight layer for handling semantics was developed. The aim of the semantic layer was to allow for directed integration and querying with reference to an ontological representation of semantic types captured in any subsequent integrated data sets for drug repositioning.

3.5.3 BioSSIP: a light-weight semantic handling module

The BioSSIP module enables users to build and analyse the same data set using different data storage solutions with the same code, encouraging reproducibility. Having one code base to learn to produce, and analyse, integrated semantic graphs using a multitude of datastores can dramatically reduce the upfront learning costs associated with moving an integration project to another data storage solution. BioSSIP makes use of the inherent scalability of integration platforms, such as Neo4j, while also allowing for the semantics to be controlled by a high-level, application-specific ontology.

At present, there are only data access layers for the *in-memory* JVM and the NoSQL Neo4j graph database. Although it would be relatively simple to develop alternative data access layers for alternative 'backends', such as a relational database, this is yet to be implemented. Through allowing alternate backends to be used, the same data set utilising a suite of parsers and a BioSSIP integration strategy may be implemented and recreated in different platforms, supporting the idea of data reproducibility. BioSSIP mining strategies are self-contained, and so providing a data access layer has been created, the same query strategies can be executed on multiple 'backends'. Ultimately, this means that a BioSSIP application may be released as a suite of parsers, an integration strategy and a mining strategy. Anyone can then recreate the data set and analyses regardless of his or her choice of 'backend'.

3.5.4 DReNInF: a data integration framework

DReNInF enables the production of integrated data sets relevant to drug repositioning. These system representative data sets allow a holistic view of drugs in mammalian systems to be realised.

Central to DReNInF is the DReNInO which provides a high-level view of data types relevant to drug repositioning. Although many ontologies exist, none captured all the data relevant to this project. DReNInO allows semantics to be controlled in a strict fashion during integration and querying alike. The ontology has been created in such a manner that it can be extended by the community and allows for both systems relevant data as well as validation (annotation) data to be combined. DReNInO was developed in order to provide a simple framework to capture entities relating to the central dogma of drug action (see Fig. 2.1).

Furthermore, it was important that the `concept_class Target` was not included (see page 76) in DReNInO. Instead, as many drug targets are indeed single proteins, the interactions between drugs (be it `Small_Molecules`, `Bio_Tech` or `Drug_Combinations`) are instead captured in `binds_to` associations from these drug types directly to a `Protein` or a `Gene`. Rare diseases affect a small percentage of the population and as such there exists a number of economic incentives for the development of treatments for rare diseases (see Section 2.1). Furthermore, they tend to be genetic in origin and monogenic— resulting from modifications of a single gene. Although not always the case, more common diseases, such as Crohn's disease, tend to have a more complex genetic causation and are said to be multigenic. For these reasons it was decided that the data set would make a clear semantic difference between these and so two disease types are defined in DReNInO; `Common_Disease` and `Rare_Disease`.

The ultimate goal of any drug repositioning data set is to aid in the inference of novel associations between types of data in the network (such as a `has_indication` association between a `Small_Molecule` and a `Rare_Disease`). It was, therefore, important to include data types in DReNInO that would allow for some form of validation, or indeed filtering, to be achieved. For this reason, `has_side_effect` associations between drugs and diseases were also included. Also included were `Clinical_Trials`, with this data allowing for any inferences to be systematically checked to see if any studies are currently investigating the drug for the inferred use.

The `Pathway` `concept_class` and ConceptWiki URLs for both the `Protein` and `Gene` concepts were taken from Open PHACTS. ConceptWiki, included in Open PHACTS, offers a more thorough mapping of elements to alternative accessions and is far beyond what could have been achieved during this project. However, ConceptWiki mapping to genes was not included. ConceptWiki mappings to genes are relatively few, while some that have been mapped appear to have been done so incorrectly (see Fig. A.1).

As an Associated Partner of the Open PHACTS project, many webinars and workshops were attended during this project to identify any overlap between the hugely funded consortium and the project presented in this thesis. Although the project offers a great deal of interesting data, until recently this data was only accessible via its RESTful API (it has recently enabled local download and a Docker instance of the RDF data set in a Virtuoso triplestore). Using REST requests meant exhaustive querying of the data was intractable— something that would have been necessary for extending the data set. At the start of this project, Open PHACTS did not include any side-effect data, indication infomation nor clinical trial data, although side-effects are now included— the other data types are still missing. The timeline of the project co-incided with this project. With a massive team of support staff, the project quickly gained momentum which made identifying an area for collaboration difficult— new releases were never far away. Furthermore, the focus of this project was pharmacologial annotation of entities represented and not so much the more abstract concept of the topological properties of highly connected data. It was for these reasons, as well as political factors (Open PHACTS had commercial partners whose requirements were more pressing than others), that it was decided that Open PHACTS would not be utilised as the sole data source for this project and that collaboration was not pursued.

During the development of DReNInF challenges of data integration, as described in Section 3.2.2, were addressed. Although there are certain limitations to open source data in terms of quality, by using only sources that are heavily utilised within the field one would assume that data sets created using DReNInF provide a relatively accurate representation of drugs and their interactions with other biological entities. Through the development of novel mining methods, data sets created using the DReNInF will ultimately be used to aid the identification of novel uses for existing drugs.

3.5.5 DReNIn: an RDF exposed data set

DReNIn, a homogeneous syntactic and semantic RDF representation of multiple, heterogeneous data sources with a single, unified query interface, was developed using DReNInF. Although large projects exist, such as Open PHACTS, they do not capture as many types of data as DReNIn. Open PHACTS does not include information that may be beneficial to the validation of predicted Drug-Disease (Dr-D) inferences, with no indication information included, a vital piece of information in any drug repositioning prediction task. As well as the lack of indication data, Open PHACTS does not contain any clinical trial information and, because of the size of the project, struggles with the mapping between particular entities (such as those described in A.1). Equivalence in DReNIn is inferred using sources listed in Table. 3.2. Four of these sources are heavily utilised in the field, and DReNIn is also able to extend mappings using manually created equivalences. Provenance, recognised as a vital aspect of a task in data integration is provided for every edge and node in the data set, with the source as well as version, date accessed and even the data format from which it was extracted is provided.

Due to the fact that the data set is hosted locally on a university server (unlike commercial projects that would have access to dedicated software), large SPARQL queries may prove troublesome. For this reason, it may be useful to implement a query 'checker' before accepting queries. One caveat to the data set is, due to the large number of data sources included, updates must be completed on a regular basis, to ensure a state-of-the-art view is provided. It takes a relatively long time to pull all clinical trials locally, and extracting data from Open PHACTS is also a time-consuming task. To ensure the future of the project, an automated workflow will need to be implemented which will also need to inform of any changes to data formats included in the data set.

The fact that DReNIn allows for queries that are not possible in other open source drug repositioning data sets (such as those presented in Fig. 3.11, Fig. 3.12 and Fig. 3.13), means it has the potential to become an important source of information for the research field in the future.

3.6 Conclusion

This Chapter has introduced data integration for drug repositioning, various data sets, a platform for integration and the application of that platform to the generation of an integrated data set for drug repositioning. A historical Ondex data set for *in silico* drug discovery was extended for the purpose of drug repositioning. It was shown that the limited scalability of the Ondex platform was not present in other data integration platforms, including Neo4j. However, Neo4j lacks a means for strictly controlling semantics. With this in mind, a lightweight framework to handle semantics, BioSSIP was developed, which utilises Neo4j as the underlying datastore. A drug repositioning data integration framework (DReNInF) was then described. DReNInF is made up of a high-level ontology (DReNInO), a suite of parsers and an integration strategy. DReNInF was used to produce an up-to-date semantically-rich drug repositioning data set, DReNIn. Finally, this data set was exposed as an RDF data set, and made accessible through a dedicated web front end, with a SPARQL endpoint, allowing for the data set to be queried.

It is hoped that in the future BioSSIP can gain some traction in the wider integration field. For this to happen it will be necessary for more ‘backends’ to be implemented as well as a wider range of parsers, covering more than just the specialised area of drug repositioning. Although the DReNIn RDF data set has enabled questions to be asked that were previously beyond current integrative efforts in the field, the interface could be improved. Improvements would centre around the creation of SPARQL queries, which require a certain level of expertise from the user. Integrating work such as that completed by General Electric Company in SPARQLgraph[47] would allow for SPARQL queries to be developed visually, removing any barriers that may currently be present.

The framework introduced in this Chapter has been used to create subsequent drug

[47] www.semtk.research.ge.com

repositioning networks as described in Chapter 6. However, in order for integrated data sets to be useful to the area of drug repositioning, systematic approaches to mining for sub-components that may be indicative of repositioning opportunities must be identified. To this regard, in the next Chapter, semantic subgraphs are formally introduced. These semantic subgraphs are sub-components of an integrated network that can be used to capture the functionality that describes a potential drug repositioning opportunity. Furthermore, Drug Repositioning Semantic Mining (DReSMin), an exact, exhaustive algorithm for the identification of semantic subgraphs is also detailed.

4

A novel algorithm for mining integrated data sets

4.1 Introduction

In graph theory, a graph is defined as a set of ordered pairs $G = (V, E)$ where V is a set of vertices, or nodes, and E, a set of edges. The order of a graph is $|V|$ (number of nodes), and the size is $|E|$ (the number of edges). All graphs share these essential structural elements, while other structural and semantic graph properties vary amongst applications [137]. A graph can be constructed in various ways by allowing or disallowing: (i) directed or undirected edges; (ii) weights on edges (simply a number assigned to each edge); (iii) multiple edges between a pair of nodes (multigraphs); (iv) edges that connect a single node to itself (self-loops) [137].

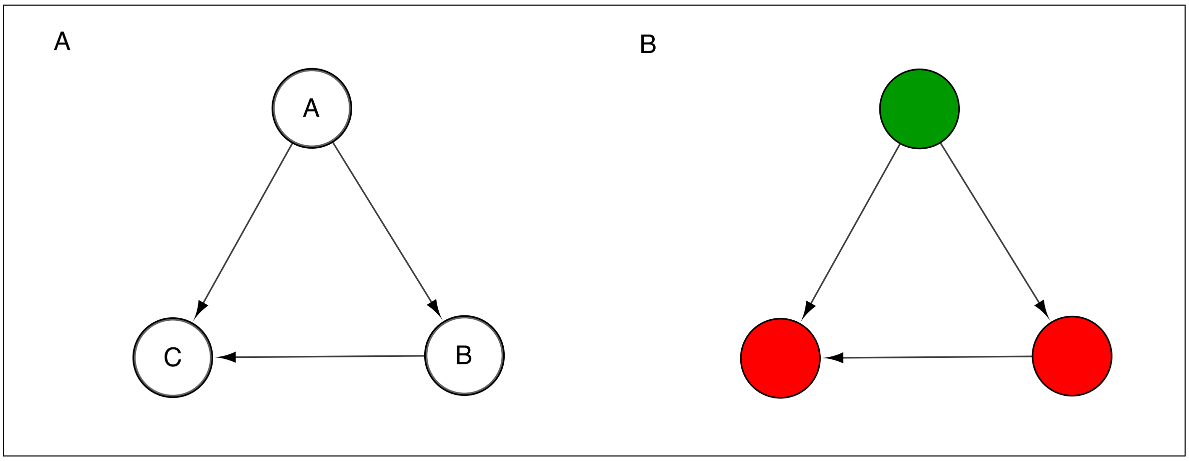


Figure 4.1: **Network motifs and coloured motifs.** (A) An example of a feed-forward loop network motif and (B) One of the most over-represented coloured motifs of $|V| = 3$ in the *Caenorhabditis elegans* neuronal network [138]. Although A and B share the same topology, through the inclusion of semantics, understanding of functionality can also be achieved. *Note:* In B, Green: sensory neurone; red: interneuron. Arrows represent the direction that the signal travels between the two cells.

All graphs, whether representing biological, social or technological data can be decomposed into subgraphs. Formally, a subgraph G', of a graph G, is a graph $G' = (V', E')$ where $G'|V'| \subset G|V|$ and $G'|E'| \subset G|E|$. These subgraphs formally capture local relationships between nodes in a graph. Often, in the life sciences, the relationships and nodes in a given subgraph are indicative of a particular biological phenomenon or function [139]. Analysis of these subgraphs allows for a greater understanding of the network as a whole and the roles that specific concepts may play. Subgraphs may capture topological information, i.e. the way in which constituent parts are interrelated or arranged structurally, and/or semantics, i.e. relating to the meaning of the sub-structure. From a topological perspective, Network Motifs (NMs), for example, are subgraphs that are statistically overrepresented in comparison to their prevalence in randomised networks [140, 141]. NMs are recognised as the simple building blocks of complex networks. A biological example, often captured in gene regulatory networks, are feed-forward loops (Fig. 4.1A) [142]. A feed-forward loop is a three gene pattern, composed of two input genes, one of which regulates the other, both jointly regulating a target gene [142]. The application of NMs is acceptable when analysing networks containing limited semantic types such as gene regulatory networks or topological analysis of the Internet, in particular, social network analysis [143]. NMs, however, are not suitable for the analysis of more semantically complex networks, such as metabolic networks or, of more relevance to this work, drug networks. This limitation is due to the fact that similar topologies can give rise to very different functions.

Extensions of NMs approaches have been described which allow the inclusion of semantic information, such as the *list coloured motif problem* [144, 145]. In this case, a motif (M) is defined as a multiset of colours (types). An occurrence of M is a subset of nodes that forms a connected subgraph whose multiset of colours, matches that of M, exactly [145]. An example of a coloured motif is shown in Fig. 4.1B, which illustrates one of the most over-represented coloured motifs captured in the *Caenorhabditis elegans* neuronal network [138]. Although this coloured motif shares the same topology as the NMs presented in Fig. 4.1A, the introduction of semantic types allows for a more detailed description of the functionality that is captured. This approach demonstrates how NMs, the simple building blocks of complex networks, may be extended to incorporate semantic information, allowing for statistically over-represented semantic patterns to be identified. Although this extension highlights the importance of including semantics in subgraphs, there is no reason to assume that semantic subgraphs capturing potential drug repositioning opportunities will be statistically overrepresented in the target network. Therefore, an alternative approach to identifying subgraphs that also incorporate topology and semantics is required for the drug repositioning setting.

As described in Chapter 3, in the case of drug repositioning networks, the types of relationships include amongst others: interactions between drugs and their targets, interactions between targets, and the diseases associated with particular targets. Furthermore, multiple types of interactions can exist between the same entity types (e.g. between drug and disease there may be side effect associations as well as indication associations) capturing a plethora of biological functions. Subgraphs that describe repositioning opportunities as a result of their semantic and topological properties have been previously introduced as semantic subgraphs by Cockell *et al.* [101].

In this Chapter, a formal definition for semantic subgraphs is first provided. An exhaustive algorithm for the identification of these semantic subgraphs is then introduced. This algorithm, Drug Repositioning Semantic Mining (DReSMin), allows for mappings of a semantic subgraph that are topologically exact and semantically 'similar' to the query to be identified; meaning that functionally similar sub-components can be identified from a target network. DReSMin is made up of novel components, in particular a graph pruning step and a graph splitting component, that enable previously intractable subgraphs to be identified. The work presented in this Chapter has also been described in Mullen *et al.* [146]. A Java implementation of the algorithm is available at `https://bitbucket.org/jmullen/dresminalgorithm`.

4.2 Background

4.2.1 Semantic subgraphs

As part of this work, building on the work Cockell *et al.* [101], a formal representation of semantic subgraphs has been defined and can be found in [146]. In this paper a semantic subgraph is defined as: $Q = (V, E, T_v, f_v, T_e, f_e)$, where V is a set of nodes, E is a set of edges, T_v is a set of node types and T_e is a set of edge types. $f_v : V \twoheadrightarrow T_v$ and $f_e : E \twoheadrightarrow T_e$ are surjective (every element of the codomain is mapped to by at least one element of the domain) functions; each node is assigned a node type, and each edge has an edge type from T_v and T_e respectively. A semantic subgraph, Q, may be designed in such a manner that mappings, or occurrences, of Q in G can aid in the inference of potential relations between nodes of particular types, where a relation does not exist. Semantic subgraphs may be manually created through the identification of known repositioning examples within an integrated network. For example, the semantic subgraph presented in Fig. 4.2A was developed by identifying

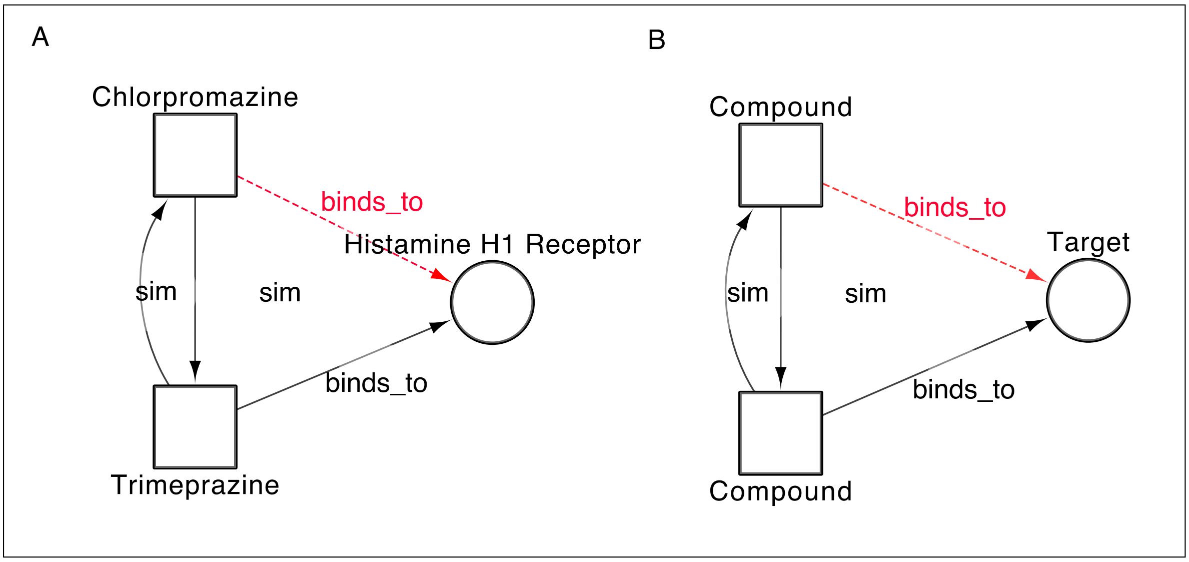


Figure 4.2: **The drug repositioning of chlorpromazine as captured in an *in silico* data set for drug development [101] (A) and the subsequent semantic subgraph representation (B).** Chlorpromazine is marketed as a non-sedating tranquilliser [147]. In DrugBankv2.5 no association between chlorpromazine and the Histamine H1 receptor is curated despite the drug also being an effective antihistamine [147]. Trimeprazine has a similar 2D structure to chlorpromazine (shown in the `sim` relations) and an association between the drug, and the Histamine H1 receptor is captured in DrugBankv2.5. Due to the similar structures of the drugs, it may be inferred that chlorpromazine is also likely to bind to the target and this inferred edge is captured as a dashed line. Figure is adapted from [101].

the known repositioning example of chlorpromazine (generic medication) within the integrated network presented in [101]. The semantic subgraph depicted in Fig. 4.2B was then derived from the chlorpromazine example in Fig. 4.2A and can be used to infer interactions between a node of type compound and a node of type target.

Chlorpromazine is an anti-psychotic drug that is also approved as an antihistamine [148]. The interactions of chlorpromazine can be captured in an integrated network (Fig. 4.2A). Data from DrugBank version 2.5 (DBv2.5) [113] provides three interactions between chlorpromazine and single protein targets; none of these interactions explain the antihistaminic affects of the drug. Structurally, chlorpromazine is very similar to the antiemetic trimeprazine. DBv2.5 captures an interaction between trimeprazine and the Histamine H1 receptor, a known target for antihistamines. Through guilt-by-association, we can, therefore, predict the Histamine H1 receptor as a target for chlorpromazine, an interaction captured in the latest editions of the DrugBank database. Fig. 4.2B describes a situation whereby a compound, structurally similar to a compound with a known target, may also bind to the same target (the inference is represented by the dashed red line). The topological and semantic properties of the depicted subgraph describe a repositioning relationship that could be generically applicable to any two drugs and a target. This real example can, therefore, be used to derive a template semantic subgraph that can be used to search Drug-Target (D-T) associations involving different drugs and targets. The template semantic subgraph, therefore, describes a pattern indicative of a drugs interaction with a target, highlighting potential new indications for the drug.

Fig. 4.2 shows a simple triad, however, semantic subgraphs that represent potential repositioning opportunities are likely to be highly complex. In the context of drug repositioning, manual identification of potential repositioning opportunities from large target networks is possible, though not efficient for systematic analysis. Instead, automated approaches to the identification of mappings of a particular semantic subgraph can allow for large-scale exploration of target networks.

To identify instances of semantic subgraphs, a means of identifying all topological mappings of the subgraph is first required. Identifying mappings which maintain the topological properties of a predefined subgraph in a target network is known as the *graph matching problem* [137].

4.2.2 The graph matching problem

There are different variations of the graph-matching problem. For example, *exact matching* occurs when the mapping between the nodes of the two graphs is *edge-preserving*; a mapping contains all edges defined by the query. One of the most stringent forms of exact matching is *subgraph isomorphism* [149] which aims to find all occurrences of a query graph and is a Nondeterministic Polynomial Time (NP)-complete problem [150]. A decision problem is defined as NP-complete when it is both NP and NP-hard, which essentially means that no fast solution is known.

Subgraph isomorphism is a task in which two graphs, G & Q are given as input, and one must determine whether G contains a subgraph that is isomorphic to Q: is there a subgraph $G'(V', E') : V' \subseteq V, E' \subseteq E$? During the search of a query graph, a mapping

(M) is expressed as the set of ordered pairs (v,m) (with $v \in V(G)$ and $m \in V(Q)$) and so $M = \{(v, m) \in V(G) \times V(Q) | v \text{ is mapped onto } m\}$; that is $M : G' \mapsto Q$.

Currently, algorithms addressing this problem are exponential in performance relative to the size of the input graphs [137]. Many algorithms which have been developed to address the subgraph isomorphism problem are based on the exhaustive algorithm developed by Ullman (1976) [151]. Applying an exhaustive method to the identification of drug repositioning opportunities is important to ensure all possible novel applications for a drug are investigated. Using a backtracking approach, Ullman's algorithm finds solutions by incrementing partial solutions or abandoning them when determining they cannot be completed [151]. Many algorithms addressing the problem of subgraph isomorphism build on Ullman's work [151]. These applications include: GraphQL [152], GADDI [153] and, one of the most efficient, the VF algorithm [17]. Performance is increased in these algorithms by exploiting different join orders, pruning rules and auxiliary information to prune negative candidate subgraphs as early as possible. Furthermore, extensions of the Ullman approach which incorporate the semantics of a graph have also been developed. These extensions have been implemented using inexact [154], as well as exact approaches [155, 156]. However, none of these approaches has been adapted to aid the automated identification of drug repositioning opportunities.

Since subgraph isomorphism algorithms are only concerned with matching topology, they must be extended to also consider semantics, if they are to be used to identify mappings of a semantic subgraph successfully.

4.2.3 Semantic similarity

A simple semantic subgraph, Q, made up of two nodes and a single edge can be used to describe a drug that is marketed to treat a particular disease, consisting of a `Compound` node and an outgoing edge of type `may_treat` linked to a node of type `Disease`. Without considering semantic similarity, an exact search for Q in G would return all exact mappings, i.e. all instances whereby a `Compound` and a `Disease` share a `may_treat` edge. However, if one were able to apply a less stringent semantic matching the same search could also return instances whereby a node of type `Compound` and a node of type `Disease` share a `may_prevent` edge. Therefore, we can see that when searching for semantic subgraphs, it is important to consider the semantic similarity between the query subgraph and the potential mapping, as well as topological similarity. By introducing a semantic similarity measure, mappings of a semantic subgraph whose semantics are similar but not necessarily exact may be achieved. Semantic similarity becomes even more valuable as semantic subgraphs increase in size. Mappings of larger semantic subgraphs that contain a single entity with a differing type (to that described in the semantic subgraph being searched for) are less likely to produce a subgraph representing a different functionality to that captured in Q.

Therefore, a measurement of semantic similarity between elements of a mapping and the equivalent element in a query must be introduced to the search. A degree of similarity can be expressed as a semantic distance. Numerous measures have been developed to score the semantic similarity between two ontological concepts [157, 158]. Previous work in the area of intelligence link analysis has used ontology-based semantic similarity scoring methods for pattern matching [159]. In Seid and Mehrotra's algorithm, an inexact topological search is carried out with matches semantically scored based on their Least Common Ancestor (LCA) within an ontology. Topological and semantic scores are then combined and k ranked matches returned. Although these measures are fine when considering approaches that utilise standardised ontologies it is likely that non-obvious semantic types may be included in the network— these types may not be captured in a standard ontology. It is, therefore, important to enable semantic similarities scores to be manually curated via a domain expert. In this instance the semantic similarity scores can be represented, in the simplist form, as an $n \times n$ matrix.

While approaches described are adequate for their particular setting, here a new exhaustive graph matching approach for the identification of semantic subgraphs relevant to drug repositioning is presented.

4.3 Materials and methods

DReSMin was implemented as a Maven project in Java and is available for download from www.bitbucket.org/ncl-intbio/dresmin. The project makes use of the JGraphT library[1] v0.8.3 for *in-memory* representation of the target network and query semantic networks. As such, a target graph is first converted to a JGraphT format before querying. JGraphT uses a bespoke Java representation of graph objects and can, therefore, be serialised and deserialized using the native Java serializable interface.

During semantic subgraph splitting, the shortest path is calculated using the implementation of Dijkstra's shortest path algorithm [160] provided by the JGraphT library. DReSMin also provides a simple visualisation framework for: presentation of the underlying data structure during a search; semantic subgraphs; as well as user specified subcomponents of the target network (e.g. nearest neighbourhood), visualisation is achieved using the GraphStream library[2] v1.2.

4.4 Results

An algorithm for the detection of semantic subgraphs, DReSMin, was developed to allow integrative networks to be searched for semantic subgraphs. The algorithm returns all mappings of a semantic subgraph that match at a level equal to, or above a given semantic threshold. DReSMin was specifically designed to work with semantically-rich target networks which possess particular properties. Specifically, DReSMin requires a *directed* graph (edges have a direction associated with them) where nodes and edges are given types from T_v and T_e respectively. T_v and T_e may be drawn from a finite hierarchy of types H, and can be annotated with attributes. The algorithm allows for multigraphs and for nodes to contain self-loops.

4.4.1 The DReSMin algorithm

DReSMin is an exhaustive algorithm for the detection of mappings of a predefined query semantic subgraph in an integrated target network. Mappings that are identified are topologically identical to the query graph, but may differ semantically depending on a user defined threshold. DReSMin requires three inputs: a target graph (G); a semantic subgraph (Q), that will be searched for in G; and a Semantic Threshold (ST),

[1] www.jgrapht.org

[2] www.graphstream-project.org

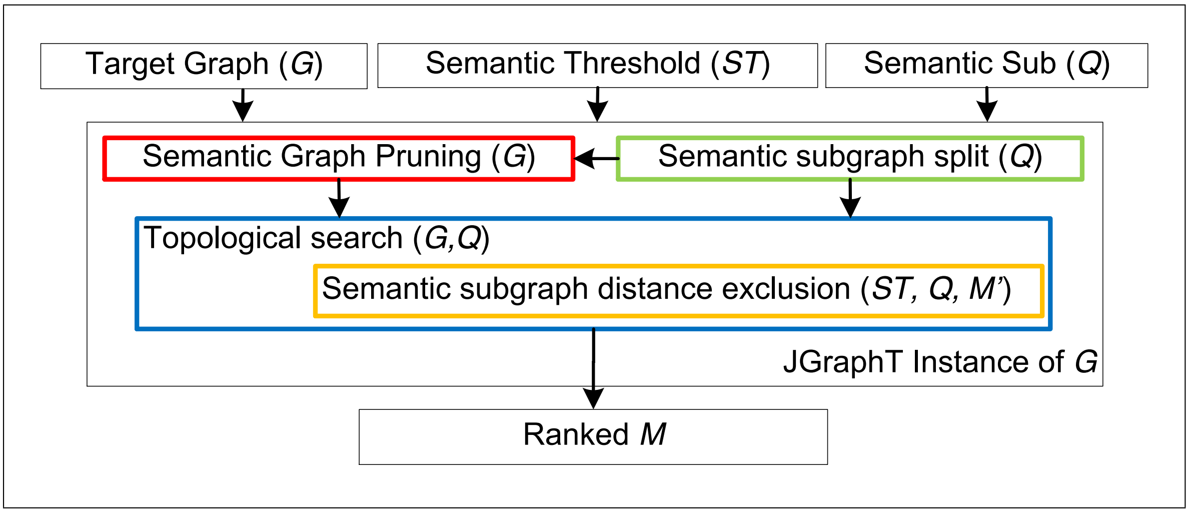

Target Graph (*G*)
Semantic Threshold (*ST*)
Semantic Sub (*Q*)
Semantic Graph Pruning (*G*)
Semantic subgraph split (*Q*)
Topological search (*G,Q*)
Semantic subgraph distance exclusion (*ST, Q, M'*)
JGraphT Instance of *G*
Ranked *M*

Figure 4.3: **Overview of the DReSMin algorithm developed for the detection of semantic subgraphs.** DReSMin requires three inputs: a target graph (G); a query semantic subgraph (Q); and a semantic threshold (ST). The DReSMin algorithm is made up of four main components: semantic graph prune (red); topological search (blue); semantic subgraph distance exclusion (yellow); and semantic subgraph splitting (green). The output of DReSMin is a set of ranked inferences.

ranging from 0-1. DReSMin is made up of four main components which are described in Fig. 4.3. These components comprise: (i) Semantic graph pruning (ii) Topological search (iii) Semantic subgraph distance exclusion (iv) Semantic subgraph splitting. DReSMin may be executed in two modes. The first requires that every element that is added to a mapping must be greater than the ST. The second allows for elements that are lower than the ST to be added to potential mappings, as long as the cumulative score for a mapping is greater than the ST.

4.4.1.1 Semantic graph pruning

The approach presented is concerned with identifying semantic subgraphs that match, semantically, at a level equal to, or above a semantic threshold. (Note: The semantic distance between two graph entities is calculated using the semantic distance calculator described in the 'Semantic subgraph distance exclusion' Section on page 96). In the graph pruning component of the algorithm, any nodes (and their associated edges) in G, which are above a certain semantic distance from all of those in Q, are removed from G. This step allows any nodes that are semantically distant from the query to be removed from the target graph prior to a search, cutting down the initial search space. Taking G, Q and an ST each $t_v \in T_v(Q)$ are sent to the semantic subgraph distance calculator (termed Semantic Distance Calculator (SDC) and described on page 96), and scored against every $t_v \in T_v(G)$. If $SDC(t_v(Q), t_v(G)) < ST$ then all $v \in V(G)$ of type t_v are removed from G as well as any $e \in E$ where $v = v_i$ or $v = v_j$. Finally, after all semantically insignificant elements are removed from G, all disconnected $v \in V(G)$ that may have resulted from the edge pruning step are also removed.

4.4.1.2 Topological search

Here, topological matching is carried out using a variation of the VF algorithm [17]. The VF algorithm is exhaustive and suitable for working with 'large' graphs (up to 3×10^4 nodes), employing a depth-first strategy implemented in a recursive fashion [17]. During a search using the VF algorithm, the search space is minimised via the introduction of topological pruning rules [17]. Integrated networks typically surpass the aforementioned 'large' graphs in size, particularly true within the biological and pharmaceutical settings. As data volumes continue to grow (e.g. omics technologies continue to mature) it is important to develop exhaustive algorithms capable of scaling with the data.

An initial implementation of the VF algorithm showed poor scalability and so, as an enhancement to the VF algorithm, three steps to improve the efficiency of searching for topological subgraphs were developed. These three steps are: a set of rules used to determine the appropriate nodes at which an instance of the search is started (initial candidate set), as described in (1) below; a topological pruning rule, based on a closed world assumption, as described in (2) below; and a semantic thresholding step (described in Section 4.4.1.3). Although DReSMin can be applied to any setting, the focus of this thesis is on the identification of new indications for existing compounds. It is, therefore, essential that mappings of semantic subgraphs, in this work, contain a compound.

1. When considering an initial candidate set of nodes from the target graph G at which to initiate the search, it is desirable to ensure that the set consists of nodes of a type, X, such as `Compound`; ensuring relevant portions of the graph are being searched. Therefore, to identify an initial candidate set for the search,

the highest connected node, v, of type X, from Q, is first identified. Next, all nodes from G that have a degree greater than or equal to v, whilst also being of type X (an exact semantic type match), are included in the initial candidate set.

2. When mining with a given semantic subgraph that describes a potential repositioning situation it is assumed that the lack of a relationship between nodes indicates the absence of a relationship between the two nodes (a closed world assumption). As a result, when searching for a given semantic subgraph, Q, we only consider a match if there exists no additional edges between the nodes in a mapping M from the target graph G and their equivalent nodes in Q. Therefore, a mapping M is expressed as a set of ordered pairs and the closed world assumption requires $(M = match) \vee (deg(v) \in V(G) \equiv deg(m) \in V(Q))$.

4.4.1.3 Semantic subgraph distance exclusion

Semantic thresholding is used to exclude matches found in G that are below a given semantic distance from Q. This process is achieved through a semantic subgraph distance calculator (SDC). An SDC is graph specific and comprises of two distance matrices, one for the nodes types captured in G ($t_v \in T_v(G)$) and one for the edge types captured in G ($t_e \in T_e(G)$). The $n \times n$ matrix for the node types ($n = |T_v|$) and the $m \times m$ for the edge types ($m = |T_e|$) are each represented as $P' = [p_{ij}]$. Matrices contain manually curated scores ranging from -1 to 1, with a score of 1 returned when both entities are of the same type, 0 returned if the entities have unrelated semantic types (such as a `Protein` and a `Publication`), and -1 returned if the entities are semantically opposite (such as the edge types `has_function` and `has_not_function`). The values used in the matrices are defined in Equation 4.1.

$$p_{ij} = \begin{cases} 1 & \text{if } p_i \text{ is semantically identical to } p_j; \\ 0 & \text{if } p_i \text{ is semantically unrelated to } p_j; \\ -1 & \text{if } p_i \text{ is semantically opposite to } p_j. \end{cases} \tag{4.1}$$

During the matching process each element of $M = (V_m, E_m)$ is scored against its equivalent in $Q = (V_q, E_q)$ using the two matrices via a call to the SDC. The resulting

Semantic Score (SS) of M is defined in Equation 4.2.

$$SS(M,Q) = \frac{\sum_{i=1}^{n} SDC(m_i, q_i)}{n} \tag{4.2}$$

An ST is defined by the user prior to a search as a value ranging from 0 (semantics in mappings are not required to be similar to those in the query— essentially a purely topological search) to 1 (semantics in mappings must be identical to those in the query; essentially a trivial task).

4.4.1.4 Semantic subgraph splitting

This component of DReSMin takes a semantic subgraph, Q, and returns a set of semantic subgraphs, D, whose $|V| < 4$. Semantic subgraphs in D share an overlapping node (ON). Semantic subgraph splitting aims to address, or at least bypass, the exponential increase in the time taken for subgraph isomorphism algorithms to search for larger subgraphs. This step of DReSMin is described in detail in Algorithm 1.

In Fig. 4.3, we see how this step is applied in context with the other components of DReSMin. The graph splitting component allows smaller subgraphs to be searched, before post-searching steps merge mappings that share a common ON. The most connected node, $v_{max}(Q)$, is first identified and used as ON. Q is then checked to see if it is either a clique (fully connected network) in which case the graph is randomly partitioned. Next, Q is converted to an undirected graph. Of all the remaining $v \in V(Q)$, the two most distant nodes (v_1, v_2) from Q are selected. Two new graphs $(D_1 \& D_2)$ are then created and populated with nodes as such: $V(D_1) \cup v \in \delta(v_1, ON), V(D_2) \cup v \in \delta(v_2, ON)$, that is every node in the shortest path from v_1 to ON is included in D_1 and every node in the shortest path from v_2 to ON is included in D_2. Remaining nodes are then allocated using the following rules:

1. If a node shares an edge with only one of D_1 or D_2, it will be allocated to that subgraph.
2. If a node shares an edge with nodes from both D_1 and D_2, it will be assigned to the subgraph with which it shares the greatest number of edges.

Algorithm 1 Graph Splitting. Algorithm takes a semantic subgraph (Q) and returns a set of semantic subgraphs (D) whose node set is less than four. Q is first converted to an undirected graph before it is checked for a clique, if Q is a clique then splitting is completed randomly. Next, an overlapping node (ON) is identified. The two most distant nodes in Q are then identified (v_1 and v_2) and new subgraphs (D_1 and D_2) are populated with all nodes that fall between ON and v_1 and v_2, respectively. Finally, all left over nodes are allocated, and D returned. If the node set of either D_1 and D_2 are still greater than four, then the algorithm may be called iteratively.

```
Input: Semantic Subgraph, $Q$
if $|V(Q)| > 3$ then
    $D = \{Q\}$
else
    if cliqueCheck ($Q$)  then
        cliqueSplit ($Q$)
    else
        $ON = v_{max}(Q); v_1; v_2; max = -1$
        $Q$ is converted to an undirected graph
        for $i \in V(\forall j(j \in V \land i \neq j))$ do
            if shortestPath $(i, j) > max$ then
                $max =$ shortestPath $(i, j); v_1 = i; v_2 = j$
            end if
        end for
        $D_1 =$ all nodes in $\delta(v_1, ON)$
        $D_2 =$ all nodes in $\delta(v_2, ON)$
        for $v \in V$ do
            if $v \notin (V(D_1) \cup V(D_2))$ then
                allocate left over nodes
            end if
        end for
        $D = \{D_1, D_2\}$
        while $|(V)d| \in D > 3$ do
            return GRAPH SPLITTING($d$)
        end while
    end if
end if
Output: $D$
```

3. If a node shares an equal number of edges with nodes in D_1 and D_2, it will be designated to the subgraph containing the fewest nodes.

4. If a node shares an equal number of edges with nodes in D_1 and D_2 and D_1 and D_2 have an equal number of nodes, it will be allocated to D_1.

Edges are then distributed as such: $\forall e \in E(Q)$ if either $V(D_1)$ or $V(D_2)$ contains both (v_i, v_j) of e; e is allocated to that graph. When v_i, v_j are not present in the same graph then e is not included in the split graphs. Instead, e is checked for during the post-searching steps, which are described below. Splitting may be called iteratively if either D_1 or D_2 still possess a $|V| > 3$ after the first round of splitting, as demonstrated in Fig. 4.4.

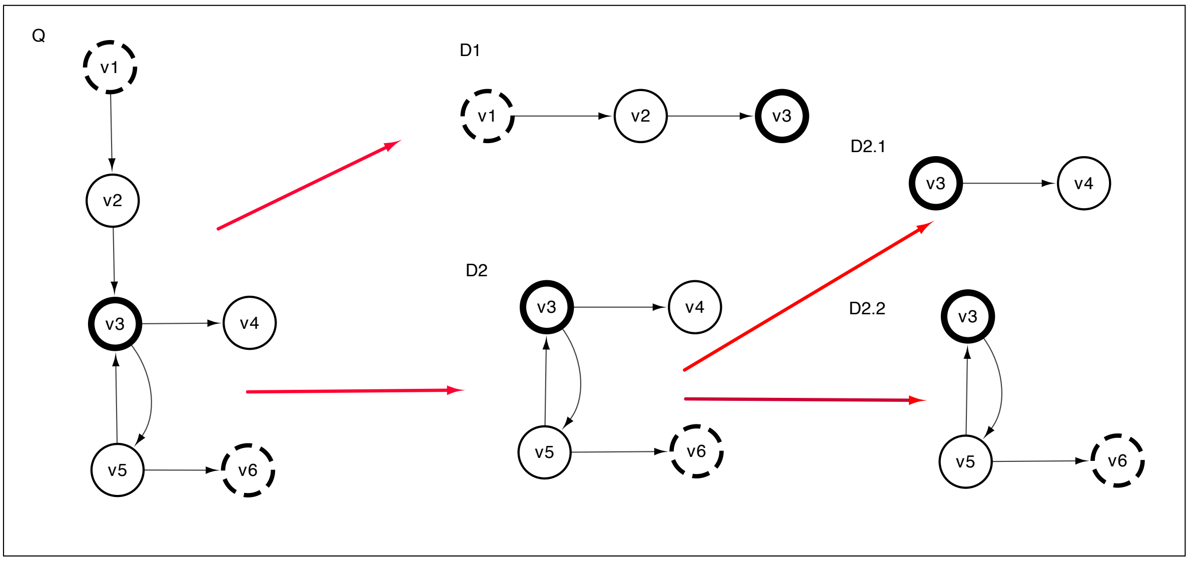


Figure 4.4: **Subgraph split procedure takes an initial semantic subgraph (Q) and produces two smaller semantic subgraphs (D_1 and D_2) using all nodes ($\boldsymbol{v}$) from (Q).** The overlapping node (ON) is identified in Q (v_3) and used as the overlapping node in both D_1 and D_2. The two most distant nodes in Q are then identified (v_1 and v_6) and nodes in the path between these and ON added to the corresponding graphs (D_1 and D_2). We also see that $|V(D_2)| > 3$ and so a second call is made to graph split giving us D_21 and D_22.

As a result of this process two graphs are produced, D_1 and D_2 as well as the original semantic subgraph, Q. A search is then initiated with D_1 or D_2, depending on which has the smallest $|V|$ (ensuring the quickest search is completed first). The first search is started using ON, maintaining the edgeset it possessed in Q; reducing the initial candidate set. The edgeset ON possesses in Q will be greater than the edgeset it possesses in the split graphs (D_1 and D_2), as Q is the larger subgraph, and ON is the highest connected node from this subgraph. Therefore, by using the edgeset ON possesses in Q, fewer candidate nodes will be identified in the target graph, resulting in the search being initiated at fewer locations. All starting nodes that lead to a mapping being identified in the first search are then passed to the second search; reducing the initial candidate set once more by excluding nodes from the starting set that did not lead to a mapping in the first search.

After a split search has been completed, the mappings from D_1 and D_2 must be consolidated in order to return mappings reflective of the original query graph, Q. This process is done using two post-searching steps:

1. The first post-searching step merges all mappings of D_1 and D_2 that share an *ON* and returns M, the set of potential mappings of Q found in target graph, G.

2. The second post-searching step involves re-scoring the merged mappings and making some final checks. This post-searching step requires: (1) all merged mappings in M from the first post-searching step (2) the original query graph, Q, and (3) $\forall\ m \in M$, the edgeset that each node in m possesses in the target graph, G. In this step, $\forall\ m \in M$, some final checks are carried out. First, a check is made to ensure any $e \in E(Q)$ that were not allocated to either D_1 or D_2 in the graph split step are present in G; done using the edgesets from G. m is also checked to ensure that the closed world assumption still holds after merging (described on page 2), again using the edgesets from G. If m passes these initial tests it is then re-scored, using the SDC (Section 4.4.1.3).

4.4.2 Defining a semantic threshold

To identify an ST that would allow for semantically similar mappings of a semantic subgraph to be returned during a search, while also limiting this set to those that still capture the desired functionality represented in the semantic subgraph. To do this, first of all, a means of measuring the Semantic Simplicity (SE) of a semantic subgraph Q, was developed:

$$SE(Q) = \frac{\sum_{v_i,v_j \in V(Q), v_i \neq v_j} SDC(v_i, v_j) + \sum_{e_i,e_j \in E(Q), e_i \neq e_j} SDC(e_i, e_j)}{|V(Q)| + |E(Q)|} \tag{4.3}$$

The SE (Equation 4.3) utilises the SDC described previously (Section 4.4.1.3) and allows for a measure from 0-1 to be calculated for a semantic subgraph. This score is based on the cumulative distance of all node types against all node types ($\sum_{v_i,v_j \in V(Q), v_i \neq v_j} SDC(v_i, v_j)$) plus the cumulative distance of all edge types against

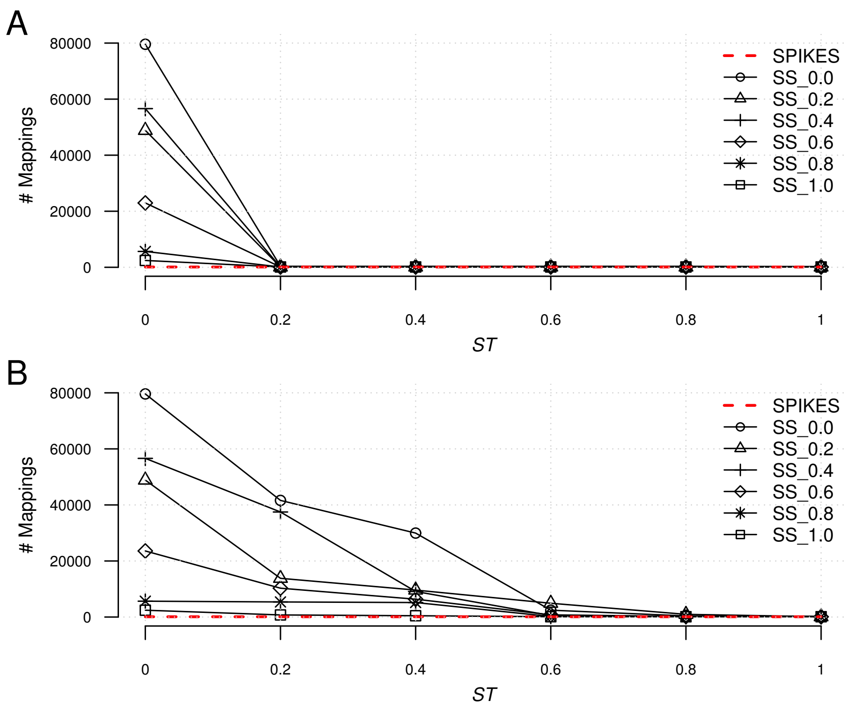


Figure 4.5: **Calculating semantic threshold (ST) to be used during searches.** *Note:* Semantic subgraphs with semantic simplicities ranging from 0-1 were created with five replicates for each point. The ST at which only the 100 spikes were returned was determined. The algorithm was executed on random graphs with a node set size of 1 x 10^3 using two alternate parameters where (A) every element of the match needed to pass the ST, and (B) all elements had to pass the ST cumulatively. When ST reached 0.8 the # mappings, not members of the spiked subgraphs, was reduced dramatically. Subgraphs were created at random with node sets between three and six.

all edge types ($\sum_{e_i,e_j \in E(Q), e_i \neq e_j} SDC(e_i, e_j)$) divided by the total number of elements (nodes plus edges) in the subgraph ($|V(Q)| + |E(Q)|$). For example, if Q contained three nodes, of which all were of type `Protein` and six edges, of which all were `similar_to` then the SE score would be 1; the semantic types of elements in the subgraph are the same for all nodes and the same for all edges. If however, the three node semantic subgraph were made up `Protein`, `Compound` and `Disease` with three edges, `binds_to`, `involved_in` and `has_indication` the SE score of this semantic subgraph would be 0; all elements are semantically different and return scores of 0 when compared using the SDC.

The SE allowed for the effects of altering the ST before a search to be investigated. In Fig. 4.5 we can see that an ST of 0.8 allows for semantically similar mappings to be returned while also limiting these dramatically when comparing to other SEs. Therefore, it was decided that an ST of 0.8 would be used as the default for searches using DReSMin. It is also noted that the ST becomes increasingly important as the SE is reduced; less spurious mappings are evident with an increased SE. An ST of 0.8 was, therefore, used during the characterisation and performance section.

4.4.3 Characterisation and performance

To characterise the performance of DReSMin random semantic target graphs (Ran) as well as random semantic subgraphs were produced in order to evaluate the performance of the semantic subgraph search strategy. These random graphs were formulated using an approach that attempted to replicate the semantic and topological properties of the integrated drug repositioning data set, Dat, described in Chapter 3. In these random target graphs $\forall v \in V(Ran)$ of type t_v, the average $deg^{-}(t_v)$ and the average $deg^{+}(t_v)$ were maintained $\forall t_v \in T_v(Dat)$.

For performance analysis, DReSMin was implemented on a 20 node Ivy-Bridge bioinformatics cluster. The SDC and graph-pruning step display linear running times of $O(n)$. The SDC was found to be capable of scoring 8×10^4 concept pairs per second and the graph pruning step took < 1 second to prune a graph G, with $|V(G)|$ of 1×10^6. Furthermore, the effectiveness of each step of DReSMin was calculated by adding each step (initial candidate set selection, topological pruning and semantic distance thresholding) sequentially to the basic topological search algorithm and then comparing the efficiency of each modified version to the VF2 topological search. The performance was measured as the time taken for a complete search for a semantic subgraph, Q, within a given target graph, G. Experiments were repeated 10 times and presented in Fig. 4.9.

A comparison of the number of initial nodes used during a standard DReSMin search and a semantic subgraph split search is presented in Fig. 4.6. In Fig. 4.6 it is shown that when searching for subgraphs with a $|V(Q)|$ of 3, the initial candidate set size is around 36% lower when using a graph split approach as opposed to a standard DReSMin search, when looking for the first split graph. It is also shown that the initial candidate set is reduced, on average, by 5% after the first search in a split search is completed (the second search in a split search is started on 5% fewer nodes than the first search). An analysis of search time and the number of mappings returned, when using a standard DReSMin search compared to a semantic subgraph split search, is also presented in Fig. 4.7, showing that the number of mappings returned using a graph split search is identical to the number of mappings identified when using a standard DReSMin search while a reduced search time is achieved. Furthermore, a comparison of search time when using a topological, a standard DReSMin search and the semantic subgraph split search for semantic subgraphs with a $|V(Q)|$ of 7 is presented in Fig. 4.8.

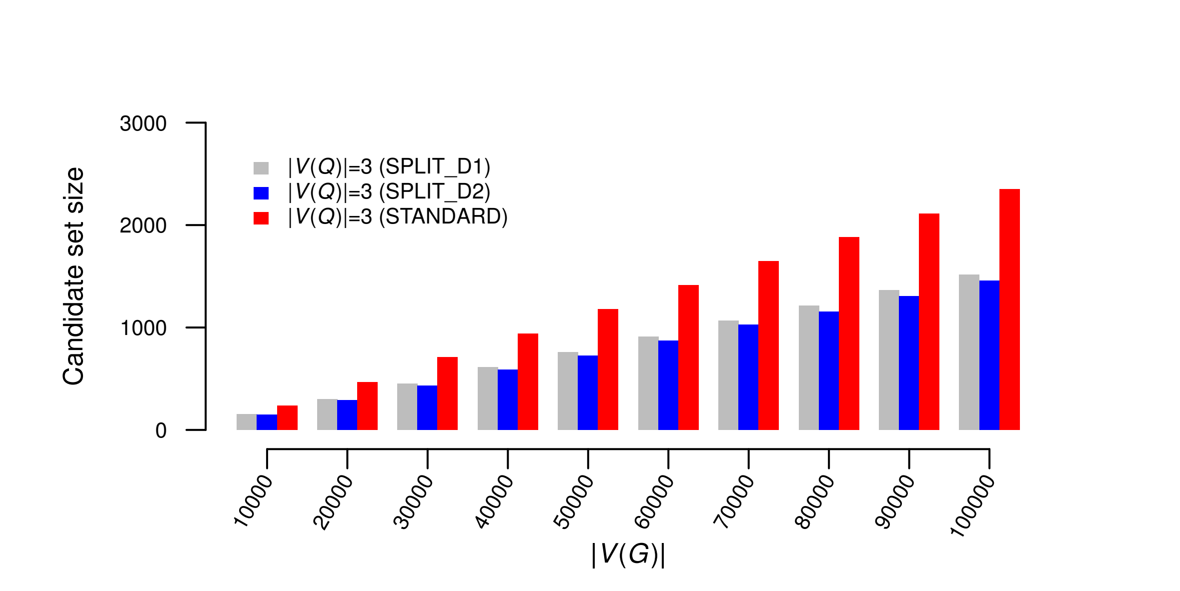


Figure 4.6: **Initial candidate set sizes when using the graph split algorithm compared to a standard DReSMin search.** Target graphs were created at random where $|V(G)|$ ranged from 1×10^4 and 1×10^5. Semantic subgraphs with a $|V(Q)|$ of 5 were created at random. A graph split search was then completed, which resulted in two subgraphs with a $|V(Q)|$ of 3 (D_1 and D_2). The initial candidate set size for both of the split graphs were counted during the split graph search ('SPLIT_D1' and 'SPLIT_D2'). A standard DReSMin search was then completed for both D_1 and D_2 and the initial candidate set counted and the average taken ('STANDARD'). Experiments were repeated 5 times.

The effect on search time when altering semantic subgraph edgeset size was also examined and is presented in Fig. B.1. It is shown that when searching for a semantic subgraph with a $|V(G)|$ of 4 and a $|V(E)|$ edgeset of 6 DReSMin performs better than when searching for a semantic subgraph with the same $|V(G)|$ but with fewer edges.

To test the impact of target graph connectivity and target graph size on performance, semantic subgraphs were created at random with a $|V(Q)|$ of between 3 and 6 and

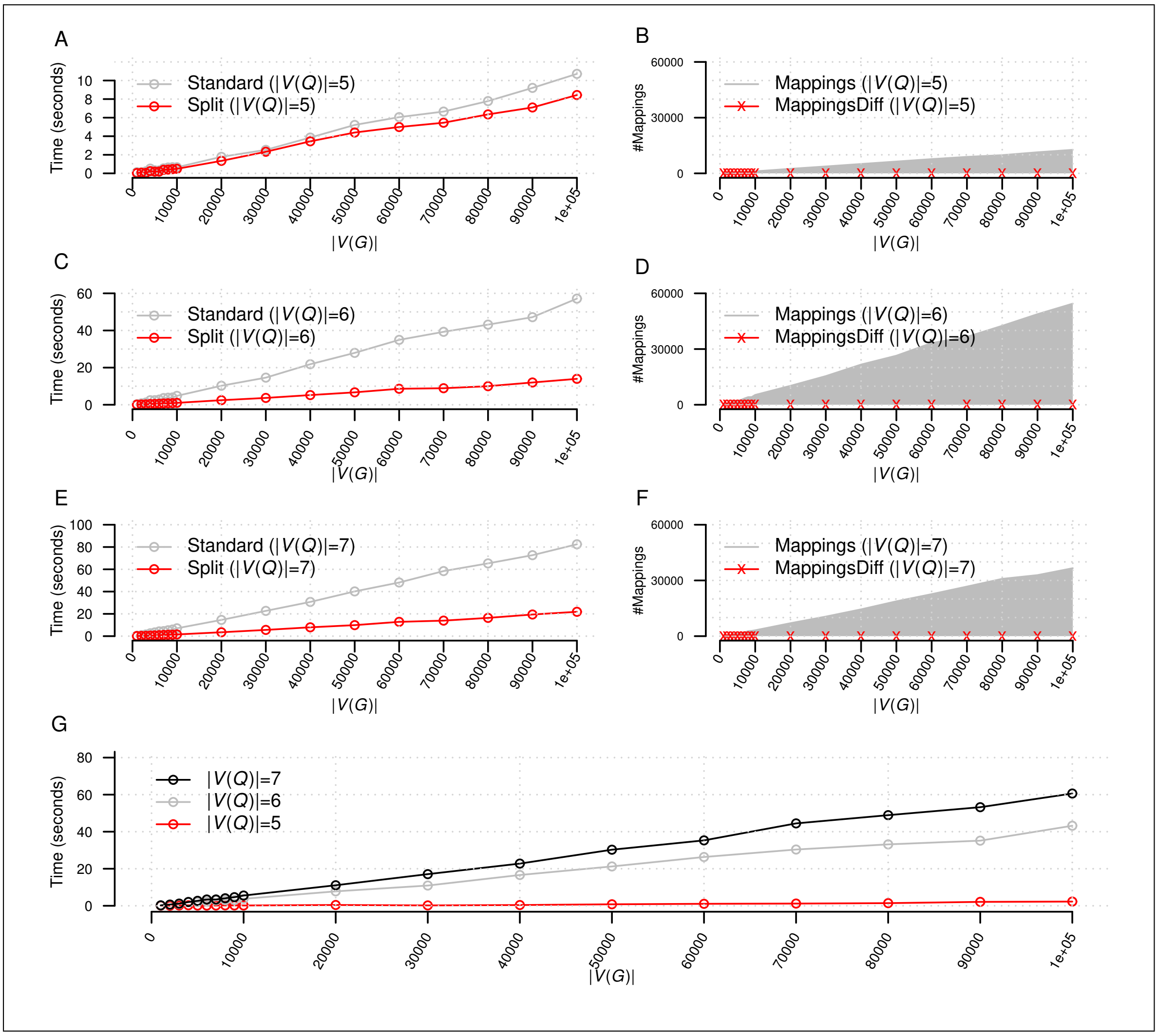


Figure 4.7: **Overview of the performance and accuracy of the graph split algorithm compared to a standard DReSMin search.** Semantic subgraphs were created at random with a $|V(Q)|$ of 5, 6 and 7. Target graphs were created at random where $|V(G)|$ ranged from 1×10^3 and 1×10^5. Semantic searches for the random subgraphs were then completed using a standard DReSMin search and a graph split search (both using an ST of 0.8). (A, C and E) The average time taken for a normal search (grey) and a graph split search (red) to be completed. (B, D and F) The average number of mappings found for each random subgraph (grey) and the difference in the number of mappings returned by a normal search and a graph split search (red). (G) Performance increase observed when using a graph split search in comparison to a normal search for semantic subgraphs with a $|V(Q)|$ of 5, 6 and 7.

searched for in the target network with i. altering graph connectivity and ii. altering target graph node set size. The target network was then spiked with 100 of the semantic subgraphs and searches repeated. It was shown that the number of mappings returned before and after spiking of the target graph differed only by the number of spikes.

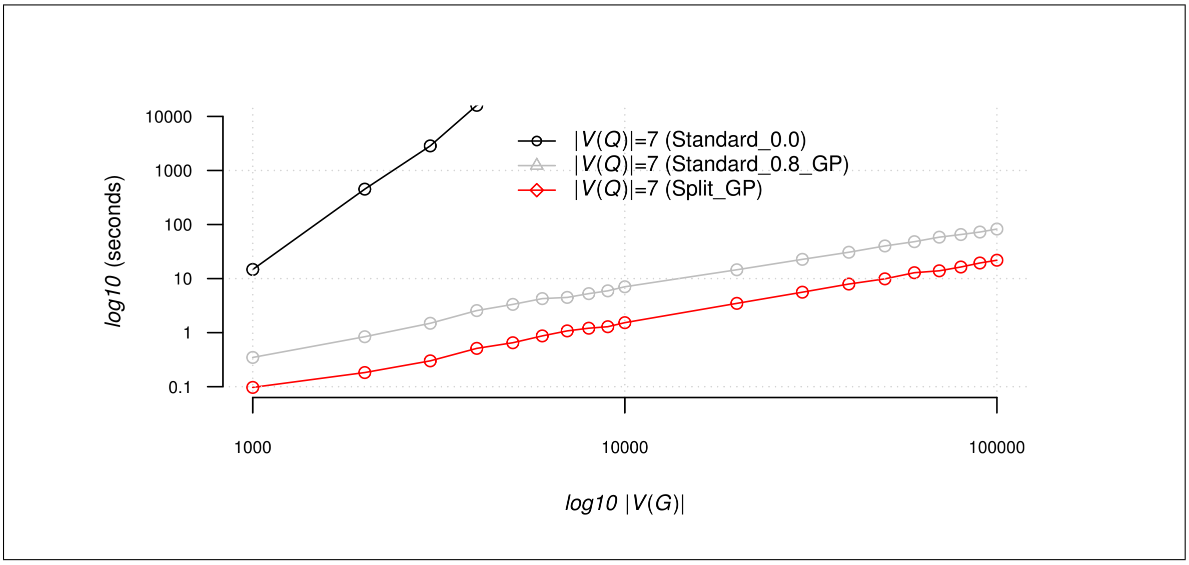


Figure 4.8: **Comparison of topological (Standard_0.0), non-split (Standard_0.8_GP) and graph splitting (Split_GP) search times.** Semantic subgraphs were created at random with a $|V(Q)|$ of 7. A purely topological search was then completed using the DReSMin algorithm (Standard_0.0). Semantic searches were then carried out using graph prune, without semantic graph splitting (Standard_0.8_GP), and finally, with the semantic graph splitting (Split_GP). Searches were carried out on random target graphs where $|V(G)|$ was between 1×10^3 and 1×10^5. Note: *GP* = graph prune was used.

4.5 Discussion

In this section the performance of DReSMin will first be discussed. Considerations for future work will then be introduced before the approach is described in the context of the drug repositioning field.

Fig. 4.9 shows each novel step of the algorithm reducing search time and highlights some key observations regarding the performance of DReSMin. First, when looking at the initial candidate set it is seen that once semantic subgraphs reach a $|V(Q)|$ of 4 then restricting the initial candidate set to include only `Compounds` improves performance. It is at this point the benefits of reducing the initial candidate set successfully reduce the search space, concomitantly increasing performance. A similar phenomenon was observed with the introduction of the closed world check, whereby the real performance benefits were apparent when semantic subgraphs reached a $|V(Q)|$ of 4. By restricting the initial candidate set, as well as using the closed world assumption a two-fold increase in performance in comparison to a purely topological approach was

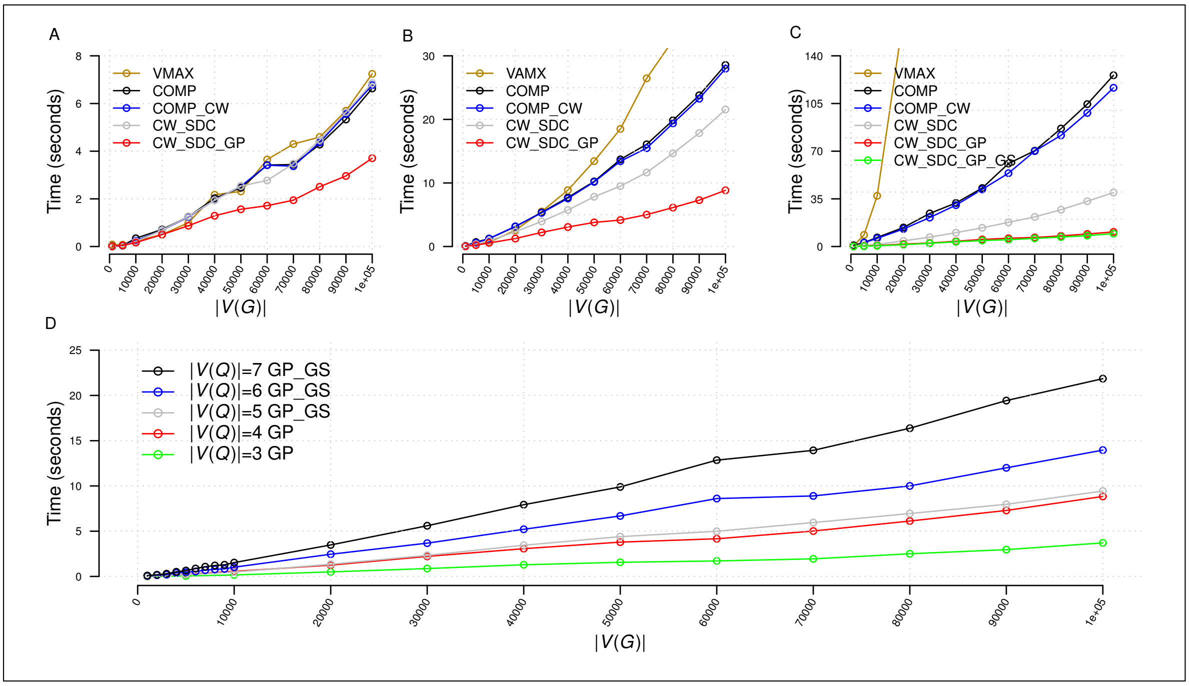


Figure 4.9: **Overview of algorithm performance with semantic subgraph (Q) queries node set $|V(Q)|$ ranging from 3-5.** (A) DReSMin performance when searching for a semantic subgraph whose $|V(Q)|$=3. (B) DReSMin performance when searching for a semantic subgraph whose $|V(Q)|$=4. (C) DReSMin performance when searching for a semantic subgraph whose $|V(Q)|$=5. (D) Best performance of an exact, exhaustive search using DReSMin for semantic subgraphs with a $|V(G)|$ of 3,4,5,6 and 7 respectively. Abbreviations: $VMAX$ = initial candidate set contained the highest connected nodes in Q, regardless of the type, COMP = `compound` makes up the initial candidate set, CW = closed world check implemented, SDC = semantic distance calculator used during search, GP = semantic graph prune step implemented, GS = graph split used.

observed. Performance was further enhanced when utilising the SDC, demonstrating an almost three-fold performance boost when compared to the topological approach. Furthermore, the semantic graph prune step introduced around a 50% increase in performance to DReSMin. However, the graph prune step also introduced a further subtle cost; any potential matches containing an element that scores $< ST$ when passed to the SDC are not returned. Semantic graph pruning is most useful when one wishes to return matches that are semantically exact to the semantic subgraph being used as Q, or when all elements of the match need to be above the ST (as opposed to the average score of all elements having to be above the ST).

When using all of the above steps as well as the graph split step of DReSMin the search time for a semantic subgraph, Q, with a $|V(Q)|$ of 6 can be reduced to one closer to the sum of a search for a subgraph with a $|V(Q)|$ of 3 and a subgraph with a $|V(Q)|$ of 4. For example, when using the SDC to search for Q, where $|V(Q)| = 6$ in G when $|V(G)| = 1 \times 10^5$, takes 60 seconds (Fig. 4.7). Using the graph split method reduces this search time to just under 14 seconds, a 4 fold increase in performance. This observation is further supported by the data presented in Fig. 4.8 which demonstrates that searching a target graph with a node set size of 2 x 10^3 for instances of a subgraph with a $|V(Q)|$ of 7 using a purely topological approach (Standard_0.0) is intractable. The performance of the algorithm was improved when searching using a standard DReSMin search with an ST of 0.8 (Standard_0.8_GP), however, when searching for a semantic subgraph where $|V(Q)| = 7$ in a target graph where $|V(G)| = 1 \times 10^5$, still takes 100 seconds. When using the graph split method during a DReSMin semantic search with an ST of 0.8 (Split_GP), a search for semantic subgraphs with a $|V(Q)|$ of 7 could be completed on a large target graph (1 x 10^5) in just over 20 seconds. These results show that the semantic graph splitting step, when used in conjunction with the other steps in DReSMin, allows larger subgraphs to be used as queries which were previously intractable using an exact exhaustive approach.

Although making use of the graph split step of DReSMin can enable an increase in performance by dramatically reducing the search space investigated during a DReSMin search (Fig. 4.6), some caveats must be discussed. Although the results presented in Fig. 4.7 show that a graph split approach returns exactly the same mappings as those achieved using a standard DReSMin search, there are potential situations where a graph split approach would not return the same set of mappings as a standard search. Firstly, there are two settings in which the DReSMin algorithm can be executed. The first setting states all elements in a mapping must be greater than the ST (the setting that has been investigated in this Chapter), whilst the second setting simply states that the average score of all elements of a mapping must be above a the ST. The performance analysis presented in Fig. 4.7 was done using the former setting. If a graph split approach was to be run whereby the average score of elements had to be greater than the ST then the graph split approach would not accurately capture the same mappings that would be identified using a standard search, and in such a situation would become a heuristic. The reason for this discrepancy is the fact that the split subgraphs will contain fewer elements, and so potential mappings of large queries containing low scoring entities will be lost, with low scoring entities affecting the average score of smaller graphs more than larger ones. Secondly, the initial candidate set for a split search is populated with nodes from the target graph that are semantically identical to the highest connected node in the query graph (i.e. the same type). This will likely become a problem when the semantics of a graph become more complex than those of *Dat*, described in Chapter 3 and used as the test bed for the work presented in this Chapter. If the *ON* is of a type that has many semantically similar types in the network then there are potentially many different semantic types that could map to this node in any potential mappings. As a standard search starts with nodes of type `Compound`, if a split search were to start on a different node type in a more semantically complex network, mappings returned by a split search and a standard DReSMin search would not be identical.

Perhaps counter intuitively, an improvement in performance was observed as the edge-set size of a semantic subgraph also increased. The performance increase is likely due to fewer nodes satisfying the more stringent topological rules. With more stringent pruning during a run of the algorithm the search space at each state is reduced. A reduction in search space means that a search involving a semantic subgraph with more edges will take less time to complete in comparison to a semantic subgraph of the same node set size but fewer edges.

Overall, when using all of the algorithmic steps in DReSMin, the algorithm showed performance characteristics approximating a linear scale, close to $O(n)$. The performance of DReSMin is in contrast to the exponential scaling characteristics observed for the purely topological search algorithm, VF2. Using DReSMin, with the previously described hardware, it was possible to complete an exact, exhaustive search for a six-node semantic subgraph in a target graph containing $> 1.5 \times 10^5$ nodes in just over 14 seconds. It was also shown, as expected, that the accuracy of the algorithm (it is exhaustive) does not decrease as the target graph connectivity, or $|E|$, increases or as the target graph $|V|$ increases. Although the performance of DReSMin is impressive, there are still limitations to the approach that need to be discussed.

DReSMin scores semantics based on types of nodes and edges; it may also prove beneficial to include scoring metrics based on node and edge attributes as well as the data sources from which they are retrieved. For example, during the process of data integration, data sets could be assigned a quality score providing a measurement of confidence in a given interaction or attribute. Such modifications would allow the scoring of semantic subgraphs to be not only topological and semantic but also based on the reliability of the source of each element.

As well as introducing additional scoring metrics to the approach the semantic subgraphs being searched for also require considerable thought. As described, these semantic subgraphs can be drawn from real life repositioning examples via manual curation. Manually developing semantic subgraphs is time-consuming, however, they may allow for the creation of more accurate representations of a functional module representative of a drug repositioning opportunity than those automatically developed. A library of semantic subgraphs curated from real world examples of repositioned drugs would be beneficial to the approach.

Concerning the mining algorithm, as new graph mining frameworks emerge with efficient graph searching algorithms (e.g. Neo4J), it may be possible to exploit these built-in algorithms to implement sections of DReSMin. However, necessarily, the nature of these implementations will depend on the specific graph database.

Unlike other approaches to drug repositioning, semantic subgraphs may be designed to infer relations between any node types in a data set. DReSMin enables the discovery of any relations that are captured in the abstracted drug-disease connection (as presented in Fig. 2.1) so long as suitable semantic subgraphs are identified. This is in contrast to many of the computational approaches to drug repositioning described in Section 2.2.2. For example, ligand structure-based approaches as well as protein structure-based are limited to the inference of drug-target associations; gene expression-based approaches along with genetic variation-based approaches have been limited to the application of the inference of drug-disease associations; similarly phenotype-based approaches (including disease and side-effect) tend to be used for the inference of drug-disease associations. Although machine learning based approaches, such as those described in Section 2.2.2, tend to take a more integrative approach to data and, therefore, have the ability to infer a wider range of associations and even properties, the fact that this is done using statistical 'black boxes' means that interpreting the repositioning hypothesis is a difficult task. Using semantic subgraphs and DReSMin allows for human interpretable hypotheses to be derived and for all evidence supporting a claim to be judged. It is hoped that the systems approach described here will allow for a more accurate, holistic, systematic approach to drug repositioning.

4.6 Conclusion

In this Chapter, a formalised framework for the definition of semantic subgraphs, connected sub-components of a semantically-rich target network was developed. Identifying these subgraphs may aid in the inference of novel interactions not immediately evident in the target network. DReSMin, an algorithm for searching integrated networks for occurrences of a given semantic subgraph using semantic distance thresholds was also presented. DReSMin optimises the search time for larger subgraphs by including a novel semantic graph pruning step and applying a method for splitting large semantic subgraphs into a set of smaller subgraphs before searching. The optimisations presented make searching for instances of complicated semantic subgraphs computationally tractable and scalable. Furthermore, the approach presented in this Chapter is not limited to the field of drug repositioning. DReSMin can be used to search for semantic subgraphs representing functional modules from any area of research, so long as the graph properties described in Section 4.4 are maintained.

As described, future work would involve extending the algorithm to consider more than the abstract semantic 'types' that the work presented supports. Furthermore, as graph-based algorithms are implemented in well supported projects, it may prove beneficial to utilise these, for example as a replacement for the topological searching steps in DReSMin. Unlike other computational approaches to drug repositioning, DReSMin is capable of inferring edges of any semantic type from an integrated semantic network.

To investigate the usefulness and utility of DReSMin in the field of drug repositioning, the algorithm was used to search for relevant semantic subgraphs. First of all, in Chapter 5, DReSMin is applied to the task of inferring novel D-T interactions; with the top ranked associations used to infer novel uses for the drug. Secondly, in Chapter 6, DReSMin is used to enable the inference of novel Drug-Disease (Dr-D) associations from a Gene-Disease (G-D) centric integrated network. It is shown in these exemplars that DReSMin can be used successfully to identify and rank novel interactions. Furthermore, in the case of D-T interaction inference it is shown that DReSMin is capable of outperforming a state-of-the-art methodology developed purely for this purpose (as opposed to DReSMin which can be applied to any area of interest).

5

Application of DReSMin to the identification of novel drug-target interactions

5.1 Introduction

Drug discovery has moved on from the once accepted 'one drug, one target' paradigm. In fact, the therapeutic efficacy of many drugs, including psychiatric and modern anticancer therapies are reliant on their 'promiscuity' [39, 161]. Drug promiscuity, or polypharmacology, describes the property of a drug to act on multiple molecular targets and exhibit distinct pharmacological effects [39].

The first association in the drug-target-phenotype-disease pathway (described in Fig. 2.1) is between a drug and its target. Prediction of these Drug-Target (D-T) associations plays a pivotal role in the drug discovery process [162]. Inference of such associations allows for novel uses of a drug to be identified as well as potentially adverse side-effects to be highlighted [163, 164]. *In vitro* approaches to determining D-T interactions are no different to other aspects of drug development and remain costly and time-consuming [162, 165]. It is infeasible to screen all drugs manually against all genes in the human genome (over 21,000), and so automated computational approaches are desired [39]. Using systematic *in silico* prediction methods allows for the D-T interaction search space to be reduced, highlighting areas of focus [164]. Many computational approaches have been described to aid in this task and are introduced below.

Molecular docking methodologies are heavily applied to D-T interaction prediction. These methods aim to give a prediction of the drug-target (or ligand-receptor) complex structure using computational methods [166]. However, molecular docking approaches require a significant amount of computational resources, are time-consuming and are known to return high numbers of false positives [165]. Other approaches involve the application of machine learning, which may use a feature vector approach (see background) or, more commonly, similarity-based approaches which exploit the similarity between drugs and proteins [162, 165]. Similarity-based approaches allow for the production of predictive models, and such approaches can be ligand-similarity-based, target-similarity-based, or a hybrid of the two [167]. For example, ligand information may be used to create models that learn which sub-structural features of a ligand correlate with activity against a particular target [168]. Other similarity-based approaches make use of a network, or more specifically a bipartite graph, data representation [164, 165, 169–171]. Within a bipartite graph, nodes are divided into two disjoint sets, proteins and drugs. Data from multiple publicly accessible data sets are integrated during the construction of these networks [172]. However, in most approaches to D-T interaction prediction, data is limited to the inclusion of the two data types; drugs, and their targets. Limiting data to these two types can restrict accurate prediction of D-T associations. As such, there is a belief that drug discovery may be improved by taking a systems approach, considering the interaction of existing drugs with target proteins as well as other biological molecules [93]. Protein targets, which make up the majority of drug targets, can be categorised into multiple classes.

The majority of current primary drug targets are: (1) membrane-bound proteins (including G protein-coupled receptors (GPCR) and ion channels); (2) enzymes (including kinases and proteases); and (3) nuclear hormone receptors [37, 64]. Membrane-bound proteins account for 60% of drug targets [39], with approximately 40% of all marketed drugs said to target a GPCR [173]. Although GPCR make up only 4% of the human genome, the prominent role they play in many pharmacological processes means they represent the most abundant class of validated pharmacological targets [136]. Ion channels are popular targets for pharmacological intervention with high potential. However, due to the plethora of physiological activities controlled by ion channels, drug development in this area remains challenging, with many drugs targeting this class displaying poor selectivity, suboptimal efficacy and even significant toxicity [174]. Like the membrane-bound proteins, enzymes, such as kinases and proteases, also make up a large proportion of drug targets [136]. One of the largest families of evolutionarily related proteins are the protein kinases which make up around 2% of the human genome. Kinases are highly conserved, involved in many cellular processes and comprise 20% of putative D-T associations [175]. Proteases make up 5-10% of targets being pursued in drug development [176]. Of the target classes introduced GPCR and ion channels are well investigated, with an abundance of information available to describe their actions and properties [177]. However, others, such as kinases, have less specific data available (the term specific is used here as there is an abundance of data regarding kinases because they are oncology targets, but most is general).

In the work described in this Chapter, the utility of Drug Repositioning Semantic Mining (DReSMin) was demonstrated, with a focus on the inference of novel D-T associations. Using the drug repositioning data set, *Dat*, a method for the automated identification of 194 semantic subgraphs is presented. Furthermore, the semantic subgraphs were then used to infer novel potential D-T interactions from the same data set, using DReSMin. Central to the approach described is the use of historical D-T data. *Dat* includes D-T interactions from DrugBank v2.5 and so more recent versions of DrugBank data sets were utilised in this work (version 3 for the creation of semantic subgraphs and version 4.2 for validation). It was shown that DReSMin could be successfully used to predict putative D-T interactions that were not explicitly represented in the *Dat* and that cover multiple protein target classes. Furthermore, the potential value, to the drug repositioning field, of inferences made by the approach can be investigated as and when they are made, as opposed to the future. The work presented in this Chapter has also been described in Mullen *et al.* [146].

5.2 Materials and methods

The integrated drug repositioning data set, *Dat*, introduced in Section 3.4.1 was used as the target network for this work. A Java implementation of DReSMin, as described in Chapter 4, was used for searching *Dat* for instances of semantic subgraphs. A semantic threshold of 0.8 was used during searches (see Section 4.4.2). All analysis code was written in a Java Maven project and is available for download, along with scored D-T associations and scored semantic subgraphs from https://bitbucket.org/ncl-intbio/dresmin. To determine the shortest semantic paths, *Dat* was converted to an undirected graph and a Java implementation of Dijkstra's shortest path algorithm [160], from the JGraphT[1] library used. Three versions of DrugBank were used in this work; 2.5, 3.0; and 4.2[2].

Mappings between DrugBank and ChEMBL compounds were retrieved from UniChem [133] via whole source mapping[3]. UniChem maps chemical identifiers from

[1] www.jgrapht.org

[2] www.drugbank.ca/downloads/archived

[3] www.ebi.ac.uk/unichem Accessed:22-06-2015

multiple sources if they share the same Standard InChI. The mapping from UniChem provides a set of 3,765 drugs that are contained in both data sets, of which 57 of the ChEMBL ids mapped to >1 DrugBank ID (one to four, five to three, and 51 to two). For each drug that DReSMin inferred D-T associations for, the top 100 protein target predictions were extracted from the ChEMBL Web resource client[4].

As the approach described in this chapter is exhaustive, it is not enough to simply count the number of true positives identified as a means of validation. It was, therefore, important to identify a statistical test that evaluated the ability of the approach to 'prioritise' these knowns. One such approach is the hypergeometric distribution, a discrete probability distribution. The hypergeometric distribution describes the probability of k successes in n draws without replacement, from a finite population of size N that contains exactly K successes, wherein each draw is either a success or a failure. During validation, inferred interactions were first ranked based on the scores that they achieved. Using a sliding window of x interactions, the `phyper` function (from the statistical computing language 'R') was then used (lower.tail set to false) to calculate the $P[X > x]$ at each position of the sliding window. The $P[X > x]$ represents the probability of identifying more true positives than were actually observed at a given position in the ranked list. By using this measure, it was possible to evaluate the ability of the approach to consistently score known D-T interactions more favourably.

For target class comparison, five human protein target classes were identified based on their sizes and importance, as described by [136]. Proteins were classified as one of the following: GPCR; ion channels; kinases; proteases; and other. To do this, the same approach described by [136] was used. Protein family membership was determined using multiple protein sources. The first is the ID attribute of a `keyword` (k) element within a UniProt[5] entry, E. All keywords assigned to E are captured in the set K. If "KW-0297" in $E(K)$ then E is classed as a GPCR; if "KW-1071", "KW-0851", "KW-0107", "KW-0869", "KW-0407", "KW-0631" or "KW-0894" is in $E(K)$ then E is classed as an ion channel; if "KW-0418", "KW-0723" or "KW-0829" is in $E(K)$ then E is classed as a kinase; if "KW-0031", "KW-0064", "KW-0121", "KW-0224", "KW-0482",

[4] www.github.com/chembl/chembl_webresource_client Accessed:22-06-2015

[5] www.uniprot.org Accessed:30-07-2015

"KW-0645", "KW-0720", "KW-0788" or "KW-0888" is in $E(K)$ then E is classed as a protease; and finally all other proteins are classed as 'other'. A protein is also classified as a GPCR, kinase or protease if it appears in the UniProt GPCR[6], kinase[7] or protease[8] family files, respectively.

5.3 Results

An approach to identify novel D-T associations from an integrated target network was developed as part of this project. The approach is made up of three main components which are: (i) Automating the development of semantic subgraphs (ii) Mining an integrated net drug repositioning network, *Dat*, for instances of the semantic subgraphs and (iii) Ranking inferred interactions. Furthermore, the ability of the approach to identify novel D-T associations was evaluated in four aspects, comprising: (i) A comparison of the inferences made to those made by a state-of-the-art method for D-T association prediction (ii) A comparison of the ability to infer associations involving numerous target (protein) classes was carried out (iii) One of the highest ranked D-T inferences is presented and used to complete the drug-target-disease pathway to enable repositioning hypotheses to be identified and finally (iv) The effect instance information had on the approach was also investigated. Central to the approach described in this section is the availability of historical DrugBank data.

5.3.1 Managing DrugBank data

DBv2.5 was used to construct *Dat* even though later releases of DrugBank are available; v3.0 (DBv3) and v4.2 (DBv4.2) [127]. D-T interactions from DBv2.5 that were integrated into *Dat* were retrieved and captured in the set *DatRel*. DBv3 contains additional drugs, targets and their interactions to those contained in *Dat* (Table 5.1) with 8,768 additional D-T interactions. Of these interactions, 2,919 involve drugs and targets that are present in *Dat*, but the interaction relationship had not yet been defined (i.e. the D-T interaction had not been annotated in DBv2.5).

[6] www.uniprot.org/docs/7tmrlist.txt Accessed:11-11-2015

[7] www.uniprot.org/docs/pkinfam.txt Accessed:11-11-2015

[8] www.uniprot.org/docs/peptides.txt Accessed:11-11-2015

Table 5.1: Drug, target and drug-target (D-T) interactions present in *Dat*, DBv3 and DBv4.2.

	Drug	**Target**	**D-T Interaction**	**Unique**	**Relevant**
***Dat* (DBv2.5)**	4,772	3,037	9,227	-	-
DBv3	6,180	4,080	14,570	8,768	2,919
DBv4.2	6,377	3,601	14,157	8,673	2,940*

Note: Unique: refers to interactions not found in *Dat*, Relevant: subset of Unique interactions, whereby both the drug and target can be found in *Dat*. *Of these 333 are unique to DBv4.2 (i.e they are not captured in DBv3).

The 2,919 interactions from DBv3 are referred to as being 'relevant' as they are of interest to the work. These relevant interactions are represented in the set *DBv3Rel* (Equation 5.1) and were used to derive a query set of semantic subgraphs that were in turn used to mine *Dat*. DBv4.2 was then used as a reference to validate the new repositioning opportunities identified through the mining of *Dat*.

$$DBv3Rel = \{DatRel \cup Unique(DBv3) \mid d \in DatRel(d) \wedge t \in DatRel(t)\} \tag{5.1}$$

Table 5.1 shows 2,940 D-T associations from DBv4.2 as being 'relevant'. These relevant associations are unique to DBv4.2 (in comparison to DBv2.5) whilst also being made up of a drug and a target that is captured in *Dat*. Furthermore, of these 2,940 D-T associations, 333 interactions are not captured in DBv3. These 333 associations are represented in the set *DBv4Rel* (Equation 5.2) and are used in this work to validate inferred D-T associations.

$$DBv4Rel = \{(DatRel \cup Unique(DBv4.2)) \cap DBv3Rel \mid d \in DatRel(d) \wedge t \in DatRel(t)\} \tag{5.2}$$

5.3.2 Automating the development of semantic subgraphs

Semantic subgraphs can be derived through manual exploration of the graph and by reference to known repositioning examples. However, in this work, semantic subgraphs were derived using an automated method. These semantic subgraphs are appropriate

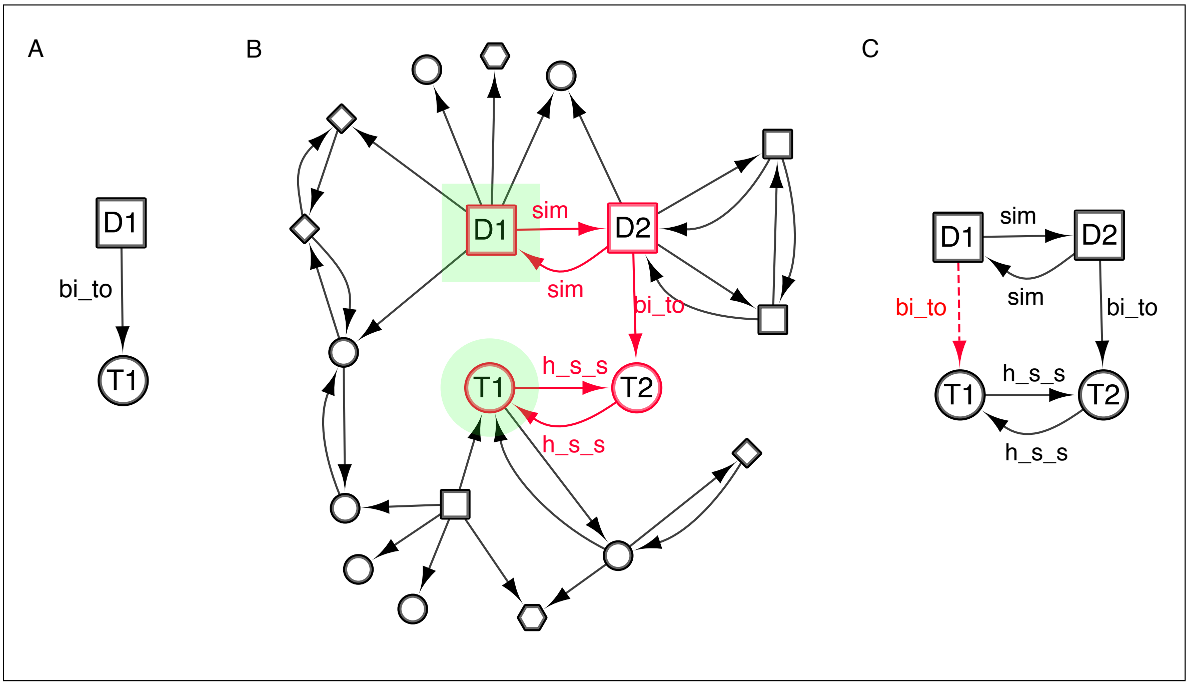


Figure 5.1: **Semantic subgraphs were derived from the semantic shortest paths between a drug and a target pair captured in *DBv3Rel*** (A) A drug-target interaction captured in *DBv3Rel* made up of a drug (D1) and a target (T1) that are both present in the network, *Dat*. To create semantic subgraphs D1 and T1 are located in *Dat* (highlighted in green in (B)) and the semantic shortest paths between the two nodes calculated (highlighted in red in (B)). Finally, all semantic node types and edge types that fall on the semantic shortest path are used to create a query graph (C). *Note:* Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, circle a `Target`, diamond a `Protein` and hectagon a `Disease`. For relation types: bi_to = `binds_to`, sim = `similar_to`, h_s_s = `has_similar_sequence`.

for systematic mining for new D-T interactions. To produce such a set, portions of the *Dat* network that contained drugs and targets from the 2,919 D-T interactions captured in *DBv3Rel* were extracted. To extract the subnetworks, each drug and target pair was identified in *Dat* and the subnetwork represented by the shortest path between them was extracted as a semantic subgraph (Fig. 5.1).

On carrying out the semantic subgraph identification exercise, 194 different subgraphs with a $|V|$ >10 were found cumulatively to identify more than 95% of the D-T interactions in *DBv3Rel*. The 194 semantic subgraphs were used as queries to search *Dat* using DReSMin to test the ability of the algorithm to identify D-T interactions in *Dat* that had not yet been annotated in DBv2.5 (but are present in DBv3). The ten subgraphs that represent the shortest path between the most D-T associations in *DB3Rel*, as well a larger illustrative subgraph, are shown in Fig. 5.2. Furthermore,

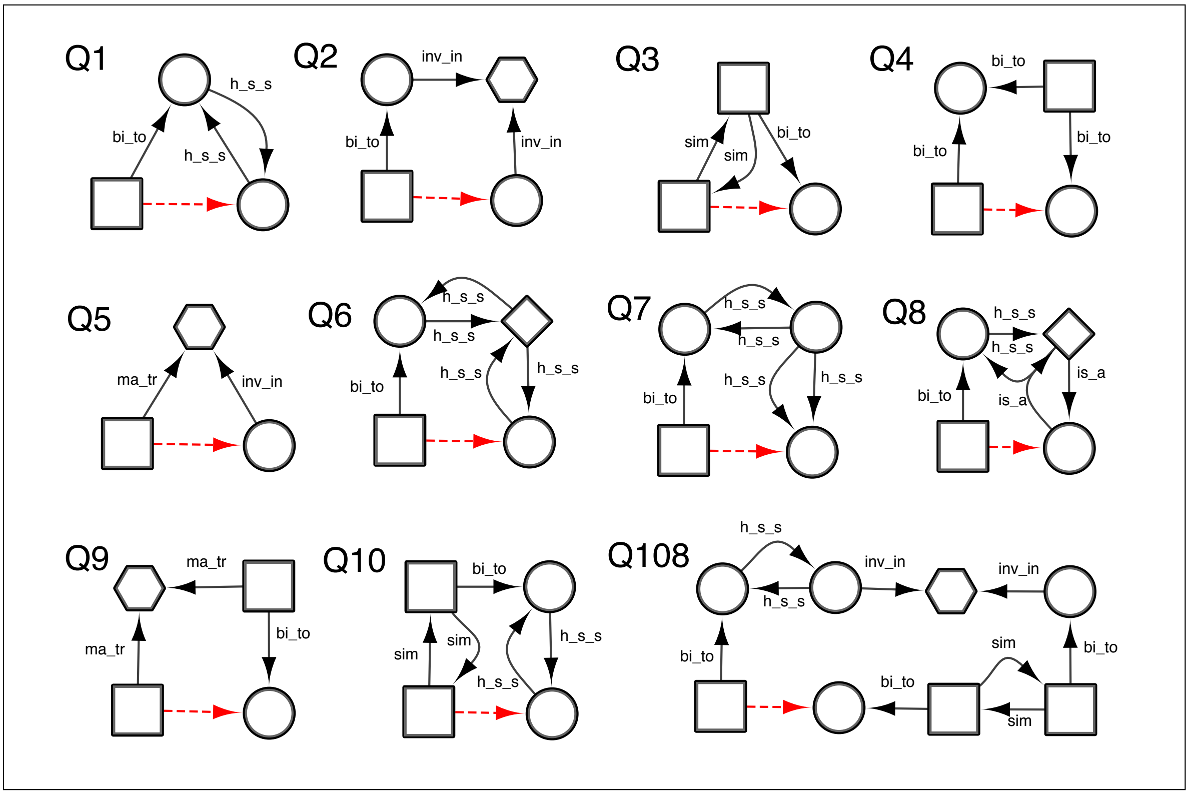


Figure 5.2: **Examples of semantic subgraphs drawn from the semantic shortest paths.** Q1 -Q10 are drawn from the semantic shortest paths that represented the shortest path between the greatest number of D-T interactions in *DBv3Rel* and Q108 is an example of a more complex semantic subgraph. *Note:* Dashed red lines represent the inferred `binds_to` relations, a square represents a `Compound`, circle a `Target`, diamond a `Protein` and octagon a `Diseases`. For relation types: bi_to = `binds_to`, sim = `similar_to`, h_s_s = `has_similar_sequence`, ma_tr = `may_treat`, inv_in = `involved_in` and is_a = `is_a`.

eight randomly selected subgraphs from the set of 194 are introduced and described below.

5.3.2.1 Automated semantic subgraphs in a biological context

In this section, eight randomly selected semantic subgraphs (created using the semantic shortest path approach described in Section 5.3.2) will be presented and analysed in a biological context. Each semantic subgraph will be evaluated based on their likely ability to capture a functional semantic sub-component containing potential D-T associations. Semantic subgraphs were ranked based on the number of D-T pairs in *DB3Rel* (Equation 5.1) for which they represent the semantic shortest path. It is this ranking position that is used to name the semantic subgraphs. For example, subgraph *Q1* represented the semantic shortest path for the most D-T pairs in *DB3Rel* (687) and subgraph *Q194* represented the semantic shortest path for the least D-T pairs in *DB3Rel* (1).

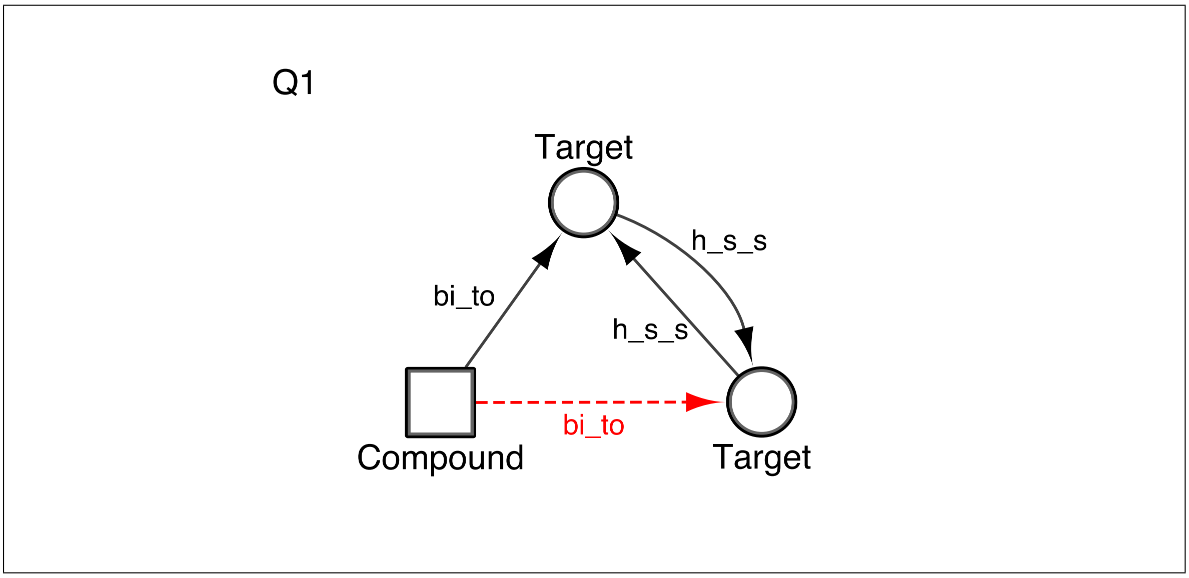


Figure 5.3: **Q1, a three node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound` and circle a `Target`. For relation types: bi_to = `binds_to` and h_s_s = `has_similar_sequence`.

Subgraph Q1, shown in Fig. 5.3, shows a three node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for 687 D-T pairs in *DB3Rel*. This simple triad is similar to the subgraph described in Fig. 4.2, which makes use of the 'similar properties principle' (SPP) [52]. The SPP states that similar molecules should have similar biological activities, affecting proteins and biological systems in similar ways [52]. Instead of looking at the similarity between compounds, subgraph Q1, makes inferences using the assumption that targets with similar sequences will bind the same compound. There exists a multitude of computational approaches to drug repositioning that make use of protein similarity to make novel drug inferences (described in Section 2.2.2.3). It is, therefore, quite reasonable to assume that similarity measures between targets can be used as a means of inferring novel D-T associations.

Subgraph Q4, shown in Fig. 5.4, shows a four node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for 252 D-T pairs in *DB3Rel*. This subgraph uses the assumption that if a drug

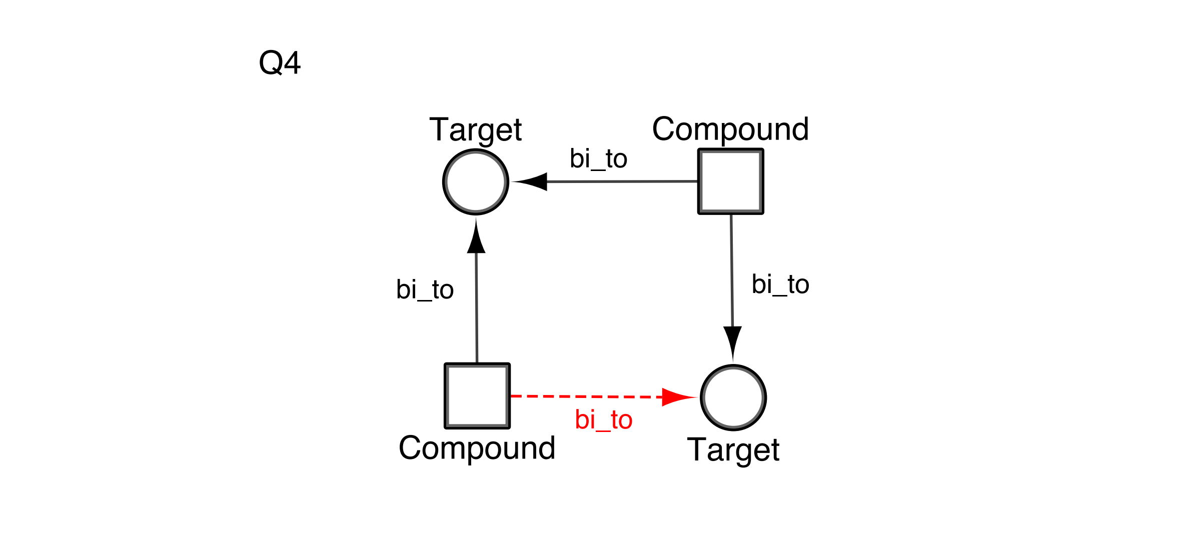


Figure 5.4: **Q4, a four node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound` and circle a `Target`. For relation types: bi_to = `binds_to`.

binds to two targets then these targets are likely similar, or at least possess similar binding sites. Although target proteins will likely possess more than one small molecule binding site, the number of these that are 'useful' in a therapeutic context is often one. These 'useful' binding sites, however, are likely to interact with several compounds. Although the assumption made by this subgraph may not hold true in protein targets with a great number of binding sites, it would for those that have a single target site. Therefore, inferences made using this subgraph have the potential to unveil some interesting potential D-T associations.

Subgraph Q6, shown in Fig. 5.5, shows a four node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for 130 D-T pairs in *DB3Rel*. Like, subgraph Q1 (Fig. 5.3), this subgraph makes inferences using the assumption that targets with similar structures will bind the same compounds. Unlike Q1, however, which makes one similarity 'step', Q6 assumes that this assumption will hold as more similarity steps are added, in this instance two. As the number of similarity steps from the original target to the novel target increases, the likelihood that these two targets share a similar sequence decreases, making the assumption that the sequences are still highly similar less likely. It may be, however, that binding site domains, or other structurally important domains relevant to a proteins ability to bind a compound are maintained as extra similarity steps are

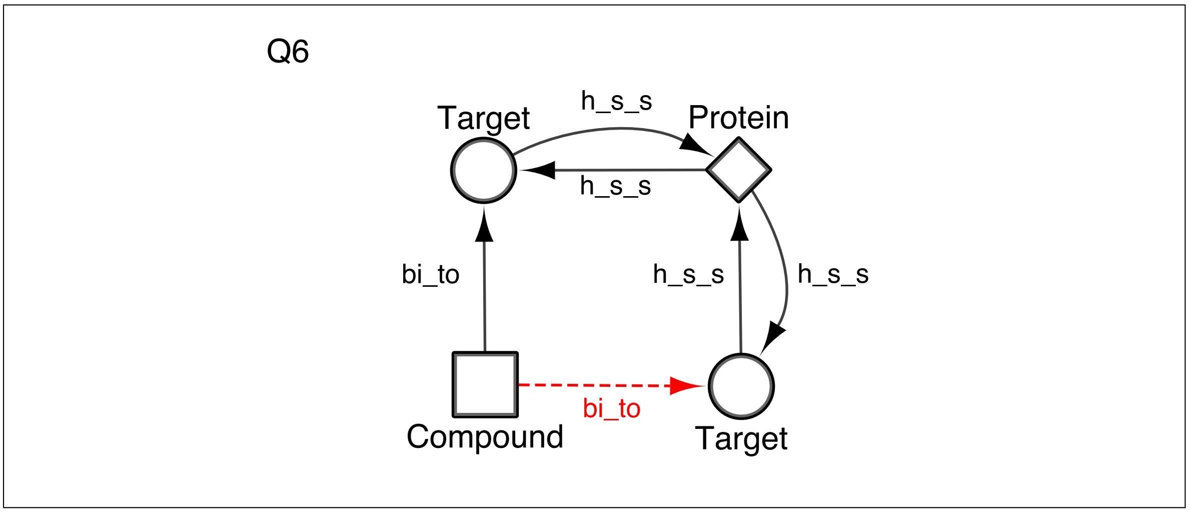


Figure 5.5: **Q6, a four node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, circle a `Target` and diamond a `Protein`. For relation types: bi_to = `binds_to` and h_s_s = `has_similar_sequence`.

added, meaning that subgraphs that make use of these extra steps may still be able to make some interesting inferences that may be missed by semantic subgraphs that use a single similarity step.

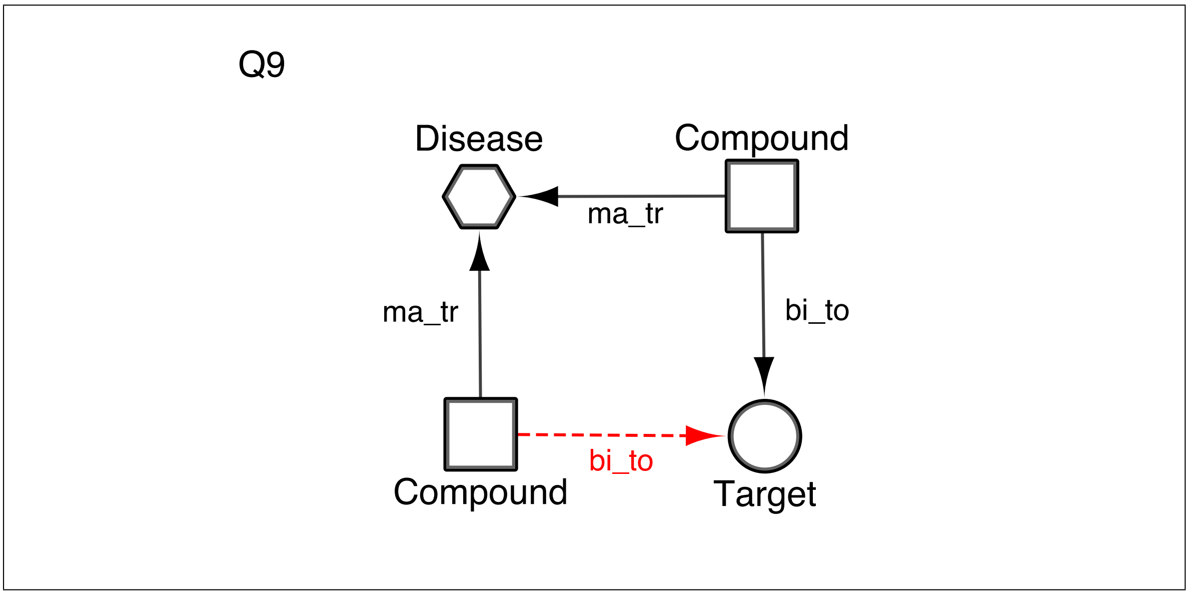


Figure 5.6: **Q9, a four node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, circle a `Target` and hectagon a `Disease`. For relation types: bi_to = `binds_to` and ma_tr = `may_treat`.

Subgraph Q9, shown in Fig. 5.6, shows a four node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for 70 D-T pairs in *DB3Rel*. This subgraph makes D-T inferences using the assumption that drugs used to treat the same disease also bind to the same target. This is a bold assumption, and likely not true, especially in multigenic disorders that tend to have a more complex genetic causation, such as Crohn's. The assumption, may, however, hold true in less complex disorders, that tend to be more monogenic in causation. Cystic fibrosis, for example, is caused by the presence of mutations in both copies of the gene for the cystic fibrosis transmembrane conductance regulator (CFTR) protein. Treatments for Cystic Fibrosis, tend to focus on the CFTR, such as ivacaftor plus lumacaftor (Orkambi; Vertex) [10]. Therefore, in certain situations, this semantic subgraph may be useful for the purpose of D-T interaction prediction.

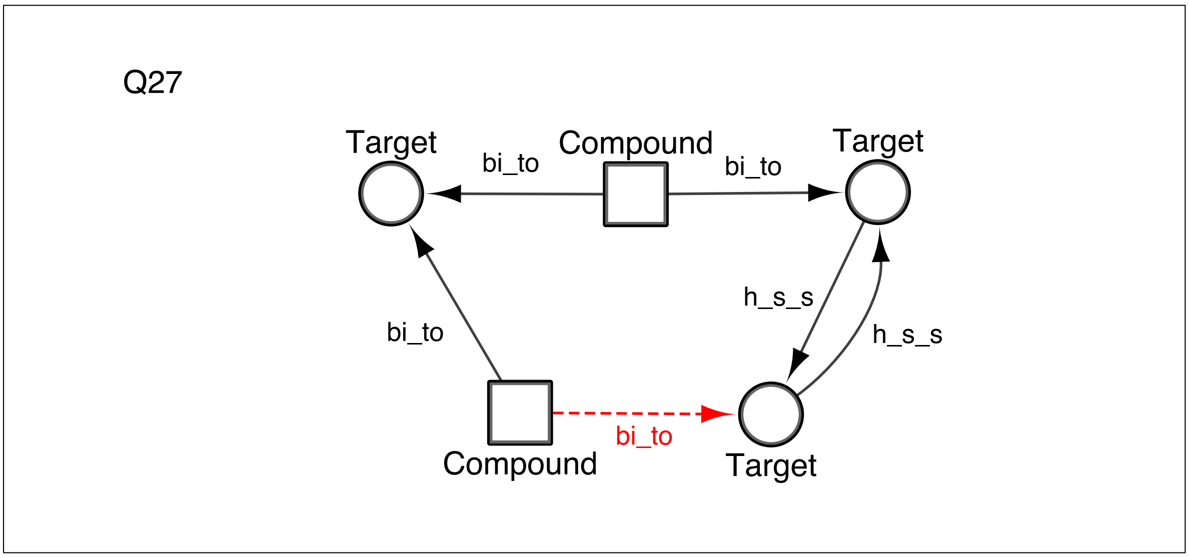


Figure 5.7: **Q27, a five node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound` and circle a `Target`. For relation types: bi_to = `binds_to` and h_s_s = `has_similar_sequence`.

Subgraph Q27, shown in Fig. 5.7, shows a five node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for 11 D-T pairs in *DB3Rel*. This subgraph makes assumptions using the logic captured in two previously introduced subgraphs: (1) the assumption that targets or proteins with similar sequences are likely to bind the same compounds, as described in subgraph Q1 (Fig. 5.3); and (2) the assumption that if a drug binds to two targets they are similar, as captured in subgraph Q4 (Fig. 5.4). Like subgraph Q4 (Fig. 5.4), the serious limitation of this subgraph is the fact that some protein targets are known to have multiple ligand binding sites. It is possible, therefore, that `CompoundX` may bind `TargetY` at `BindingSiteA`, whilst `CompoundZ` may bind the same protein at `BindingSiteB`, in which case the inference made would be incorrect. In the instances where this subgraph identifies protein targets with single binding sites, or at least where compounds are binding to the same binding site, inferences are more likely to be meaningful. It can, therefore, be seen that this subgraph may be able to unveil novel D-T associations.

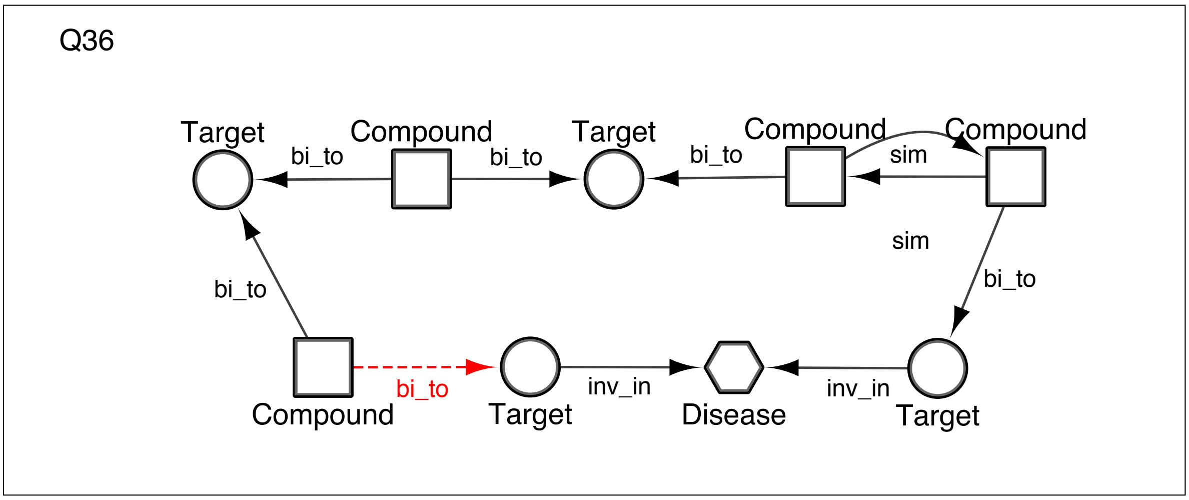


Figure 5.8: **Q36, a nine node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, hectagon a `Disease` and circle a `Target`. For relation types: bi_to = `binds_to`, sim = `similar_to` and inv_in = `involved_in`.

Subgraph Q36, shown in Fig. 5.8, shows a nine node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for six D-T pairs in *DB3Rel*. Like subgraph Q4 (Fig. 5.4), this subgraph makes the assumption that if a drug binds to two targets these targets are likely similar. It also uses the SPP, which states compounds with similar structure will affect proteins in similar ways [52]. Finally, the subgraph makes the rather adventurous assumption that if two targets are involved in the same disease then they are likely similar (sharing properties associated with their ability to bind compounds). Although this may hold true somewhat in certain cancers, where numerous kinases (a target class known to be highly conserved) have been linked to their pathogenesis and progression [178], it is simply not the case in many other diseases. Diseases such as the autoimmune disorder, multiple sclerosis, have a variety of different target proteins associated with their development, such as proteins that make up the major histocompatability complex and various interleukin receptors, to name but a few [179]. Due to the fact that this semantic subgraph makes so many assumptions, it is probably not the most accurate representation of a sub-component containing putative D-T interactions.

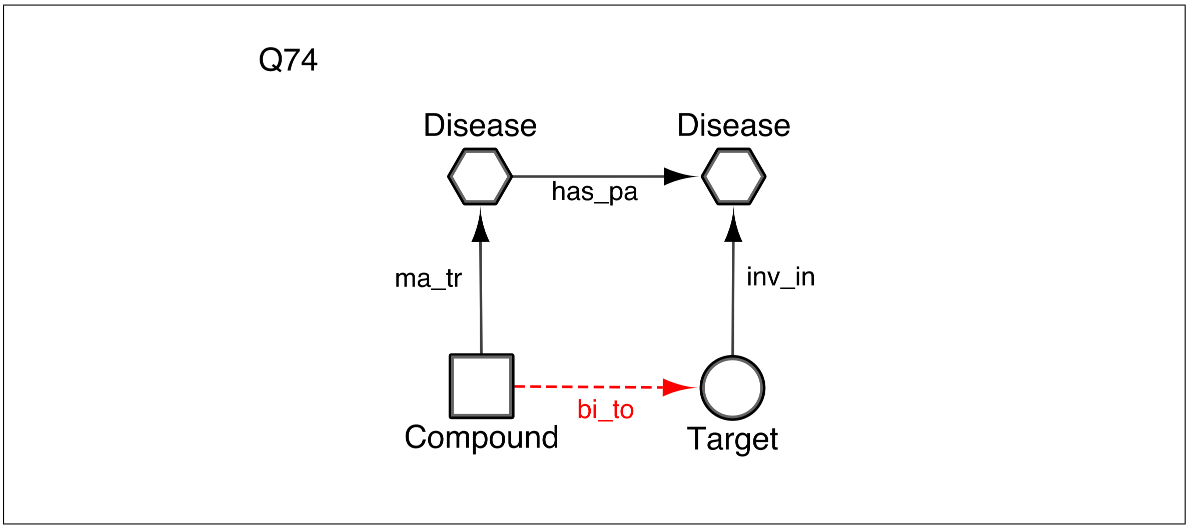


Figure 5.9: **Q74, a four node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, hectagon a `Disease` and circle a `Target`. For relation types: bi_to = `binds_to`, ma_tr = `may_treat`, has_pa = `has_parent` and inv_in = `involved_in`.

Subgraph Q74, shown in Fig. 5.9, shows a four node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for two D-T pairs in *DB3Rel*. Q74 makes D-T association inferences using the assumption that parent diseases have similar targets associated with them when compared to their child diseases (parent and child in terms of their position in the MeSH hierarchy). Although this may be the case in variations of similar, less complex, disorders, it will very rarely be the case in more complex multigenic disorders. It can be seen that the logic behind this semantic subgraph will hold true only in rare situations.

Finally, subgraph Q140, shown in Fig. 5.10, shows a four node semantic subgraph derived using the semantic shortest path approach. This subgraph represented the semantic shortest path for one D-T pair in *DB3Rel*. Q140 makes D-T association inferences by combining multiple assumptions. First of all, like Q1 (Fig. 5.3), this subgraph uses the assumption that targets with similar structures will bind the same

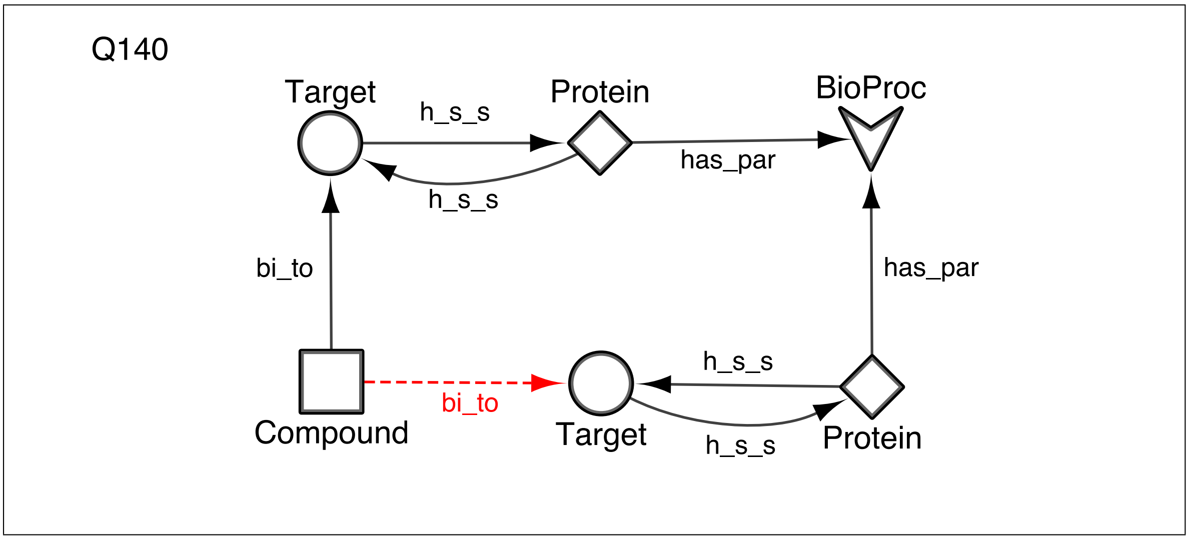


Figure 5.10: **Q140, a four node semantic subgraph created using the semantic shortest path.** Dashed red line represents the inferred `binds_to` relations, square represents a `Compound`, diamond a `Protein`, circle a `Target` and V a `Biological_Process`. For relation types: bi_to = `binds_to`, h_s_s = `has_similar_sequence` and has_par = `has_participant`.

compounds. It then makes the valorous assumption that proteins involved in the same biological process will be structurally similar, before making use of the assumption that targets with similar structures will bind the same compounds once more. Being involved in the same biological process does not necessarily mean structural similarity between proteins. Although, functionally, the proteins will be similar it is a large assumption to state that because of this proteins will likely bind the same compounds.

Individually, some of the semantic subgraphs created using the automated approach described here may not, upon closer inspection, appear to be the best biological representation of sub-components relevant to D-T interaction prediction. It can be seen, however, that in certain situations they may be able to make interesting inferences with varying numbers of false positives. It was, therefore, important to develop scoring metrics that considered the inferences made by all semantic subgraphs, but that also penalised inferences that were made by semantic subgraphs that identified great numbers of false positives. The metrics used during this approach are described next.

5.3.3 Ranking inferred interactions

After an exhaustive search of *Dat* with the 194 semantic subgraphs a set of mappings (or instances) of each subgraph was identified. It is in these mappings that inferred D-T interactions are captured. To score inferred interactions, it was first necessary to score the semantic subgraphs whose mappings capture the potential interactions. Scoring of a semantic subgraph, Q, was achieved by determining the number of known D-T interactions in the predicted total set of D-T interactions inferred by Q. The complete set of inferred interactions was referred to as $Q(I)$. A score, RS, was calculated based on the ability of Q to identify D-T interactions captured in *DBv3Rel* (Equation 5.1).

$$RS(Q) = \frac{|Q(I) \cap DBv3Rel|}{|Q(I)|} \tag{5.3}$$

Once RS is calculated for each semantic subgraph, individual D-T interactions, i, are scored based on the cumulative score of all semantic subgraphs whose mappings predicted i. This score, $RD(i)$, allows for interactions to be ranked.

$$RD(i) = \sum_{i \in Q'(I)} RS(Q') \tag{5.4}$$

DReSMin is an exhaustive algorithm, as such, scoring inferred interactions allows for ranking, with those ranked higher inferred with greater confidence than those ranked lower.

5.3.4 Mining using DReSMin

A search of *Dat* with the set of 194 semantic subgraphs described above resulted in 906,152,721 mappings. These mappings capture the potential drug target interactions in the structure of the mapping subgraph. The 906,152,721 mappings predicted 9,643,061 D-T interactions. Semantic subgraphs were scored on their ability to identify D-T interactions captured in *DBv3Rel* using Equation 5.3 (see Fig. 5.11A), with these scores ranging from 0.0 to 0.06589 (Fig. 5.12). A single D-T interaction, i, can be inferred by mappings of more than one query semantic subgraph, thus adding confidence to the prediction that a D-T interaction exists. Therefore, to rank the D-T

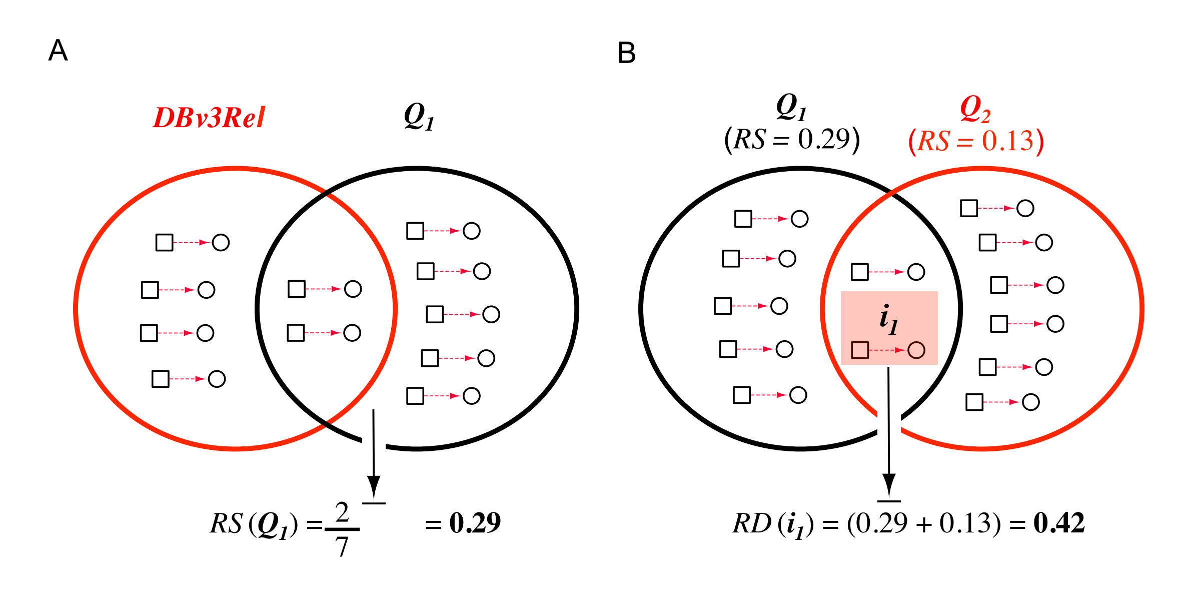

A
DBv3Rel
Q_1
$RS\,(Q_1) = \frac{2}{7} = \mathbf{0.29}$
B
Q_1
(RS = 0.29)
Q_2
(RS = 0.13)
i_1
$RD\,(i_1) = (0.29 + 0.13) = \mathbf{0.42}$

Figure 5.11: **Semantic subgraph (A) and individual D-T interaction (B) scoring.** (A) A semantic subgraph, Q_1, is scored by dividing the size of the intersection of its inferred associations and those captured in *DB3Rel* (2) by the total number of inferences captured in mappings of Q_1 (using Equation 5.3). (B) An individual D-T interaction, i, is scored by summing the scores of all semantic subgraphs whose mappings inferred i_1 (using Equation 5.4). *Note:* squares represent drugs, circles represent proteins and dashed red lines represent inferred D-T interactions.

interactions in terms of confidence, the scores assigned by all query semantic subgraphs that produced a mapping containing a potential D-T interaction were summed using Equation 5.4 (see Fig. 5.11B). The $\sum R_q$ of the scores of all 194 query semantic subgraphs was 0.9499 (Fig. 5.12) and so inferred D-T interaction scores contained within mappings could potentially, range from 0.0 to 0.9499, with the higher the score, the greater the confidence in the prediction.

Unsurprisingly, the interactions from *DBv3Rel* that were used to create the semantic subgraphs are identified. Importantly, however, these interactions score highly, which indicates that a single interaction was identified by multiple semantic subgraphs. The D-T interactions from *DBv3Rel* consistently scored better and ranked higher than the unsupported inferred associations (Fig. 5.13A and Fig. 5.13B). It was also observed that the D-T interactions subsequently annotated and captured in *DBv3Rel* are identified, on average, by two-fold the number of semantic subgraphs that infer D-T associations not present in *DBv3Rel* (Fig. 5.13C).

However, to quantify the predictive power of DReSMin the number of high scoring

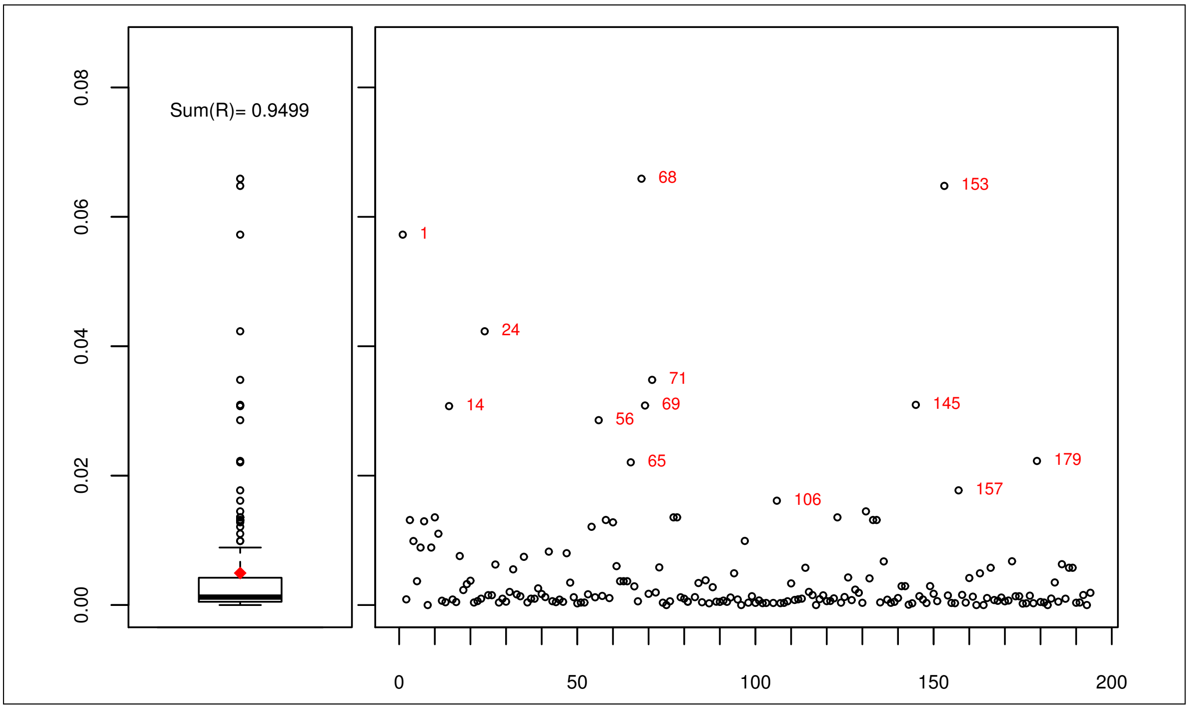


Figure 5.12: **Distribution of semantic subgraph scores.** Boxplot shows the distribution of semantic subgraph scores, with the mean score captured as a red diamond (0.005). The scatterplot shows individual scores, with the higher scoring subgraphs (those with a score above 0.15) labelled in red; these labels correspond to the name assigned to the semantic subgraph.

D-T predictions that were subsequently annotated in DBv4.2 was investigated. These 333 new interactions had not been used to construct the semantic subgraphs used for searching *Dat*. Of the 333 D-T interactions captured in *DBv4Rel*, 309 were successfully identified in the set of inferences made by DReSMin (94%). A high ranking and scoring of the 309 D-T interactions from *DBv4Rel* that were successfully identified by DReSMin is also observed (Fig. 5.13D and Fig. 5.13E). The average number of semantic subgraphs that have mappings inferring the 309 annotated D-T associations captured in *DBv4Rel* is increased >4 fold in comparison to the number of semantic subgraphs that produce mappings that infer interactions not captured in *DBv4Rel* (Fig. 5.13F).

Looking in more detail at the top 20 inferred D-T interactions (Table 5.2) 12 different drugs and eight targets are presented. Drugs include: three antiarrhythmic calcium channel blockers (verapamil, mibefradil and bepridil); three phenothiazine antipsychotic agents (promazine, perphenazine and thioridazine); three atypical antipyschotic agents (propiomazine, clozapine and quetiapine); two anticonvulsants (zonisamide and levetiracetam) and one antiarrhythmic adrenergic beta-antagonist (propranolol). It can be seen that in the top ranked D-T interaction inferences, the 12 drugs captured are involved in, on average, 13 D-T interactions in *Dat*. Such a high number of D-T interactions is in contrast to the average for all drugs in *Dat*, which is closer to three. The contrast highlights the fact that the top 20 inferred D-T interactions involve drugs that are well studied and highly annotated.

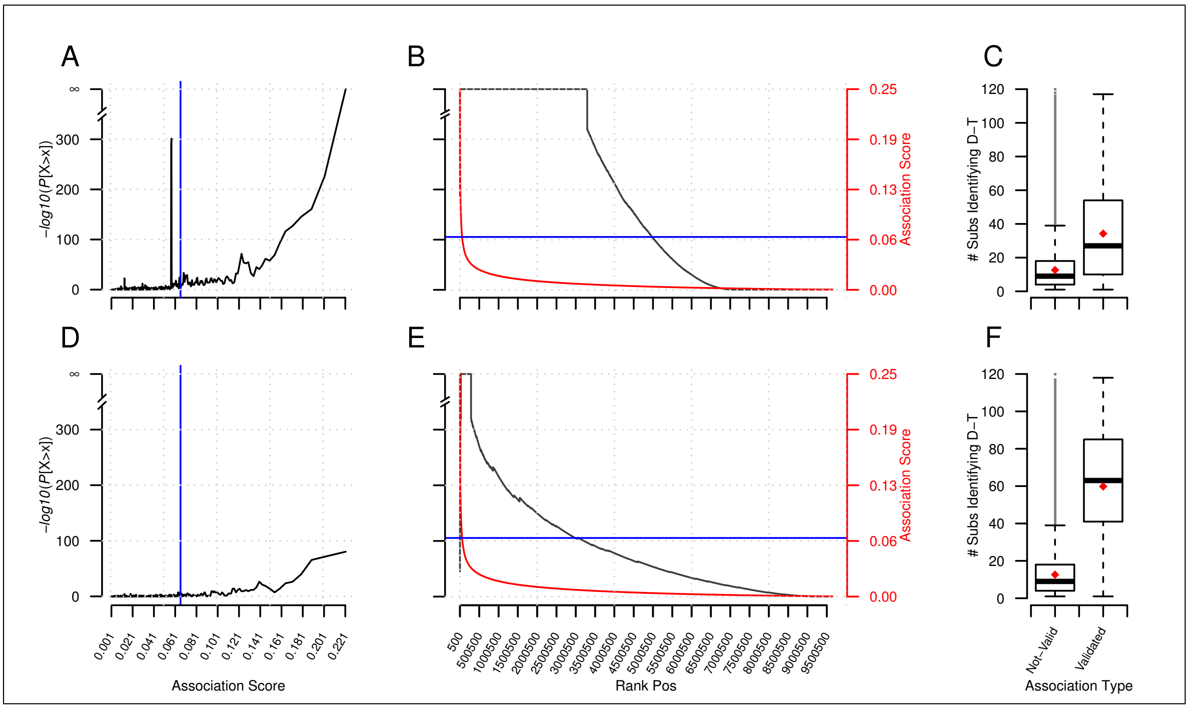


Figure 5.13: **Validation of inferred D-T associations with known D-T associations from DBv3 and DBv4.2.** (A), (B) and (C) show how DReSMin identifies and ranks the 2,919 known interactions from DBv3 when searching *Dat*. (D), (E) and (F) show how DReSMin identifies and ranks the 333 known interactions from DBv4.2. For (A) and (D) hypergeometric distribution of inferred knowns was calculated using the scores of the validated associations. All inferred D-T interactions were grouped based on their association scores and association scores were ranked in ascending order, shown on the x axis. Hypergeometric calculations were made at intervals of 1,000 association scores. The *-log10* probability of identifying more knowns than were observed at this point ($P[x > X]$) is shown on the y axis. For (B) and (E) hypergeometric distribution of inferred knowns was calculated using the ranked position of the validated interactions. All inferred D-T interactions were ranked from highest scoring interaction (Rank pos = 1) to lowest scoring interaction (Rank Pos = 9,643,061), shown on the x axis. Hypergeometric calculations were made at intervals of 10,000 associations. The black line shows the *-log10* probability of identifying more knowns than were observed at this point ($P[x > X]$) (left y axis) whilst the red line shows the score of the associations (right y axis). (C) and (F) show the number of semantic subgraphs that inferred knowns in comparison to the number of semantic subgraphs that inferred novel interactions. *Note:* Blue line shows the highest scoring semantic subgraph; all scores above this line are inferred by more than one semantic subgraph.

Table 5.2: Top 20 ranked DReSMin inferred D-T associations

Drug (*DrugBank ID*)	**Type, *Cat***	**Target (*UniProt ID*)**	**Ev**	**#Subs**	**Score**
Verapamil (*DB00661*)	SM, *AP*	Voltage-dependent L-type calcium channel subunit alpha-1C (*Q13936*)	Y	85	0.49211
Mibefradil (*DB01388*)	SM, *WI*	Voltage-dependent P/Q-type calcium channel subunit alpha-1A (*O00555*)		74	0.44378
Mibefradil (*DB01388*)	SM, *WI*	Voltage-dependent N-type calcium channel subunit alpha-1B (*Q00975*)		59	0.43097
Promazine (*DB00420*)	SM, *AP*	Alpha-1A adrenergic receptor(*P35348*)	Y	117	0.39090
Quetiapine (*DB01224*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)		61	0.38779
Propiomazine (*DB00777*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)		69	0.38774
Verapamil (*DB00661*)	SM, *AP*	Voltage-dependent P/Q-type calcium channel subunit alpha-1A (*O00555*)	Y	78	0.38436
Verapamil (*DB00661*)	SM, *AP*	Voltage-dependent N-type calcium channel subunit alpha-1B (*Q00975*)	Y	64	0.38180
Mibefradil (*DB01388*)	SM, *WI*	Voltage-dependent L-type calcium channel subunit alpha-1D (*Q5SQC4*)*		52	0.37525
Perphenazine (*DB00850*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)		86	0.37383
Thioridazine (*DB00679*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)		75	0.36830
Promazine (*DB00420*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)		75	0.36824
Propranolol (*DB00571*)	SM, *AP*, *IN*	D(1A) dopamine receptor (*P21728*)		96	0.36084
Zonisamide (*DB00909*)	SM, *AP*, *IN*	Voltage-dependent N-type calcium channel subunit alpha-1B (*Q00975*)		50	0.35478
Levetiracetam (*DB01202*)	SM, *AP*, *IN*	Voltage-dependent N-type calcium channel subunit alpha-1B (*Q00975*)	Y	50	0.35478
Thioridazine (*DB00679*)	SM, *AP*	5-hydroxytryptamine receptor 2B (*P41595*)		107	0.35036
Clozapine (*DB00363*)	SM, *AP*	5-hydroxytryptamine receptor 7 (*P34969*)	Y	64	0.34799
Propranolol (*DB00571*)	SM, *AP*, *IN*	Alpha-1A adrenergic receptor (*P35348*)		84	0.34663
Bepridil (*DB01244*)	SM, *AP*, *WI*	Voltage-dependent L-type calcium channel subunit alpha-1C (*Q13936*)		77	0.34610
Levetiracetam (*DB01202*)	SM, *AP*, *IN*	Voltage-dependent P/Q-type calcium channel subunit alpha-1A (*O00555*)		63	0.34605

Note: Of the 20 interactions ranked highest by DReSMin, six were found in DBv3 (*Ev*); having literature supporting their existence. For drug Type and *Category*: SM = small molecule, *AP* = approved, *IN* = investigational and *WI* = withdrawn. Scores are to 5 decimal places. # Subs refers to the number of semantic subgraphs that inferred the D-T interaction, with the maximum being 194. * Accession since deprecated- now *Q01668*.

Targets include four voltage-dependant calcium channels (VDCCs) and four G-Protein coupled receptors (GPCRs). VDCCs display selective permeability to calcium (Ca^{2+}) ions which enter a cell through the pore formed by the a_1 subunit when the receptor is activated. Three sub-types of VDCC are represented in Table 5.2, being: L-type (CAC1C and CAC1D); P/Q Type (CAC1A) and N-type (CAC1B). Members of the GPCR superfamily in Table 5.2 include receptors activated by the neurotransmitters: serotonin (5HT7R and 5HT2B); epinephrine (ADA1A) and dopamine (DRD1).

5.3.5 Drug-target interaction prediction evaluation

The ranked set of predicted D-T interactions were then compared to those produced by another state-of-the-art method for drug target interaction prediction - a ligand-based method. One implementation of such an approach is provided by ChEMBL[9] [180, 181]. ChEMBL provides two models for target prediction, using bioactivity data with a cut-off of 1 micromole (μM) and 10μM respectively. These models allow for n predicted interactions to be made for a given drug. Inferred interactions are also scored and ranked, allowing for a direct comparison to DReSMin. Predictions using the ChEMBL models can be found in compound report cards, accessed via their website.

DReSMin inferred D-T associations for 2,223 drugs common to both DrugBank and ChEMBL. In the comparison presented in Fig. 5.14, only D-T interaction inferences involving this set of 2,223 drugs are considered. For each of the drugs, the top 100 D-T associations involving single proteins were identified using the ChEMBL Web resource client[10]. The set of inferences from DReSMin contained a total of 2,456 protein targets (of which 1,133 are from *Homo sapiens* and 1,323 from other organisms). The set of ChEMBL inferences involves 870 human protein targets, of which 362 are also captured in DReSMin inferred D-T associations.

Any interactions which were already caught in *Dat*, involved targets from organisms other than humans or were not captured in the overlapping 362 protein targets, were excluded from the analysis. This process was repeated for both the 1μM and the 10μM ChEMBL models, giving two sets of predicted D-T associations. For a fair comparison

[9] www.ebi.ac.uk/chembl

[10] www.github.com/chembl/chembl_webresource_client Accessed:22-06-2015

to be made for each of the 2,223 drugs the top x[11] inferred single protein targets were collated and ranked. This process resulted in three sets of 215,075 ranked D-T interactions; *DReS*, *Chem1* and *Chem10*.

First, the number of D-T interactions predicted by both methods (co-prediction) was examined using interactions captured in the sets *DReS*, *Chem1* and *Chem10*. 35% of the top x D-T interactions inferred by DReSMin are found in the top x D-T interactions predicted by ChEMBL models (Fig.5.14A and Fig. 5.14B). More interestingly DReSMin successfully infers 10% more of the knowns from *DBv*4*Rel* than ChEMBL, for both models (Fig.5.14C and Fig. 5.14F). It was found that DReSMin was able to rank the known D-T interactions more effectively than ChEMBL, with a mean ranking position of known D-T interactions from *DBv4rel* of 16,977, as opposed to the 47,618 achieved by ChEMBL (47,746 for 1μM model and 47,490 for 10μM model). It is important, however, to point out that the semantic subgraphs used during this work were derived using DrugBank data and the ChEMBL models trained on ChEMBL data.

5.3.6 Target class comparison

After classifying all human proteins in *Dat*, the following were identified: 826 GPCRs; 343 ion channels; 638 kinases; and 560 proteases (Table 5.3). Interestingly, when analysing the average connectivity of the target classes in *Dat* (Table 5.3), it can be seen that the kinase class has, on average, the most associated edges. Not surprising given their evolution, kinases have by far the greatest number of *has_similar_sequence* edge types in *Dat*. More interesting is the fact that kinases have the least *binds_to* edges in *Dat* when compared to the other target classes.

Of the 9,643,061 D-T interactions inferred by DReSMin, 4,780,935 (49.6%) involve human protein targets. Within these associations, there are 103 GPCRs; 85 ion channels; 89 kinases; 60 proteases; and 782 ‘others’. Fig. 5.15A shows how the target classes defined are represented in the inferences made by DReSMin. 89% of proteins are classed as ‘other’, and make up the targets in 69% of the human protein inferences. GPCR’s make up only 4% of all human proteins and yet are involved in nearly 10% of all human inferences. Only 3.2% of human proteins are classed as kinases, yet these are captured

[11] x=100 or, if DReSMin inferred <100 targets for this drug, x=number of DReSMin inferred targets

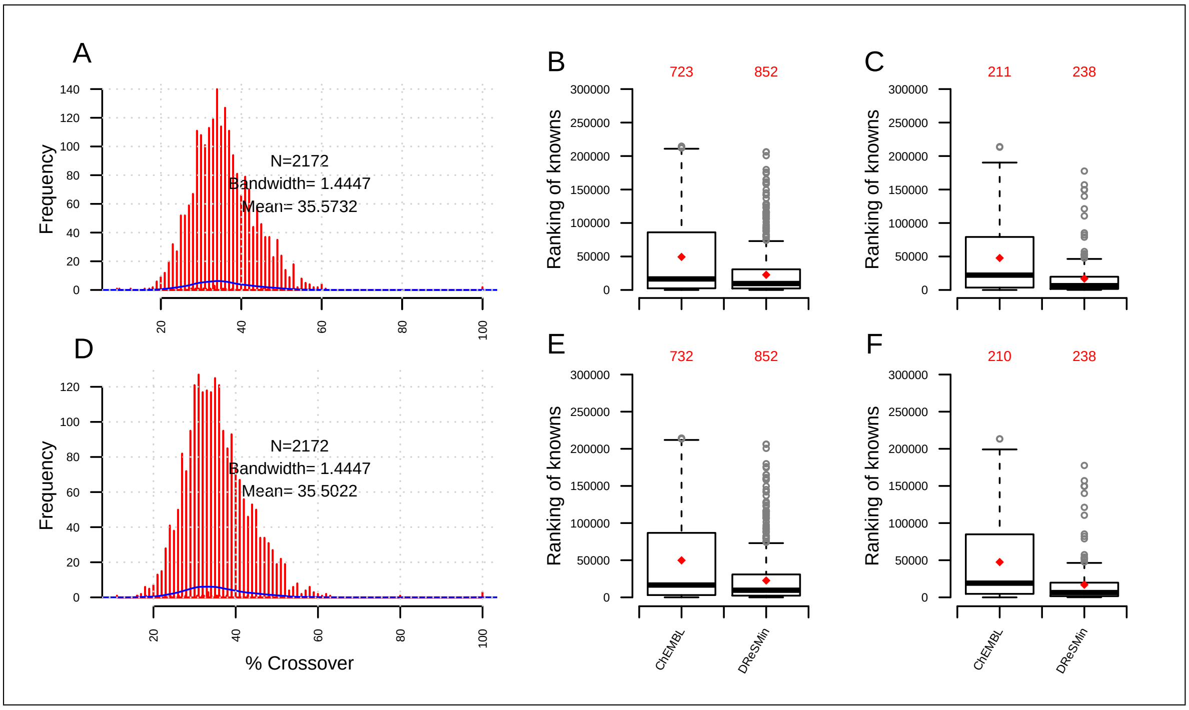


Figure 5.14: **DReSMin inferred D-T associations in comparison to those inferred using the ligand-based similarity models provided by ChEMBL.** Top graphs, (A), (B) and (C), show comparison to those using the 1μM model from ChEMBL and bottom graphs, (D), (E) and (F), show comparison with the 10μM model. (A) and (D) show the % crossover between the top-ranked x associations from each method for each drug. (B) and (E) show the comparative ranking of the 2,919 known D-T interactions from *DBv3Rel*. (C) and (F) show the comparative ranking of the 333 known D-T interactions from *DBv4Rel*. In (B), (C), (E) and (F) red diamonds show the mean ranking and numbers in red show the number of knowns captured by each method. Only associations inferred by the 2,223 drugs with a mapping between DrugBank and ChEMBL and those that contain the overlapping set of target proteins are included in the comparison.

in 7.9% of all human inferences. Proteases make up 2.8% of human proteins and are shown to be part of 5.4% of all DReSMin inferences involving human proteins. Finally, ion channels make up 1.7% of human proteins and are contained in 7.4% of all human predicted D-T associations.

Ion channels make up the second smallest set of comparative inferences, behind only proteases. However, 5.15A shows that this class is inferred by, on average: the highest scoring predictions (using Equation 5.4); the highest ranked inferences; and the most semantic subgraphs per inference. Second are the GPCRs, followed by proteases, 'others' and, finally, kinases.

5.15B shows how highly ranked the D-T interactions captured in *DBv3Rel* are ranked

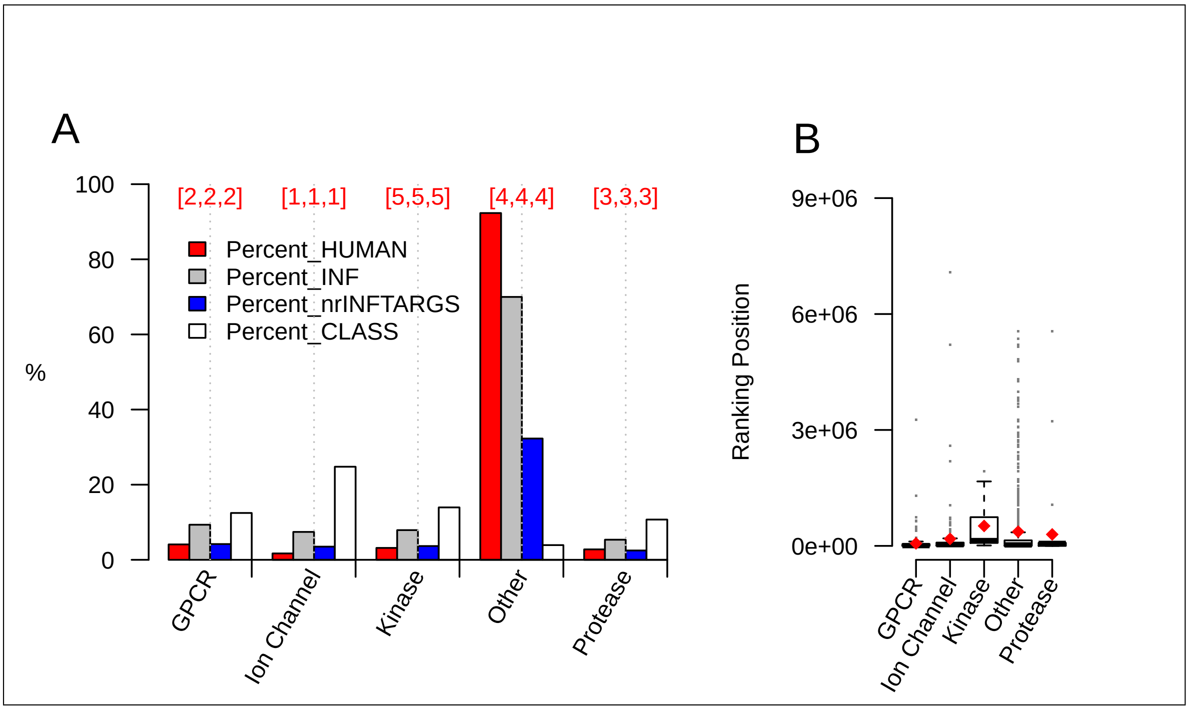


Figure 5.15: **Target class distribution in *Dat*.** Human proteins in *Dat* were assigned to one of five target classes: kinases; ion channels; G protein-coupled receptors (GPCR); proteases; or other. (A) % of the total set of human proteins assigned to each class (Percent_HUMAN), what % of all DReSMin inferred associations contained a protein target from that class (Percent_INF), the % of the unique targets inferred by DReSMin that were from that class (Percent_nrINFTARGS) and the % of the human target class for which an inference was made (Percent_CLASS). (B) How the known associations captured in *DBv3Rel* were ranked in the DReSMin inferred D-T interactions, with 1 being the highest ranked association. *Note:* Sets of numbers above each class in (A) represent how the target class ranked in performance in comparison to the other classes in the following measures: the average score of inferred interactions; the average ranking of inferences for that class; and the average number of semantic subgraphs that made an individual inference for that class. One represents the best performing target class and five the worst.

in each of the classes. D-T associations captured in *DBv3Rel* are ranked highest for GPCRs, followed by: ion channels; proteases; other; and kinases. Known associations from all classes are, on average, ranked in the top 6% of all inferred associations.

5.3.7 Completing the drug-target-disease pathway

The highest ranked D-T interaction identified by DReSMin, with a score of 0.49211, is supported by the literature and, therefore, known to the scientific community [182]. This D-T interaction is between one of the antiarrhythmic calcium blockers, verapamil, and the Voltage-dependent L-type calcium channel subunit alpha-1C, CAC1C.

Table 5.3: Frequency and connectivity of target classes in *Dat*.

Target Class	***Dat* Freq.**	*Connectivity*	*bi_to*	*h_s_s*	**Mapping Freq.**
GPCR	826	33.45 (4)	8.08 (1)	11.53 (5)	103
Ion Channel	343	49.78 (2)	4.46 (2)	36.97 (2)	85
Kinase	638	74.38 (1)	2.66 (5)	54.55 (1)	89
Protease	560	41.41 (3)	3.97 (3)	14.59 (4)	60
Other	19,974	33.06 (5)	3.47 (4)	14.60 (3)	782

Note: Connectivity is rounded to 2 d.p. Brackets show ranks for the classes. *bi_to* = `binds_to`, *h_s_s* = `has_similar_sequence`.

Within *Dat* eight indications are associated with verapamil and 12 diseases associated to CAC1C. One indication, hypertension, shares a `has_Indication` edge with verapamil and an edge of type `involved_in` with CAC1C. Although verapamil is already used to treat hypertension, and the inferred D-T interaction already known, it can be observed that DReSMin may be used to help understand the molecular mechanism of a drug and thus complete the 'drug-target-disease' pathway. Understanding the molecular mechanisms of drugs can aid the identification of repositioning opportunities. In Fig. 5.16 examples of unsupported, and, therefore, novel, DReSMin inferred D-T interactions that allow the understanding of molecular mechanisms involved in a drug's ability to treat a disease are shown.

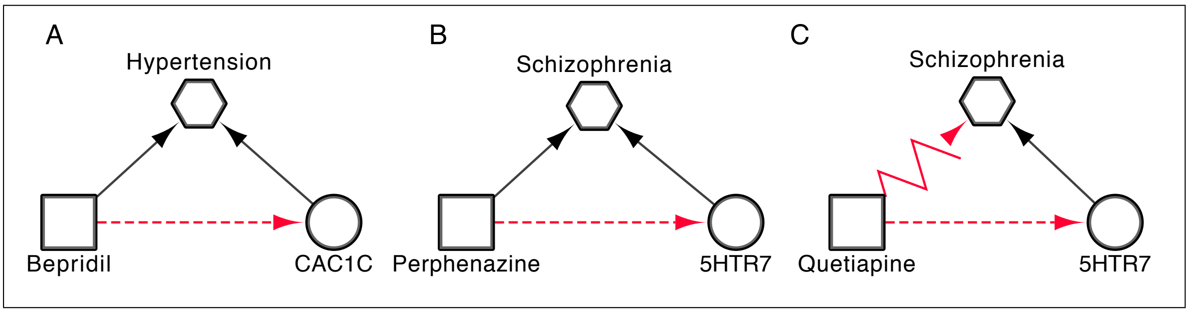


Figure 5.16: **Drug-target-disease pathways completed via inferred D-T associations**. Data presented is extracted from *Dat* with one association extracted from the literature. *Note:* Dashed lines represent the inferred `binds_to` relations, zig-zag lines represent `has_Indication` relation not captured in *Dat* and extracted from literature [183], squares represent compound, circles target and octagon diseases.

Like verapamil, bepridil is also a calcium channel blocker with known antiarrhythmic activities and has been used as a treatment for hypertension. Table 5.2 shows an inferred D-T association involving bepridil and CAC1C. Bepridil is one of the two drugs from Table 5.2 that have been withdrawn from the market due to safety concerns, and as such bepridil is not a strong candidate for repositioning. However, using the inferred association involving bepridil, a better understanding of the molecular mechanism of the drugs use as a treatment for hypertension (Fig. 5.16A) can be achieved.

Dat captures three indications for quetiapine (Psychotic Disorders, Bipolar Disorders and Autistic Disorders) and three `involved_in` associations involving 5HT7R (Schizophrenia, Pain and Muscular Diseases). Although not captured in *Dat*, quetiapine is approved for the treatment of Schizophrenia. By integrating this knowledge with *Dat* and the DReSMin inferred associations another drug-target-disease pathway (Fig.5.16C) can be completed. Although Schizophrenia, along with many other diseases, is classified as a psychotic disorder, the inferred knowledge enables a better understanding of drug-target-disease pathways in more specific disease areas.

Propranolol One inferred D-T interaction in Table 5.2 involves the antiarrhythmic adrenergic beta-antagonist, propranolol, and the GPCR, D(1A) dopamine receptor, DRD1. *Dat* captures 12 indications for propranolol and 17 disease associations for DRD1, with Hypertension captured as both an indication for propranolol and an associated disease for DRD1 (Fig. 5.17). Of the remaining 16 `involved_in` edges involving DRD1 five of the diseases represent known off-label indications for propranolol: Bipolar disorders; Psychotic Disorders (particularly Schizophrenia); Alcoholism; and as a non-stimulant treatment for ADHD [184]. The remaining 11 diseases present and support some interesting repositioning opportunities for propranolol.

Three of the 11 potential indications of propranolol proposed using inferences made by DReSMin are currently being investigated by the scientific community through various clinical trials. *Dat* contains an association between DRD1 and cocaine-related disorders, with multiple clinical trials being undertaken to analyse the use of propranolol as a treatment for cocaine addiction [185] as well as cocaine cravings [186]. A trial looking at the use of propranolol as a treatment for Autism is also, at the time of writing, recruiting [187]. Finally, a clinical trial has also been undertaken to investigate the effects of using propranolol as a treatment for drug-induced movement disorders [188].

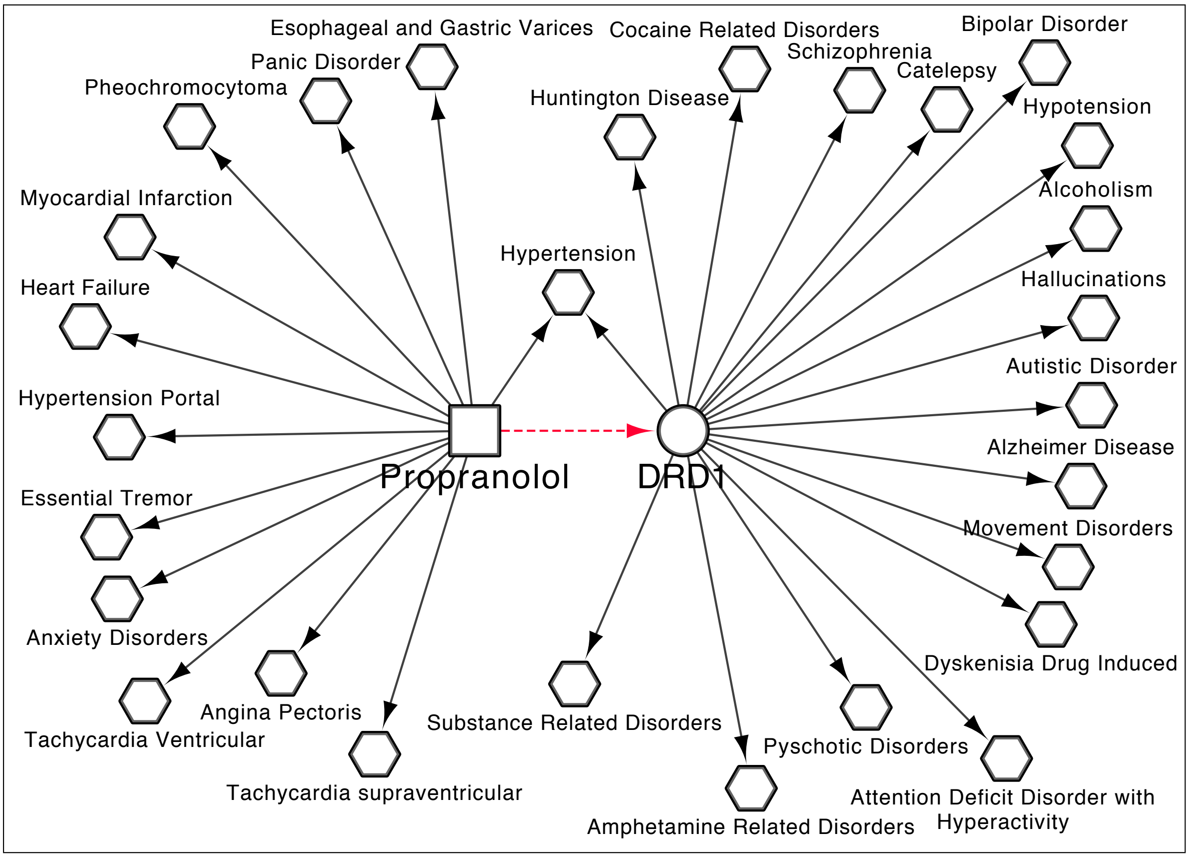


Figure 5.17: **Diseases associated with Propranolol and DRD1 in *Dat*.** Drug-disease `has_Indication` edges involving propranolol and gene-disease `involved_in` edges were extracted from *Dat*. *Note:* Dashed lines represent the inferred `binds_to` edges, squares represent compound, circles target and octagon diseases.

DReSMin inferred D-T interactions allow for the prediction of repositioning opportunities that are currently being investigated in the clinical setting.

After removing currently investigated indications, eight potentially novel uses for propranolol are inferred, including: Amphetamine-Related Disorders; Alzheimer Disease; Catalepsy; Dyskinesia Drug Induced; Hallucinations; Huntingdon Disease; Hypotension; and Substance-Related Disorders. As previously mentioned, propranolol is used to treat high blood pressure, Hypertension. It is, therefore, unlikely that propranolol will prove a useful treatment for low blood pressure or hypotension. Hypotension would only be exacerbated through the administration of propranolol. Studies have shown that propranolol may be helpful as a means of controlling the aggressive behaviour associated with patients in the advanced stages of the disease Huntingdon Disease [189]. Although no clinical trials are currently underway, work on cell cultures has investigated the effect of cardiovascular drugs, such as propranolol on the prime target for Alzheimer's prevention, amyloid β peptides [190]. These peptides are of interest as it has been shown that decades before the onset of dementia this small protein accumulates in clumps in the brain. The work carried out in [190] demonstrated that propranolol (as well as nicardipine) exert *in vitro* amyloid β- lowering activity and supports an interesting hypothesis that can be made via the approach that propranolol may be used as a treatment for Alzheimers.

5.3.8 Instance information

A single D-T association, i, can be inferred by multiple mapping instances of the same semantic subgraph Q. The number of these instances is not considered during the approach described previously, where the score of Q is added to the sum score of i regardless of the number of instances of Q that predict i. The effect of instance information on the scoring and ranking of D-T interactions inferred by DReSMin was, therefore, investigated (Fig. 5.18). Instead of scoring i using the sum of the scores of the semantic subgraphs that inferred that i, instance information also multiplies individual semantic subgraph scores by the number of mappings that inferred i. For example, if i was inferred via 34 mappings of semantic subgraph Q, which achieved a $RS(Q)$ (see equatp ion 5.3) of 0.1, Q would account for 0.1 of $RD(i)$ (using equation 5.4). Considering instance information, Q would account for 3.4 of $RD(i)$ (0.1×34).

Using instance information, D-T interaction scores ranged between 0.0 - 624.13. An unsupported interaction between the approved small molecule atorvastatin (*DB01076*) and the GPCR Histamine H1 receptor (*P35367*) identified is ranked top and identified by the greatest number of mapping instances, 1,750,194. The same interaction is ranked 60,615 when instance information is not considered. When comparing the ability of the standard approach to prioritise knowns (Fig. 5.13) in comparison to an approach that considers instance information (Fig. 5.18), some notable differences can be observed. It was noted that instance information performed poorly in comparison when looking at the probability of higher scores identifying more of the knowns. Unsurprisingly, it was also shown that when comparing the ability of the approaches to rank the knowns higher than the unknowns, the standard approach outperformed an approach considering instance information. Finally, there is little difference between

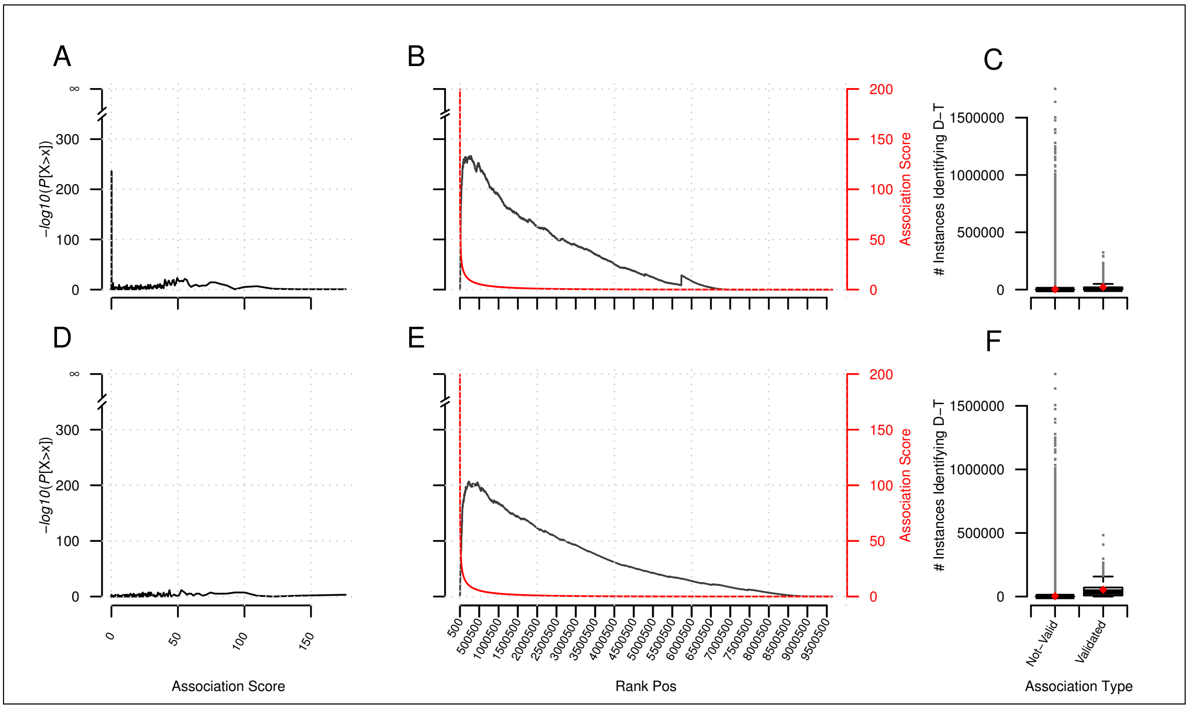


Figure 5.18: **Validation of inferred D-T associations, ranked using instance information, with known D-T associations from DBv3 and DBv4.2.** (A), (B) and (C) show how DReSMin identifies and ranks the 2,919 known interactions from DBv3 when searching *Dat*. (D), (E) and (F) show how DReSMin identifies and ranks the 333 known interactions from DBv4.2. For (A) and (D) hypergeometric distribution of inferred knowns was calculated using the scores of the validated associations. All inferred D-T interactions were grouped based on their association scores and association scores were ranked in ascending order, shown on the x axis. Hypergeometric calculations were made at intervals of 1,000 association scores. The *-log10* probability of identifying more knowns than were observed at this point ($P[x > X]$) is shown on the y axis. For (B) and (E) hypergeometric distribution of inferred knowns was calculated using the ranked position of the validated interactions. All inferred D-T interactions were ranked from highest scoring interaction (Rank pos = 1) to lowest scoring interaction (Rank Pos = 9,643,061), shown on the x axis. Hypergeometric calculations were made at intervals of 10,000 associations.The black line shows the *-log10* probability of identifying more knowns than were observed at this point ($P[x > X]$) (left y axis) whilst the red line shows the score of the associations (right y axis). (C) and (F) show the total number of instances of semantic subgraph mappings that inferred knowns in comparison to the total number of instances of semantic subgraph mappings that inferred novel interactions.

the average number of mapping instances that identify validated D-T interactions and the number that identify those that are not validated. Such a small difference between the number of mappings that identify validated and non-validated D-T interactions was not observed in the standard approach, where the number of subgraphs that infer the validated associations is significantly higher than the number inferring D-T interactions that are not validated.

5.4 Discussion

DReSMin can identify known D-T interactions regardless of the class to which the target belongs. It was shown that kinases are the highest connected class in *Dat*. Kinases represent a highly conserved protein class and as such contain a large number of associations describing their structural similarities. Conversely, kinases have the least amount of data regarding known interactions with drugs. Known D-T associations involving GPCR and ion channels fall, on average, in the top 2% of inferred D-T interactions. Although known D-T associations involving other classes, such as kinases are still, on average, captured in the top 6% of all DReSMin inferred D-T associations; the approach does favour target classes where there exists a relatively high amount of drug interaction information. The methodology presented makes use of the holistic view of an entity, and so if less is known about a target it will be captured in fewer semantic subgraphs, and thus D-T interactions that it is predicted to be involved in will obtain a lower score. As more and more data is produced for target classes, such as proteases and kinases, bias in the approach will also be reduced.

DReSMin is able to prioritise known D-T interactions, however, for inferences made to be useful to drug repositioning, there are still some limitations that must be discussed. These restrictions are introduced here with some examples provided.

First of all, DReSMin infers an interaction between dexrazoxane (*DB00380*) and dactinomycin (*DB00970*) and the Sodium channel protein type 1 subunit α (*P35498*). The Sodium channel protein type 1 subunit α target is located predominantly in the brain and is heavily associated with epilepsy [191, 192]. To reach the brain, a drug must cross the Blood-Brain Barrier (BBB). Multiple drugs, such as dexrazoxane and dactinomycin [193], are restricted by their pharmacokinetics and are unable to cross the BBB and so this inference is unlikely to highlight any realistic repositioning opportunity.

Secondly, DReSMin infers D-T interactions involving drugs from a range of marketed statuses. Examples include drugs that have been withdrawn as they are not as effective as first thought, such as drotrecogin alfa (*DB00055*) in the treatment of sepsis, or due to poor sales, such as halazepam (*DB00801*). These represent interesting candidates for repositioning. Drugs that have been withdrawn from the market for reasons involving safety concerns prove a more problematic repositioning opportunity. Some examples are included in DReSMin inferences, such as drugs that have been withdrawn from the market due to potentially fatal side-effects, such as: metamizole (*DB04817*); grepafloxacin (*DB00365*); and temafloxacin (*DB01405*).

Protein similarities are calculated by doing an all against all BLAST using `Protein` sequences captured in *Dat*. Only those similarities with an E-value of less than 1e-4 are included in *Dat* as an edge of type `has_similar_sequence`(see Table A.2). Perhaps a more stringent protein similarity cut-off could be used to ensure paralogs; as already mentioned kinases can appear very similar but may have very different functions.

Finally, DReSMin inferred an association between domperidone (*DB01184*) and the Beta-1 adrenergic receptor (*P08588*). Heart palpitations are a known side-effect of domperidone and Beta-1 adrenergic antagonists, such as propranolol have been administered to those suffering heart palpitations. It can, therefore, be deduced that domperidone may have some agonistic action upon the Beta-1 adrenergic receptor. With these examples in mind, other properties must be considered in further extensions to the approach. Drug properties, such as pharmacokinetics, in relation to the target location, must be considered, as well as the implementation of a pre-filtering step to remove all drugs from the search that are likely to be unsafe. Post-filtering of results based on the likelihood of a D-T interaction prediction leading to a potential side-effect would also be a useful addition and would need to consider a drugs action (be it an agonist, antagonist, etc.).

When comparing DReSMin to other state-of-the-art D-T prediction methods an average co-prediction of 35% is observed. Although this still leaves a large proportion of unique inferences, it is shown that DReSMin inferences identify >16% of the knowns when using DBv3, and >10% of the knowns when using *DBv4*, in comparison to ChEMBL. The ability of DReSMin to identify more of the knowns than ChEMBL supports the approach and provides evidence that it produces an improved prediction set. After directly comparing and contrasting the results it was found that DReSMin outperformed the ChEMBL models at inferring annotated DrugBank D-T interactions. Considering DReSMin is a general algorithm, not specifically developed for the inference of D-T interactions, this highlights its potential. Although the semantic subgraphs used to search *Dat* were derived from the shortest paths between a drug and target from D-T interactions in DBv3, these interactions were inferred, on average, by around 40 different semantic subgraphs; in contrast to the 15 semantic subgraphs that inferred D-T interactions not captured in DBv3. Again, the fact that known interactions are captured by a far greater number of semantic subgraphs than those not in DBv3 further validates the approach employed during this work. Annotated D-T interactions were not only captured by the semantic subgraph derived from the semantic shortest path between their drug and target but also by many more semantic subgraphs.

5.5 Conclusion

In this Chapter, an approach for the identification of novel D-T interactions using DReSMin was introduced. A set of 194 relevant semantic subgraphs were derived from the semantic shortest paths between known D-T and subsequently searched for in *Dat*. Mining resulted in the identification of 9,643,061 potential D-T interactions which were then scored and ranked before being validated against more recent DrugBank data sets. It was shown that these inferred D-T interactions can be used to identify novel drug repositioning leads, while also supporting repositioning investigations currently being undertaken. Comparison of the approach described with other dedicated, state-of-the-art D-T prediction methods also positively highlighted the potential of the work. Furthermore, it was shown that instance information does not provide as accurate an inference set as the ‘typical’ approach described.

Like the DReSMin algorithm, the approach presented in this Chapter, including semantic subgraph development and inferred edge scoring, is not limited to the drug repositioning setting. Although the data set and historical data used here were specific to drug repositioning, the approach is generic enough to enable the prediction of edges from any data set providing relevant historical data sets are available.

Regarding drug repositioning, future extensions of the work would require more focus on ensuring a satisfactory drug candidate set as well as post-filtering of likely irrational predictions. Examples of irrational predictions have been previously described and include: those that involve drugs that are unable to reach their predicted target (such as targets that require a drug to cross the BBB); predictions that involve drugs that have failed to reach the market or have been withdrawn from the market due to safety concerns; and predictions that involve a drug likely to exacerbate a condition (i.e. a disease is caused by a Loss of Function (LoF) mutation and the drug inferred is an antagonist). Furthermore, semantic subgraphs should be expanded to not only include semantic types but also include properties that may be relevant to research question being asked.

Considering the limitations of the approach described, it would be beneficial to consider filtering potential side-effects from an inference set. As well as filtering likely side-effects it would also be beneficial to include, where possible, information regarding a drug's action, i.e. is the drug an agonist or an antagonist? These limitations are considered in the approach presented in Chapter 6 where ranked Gene-Disease (G-D) associations are integrated with other relevant drug repositioning data. The resulting data set is then mined using a predefined semantic subgraphs and DReSMin to identify novel Drug-Disease (Dr-D) indications. In the approach described in Chapter 6, the DReSMin algorithm is extended to include attributes and a means of filtering potential side-effects is described.

6

Identification of novel uses for drugs with focus on gene-disease associations

6.1 Introduction

Prediction of novel Drug-Disease (Dr-D) associations can be achieved by 'filling in' all the blanks between a drug and a disease in the drug-target-phenotype-disease pathway (described in Fig. 2.1). Target-based approaches to drug discovery (see Fig. 2.1B) focus on identifying links between targets and their associated diseases, in the hope that identifying these links will allow for completion of the drug-disease pathway. As such, understanding the molecular mechanisms of diseases is vital within the field of target-based drug discovery.

A causal association between a gene target and a disease describes a situation where a gene is directly or indirectly responsible for disease risk via one or more mechanisms [194]. Monogenic disorders, such as Huntington's disease, are identified simply through the presence, or absence, of single gene mutations; in this case a mutation in the Huntingtin protein, HTT [195]. Conversely, multigenic, or complex, disorders are caused by multiple genetic variants, which may affect pleiotropic genes and be influenced by various environmental factors [196]. Due to the complexity of multigenic diseases, allele associations are more probabilistic and less deterministic; the presence of a high-risk allele may only mildly increase the chance of disease [196] [197]. For these reasons identifying causal links between a gene and disease experimentally is expensive and time-consuming. Association studies, however, identify disease susceptibility variants that do not necessarily mean the variant is important in disease causation and instead are associative— present as a consequence of the disease state as opposed to being responsible for causing the disease state. It is an easier task to identify susceptibility Gene-Disease (G-D) associations as opposed to causal G-D associations [198].

The shift to large-scale sequencing of individual genomes and the availability of new techniques for probing thousands of genes provide new means for identifying these susceptibility G-D associations (from here in G-D associations will refer to associative associations as opposed to causal associations). Experimental techniques such as positional cloning and/or microarray analysis can return tens to hundreds of candidate genes [98]. Managing and integrating these data has thus become an important task within bioinformatics, and numerous G-D databases have been developed to aid this. Entries in databases are mainly obtained through manual curation of the biomedical literature [199]. To capture data that may have been missed by manual curation of the literature, automated text mining approaches can also be used [124]. Although automated text mining approaches improve recall, precision is drastically reduced in comparison to manual extraction. Genetic associations can also be extracted directly from experimental data, such as Genome-Wide Association Studies (GWAS), and stored in dedicated databases. Furthermore, predictive methods may also be used to populate databases identifying associations through statistical inference, including cross-species inferences derived from animal models. Mouse and rat models have been used to predict human G-D associations for many years, and there exists a wealth of cross-species G-D association data available [200–202]. Cross-species models can be complicated by diverse types of phenotype representations in terms of physiological and anatomical differences between species. However, this knowledge cannot be ignored [203]. To create a state-of-the-art view of current knowledge regarding G-D associations, integration of these heterogeneous data sources is required.

A holistic view of the field allows for emergent properties that would otherwise be invisible to be realised [204]. Efforts such as DisGeNET [205, 206] and MalaCards [207], already integrate associations from multiple primary resources that have been curated, predicted and derived computationally from the text. DisGeNET applies a systematic scoring to these associations. However, the chosen metric fails to give a relative view of known G-D associations. A complete ranking of G-D associations from primary resources, taking into consideration the reliability of each data set using current knowledge, would allow inevitable bias present in data sets, that were all developed for different purposes, to be reduced. Furthermore, an exhaustive ranking of G-D associations would also aid tasks such as computational target-based drug discovery [208],

Despite historically being discovered via phenotypic approaches [5], target-based approaches to drug discovery came to prominence after sequencing of the human genome. It was believed that target-based drug discovery would allow for a more rational approach to drug design, and thus increase research and development (R&D) success and productivity [5, 209]. Target-based approaches are still heavily prominent and extensively used in the pharmaceutical industry [210], with successes including the tyrosine kinase inhibitors imatinib (Glivec; Novartis) and gefitinib (Iressa; AstraZeneca) [211]. Overall, due to increased costs and reduced productivity, there is a general acceptance that the current state of R&D needs to change [14]. Part of the solution, in the short term, is drug repositioning which is described in detail in Chapter 2.

In this Chapter, an exhaustive, novel approach for identifying new uses for existing drugs, with a focus on G-D associations is introduced. A Bayesian statistics approach, developed by Lee and colleagues [212] is utilised for the purpose of integrating and ranking G-D associations captured in 10 primary data sources. These scored G-D associations, which provide a state-of-the-art view of G-D knowledge, are then integrated with other biological entities to produce a semantic network, *GenDat*, for target-driven drug repositioning. A method for the automated detection of therapeutic areas of interest is also introduced. Finally, a four-node semantic subgraph (semantic subgraphs are formalised in Section 4.2.1) is introduced, and instances of this are identified in *GenDat*, using Drug Repositioning Semantic Mining (DReSMin), described in Chapter 4. Novel Dr-D interactions inferred from the integrated network are then ranked, with those involving diseases from the therapeutic area of interest discussed in more detail. The work presented in this Chapter has also been described in Mullen *et al.* [213].

6.2 Background

6.2.1 Gene-disease association databases

Several existing primary databases focus on G-D associations. These databases typically contain associations obtained through manual curation of the biomedical literature. One well-established source of G-D associations is the Online Mendelian Inheritance in Man (OMIM) database [121]. More recent projects include the Comparative Toxicogenomics Database (CTD) [131] and UniProtKB [214]. Another source, Orphanet [215], focusses primarily on rare diseases and orphan drugs. Databases populated with associations extracted directly from the literature, using text mining approaches, also exist [199], such as BeFree [124] and SemRep [216]. Although the accuracy of automatically extracted associations is lower than manually curated data, the systematic approach to their construction means they are more inclusive of true positives.

BeFree [124] provides a good example of a text mining resource. BeFree, along with supporting statements and provenance is available for download and uses the EU-ADT and GAD corpora to extract associations from the text. Focussing on a subset of abstracts returned from PubMed, BeFree uses their own query (only querying about 3 % of current MEDLINE databases). After applying filtering, BeFree captures 330,888 associations involving 13,402 genes and 10,557 diseases [124]. SemRep [216] also provides text mined associations. Like BeFree, SemRep provides G-D, drug-disease and drug-target associations, but unlike BeFree has been designed to identify a large variety of semantic predictions. When using the same corpus as BeFree, SemRep has a higher precision but a lower recall [124]. Other approaches to collecting G-D associations involve cataloguing data directly from genetic experiments (such as GWAS), or inferring associations from animal models.

Over the last decade, GWAS have produced data on thousands of Single-Nucleotide Polymorphisms (SNPs). These SNPs are associated with the risk of hundreds of diseases. Originally developed as a means to identify causal SNPs, GWAS data are non-trivial to work with; they identify marker SNPs that are often simply associative SNPs as opposed to causal SNPs. The assignment of SNPs to associative genes is often difficult due to factors such as the SNPs being present in regulatory regions. Furthermore, GWAS data only contains associations derived from a subset of diseases for which genetic studies have been conducted. As with any exercise in data collection, the data captured in data sources may be biased, depending on the intended purpose of the data. Bias is especially true of GWAS data [217], which is particularly biased to diseases such as Crohn's disease that are of interest to the industry. Nevertheless, GWAS data are available for download via the GWAS catalogue [122]. The Rat Genome Database (RGD) [201] and the Mouse Genome Database (MGD) [200] provide G-D associations that have been identified in animal models but are statistically inferred to represent human associations. One limitation of the various G-D association data sets is the lack of associations annotated with the gene functionality associated with the disease state (i.e. Gain of Function (GoF) or Loss of Function (LoF)).

6.2.2 Controlled vocabulary of diseases

Before working with G-D associations, it is important to determine a standardised representation of both genes and diseases. Due to work completed by the Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC), it is a fairly straightforward task to identify a strict representation of human genes [218]. To identify a satisfactory disease representation is more complicated since there are numerous disease classifications and ontologies competing with one another. These disease classifications are designed for different purposes and are mutually inconsistent. Consequently these are poorly integrated with each other. One example is the Disease Ontology (DO) [219] which is part of the Open Biomedical Ontologies (OBO) Foundry Initiative. The DO has extensive cross-referencing. However, the DO maps poorly to diseases captured in data sets such as DisGeNET. Another example comes is the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) which cross maps to several revisions of the International Classification of Diseases— used in the clinical setting [208]. SNOMED-CT is one of the many terminologies that is combined with the even broader Unified Medical Language System (UMLS) Metathesaurus. UMLS contains many distinct concepts that are very close in meaning, and as a result, even human annotation using UMLS concepts is problematic [220]. Finally, the Medical Subject Headings (MeSH) is a comprehensive controlled vocabulary displaying a hierarchical data structure [120]. MeSH was developed for the purpose of indexing journal articles and books in the life sciences and like SNOMED-CT is also combined with the UMLS Metathesaurus.

Identifying a standardised disease representation is a current challenge for large-scale disease data integration which aims to gather a comprehensive coverage of disease to enable systematic interoperability across biomedical domains [124].

6.2.2.1 MeSH therapeutic areas

The MeSH hierarchy contains 16 top-level categories, each assigned a letter descriptor. For example, the *Disease* top-level category is assigned the letter C. Underneath these top-level descriptors are more specific MeSH codes which describe more specific areas in the categories. For example, there are 26 MeSH codes in the C category. C06, for example, is used to capture *digestive system diseases.* Terms become more specific as depth in the hierarchy increases— C06.405.205.731.500, for example, is the entry for *Crohn Disease.* Furthermore, three concepts of the F top-level category (*Psychiatry and Psychology*) also describe diseases; F01 (*behaviour and behaviour mechanisms*), F02 (*psychological phenomena and processes*) and F03 (*mental disorders*). This leaves 29 top-level categories in MeSH that can be used to define therapeutic areas.

6.2.3 Calculating disease similarities

Although alot 'cleaner' than UMLS, the MeSH hierarchy is more verbose than disease representations such as the DO. As such, the specificity of terms is a potential problem. Two diseases may be synonymous yet captured in multiple parts of the taxonomy. In order to quantify the semantic similarities of two terms in the hierarchy ontology-based similarity measures may be used. Ontology-based similarity measures may be structure-based (e.g. path length, depth of concept or LCS (lowest common subsumer)) or content-based, whereby a corpus of terms is used, and information content is analysed). One example of a structure-based similarity measure is that developed by Leacock and Chodorow [221]. This measure is used to quantify the distance between two instances in an ontology or hierarchy. Although originally developed to measure the distance between nouns in WordNet, an electronic lexical database [221], the method has previously been applied to MeSH [222].

6.2.4 Ranking disparate data

Gold standards are used as a reference point for many approaches to predictive and scoring methodologies due to the belief that they offer superior quality to other sources of similar data. For areas whereby a gold standard does not exist, such as the G-D setting, this set becomes subjective to the field of use and the task at hand. Different approaches exist for ranking disparate data; some do not use a gold standard, such as that described by Weile and colleagues [223], and some that do utilise a gold standard, like the work completed by Lee and co-workers [212]. The approach described by

Lee and co-workers make use of a log-likelihood score (LLS), which is calculated for each data source that is to be integrated using the gold standard. The LLS score can be interpreted as the likelihood of a linkage (or association) existing conditioned on the given evidence and corrected for background expectations of linkages. By cumulatively summing the LLS score from each data set that contains an association a confidence score can be assigned to an association. Furthermore, these LLS scores can be summed using a D parameter. This parameter allows for higher weighting to be given to data sets with a higher LLS— facilitating dependencies between the data sets. Division of the score by a computed D parameter means that, while the highest score is integrated unchanged, subsequent LLS scores are progressively down-weighted. Down-weighting subsequent LLS scores are particularly relevant to G-D associations, whereby it is standard practice primarily to populate a database with associations from other curated sources before extending it (CTD, Orphanet and UniProt all collect a subset of associations captured in OMIM).

6.2.5 Graph model

To view G-D associations in the biological context, it is important to define a data structure that will aid in this task. Graph representations of complex systems are widely used in computer science, social and technological network analysis, and is especially relevant to many studies in bioinformatics [96]. Semantically-rich networks, which implement a graph-based representation, are ideal for representing integrated data [144]. In *semantic graphs*, each edge (or relation) and vertex (or node) are assigned a single type from a predefined set to describe their meaning semantically. In such a representation, node v_1 may represent cGMP-specific 3',5'-cyclic phosphodiesterase and is assigned the type `Protein`, while node v_2 represents sildenafil and is thus assigned the type `Small_Molecule`. If v_1 is a known target of v_2, this interaction is captured in a directed edge, e_1, of type `binds_to`. Nodes and edges of semantic graphs may also be annotated with attributes. For example, a node of type `Small_Molecule` may be annotated with a *Group* attribute that would be used to describe whether the drug was approved, investigational or withdrawn. Furthermore, an edge of type `binds_to` could indicate that a drug binds to a particular target with attributes of type *Activity Type* and *Activity Value* used to describe the type of activity, such as IC_{50}, K_d, K_i or Potency and the associated activity value, such as 1nM, respectively.

6.3 Materials and methods

An implementation of DReSMin, described by Mullen *et al.* [146] (and in Chapter 4), was used in the work presented in this Chapter. The algorithm was extended to allow for attribute comparison and implemented. *GenDat* was built using the Drug Repositioning Network Integration Framework (DReNInF) (as described in Chapter 3). A Java implementation of the structure-based ontology measure developed by Lee and colleagues [212] was completed as part of this work.

G-D associations included were taken from all four database types as previously described (curated, experimentally derived, literature derived, and those inferred from animal models), to reduce bias in the integrated set. G-D associations were extracted from the sources listed in Table 6.1. Only G-D associations that contained diseases mappable to MeSH Unique Identifiers (UIs) were included in the work. The mapping between UMLS ® Concept Unique Identifiers (CUIs) and MeSH was done using the Metathesaurus ® and was used for associations captured in BeFree, CTD and SemRep. In SemRep associations were extracted between gene and disease that were of the following predicates: AFFECTS; ASSOCIATED_WITH; AUGMENTS; CAUSES; PREDISPOSES; COEXISTS_WITH and NEG_ASSOCIATED_WITH as described in [124]. Next, all 2,208 mappings present between OMIM and MeSH identifiers were extracted from ORDO. This set of mappings was extended to 3,967 using a manually curated mapping set of 3,029 (with overlap). The mapping between OMIM and MeSH was then used to parse associations captured in MGD, OMIM and UniProt. For G-D associations in GWAS, a manually curated mapping between 1,131 GWAS traits and MeSH UI was used with a threshold of 1e-7 [224]. G-D associations from RGD required no mapping as diseases were already mapped to MeSH UIs. Associations from MGD and RGD were done so using the same parameters used by DisGeNET.

GoF and LoF searches were performed across Medline abstracts using automatic text mining with Linguamatics I2E. Sentences needed to contain three semantic entities: a

Table 6.1: Summary of integrated gene-disease associations.

Source	**Version/Accessed**	**Type**	**#Ass.**	**#Map**	**% Map**
CTD [225]	Jul_02_2015/Aug '15	Curated	24,346	23,813	97.8
OMIM ® [121]	18-08-2015/Aug '15	Curated	5,143⋆	3,375	65.6
Orphanet [215]	2015_07_31/Jul '15	Curated	6,094	1,744	28.6
UniProt [214]	2015_08	Curated	4,679	3,203	68.5
GWAS Cat [122]	24_08_2015/Aug '15	Experimental	13,326	5,112	38.4
BeFree [124]	24-Aug-2015/Aug '15	Literature	330,888	233,264	70.5
GoF/LoF◃	-/Oct '15	Literature	4,793	3,459	72.2
SemRep [216]	25/Feb '15	Literature	96,024	72,908	75.9
MGD[200, 202]	24_08_2015/Aug '15	Predicted	1,943	1,577	81.2
RGD[201]	21_08_2015/Aug '15	Predicted	7,667	7,667	100

Note: Data sources used for G-D associations. *Note:* '# Ass.' = Associations from source, '#Map' = number mapped to MeSH, '% Map' = percentage mapped to MeSH. 'Curated' refers to manually curated associations, 'Experimental' refers to associations drawn directly from genetic experimental observations, 'Literature' refers to associations automatically mined from literature and 'Predicted' refers to associations statistically inferred from animal models. ⋆ Not including 1,397 associations for which the molecule basis is unknown. ◃ See Table. C.2 for a breakdown of these.

disease, a gene and a loss or gain of function phrase. For the gain of function any of the following phrases were sought: "gain-of-function"; "gain of function"; and "activating mutation" (and morphological variants of activating). Similar phrases were sought for the loss of function. To increase accuracy, disease and gene semantic terms were filtered to exclude the most ambiguous terms (by using Linguamatics I2E's disambiguation score $\leq$ 75). The disease terms were automatically mapped to MeSH Tree Terms and the genes to NCBI Entrez Gene Id. Data was prepared as a triplet: disease, gene, GoF/LoF[1]. MeSH tree terms had to be mapped to MeSH UIs.

All sources included in the integrated data set as well as the data types included in the network are detailed in Table C.1.

6.4 Results

An approach to identify novel Dr-D associations from an integrated target network was developed as part of this project. Moving away from typical manual hypothesis generation the approach described uses data to inspire hypothesis generation. The

[1] These associations were produced on 16th October 2015.

approach is made up of five main components which are described in Fig. 6.1.

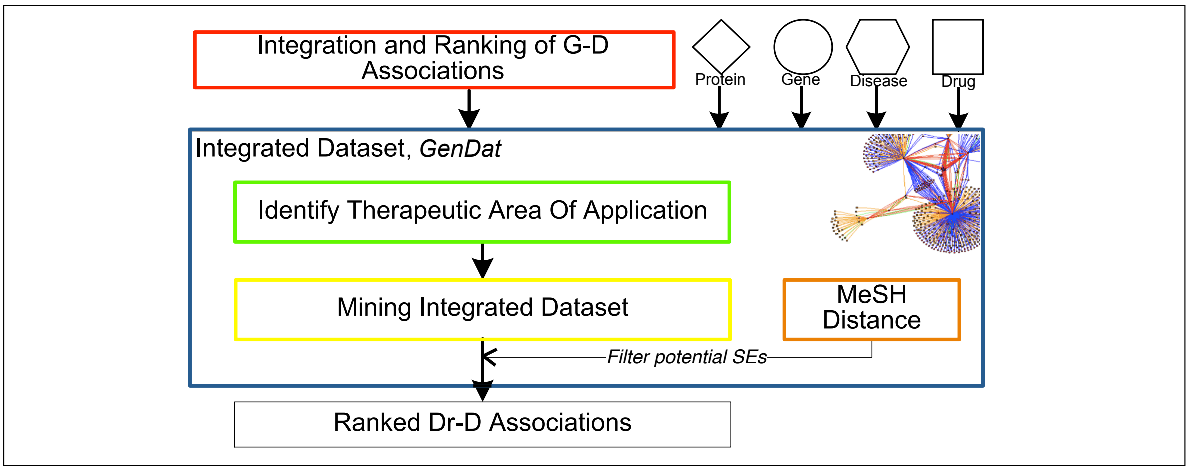

Integration and Ranking of G-D Associations
Protein
Gene
Disease
Drug
Integrated Dataset, *GenDat*
Identify Therapeutic Area Of Application
Mining Integrated Dataset
MeSH Distance
Filter potential SEs
Ranked Dr-D Associations

Figure 6.1: **Overview of approach to identify novel drug-disease (Dr-D) associations.** Gene-disease associations from 10 sources are first integrated and ranked (red). These scored associations are then integrated with protein, gene, disease and drug data to produce an integrated drug repositioning data set, *GenDat* (blue). A therapeutic area of application is then identified (green) before the data set is mined for instances of a semantic subgraph whose mappings contain inferred Dr-D associations (yellow). Finally, any Dr-D associations that are likely side-effects (SEs) are filtered using a MeSH distance measure (orange), before all ranked Dr-D associations are returned.

These components comprise: (i) Integration and ranking of G-D associations (ii) Creation of a semantically integrated data set for target-based drug repositioning, using scored G-D associations, protein, gene, disease, and drug data (iii) Identifying a therapeutic area of application, this method uses only the integrated network (iv) Mining of a semantic subgraph whose mappings in the integrated network allows novel uses for existing drugs to be inferred (v) Calculating MeSH distance, a method for calculating the semantic equivalence between two diseases within the MeSH hierarchy is also described and is used during the filtering of results.

6.4.1 Integrating and ranking of gene-disease associations

This component of the approach is shown in red in Fig. 6.1. After mapping all G-D associations to MeSH, a relatively even spread of G-D associations across all 29 therapeutic areas of the MeSH hierarchy was observed, with C04, C10, C16 and C23 being slightly overrepresented (Fig. 6.2A.). Associations from OMIM and UniProt are, on average, captured in more than three of the other data sources, while, on average, there is little crossover between associations captured in BeFree, GoF/LoF, RGD and SemRep (Fig. 6.2B.).

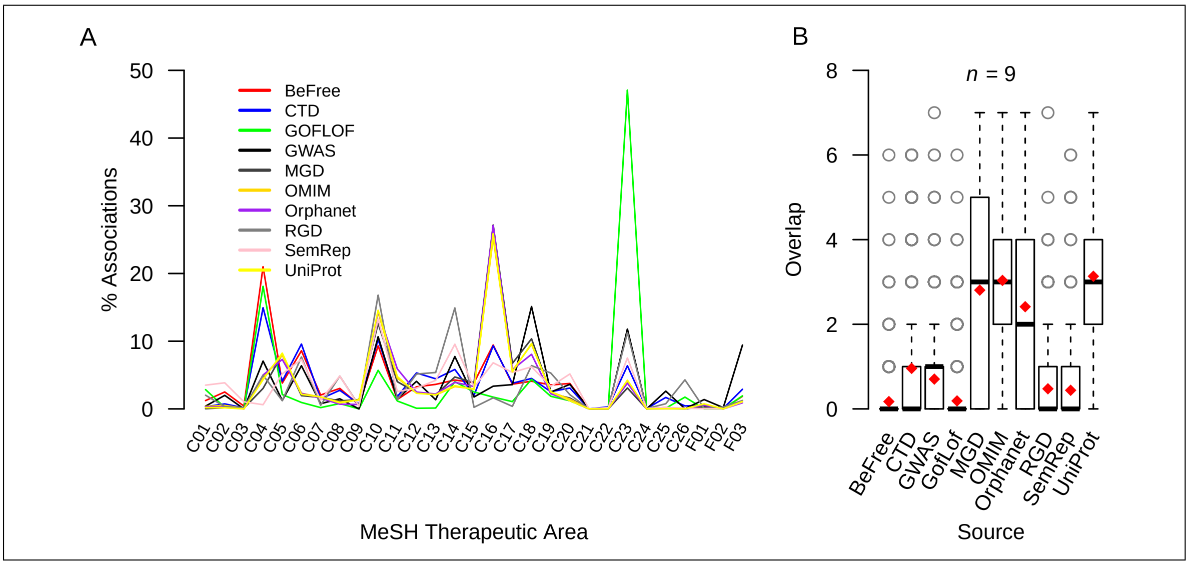


Figure 6.2: **Comparison of gene-disease (G-D) sources.** (A) Shows the percentage spread of G-D associations from each integrated data source across the 29 MeSH therapeutic areas. (B) A boxplot showing the overlap of G-D associations between the ten data sources. 1,000 associations were picked at random from each data source, listed on the x axis. For each of these associations the number of the remaining nine data sources that also contained this association were counted. Overlap could therefore be between zero (the association was not captured in any other dataset) and nine (the association was present in all data sources), shown on the y axis. NOTE: n = number of data sources checked, red diamonds show the mean, open circles are outliers and the median is represented by the thick horizontal black lines.

Once G-D associations were standardised in the format '[HUGO gene ID][MeSH Disease UI]', the next step was to rank these associations based on the evidence that supported them. First a log-likelihood score (LLS) used to score each data set was defined and is shown in Equation 6.1:

$$lls^L(E) = log\left(\frac{P(L|E)/\neg P(L|E)}{P(L)/\neg P(L)}\right) \tag{6.1}$$

In Bayesian terms, the ratios $P(L)/\neg P(L)$ represents the *prior* odds ratio, which is the ratio of the probability of the G-D linkage (L) and its negation before the evidence (E) is seen. The ratio $P(L|E)/\neg P(L|E)$ represents the *posterior* odds ratio, that is the probability of the G-D linkage in the light of the new data. For the *prior* odds ratio $P(L)$ was calculated by dividing the total number of true positives in E

(the number of G-D associations in both E and the gold standard) over all possible G-D combinations (calculated using all genes and all diseases captured in the gold standard) whilst $\neg P(L)$ was calculated by dividing the total number of false positives in E (the number of G-D associations in E but not in the gold standard) over all possible G-D combinations. For the *posterior* odds ratio $P(L|E)$ was calculated by dividing the number of true positives in E over the total number of associations in E whilst $\neg P(L|E)$ was calculated by dividing the total number of false positives in E over the total number of associations in E. The LLS is, therefore, proportional to the accuracy of the data source and is estimated by counting the number of G-D pairs with a known interaction and those without any shared annotation among all possible G-D pairs captured in the data.

The confidence scores were then integrated using the weighted sum (WS) as described by [212] and summarised in Equation 6.2:

$$WS = \sum_{i=1}^{n} \frac{C_i}{(D^{i-1})} \tag{6.2}$$

where C_1 is the highest confidence score and C_n the lowest confidence score computed from a set of n data sets and allows for the integration of data sets in order of their confidence scores.

Due to the fact that there exists no gold standard for G-D associations, each of the ten G-D association data sources listed in Table. 6.1 were investigated to evaluate their suitability for use as a 'gold standard' during this work. To do this, each data set in turn was selected as the gold standard and an LLS was calculated for each of the other nine data sources. The nine data sources that received an LLS were then ranked in descending order (the highest scoring data source was ranked number one, whilst the lowest scoring data source received a rank of nine). When using each G-D data source as the gold standard UniProt, on average, ranked first for the score attributed by the LLS (Table. C.3). Because of its consistently high ranking, G-D associations from UniProt were used as the gold standard for the scoring of G-D associations. Using UniProt, LLS scores for the data sets ranged from 16.57 for OMIM to 10.95 for GWAS. After testing a range of D parameters, a D value of 5.0 was used for this work as it was deemed to optimise the area under the curve (AUC) value (see Fig. C.1). Using a D value of 5.0 resulted in a total of 309,885 unique scored G-D associations, with scores ranging from 10.95 to 20.29.

6.4.2 Calculating MeSH distance

The MeSH hierarchy is rather verbose, and thus the specificity of terms is a potential problem. Due to the complexity and coverage of the MeSH hierarchy, two diseases may be synonymous yet captured in multiple parts of the taxonomy. Therefore, an implementation of the structure-based approach described by Leacock and Chodorow [221] was used to measure the distance between two diseases in the MeSH hierarchy. This component of the approach is shown in orange in Fig. 6.1. The formula used to calculate the difference is shown below in Equation. 6.3.

$$Sim(C_i, C_j) = \left(\frac{1}{MAX}\right) \times \left(-log\frac{Dist(C_i, C_j)}{2depth}\right) \tag{6.3}$$

Where MAX is the maximum mapping score, depth is the max depth of the hierarchy and Dist is the shortest path length between the two concepts, C_i and C_j. Reducing the stringency at which diseases are mapped to those in the MeSH hierarchy allows better filtering of potential noise caused by `has_side_effect` edges. For example, an inferred `has_indication` edge is made between `drugX` and `diseaseY`, while a known side-effect of `drugX` is `diseaseZ`, a child term of `diseaseY` in the MeSH hierarchy. As one of `drugX`s known side-effects is semantically similar to the inferred indication, it is fair to assume that `drugX` is not a reasonable candidate for the treatment of `diseaseY`. In this instance $Sim(\texttt{diseaseY}, \texttt{diseaseZ})$ would give a value of 0.768. Using 0.768 as the equivalence threshold (ET) during filtering means all inferred associations that are one node away, in the MeSH hierarchy, from known side-effects will be removed. Therefore, the Sim value allows for the identification of semantic ‘equivalence’ using a certain threshold or leniency, the ET.

6.4.3 *Integrated data set,* GenDat

This component of the approach is shown in blue in Fig. 6.1. Ranked and scored G-D associations were then integrated with protein, gene, disease and drug data to create a semantically-rich network, *GenDat* to aid in the identification of potential drug repositioning opportunities. *GenDat* contains 57,453 nodes and 528,930 edges. To distinguish between rare (generally monogenic) and common (often complex) diseases the Orphanet Rare Disease Ontology (ORDO) was also included. 1,779 MeSH UI were captured as synonyms within the ORDO. Wherever a `Rare_Disease` node contained a MeSH UI, the MeSH node was integrated with the ORDO disease and resulted in a `Rare_Disease` node with the synonymous MeSH UI becoming attributes. The metagraph for *GenDat* is shown in Fig. 6.3.

GenDat contains approved drugs, `Small_Molecules`, and `binds_to` interactions from these to single `Protein` targets. Wherever possible these `binds_to` associations are annotated with activity types (IC_{50}, K_d, K_i and *Potency*) and the corresponding activity values (nM). For each `Protein`, the `Gene` which it is `encoded_by` is also included. A `Gene` may also be linked to diseases, either a `Rare_Disease` or a `Common_Disease`, via `involved_in` edges. These `involved_in` edges are annotated with values produced during the G-D association ranking described previously. Finally, diseases and drugs may share `has_indication` and `has_side_effect` edges. *GenDat* includes data types to enable target-based drug repositioning opportunities to be identified.

6.4.4 *Identifying a therapeutic area of application*

Using both G-D associations and Dr-D associations from *GenDat* a therapeutic area unmet score Therapeutic Area Unmet Score (TAU) was calculated, using the formula in Equation 6.4;

$$TAU(ta) = \neg P(DD) \times P(GD) \times \left(1 - \frac{1}{MAX} \times |ta|\right) \tag{6.4}$$

Where *ta* is the therapeutic area being looked at, e.g. C01, *P(GD)* is the probability that the data contains a G-D association for a disease in that *ta*, ¬ *P(DD)* is the

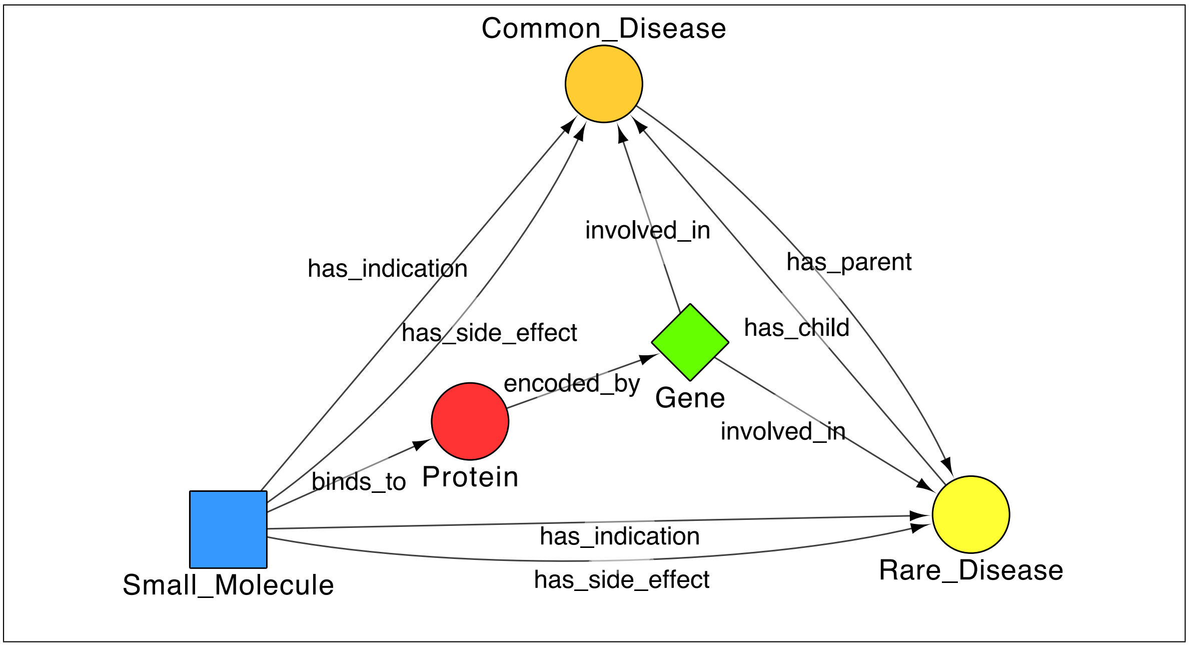


Figure 6.3: **Metagraph of *GenDat*.** Metagraph shows the node types and the edge types used in *GenDat* and how they interact to one another.

probability that the data does not contain a Dr-D association for a disease in that *ta* and MAX represents the size of the greatest *ta*. The TAU, in theory, can range from 0 to 1. A score of 0 represents a therapeutic area that contains few, highly drugged diseases, with little knowledge captured in *GenDat*. A score of 1 represents a therapeutic area that contains a great number of diseases that have relatively few marketed treatments and have high levels of knowledge describing them in *GenDat*.

A simple equation to calculate a Rich Therapeutic Area (RTA) score was defined and is shown in Equation 6.5. Equation 6.5 uses the same notation as Equation 6.4 and can also produce scores from 0 (areas with little knowledge captured in the data set) to 1 (areas with a lot of knowledge captured in *GenDat*), using the following:

$$RTA(ta) = P(DD) \times P(GD) \tag{6.5}$$

In order to identify a therapeutic area of focus for this work a therapeutic area of unmet need first needed to be identified. To do this two relevant data types from *GenDat*, G-D associations and Dr-D associations were utilised. As the approach to inferring novel Dr-D associations described in this Chapter utilises G-D associations for predicting target-based drug repositioning, it is important to target therapeutic areas for which a large proportion of the contained diseases have data supporting their genotypic mechanisms. Indications cannot be inferred for areas where there is no data describing them in the network. The percentage of each therapeutic area that was involved in at least one `is_involved_in` edge was, therefore, calculated; this is shown in Fig. 6.4A. It was then necessary to identify a therapeutic area that had G-D associations describing the diseases, but also had fewer small therapeutic molecules. The percentage of each area for which there already exists a marketed small therapeutic molecule was calculated. The percentage was calculated using the `has_indication` relations present in the network (see Fig. 6.4B).

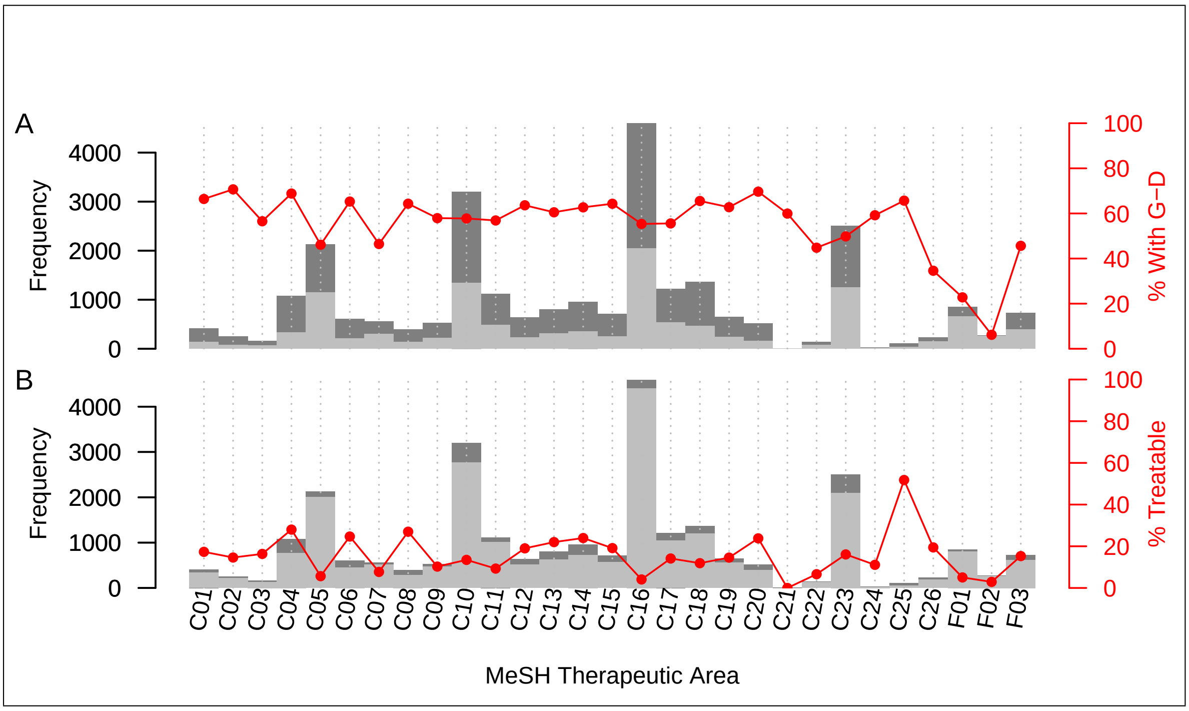


Figure 6.4: **Gene-disease association and indications data captured in *GenDat* for each therapeutic area.** (A) Dark grey bars show the frequency of diseases in each therapeutic area of the MeSH hierarchy. Light grey portions of the bar show the number of those diseases that are not involved in any of the gene-disease associations captured in *GenDat*. Red shows the percentage of diseases in that therapeutic area involved in one or more gene-disease association(s). (B) Dark grey bars shows the number of diseases in each therapeutic area of the MeSH hierarchy. Light grey portions of the bar show the number of those diseases that currently do not have a small molecule treatment on the market. Red shows the percentage of diseases in that therapeutic area that has one or more treatment(s) on the market, as captured in *GenDat*. *Note:* please see Table C.4 for therapeutic area descriptions.

When calculating a therapeutic area which had a relatively large amount of knowledge captured in *GenDat*, its size, or cardinality, also has to be considered. For example, the

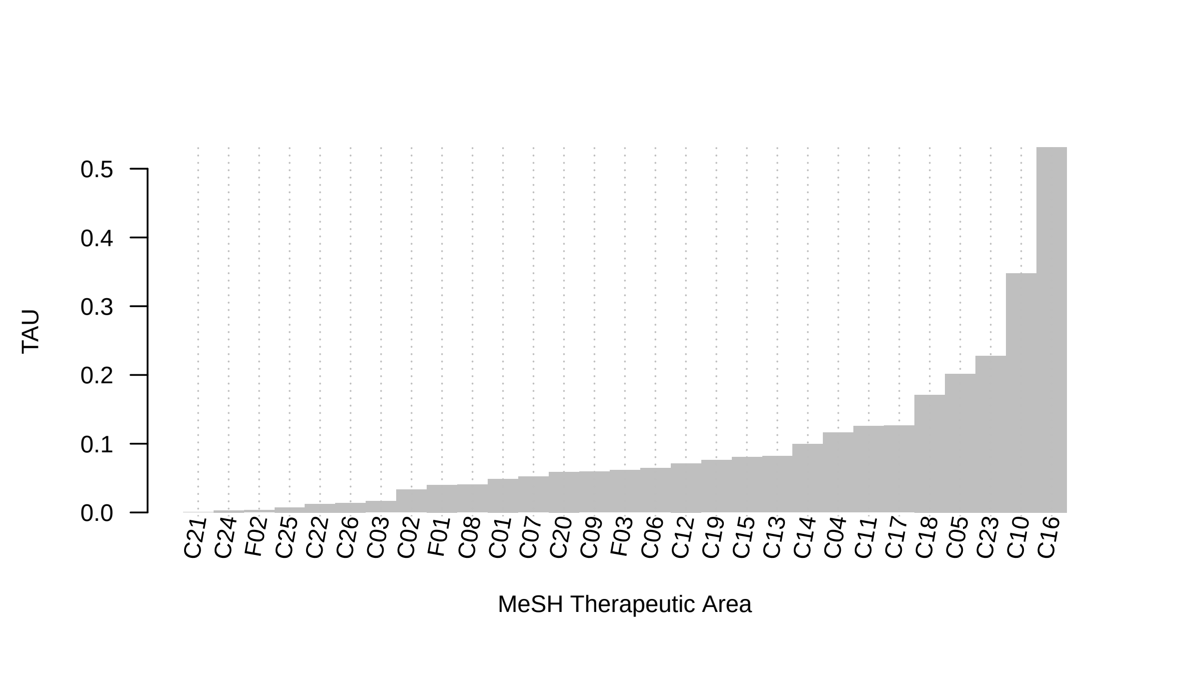


Figure 6.5: **Ranking therapeutic areas in terms of their TAU score.** Using Equation 6.4 each therapeutic area in the MeSH hierarchy was scored. The TAU score considers how much data is captured in *GenDat* and the percentage of diseases in that therapeutic area that do not have a marketed small therapeutic molecule. The higher the TAU the more likely the therapeutic area is to be an area of unmet need.

therapeutic area C25 (chemically-induced disorders) has Dr-D associations for 52 % of the disorders which are contained within the term, as well as G-D associations for 65 %. Looking at these two values alone it can be seen that there is a relatively large amount of data for this area. However, it is made up of only 108 diseases and makes up only 0.4 % of diseases in *GenDat.* To avoid identifying such small areas, the focus was placed only on therapeutic areas that represent over 3.44 % (there are 29 therapeutic areas and so 100/29) of the total diseases captured in *GenDat* to identify a rich therapeutic area.

With a TAU (Equation 6.4) of 0.53, it is shown that C16 (hereditary diseases) is the largest unmet therapeutic area (Fig. 6.5) and C10 (diseases of the central nervous system) achieves the second highest TAU of 0.38. The work here focusses on approved small molecules, as these drugs have already passed safety tests and are easier to reposition. Many genetically simple hereditary diseases, such as those captured in the MeSH therapeutic area C16 are not suited to this type of treatment, as some are untreatable, and others are caused by gene knockouts. Instead, hereditary diseases

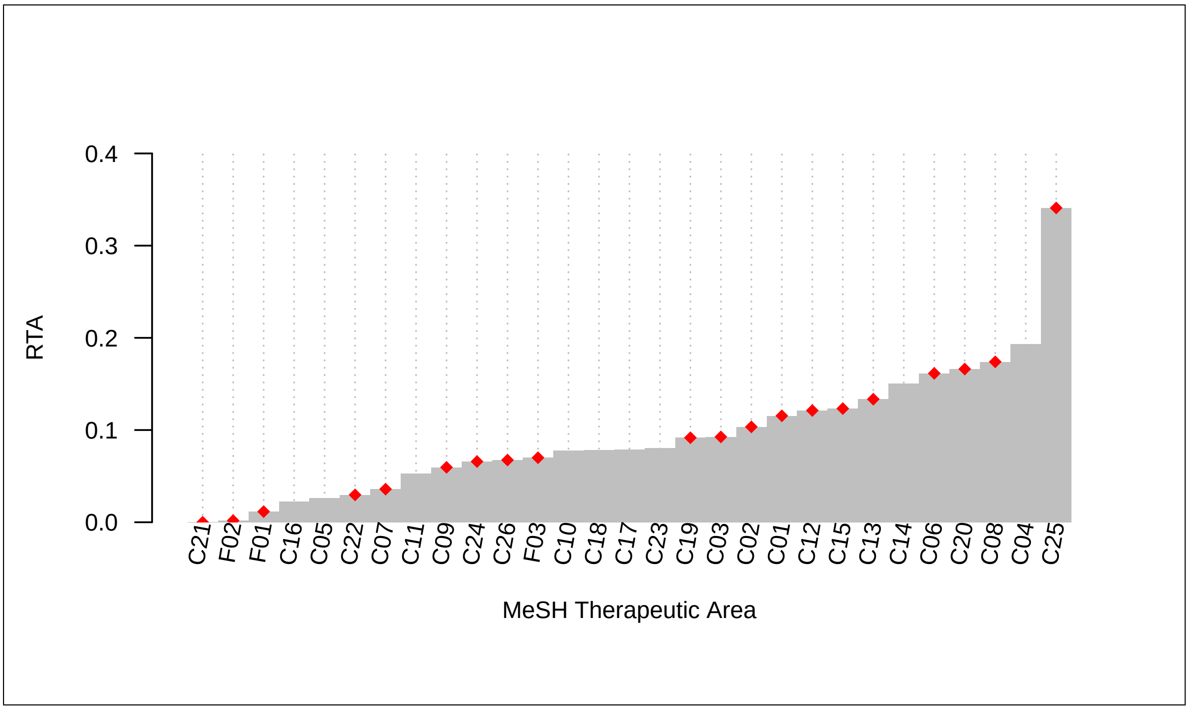


Figure 6.6: **Ranking therapeutic areas in terms of their RTA score.** Using Equation 6.5 each therapeutic area in the MeSH hierarchy was scored. The RTA score considers how much data is captured in *GenDat* for each therapeutic area. *Note:* red diamonds show the therapeutic areas which include <3.44 % (100/29) of all diseases. For the purpose of this exercise, they will not be considered for analysis as they do not offer a fair representation of the data included in the work. The higher the RTA the more likely the therapeutic area is to be well 'drugged'.

tend to be better treated using metabolic manipulation, protein augmentation and gene therapy [226]. It is for this reason that the approach here is not applied to hereditary diseases and instead to the therapeutic branch C10. The therapeutic area C04 (Neoplasms) is identified as having the greatest RTA (Equation 6.5) of therapeutic areas containing more than 3.44 % of the total number of diseases in *GenDat* (Fig. 6.6). Therefore, the approach is applied to C10, an area of unmet need, and C04, an area relatively rich in data.

6.4.5 Mining integrated data set

GenDat took 64 minutes to build on a local machine (8Gb RAM and 1.8GHz Intel Core i5). Mining used an initial candidate set of 1,188 nodes (approved, small molecule drugs that target humans or other mammals) and took 13 minutes to complete. An exhaustive search returned 539,162 mappings.

The subgraph depicted in Fig. 6.7 was used as it is the most simple schematic representation of a drug-disease pathway. Searching for instances of the four-node subgraph will allow novel drug-disease associations to be identified, essentially by 'filling in the blanks'. This component of the approach is shown in yellow in Fig. 6.1.

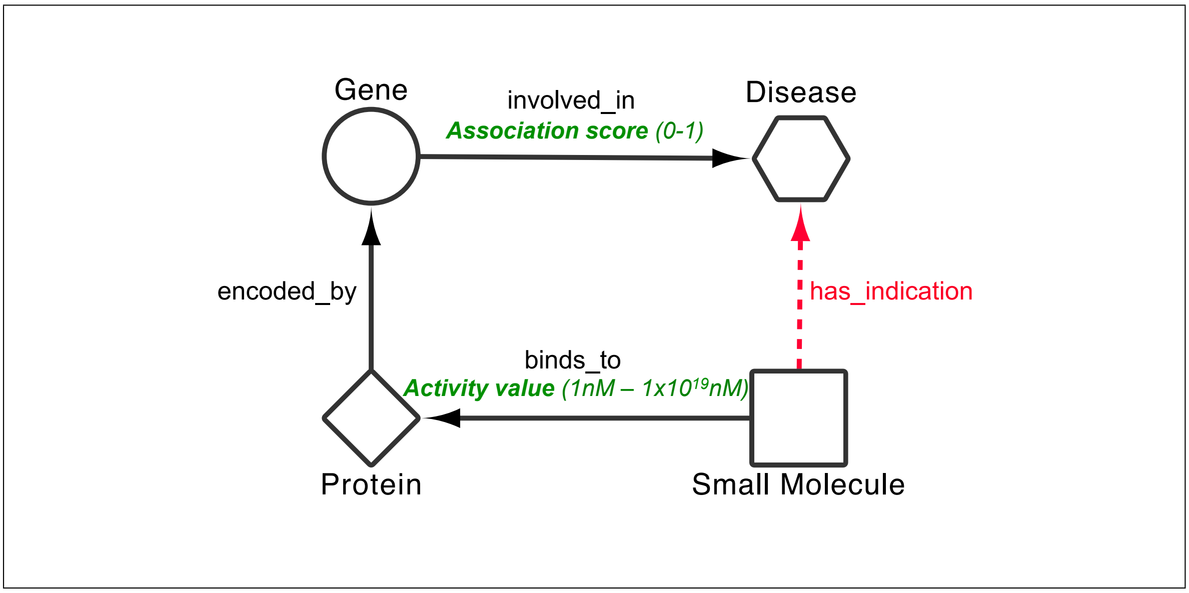


Figure 6.7: **Semantic subgraph used to infer novel drug-disease associations from the *GenDat*.** Subgraph represents the simplest approach to schematically representing the route from drug to disease using target-based approaches to drug repositioning. Through identifying mappings of the subgraph in the *GenDat* the aim is to infer the red `has_indication` relations. Mappings are scored using the values captured in the *Activity value* and *Association score* attributes (shown in green) found on the `binds_to` and the `involved_in` relations, respectively. *Note:* in mappings 'Disease' can be either a `Common_Disease` or a `Rare_Disease` and a 'Drug' is an approved `Small_Molecule`.

Mappings, M, were scored and ranked using the *Activity value* and *Association score* values attached to the `involved_in` and `binds_to` relations respectively, using equation 6.6:

$$Score(M) = \frac{\diamond Activity\ value\ (M) + Association\ score\ (M)}{2} \tag{6.6}$$

The *Association score* captured on the `involved_in` relations were created during the G-D ranking section of the approach, and ranged between 0-1. The *Activity value*, attached to the `binds_to` relation was extracted from ChEMBL and included values associated to: IC_{50}; K_i; K_d and potency, all of which had values ranging from 1nM - 1×10^{19} nM. The *Activity value* for each `binds_to` association was normalised

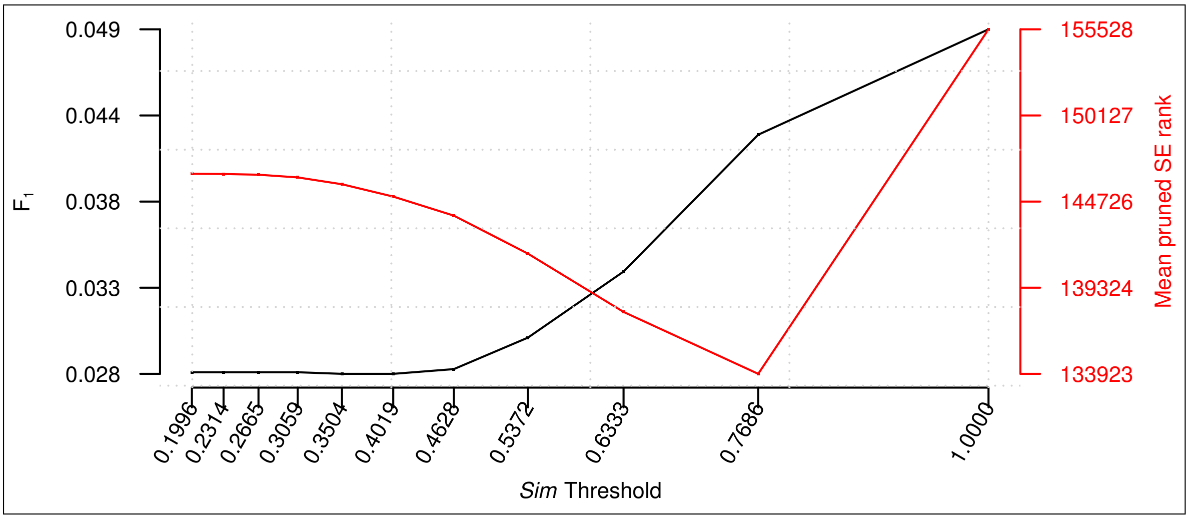


Figure 6.8: **Calculating *Sim* threshold for pruning potential side-effects from inferred indications.** This figure provides a graphical representation of the data captured in Table C.5. For each threshold, the F-Measure, F_1 (using precision and recall of known indications captured in the network), shown in black, as well as the average ranking position of the excluded potential side-effects, shown in red, were calculated. To calculate the ranking positions of those excluded, all associations inferred by the methodology were ranked before any filtering, and it was these rankings used throughout all subsequent analysis. The aim of filtering out potential side-effects was to reduce noise in the results while also ensuring potential indications were not excluded. It was assumed that the associations scoring higher, and thus rank higher (highest being 1), are predicted with more confidence, and thus potential side-effects are excluded from the highest ranking associations.

to the same range as the *Association score*, to give $\diamond$*Activity value*; this was done simply by subtracting $\log_{10}(Activity\ value) \times 0.1$ from 1, where 1 is the maximum $\log_{10}(Activity\ value)$ captured in *GenDat*. The mean *Activity value* from ChEMBL was 0.8 and this value was assigned to `binds_to` relations taken from DrugBank which do not come with *Activity value* data.

Steps were then taken to filter results in order to remove as much 'noise' as possible. Mappings containing predicted `has_indication` edges that were known side-effects (captured as `has_side_effect` relations in *GenDat*) were removed. Mappings that predicted `has_indication` edges with a *Sim* value $\geq$ 0.768 to known `has_side_effect` edges were also dismissed as potential side-effects. An equivalence threshold of 0.768 was used as it gave the best balance between precision and recall of the known `has_indication` edges while also pruning, on average, the highest ranked inferred associations (Fig. 6.8 and Table C.5). Of the 539,162 mappings, 42,689 were classed as potential side-effects. A further 4,947 mappings were removed as the mechanism of the drug, and the G-D association directionality (LoF or GoF data) contradicted one another (e.g. the drug was an agonist and the gene was associated to a disease via a GoF relation). Finally, 41,798 mappings containing one of the 298 absorption, distribution, metabolism, and excretion (ADME) genes [227] were dismissed.

After filtering, 451,269 mappings inferring 275,934 unique(some associations were identified by more than one mapping) `has_indication` edges were left. Of all the mappings that inferred the same `has_indication` edge, the mapping that achieved the highest score was kept and used for all analysis. Inferred indications covered every therapeutic area of the MeSH hierarchy, ranging from 72,613 for neoplasms (C04) to 2 for disorders of environmental origin (C21) (Table C.4). 219,623 unique associations involved `Common_Disease` (inferred from 369,124 mappings) whilst 56,311 associations involved `Rare_Diseases` (inferred from 82,145 mappings) (see Table 6.2).

Table 6.2: Number of mappings for each disease type and therapeutic area post filtering.

	All Diseases	**Common Disease**	**Rare Disease**
All Therapeutic Areas	275,934 (451,269)	219,623 (369,124)	56,311 (82,145)
C04: Neoplasms	55,875 (102,832)	39,383 (73,501)	16,492 (29,331)
C10: Nervous System	54,635 (84,213)	41,241 (66,536)	13,394 (17,677)

Note: After applying filtering a set of mappings that inferred unique (no repeats) drug-disease associations are left. Numbers in brackets denote how many mappings inferred the unique associations.

The ability of the approach to identify known `has_indication` edges captured in *GenDat* was then investigated. All `has_indication` edges (from the four sources listed in Table C.6) that involved the 1,188 approved small molecules used during the search were extracted. Fig. 6.9 shows how the approach performs in identifying known `has_indication` edges for different therapeutic areas (all, C04 and C10) and different disease types (`Common_Diseases` and `Rare_Diseases`). Of the 18,889, known `has_indication` edges, 1,006 involved 63 drugs that although were part of the 1,188 investigated, returned no mappings, leaving 17,883 that could potentially be validated. For mapping, known `has_indication` edges to those inferred by the approach a *Sim* threshold of 0.633 was used, the equivalent of a two-node distance within the MeSH hierarchy. It is believed a *Sim* threshold of 0.633 provides the best trade-off between the verbosity of the MeSH hierarchy while also ensuring inferred diseases are close enough in disease mechanism for the proposed therapeutic small molecule to be relevant. Using the *Sim* threshold of 0.633, the approach identifies 12,955 of the known `has_indication` edges (72.65 %) (Table C.7). An AUC of 0.73 was achieved when looking at all of the inferred Dr-D associations (Fig. 6.9). The number of knowns identified by the approach can be increased to 97.6 % if the *Sim* value is relaxed to 0.231, which represents a node distance of nine in the MeSH hierarchy (Table C.7).

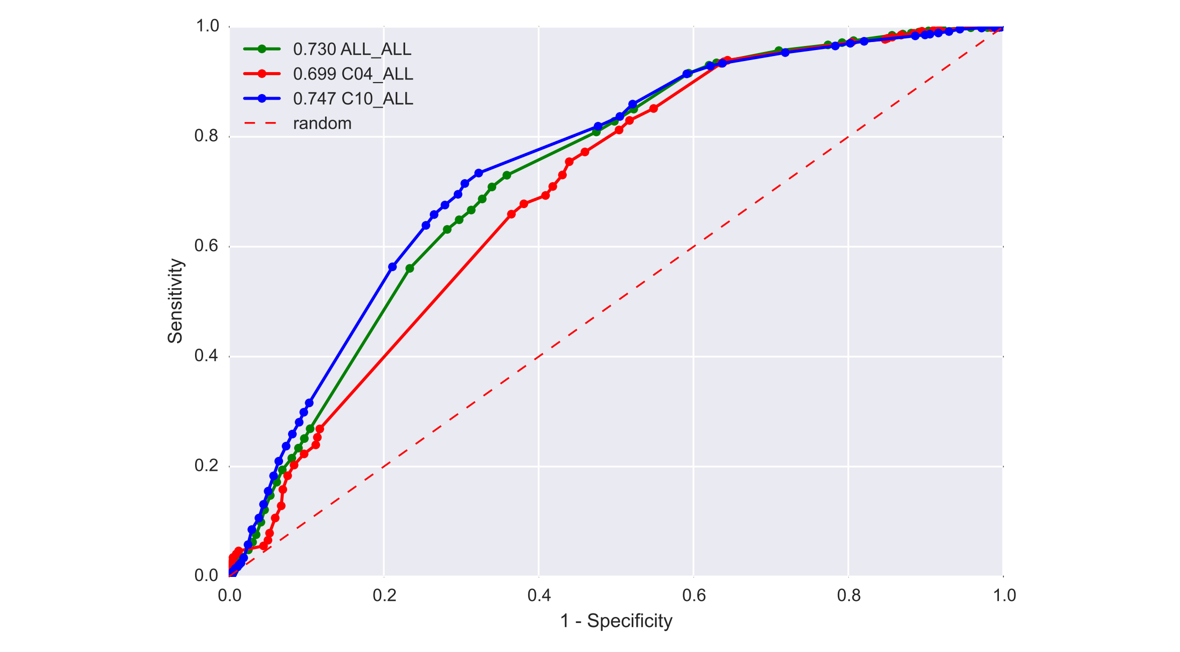


Figure 6.9: **Validating inferred `has_indication` edges.** All 18,889 `has_indication` edges captured in *GenDat* were extracted. These associations were used as means of validating the ability of the approach to identify known `has_indication` edges. *Note:* For each disease category (ALL, C04 and C10) the set of known indications were pruned to only include those relevant, mapping was done using a *Sim* value of 0.633, this is equivalent to a distance of two nodes in the MeSH hierarchy.

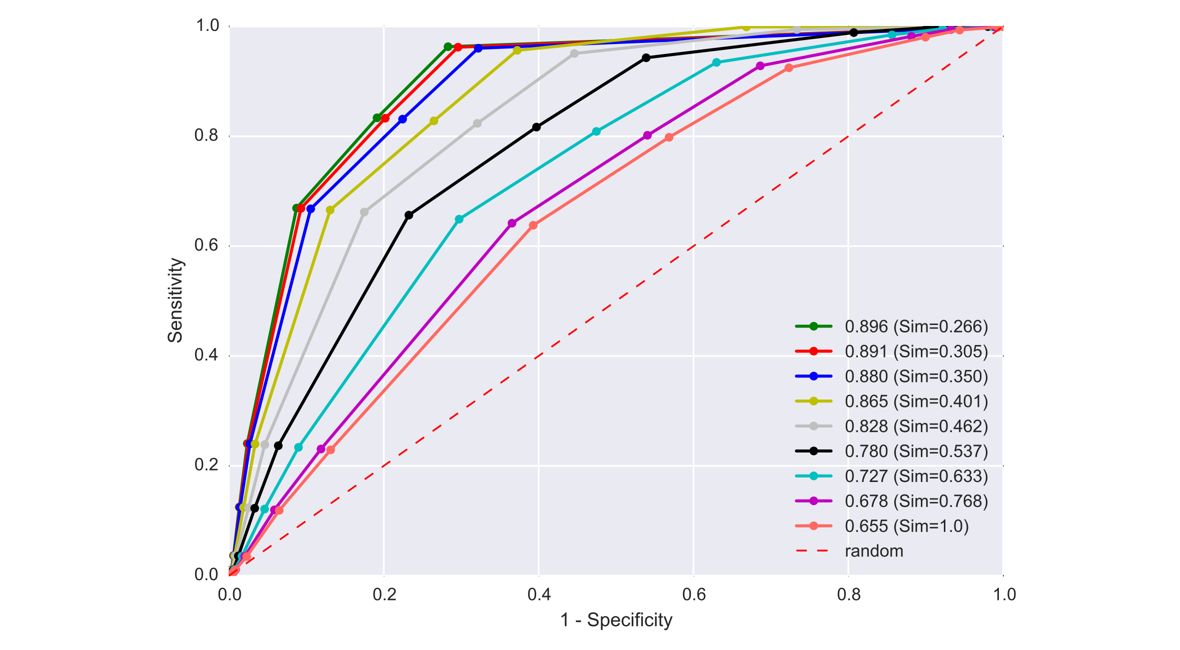


Figure 6.10: **Validating inferred `has_indication` edges using different MeSH distance measures.** The *Sim* threshold used to map inferred `has_indication` edges to the 18,889 known `has_indication` edges was altered. For each of the *Sim* values investigated the number of known `has_indication` edges they manage to identify is shown in Table C.7.

Table 6.3: Top 10 inferred associations involving neoplasm diseases.

Drug (*DrugBank ID*)	**Gene**	**Disease (*MeSH UI*)**	**Type (*ORDO*)**	**Evidence**	**Score**
Sunitinib (*DB01268*)	*KIT*	Gastrointestinal Stromal Tumors (*D046152*)	R (*44890*)	M(1.0)	0.999
Ponatinib (*DB08901*)	*FLT3*	Acute myeloid leukemia *D015470*	R (*519*)	A	0.998
Dasatinib (*DB01254*)	*EPHB2*	Familial prostate cancer (*C537243*)	R (*1331*)	C [228]	0.996
Ethinyl Estradiol (*DB00977*)	*ESR1*	Breast Neoplasms (*D001943*)	C	M(1.0)	0.988
Dasatinib (*DB01254*)	*BCR*	Myelogenous, Chronic, BCR-ABL Positive (*D015464*)	C	M(1.0)	0.988
Pazopanib (*DB08901*)	*KIT*	Mastocytosis (*D008415*)	R (*98292*)	-	0.984
Afatinib (*DB08916*)	*ERBB2*	Stomach Neoplasms (*D013274*)	C	-	0.973
Sunitinib (*DB01268*)	*RET*	Multiple endocrine neoplasia type 2B (*D018814*)	R (*247709*)	-	0.961
Sunitinib (*DB01268*)	*RET*	Pheochromocytoma (*D010673*)	C	C [229]	0.960
Sunitinib (*DB01268*)	*NTRK1*	Familial medullary thyroid carcinoma (*C536911*)	R (*99361*)	P [230]	0.958

Note: The top ranked 10 inferred `has_indication` edges involving neoplasms are presented. All ranked associations are available for download. A disease is classed as Rare (R) if it maps to ORDO and Common (C) if it is only in MeSH and not mappable to an ORDO concept. Evidence: M = maps to indications in data set with *Sim* 0.66 or above; A = approved; C = clinical trial; and P = scientific paper.

Table 6.4: Top 10 inferred associations involving diseases of the nervous system.

Drug (*DrugBank ID*)	**Gene**	**Disease (*MeSH UI*)**	**Type (*ORDO*)**	**Evidence**	**Score**
Nitrendipine (*DB01054*)	*CACNA1S*	Hypokalemic periodic paralysis (*D020514*)	R (*681*)	-	0.999
Clonazepam (*DB01068*)	*GABRA1*	Juvenile myoclonic epilepsy (*D020190*)	R (*307*)	M (0.76)	0.999
Mifepristone (*DB00834*)	*ESR1*	Bulbospinal neuronopathy, X-linked recessive (*C537017*)	C	-	0.999
Memantine (*DB01043*)	*GRIN2A*	Landau-Kleffner Syndrome (*D018887*)	R (*98818*)	-	0.996
Bromocriptine (*DB01200*)	*DRD2*	Myoclonus-dystonia syndrome (*C536096*)	R (*36899*)	-	0.994
Roflumilast (*DB01656*)	*PDE4D*	Acrodysostosis (*C538179*)	R (*950*)	-	0.991
Lisinopril (*DB00722*)	*ACE*	Alzheimer Disease (*D000544*)	C	-	0.991
Roflumilast (*DB01656*)	*PDE4D*	Stroke (*D020521*)	C		0.987
Clonazepam (*DB01068*)	*GABRB3*	Epilepsy, Absence (*D004832*)	C	M (1.0)	0.991
Triazolam* (*DB00897*)	*GABRG2*	Generalized Epilepsy With Febrile Seizures Plus, Type 3 (*C565811*)	C	-	0.988

Note: The top ranked 10 inferred `has_indication` edges involving unique diseases of the central nervous system are presented. All ranked associations are available for download. A disease is classed as Rare (R) if it maps to ORDO and Common (C) if it is only in MeSH and not mappable to an ORDO concept. Evidence: M = maps to indications in data set with *Sim* 0.66 or above; A = approved; C = clinical trial; and P = scientific paper. (*This drug has been withdrawn in the UK due to risk of psychiatric adverse drug reactions, but continues to be available in the U.S)

6.4.5.1 C04: Neoplasms

55,875 unique has_indication inferred edges involved neoplasms (therapeutic are C04 of the MeSH hierarchy). 16,492 of these unique associations involve Rare_Diseases (inferred from 29,331 mappings) while 39,383 unique has_indication edges involving Common_Diseases were identified (inferred from 73,501 mappings). Of the 2,856 known has_indication Dr-D associations the approach identifies 1,927 of these or 68 %. 455 of the knowns involve 28 drugs that this approach was unable to infer associations for, due to lack of data, giving an 80 % identification rate for known Dr-D associations involving neoplasms, with an AUC of 0.69 (Fig. 6.9).

Of the top 10 ranked inferred Dr-D associations involving neoplasms (Table 6.3), it can be seen that three map exactly to indications in *GenDat*. Furthermore, one is currently being investigated in a clinical trial [229], one has been previously studied in the clinic [228], one is now approved for the indication proposed and the literature supports another. Of the top 10 inferred indications, three are novel and are currently not supported by evidence. One of those indications is the use of pazopanib (Votrient; Novartis) in the treatment of Mastocytosis.

Pazopanib as a treatment for Mastocytosis? Pazopanib is a small molecule inhibitor of multiple protein tyrosine kinases and is approved for the treatment of advanced renal cell carcinoma and advanced soft tissue sarcomas. Mastocytosis, classed as a rare disease, is a mast cell activation disorder of both children and adults caused by the presence of too many mast cells (mastocytes) and CD34+ mast cell precursors. The cause of mastocytosis is not known but activating mutations in the proto-oncogene receptor tyrosine kinase, *KIT*, are found in most patients with mastocytosis [231]. The mutation makes mast cells more sensitive to stem cell factor (SCF). SCF plays an important role in stimulating the production and survival of cells such as blood cells and mast cells, inside the bone marrow. When bone marrow is exposed to SCF, it produces more mast cells than the body can cope with, leading to symptoms of mastocytosis [231]. Although no official treatment exists for mastocytosis many drugs are prescribed off-label, including the tyrosine kinase inhibitors, dasatinib, imatinib and masitinib [231]. Due to the fact that that pazopanib displays inhibitory effects on the

KIT enzyme similar to those that have been used as off-label treatments, it poses an interesting alternative in the treatment of mastocytosis.

6.4.5.2 C10: Nervous System Diseases

54,635 unique `has_indication` inferred edges involved diseases of the nervous system (therapeutic area C10 of the MeSH hierarchy). 13,394 of these unique associations involve `Rare_Diseases` (inferred from 17,677 mappings) whilst 41,241 unique `has_indication` edges involving `Common_Diseases` were identified (inferred from 66,536 mappings). Of the 4,249 known `has_indication` Dr-D associations the approach identifies 2,846 of these. 125 of the knowns involve 37 drugs that, due to holes in the data, the approach was unable to infer associations for, a 69.0 % identification rate for known Dr-D associations involving nervous system diseases, with an AUC of 0.75 (Fig. 6.9).

Of the top 10 ranked inferred Dr-D associations involving diseases of the nervous system (Table 6.4), it can be seen that only one maps exactly to an indication in *GenDat* while another maps with a *Sim* of 0.66 (MeSH distance of two nodes). Another eight are novel and are currently not supported by evidence. One of those indications is the use of lisinopril (Zestril; AstraZeneca) in the treatment of Alzheimer Disease.

Lisinopril as a treatment for Alzheimer Disease? Alzheimer Disease is a chronic neurodegenerative disease that usually starts slowly and gets worse over time and currently has no cure. Lisinopril, a potent, competitive inhibitor of angiotensin-converting enzyme (ACE), is used to treat hypertension and symptomatic congestive heart failure. There is evidence to suggest that Angiotensin-converting enzyme inhibitors and the reduced risk of Alzheimer's disease in the absence of apolipoprotein E4 allele [232]. As such, lisinopril is proposed as a potential treatment for Alzheimer's disease.

6.5 Discussion

In this Chapter, an approach for inferring novel drug repositioning leads was explored. As part of this approach diseases of the nervous system were identified as a therapeutic area in need of more treatments. G-D associations were integrated and ranked from multiple data sources and as a consequence, the need for a standard representation of G-D associations was made apparent. These ranked associations were used to create *GenDat*, a semantically-rich integrated network for drug repositioning. It was also shown how mining *GenDat* for semantic subgraphs can enable the inference of novel Dr-D interactions.

The UMLS contains over 3 million concepts (covering anatomical structure, biological function, chemical, disease or syndrome, laboratory or test result, medical device, and organism), MeSH therapeutic areas are made up of 11,735 concepts and the DO 10,905 concepts. As part of this work, calculations showed that of the diseases caught in DisGeNET, 100% mapped to UMLS, 60% mapped to MeSH and 24% mapped to the DO. At present, it appears that MeSH offers the best trade-off between interoperability and semantic clarity. It was for this reason that MeSH was used during this work.

Two MeSH therapeutic areas were identified to focus on, one, neoplasms, or C04, with a relatively rich knowledge base, and one, diseases of the nervous system, or C10, containing many diseases that are currently in need of a therapeutic molecule. It is seen, as expected, that the approach performs better when looking at C04 diseases in comparison to the less treated and less informed C10 diseases; highlighting the fact that systems approaches are limited by the data available. Limited data may become more of a problem in the long term, especially when it comes to developing treatments for diseases of the central nervous system. Clinical trials are very expensive in the area of nervous system diseases, due to the placebo effect, meaning that great numbers of trialists are needed. As a result, many companies are withdrawing their development efforts from this area, making nervous system diseases a great area of opportunity for repositioning, and in particular *in silico* approaches. The approach does not address the problems caused by the placebo effect. Rather, by bringing data together, in a similar fashion to the clinician, it is hoped that as more data becomes available, this approach can reduce the attrition rates while also improving efficacy.

The approach presented here makes use of a MeSH distance measure, *Sim* (see Equation 6.3). The *Sim* measure is used twice during the approach. A *Sim* value of 0.768 is used for filtering potential `has_side_effect` edges, equivalent to a one node path from a known side-effect. A lower *Sim* value of 0.633 is used to validate inferred `has_indication` edges against the known indications captured in the network, equivalent to a two node path. The two values vary as they are used for different purposes. If the stringency was reduced to filter potential side-effects, a lot of the true positives are quickly lost (Table C.5). Indeed by filtering potential side-effects using a *Sim* of 0.633 instead of 0.768 would result in a loss of 31% of true positive inferred `has_indication` associations. When validating inferred `has_indication` edges, the lower the *Sim* value, the greater the AUC (Fig. 6.10). It is believed that, in this instance, a *Sim* value of 0.633 gives the best trade-off between AUC and maintaining semantic ‘equivalence’, when it comes to validation. These differing *Sim* values reflect the manner in which drugs are marketed, with indications being as high level as possible for marketing reasons. On the other hand, side-effects tend to map to a greater level of granularity and so do not require less stringent mapping.

Possible extensions to this approach should include more thorough analysis regarding the identification of disease areas of interest. Instead of simply identifying a therapeutic area that appears to be relatively untreated one could consider other factors for disease prioritisation. For example, not all diseases have the same impact on society and so integrating data that considers the stress a disease places on society would be useful. For example, the WHO global burden of disease measures burden of disease using the disability-adjusted-life-year (DALY).

As well as a more thorough disease prioritisation step, more focus must be placed on directionality, both regarding the effect on function of the gene mutation and the drug functionality (e.g. agonist, antagonist). No data source details the effect of function that a gene mutation has; i.e. does it result in LoF (the gene product has less or no function) or GoF mutation (product of mutated gene gains a new and abnormal function). Although a text mining approach was used as part of this work to try and identify the effect on gene function of a mutation (see Section 6.3) this was not exhaustive. Drug functionality must also be considered if this work is truly to provide detailed inferences. It was possible to get drug functionality for around 500 drugs from ChEMBL, but this did not cover all drugs in *GenDat*. Problems arising from limited drug functionality data are highlighted by the first ranked inferred association from the diseases of the central nervous system (Table 6.4). Nitrendipine, a potent blocker of the calcium channel (CACNA1S), is proposed as a treatment for Hypokalemic periodic paralysis. Although both the `binds_to` and `involved_in` edges are correct, the lack of directionality attached to the G-D association makes this particular inference a poor one. Nitrendipine is annotated as being an inhibitor of CACNA1S in the data set, as such if the mutation involved of CACNA1S had been correctly annotated as a LoF mutation, this inference would have been filtered as contradictory. As such, the administration of Nitrendipine as a treatment for Hypokalemic periodic paralysis is likely to exacerbate the condition as opposed to treating it.

Despite this approach allowing for an initial reduction of the search space the next step would require a more robust filtering of the results. One would need to ensure that the target could indeed be reached by the drug, i.e. if a compound is unable to pass the membrane the target must be located on the surface of the cell. Looking at the cellular location of targets, which could be extracted from GOA, as well as the physiochemical properties of the compound, from DrugBank or ChEMBL, may allow for more accurate inferences to be made.

A strategy for mining for potential drug repositioning opportunities was introduced in this Chapter, however, at the moment, it can be seen that this is limited by the data available. The approach paves the way for more stringent ontological representation of G-D associations; like the Experimental Factor Ontology (EFO) work being carried out at the Centre for Therapeutic Target Validation (CTTV). Furthermore, the Open Targets Platform[2], which at the time of writing is still in its infancy, will utilise the EFO and provide systematic access to a mass set of integrated of G-D associations. It is believed that as the quality of data increases this *in silico* approach will complement target identification and validation; reducing target attrition through efficacy.

[2] https://www.opentargets.org

6.6 Conclusion

This Chapter has presented a method for integrating and ranking G-D associations from numerous data sources. Scored G-D associations were then integrated with other relevant data to produce *GenDat*, a target-based drug repositioning network. A method for the automated identification of therapeutic areas of interest was introduced, enabling data-driven hypotheses to be generated. A simple four-node semantic subgraph was described that best represents, schematically, an abstract view of the drug-disease pathway. Instances of this subgraph were searched for in *GenDat*, using the previously described DReSMin algorithm. Mappings were scored based on the evidence supporting the G-D associations and the known binding values of the drugs.

The work presented in this Chapter provides some interesting feedback to the wider community. First, a method for identifying therapeutic areas of unmet need is presented. A state-of-the-art view of G-D associations is provided and includes data that is otherwise inaccessible to the community (such as those annotated as being caused by either a GoF or LoF mutation). Furthermore, by using the disease prioritisation metrics (TAU and RTA scores), future projects can identify areas of application for drug discovery projects. Finally, the method for filtering potential side-effects means that other projects can further prune inferred Dr-D associations— something that can only increase the accuracy of the final prediction set.

Future work would need to focus on: applying extra dimensions to the prioritising of therapeutic areas; as datasources become more plentiful and the depth of coverage greater, the inclusion of directionality data is required, for both G-D associations and Drug-Target (D-T) associations. Furthermore, the differentiation between causal G-D associations and susceptibility associations. Differentiating G-D associations in such a manner would enable an extra dimension of scoring to be achieved, with those known to be causal given greater preference.

The approach described highlights disease areas of need that would benefit greatly from more data. As such, the future work discussed is heavily reliant on data availability, something that is currently lacking. There are, however, some interesting and relevant efforts, such as the public-private initiative from the CTTV; The Open Targets Platform[3]. At the time of writing this platform is still in its infancy, however, it has the potential to provide a powerful G-D association accession point. Finally, to investigate the proposed indications more thoroughly, they would need to be validated in a clinical setting. Future work is discussed in more detail in Chapter 7.

[3] https://www.opentargets.org

7

Discussion and future work

7.1 Introduction

In this thesis, methodologies that enable a systems approach (as defined in Section 2.2.3) to drug repositioning have been described. There are two major requirements to take such an approach: integrated data sets; and automated algorithms for the systematic exploration of these data sets. During this project, a Drug Repositioning Network Integration Framework (DReNInF) was developed to enable the creation of integrated data sets relevant to drug repositioning. This integration framework makes use of a high-level Drug Repositioning Network Integration Ontology (DReNInO). An application data set, Drug Repositioning Network Integration (DReNIn), created using the DReNInF was also made available as an Resource Description Framework (RDF) data set with a dedicated Web front-end. Furthermore, sub-components of a target network that represent a particular function were introduced and formalised as semantic subgraphs. Drug Repositioning Semantic Mining (DReSMin), a custom data mining algorithm, was also developed in order to allow integrated networks to be systematically mined for instances of a predefined semantic subgraph. DReSMin was then used for the inference and prioritisation of novel Drug-Target (D-T) associations as well as novel Drug-Disease (Dr-D) associations. Furthermore, DReSMin was shown to perform better than other state-of-the-art approaches when applied to D-T interaction prediction, identifying >10% more of a set of known D-T associations than the approach developed by ChEMBL[1] [180, 181].

In this Chapter, the outcomes of the project will first be described in the context of the initial objectives. For each objective, project outcomes will be assessed in terms of their contribution to the field. Limitations of the approaches developed as part of this work will then be discussed. Project outcomes will then be reviewed in the broader context of drug repositioning, focussing on recent developments in the field. Finally, future opportunities will be identified and introduced.

[1] www.ebi.ac.uk/chembl

7.2 Discussion

To fulfil the main aim of this project, three objectives were set at the beginning of the project (as described in Section 1.3). These are:

1. To extend existing data integration platforms relevant to drug repositioning.
2. To research and implement appropriate strategies for semantic data integration of network construction including Ondex, RDF and others.
3. To develop algorithms to search for topological and semantic network structures indicative of repositioning opportunities.

7.2.1 Aim 1: Extend existing data integration platforms relevant to drug repositioning

To extend existing data integration platforms relevant to drug repositioning, the *in silico* drug discovery data set described by Cockell *et al.* [101] was expanded. This expansion involved the integration of further biological entities and interactions relevant to drug repositioning (see Section 3.4.1). It was shown that Ondex was not an ideal platform for the creation of large integrated data sets required for the project. The necessity for time efficient data set builds with scalable integration led the project to move to an alternative graph-based datastore, Neo4j.

Originally developed for *in silico* drug discovery, expansion of the original data set using a single node type (`Indication`) and four edge types (`has_parent`, `has_child`, `may_treat`, `may_prevent` and `involved_in`) enabled the data set to be used for drug repositioning tasks (see Section 3.4.1). This resulted in *Dat*, the first Ondex drug repositioning data set that can be both interrogated and extended by the community. Furthermore, a greater understanding regarding the data types required to enable a holistic approach to drug repositioning was achieved and the data types included are represented by the metagraph (see Fig. 3.3). This metagraph did not only influence further data set development during this project but provided the wider community with an archetypal minimalistic representation of the data types and interaction types required to enable a systems approach to drug repositioning.

7.2.2 *Aim 2: Research and implement appropriate strategies for semantic data integration*

Various strategies for graph-based semantic data integration were researched, including Ondex, RDF and Neo4j (see Chapter 3). As mentioned, Ondex was used for the expansion of an existing drug discovery data set, and it was this data set that was used in Chapter 5 for the inference of novel D-T associations. Through benchmarking and performance testing, it was shown that Ondex was not a viable option for 'large' data set creation, i.e. highly connected graphs with over 1 million nodes. For this reason, it was decided that Neo4j would be used as the integration platform and data storage solution for further work. A typical graph database (as opposed to the triplestore) and not RDF was used for data set development due to the size of the data sets to be produced, the requirements of the data sets, and the need for exhaustive and complex graph-theory based querying. Graph databases are designed to support property graphs (graphs where properties may be assigned to either entities or their relationships, or both) and although recently some triple stores have added this capability their initial design did not make this an easy task. Graph databases, on the other hand, build upon the mathematics of graph theory. Since subgraph pattern matching was central to the mining approaches developed during this project, it was decided that a typical graph based database would be the most appropriate.

Although alot more scalable than Ondex, Neo4j provides no semantic integration and query features. Therefore, to control the semantics used during data set production, a lightweight software module, Bioinformatics Semantic Integration Platform (BioSSIP), was developed. This module sits between Neo4j and a data integration framework, enforcing strict semantic typing during data set production and later querying. Therefore, BioSSIP allows a data integration project to take advantage of the inherent scalability of Neo4j, whilst also maintaining semantic consistency.

With the underlying data storage and querying system in place, a framework for the development of drug repositioning data sets was then designed and implemented: DReNInF (see Section 3.4.4). This framework is made up of three components: a high-level ontology, DReNInO; a suite of parsers for relevant data sets; and an integration strategy. DReNInF was used to provide an RDF knowledge base (DReNIn) that can be queried by the community. DReNIn is accessible via a dedicated Web site and can also be queried systematically by using a SPARQL Protocol and RDF Query Language (SPARQL) endpoint.

DReNInF provided the first data integration framework for drug repositioning and is open to the community. Including 21 parsers and 6 mappers as well as a well-defined integration strategy, DReNInF is a rich resource for those wishing to create integrated drug repositioning data sets. To create a local integrated drug repositioning graph a user simply downloads the DReNInF source code as well as any version of the data sets to be included and runs the integration strategy. Central to the integration framework is DReNInO. DReNInO provides the field with the first high-level ontology developed specifically for system approaches to drug repositioning. The ontology is available to the community for extension and manipulation. Unlike other drug repositioning integration efforts, DReNInO supports the inclusion of data types that can be used to further assess the validity of any inferences made using a data set developed using the ontology. Data types such as `Clinical_ Trials` and edge types such as `has_indication`, `has_side_effect` and `involved_in_clinical_trial` allow for pruning of potential Dr-D associations— something that is not made possible through large consortium efforts such as Open PHACTS.

DReNIn was developed in DReNInF and contained 466,540 nodes connected via 2,688,436 edges before being exported to RDF. Made up of over 8.5 million triples the RDF data set has a SPARQL endpoint, enabling questions to be posed that were previously not possible to ask of any other integrated drug repositioning resource.

7.2.3 Aim 3: Develop algorithms to search for network structures indicative of repositioning opportunities

Semantic subgraphs were first introduced and defined (see Section 4.2.1). An algorithm, DReSMin, was then developed for the identification of mappings of semantic subgraphs from an integrated target network. Furthermore, approaches making use of this algorithm were developed to search for topological and semantic network structures indicative of repositioning opportunities. DReSMin is an exhaustive, exact topological matching algorithm to identify instances of pre-defined semantic subgraphs from a target network (see Section 4.4.1). In DReSMin, semantics can be matched to a defined threshold provided by a user. Furthermore, this search algorithm was used as a component in larger approaches for the automated identification of novel D-T associations and novel Dr-D associations.

Semantic subgraphs and the DReSMin algorithm are not restricted to edge inferences within the field of drug repositioning, rather they can be applied to any area of research that can be captured in a semantically-rich graph-based representation as defined in Section 4.4. Central to the algorithm is the Semantic Distance Calculator (SDC) (see Section 4.4.1.3), which uses matrices to score the semantic similarity between target graph entities (either edges or nodes) and their equivalent entities in a query semantic subgraph. DReSMin is made up of many novel sub-components that improve searching performance, making exhaustive searching tasks computationally tractable (such as the identification of mappings of semantic subgraphs with a node set greater than 6). These steps include the semantic graph pruning step (see Section 4.4.1.1). This optional step essentially removes all entities from the target network that are deemed semantically distant from those elements represented in the query semantic subgraph, thus reducing the search space to be examined by the algorithm. Furthermore, the semantic subgraph split component (see Section 4.4.1.4) splits semantic subgraphs whose node set is greater than 3 into sets of semantic subgraphs which are searched for in separate queries. After all split semantic subgraphs are searched for instances are then mapped back together. This component, in particular, has applications outside of the drug repositioning setting and could be used to enable the exhaustive searching of either topological or semantic subgraphs that were previously not tractable. This step reduces the search time for a semantic subgraph with a nodeset of 6 from 60 seconds to 8 seconds.

A method to infer novel D-T associations was also developed by making use of historical data and this work was subsequently published [146]. This approach was shown to be better at prioritising known D-T interactions than other state-of-the-art methods developed specifically for the purpose of identifying D-T interactions, such as the approach developed by ChEMBL[2] [180, 181]. Unlike traditional methods for D-T

[2] www.ebi.ac.uk/chembl

interaction prediction, the strategy presented in this work can be used to infer any 'type' of edge captured in the integrated network, and is not limited to the exemplar application described in Chapter 5.

Computational methodologies reliant on a single data type have various limitations (as described in Chapter 2). By taking a holistic view and considering all evidence available to support a prediction more confident inferences can be made. The method described in Chapter 5 allows for the inference of associations using the data types that are captured in a network. The more relevant data that is collated, the more confident one can be that an inference is likely to be a true positive. Unlike other approaches to drug repositioning, semantic subgraphs may be designed to infer relations between any node types in a data set. DReSMin enables the discovery of any relations that are captured in the abstracted drug-disease connection (as presented in Fig. 2.1), so long as suitable semantic subgraphs are identified. This is in contrast to many of the computational approaches to drug repositioning described in Section 2.2.2. For example, ligand structure-based approaches, as well as protein structure-based algorithms are limited to the inference of drug-target associations; gene expression based approaches along with genetic variation-based approaches have been limited to the application of the inference of drug-disease associations. Similarly, phenotype-based approaches (including disease and side-effect) tend to be used for the inference of drug-disease associations. Although machine learning based approaches tend to take a more integrative approach to data and, therefore, have the ability to infer a wider range of associations and even properties, the fact that they are effectively statistical 'black boxes' means that interpreting the repositioning hypothesis is a difficult task. Using semantic subgraphs and DReSMin allows for human-interpretable hypotheses to be derived and for all evidence supporting a claim to be judged.

The work presented in Chapter 5 provides some useful feedback to the drug repositioning field. First and foremost it was shown practically how semantic subgraphs may be derived from historical data. The resulting set of 194 semantic subgraphs are available to the community and can be used to mine semantic networks with similar properties to those captured in *Dat*, the data set that was queried using these subgraphs. As well as providing a method for the development of semantic subgraphs a method for ranking inferences is provided. Using the semantic subgraphs and the ranking methodology the approach is general enough to be used to identify any associations relevant to drug repositioning, such as the associations captured in the drug action 'central dogma' (see Fig. 2.1). The approach also highlights target classes that are in need of more research, such as kinases, which the community can use to direct future investigational work. A set of 9,643,061 novel D-T associations were identified and ranked. These associations are available for download and can be used as a starting point for further analysis. Furthermore, the approach developed was shown to outperform state-of-the-art approaches dedicated to D-T association prediction, identifying >10% more of a set of known D-T associations than the approach developed by ChEMBL.

A method to infer novel Dr-D associations, which also makes use of the DReSMin algorithm, was developed and this work was subsequently published [213]. This strategy, described in Chapter 6, predicts target-driven hypotheses for novel indications. Central to any *in silico* target-driven approach to drug discovery is the integration and ranking of all known Gene-Disease (G-D) associations. In this project, G-D ranking was achieved using a Bayesian approach. Ranked G-D associations were then integrated with other relevant data to produce a target-driven drug repositioning data set. Novel strategies for identifying therapeutic areas of interest were described. Finally, this integrated data set was mined for instances of a semantic subgraph representing an abstract view of the target-driven drug-disease relationships as captured in the drug action 'central dogma' in Fig. 2.1.

Although sources such as DisGeNET [206] and Malacards [207] integrate and score G-D associations, they are scored in an arbitrary fashion. DisGeNET, for example, score G-D associations based on the evidence supporting them. In this metric, scores assigned to associations are the sum of the frequency of three 'types' of evidence supporting them: i. manually curated sources capturing the association ii. model organism databases that capture the association and iii. text-mined associations. The problem with this metric is the fact that if a G-D association (even a negative association) is well-studied it will score well based on the cumulative nature of the scoring metric and the number of papers describing the well-studied association. The Bayesian approach described in Chapter 6 allows for a more holistic, less biased view of the

G-D data. Unlike DisGeNET, using the Bayesian approach described in this work, if an association is captured in a text mining source it will receive one score from the source (even if it has been captured 350 times by the source)— reducing the bias in well-studied associations. The G-D ranking provided by this work can enable further target-driven drug repositioning projects by taking a state-of-the-art view of the G-D association landscape. The ranked set of 309,885 G-D associations are available to the community and provide a resource capturing data that have been previously inaccessible, such as the 3,407 associations annotated with either Gain of Function (GoF) or Loss of Function (LoF).

As well as the G-D associations, the approach described in Chapter 6 has many subcomponents that can enable future drug repositioning projects. The approach highlights therapeutic areas that would benefit from drug development projects (areas of therapeutic need) using two metrics for ranking therapeutic areas in terms of approved treatments available to treat them (see Section 6.4.4). The Therapeutic Area Unmet Score (TAU) ranks therapeutic areas in terms of those that would benefit from more treatments and the Rich Therapeutic Area (RTA) enables therapeutic areas to be ranked based on how rich the set of marketed treatments is for that area. These metrics can be used to drive other drug repositioning projects and can also help industry to identify areas of need for drug discovery projects. A method for scoring the similarity of diseases captured in the MeSH hierarchy is also provided. This similarity measure is used to filter novel Dr-D associations that are likely side-effects. The method for pruning inferences can be used in any *in silico* drug repositioning project and can enable for more accurate inference sets to be produced. Finally, a set of 275,934 unique novel Dr-D associations was provided, with two of the most promising inferences (pazopanib as a treatment for mastocytosis and lisinopril as a treatment for Alzheimers) discussed and made available for future analysis.

7.2.4 Limitations of the approaches described

Throughout this thesis, computational approaches to drug repositioning have been presented. To consider drugs in a systems setting, where a holistic view of their interactions with multiple entities is required, simplified representations of drugs have been used. For example, the active form of a drug may differ from the marketed compound, with many transformations occurring during drug metabolism. Like all *in silico* strategies to analysing *in vivo* and *in vitro* systems, the accuracy of the systems view is limited by the available domain knowledge. Overly-simplified reproductions of a system innately struggle to accurately represent *in vivo* and *in vitro* systems. In order to capture a more accurate representation of the drug landscape, other factors may also need to be considered. Through integrating these factors with the work presented, more confident inferences could possibly be made in the future. Some of these limiting factors are introduced in the following section.

7.2.4.1 Drug selection

Drug repositioning tends to focus on approved drugs (however drugs that have failed to reach the market for reasons other than safety concerns may also be considered), yet there are still numerous factors aside from intellectual property or cost that can limit the initial set of drugs to be investigated. These additional factors are briefly mentioned here because they have not yet been presented or discussed. In the cases presented in this work, post filtering could be used to take account of these factors.

For example, certain regulated substances, such as opiates and anabolic steroids, have restrictions on handling and distribution, and as such may not be suitable for drug repositioning [32]. Other approved drugs, such as magnesium chloride and the amino acid arginine would not be classified as 'therapeutic' in most settings aimed at drug re-use [32]. Furthermore, regulatory approval (European Medicines Agency (EMA) and Food and Drug Administration (FDA)) approval is dependent on geographical location. Therefore, opportunities to identify drugs approved in some locations but not others exists [32]. Another factor that may be somewhat more difficult to implement is the pharmacologically active form of a drug. Many drugs, such as the oestrogen antagonist, tamoxifen (generic drug), are pro-drugs, either by design or are converted *in vivo* to active metabolites [32]. It is, therefore, important to know the pharmacologically active form of a drug [32].

One source that may enable the inclusion of such data in the future is ChEMBL [128]. At the time of writing the data source is making a concerted effort to annotate the active form of the drugs captured in the resource. Although the effort is not exhaustive at present, in the future this data may prove priceless to drug selection in *in silico* approaches to drug repositioning.

7.2.4.2 Dose levels

Arguably the most complicated aspect of drug repositioning is dosing. The dosage of a drug can make a substantial difference due to the half-life of drugs varying widely. In some drugs, there is a 'binary' therapeutic threshold, whereas in others this is more graduated. Drugs are approved at well-defined dosage strength levels. The ideal scenario for drug repositioning would be to use the existing approved specific product [32] at the approved, or lower, dosage. If a drug needs to be administered at dosages above these parameters, it would require substantial development costs. For example, sildenafil, with dose strengths 25mg, 50mg, 100mg is marketed as Viagra (Pfizer) for erectile dysfunction. Sildenafil was reformulated to dose strengths of 5mg and 20mg when repositioned for the treatment of pulmonary arterial hypertension, Revatio (Pfizer). Caution should be applied where *in vitro* concentrations are significantly higher than previously reported concentrations for the same drug observed in clinical settings [32]. Therefore, the dose levels of potential repositioning opportunities presented in this work must be considered before taken any further.

One way to predict dose ranges is by using pharmacological modelling, which enables human dose-prediction; a fundamental for ranking lead-optimization in drug discovery programs and to inform design of early clinical trials [233].

7.2.4.3 Data bias

Although the strategy described in Chapter 5 for the identification of novel D-T associations allows inferences involving multiple types of target, there are definite limitations on those types that have less information available. Similarly, the approach described in Chapter 6 for the identification of novel Dr-D associations, shows predictions involving diseases such as those of the central nervous system to be limited by the amount of data that is currently available to describe them. Although the problems arising from 'missing data' or low coverage are likely to become less of a hindrance as high throughput technologies continue to produce data at scale, this is presently an obvious limitation to systems approaches.

Furthermore, every data set is biased based on the motivation for which it was developed in the first place. For example, Genome-Wide Association Studies (GWAS) data are very much limited to diseases that are of interest to a particular field [146]. As such there is a disproportionate amount of research into cancer and immuno-inflammatory areas as well as infectious diseases. Although integration of multiple data sources covering the same 'types' of data (each source with different bias) can reduce the effects of such bias, it can still be a problem.

Data bias is a common problem within the field of bioinformatics, with the less represented class, or category, often the one of interest. One method to avoid such bias is to use data sampling. Data sampling involves selecting a subset of individuals from within a population. By using data sampling techniques, a more evenly-spread subset of data can be identified. Although this approach would result in a 'fairer' representation of associations involving a particular target class, or therapeutic area, it would also remove data— something that is against the central idea of the approaches described in this thesis.

7.2.4.4 Data reliability

Many types of data from multiple sources are included in the integrated data sets developed during this project. There is always a risk that drug-related commercial or open source projects contain errors; both automated and manual data curation can introduce errors. As such there are tradeoffs in accuracy versus coverage. For example, it is a well-known problem that the curation of a correct set of structures for approved drugs is difficult, and arguably still not available at the required level of accuracy [234]. Problems with structure curation is an unavoidable problem at present.

As large pharmacological integration projects, such as Open PHACTS [130] and data standards projects such as ELIXIR[3] and eTRIKS[4] continue to encourage the improvement of data content, accessibility and format, it is hoped that the community will

[3] www.elixir-europe.org

[4] www.etriks.org

benefit from an overall improvement in the near future.

7.2.4.5 Computational validation

The lack of a structured gold standard for drug repositioning makes it difficult to compare, evaluate and validate the performance of computational methodologies [18]. Furthermore, there is little to no negative data available. The lack of a gold standard leads to various approaches to the same problem using different means of validating results. A lack of reliable validation data is highlighted in Chapter 5, where historical data was used as a means of validation. To help address this deficiency, DReNIn includes both clinical trial data, side-effects and indications. These data enable validation questions to be posed that are currently not possible to query with other resources, including large projects such as Open PHACTS. For example, DReNIn can answer a question as such: 'I have made a prediction that `smallMoleculeA` may be used to treat `commonDiseaseB`— tell me if `smallMoleculeA` is currently marketed to treat `commonDiseaseB`, furthermore tell me if `commonDiseaseB` is a side-effect of `small-MoleculeA`. Finally, tell me if there are any stage 4 clinical trials that involve both `smallMoleculeA` and `commonDiseaseB`'. It is hoped that this functionality will provide a useful resource to the community.

Recent work has focussed on curating a comprehensive, public catalogue of existing drug indications, that uses a crowd-sourcing approach [235, 236]. Although the accuracy of such approach remains to be seen, this source has the potential to be an interesting focus point for future drug repositioning approaches— a consensus set of indications will enable cross-platform comparisons as well as validation.

7.2.5 Results in the broader context of recent developments

In this section, the results of the project will be discussed in terms of recent, external, developments relevant to the work. Here, it will be shown how the work of this project may benefit from work developed externally during the course of the project.

7.2.5.1 DReNInF

The DReNInF provides a framework for the production of user specific drug repositioning data sets. This framework enables a user to create integrated data sets that make use of historic versions of data sources. It is the historic versions of data sets that enable sets of relevant semantic subgraphs to be created (see Chapter 5). Furthermore, the framework can be easily extended to include data types that are not currently accounted for.

At the time that this project commenced, these tasks were not possible to complete with Open PHACTS. At present, Open PHACTS provides a Docker[5] image that allows a user to have a local instance of the Open PHACTS database. Furthermore, at the time of writing, Open PHACTS are working on a framework for integrating user-specific data that would enable RDF sub 'graphs' (such as UniProt) to be removed from the local installation and be replaced with alternative versions of the data source. Though useful, this approach has its limitations since many historical versions of data sets are not available in RDF. Although this Open PHACTS feature is still experimental, it has the potential to provide a fantastic source for the community. However, the project still misses out on valuable data for the task of repositioning that is included in DReNInF, such as indications and clinical trial data.

Other integration projects that have commenced since the start of this project include ONCOTrack[6] for the identification of novel biomarkers in tumours, and eTOX[7] which aims to enable the production of tools to better predict the toxicological profiles of small molecules in early stages of the drug development pipeline. These data sources would provide useful additions— complementing the drug repositioning data sets created using DReNInF.

7.2.5.2 Integration and ranking of gene-disease associations

Recently, the world of G-D associations has taken a large step forward, due to the work completed by the Open Targets platform (OTP)[8]. This platform was in its in-

[5] www.docker.com

[6] http://www.oncotrack.eu/

[7] www.etoxproject.eu

[8] https://www.targetvalidation.org/

fancy when the work presented in this thesis commenced but has recently made great progress. Making use of the Experimental Factor Ontology, the OTP provides a state-of-the-art view of G-D associations taken from multiple sources and scores these using harmonic progression. Furthermore, at the time of writing, a remote API has been launched, meaning that systematic querying of the data source can now be achieved. This data source should be a great resource for the community and any extensions of this work would need to utilise the work from those at the OTP.

7.2.5.3 The drug discovery climate is changing

As well as improving data set development and accessibility, many more external factors have changed during the course of this project— to be expected in such a fast-paced area of research. For example, new scientific areas of interest are coming to the fore, such as epigenetic therapy. Epigenetics is the study of heritable changes in gene expression that do not result in an alteration in the DNA sequence itself (such as methylation) [237]. Epigenetic aberrations are particularly relevant to diseases such as cancer, heart disease, diabetes and a variety of neurological disorders [237]. Epigenetic modifications work in concert with genetic mechanisms to regulate transcriptional activity in normal tissues and are often dysregulated in disease. Such changes can be used as a diagnostic indicator— the epigentic changes often precede disease pathology. Interestingly these modifications of DNA and histones are reversible, meaning they are also good targets for therapeutic intervention [237]. Epigenetic therapies use drugs, or other epigenome-influencing techniques, to treat medical conditions by influencing pathways directly.

Furthermore, with the increased knowledge surrounding genomics data has come an increased interest in personalised medicine. Specifically, the interest in personalised medicine has come at a time when efforts such as the UK 100,000 Genomes Project and the US Precision Medicine Initiative seek to scale up population-based genome sequencing and integrate it with clinical data [30]. It is hoped that many of the approaches presented in this thesis could also be transferable to this continually developing area of interest.

7.3 Future work

This project has produced a number of novel algorithms and approaches as well as drug repositioning data resources (described in Section 7.2). It has been shown that for some tasks, such as D-T association prediction, our system is more capable than other tools that are operating within the drug repositioning domain, identifying >10% of a set of known D-T associations than the ChEMBL models[9] [180, 181]. However, there is plenty of scope for future improvements. A non-exhaustive discussion of a number of interesting future directions is provided in the following section.

7.3.1 Extend the integration framework

Structural data is limited for both proteins and ligand binding sites. A lack of accuracy also results in high levels of false positives for approaches that focus purely on these data. However, it would be interesting to import structural data from sources such as PDB [238] into DReNInF. Furthermore, PDTB is a database containing current and potential drug targets with known 3D structures [36]. Structural data would offer an alternative type of data in DReNInF. Using an holistic approach to semantic subgraph development for mining of resulting integrated data sets (such as that described in Chapter 5) it should be possible to reduce any associated false positives— inferences made using the approaches presented in this thesis use all data types and not just one.

The inclusion of other genomic profile data, such as that captured in GEO [239] and ArrayExpress [240] may also offer valuable input to drug repositioning data sets. Insights from ENCODE [241] and Fantom5 [242] provide information about the potential regulatory effects of Single-Nucleotide Polymorphisms (SNPs) in non-coding regions, which may offer interesting sources to be included in later data sets.

7.3.2 Extend applications

The approach described in Chapter 5 made use of historical data to create a set of semantic subgraphs. These subgraphs were then searched for in a target network

[9] www.ebi.ac.uk/chembl

using the DReSMin algorithm, with inferences scored and ranked using the scoring framework described in the same Chapter. Although the use case presented considered D-T associations, the approach can be used to identify any 'type' of interaction that is both captured in the integrated network and has historical data sets available. It would, therefore, be attractive to apply the strategy to the inference of interactions such as G-D associations; Dr-D associations; or even use it to predict MoA. The only limitation to the application of this methodology to the inference of alternative 'types' of associations is the requirement of historical data.

The analysis described in Chapter 6 infers Dr-D associations from an integrated data set. The data set made use of integrated and ranked G-D associations as well as drug, protein and gene data. A semantic subgraph that best represented the abstracted 'central dogma' (see Fig. 2.1) was searched for using DReSMin and Dr-D inferences ranked. Using the semantic subgraph from this approach, as well as the data types included in the network, it would be interesting to infer novel G-D associations. These associations could be scored based on instance information (i.e. the number of instances that inferred the association) and those scoring highly could then be integrated back into the network. The approach could then be repeated to infer Dr-D associations using these novel G-D associations.

7.3.3 Library of semantic subgraphs

A library of semantic subgraphs could be presented as an editable resource, open to the public. The database would contain manually curated semantic subgraphs as well as automatically generated semantic subgraphs, such as those presented in Chapter 5. Furthermore, semantic subgraphs should be annotated with categories such as 'antidepressants', 'antiemetics', 'kinases' or 'G protein-coupled receptors (GPCR)'.

Semantic subgraphs can be drawn from real-life repositioning examples via manual curation. The manual development of semantic subgraphs, such as the one described in Fig. 4.2, is time-consuming. However, manually curated semantic subgraphs may allow for more accurate representations of a functional module capturing a potential drug repositioning opportunity, as opposed to those created via automated approaches. Therefore, creating a library of semantic subgraphs curated from real world examples of repositioned drugs would be a great resource for the community. Manually developed semantic subgraphs could be submitted to the database and be editable by other members of the community whilst also being annotated manually with relevant subcategories of labels.

194 semantic subgraphs were defined in Chapter 5. It would be intriguing to examine these further. For instance, to determine whether particular semantic subgraphs are better at inferring D-T associations for different classes of drugs or indeed different classes of target. Using the D-T association inference ranking, a simple metric could be used to score a semantic subgraph based on their ability to prioritise D-T associations involving (i) all drug classes and (ii) all target classes. Ranking positions of false positives for the same class of interest would also need to be considered.

This semantic subgraph database would enable far more specific hypotheses to be investigated. For example, if a user were interested in identifying potential D-T associations involving the protein target class GPCR, the library would contain a subset of semantic subgraphs that are ‘specialised’ for this purpose.

7.3.4 In vitro / In vivo *validation*

The ultimate goal of any drug *in silico* repositioning project is to take some of the interesting hits identified into the clinic to show efficacy and to ultimately benefit patients [18]. However, it is argued that drug repositioning is vastly more complicated than typically imagined and thus many repositioning projects stop at the *in vitro* level [32]. *In vitro* and *in vivo* models (e.g. cell-based targeted assays and mouse models) are required to validate the candidate hits for preclinical drug evaluation [18]. In addition to the right model, the selection of the appropriate hits for validation is also critical. Physicians or biologists may not favour some drugs due to reasons such as high toxicity, high cost and low bioavailability. Also important is the cost of clinical trials to show efficacy and companies are not likely to pay if a drug is off-patent. Once these factors have been considered, it would be necessary to identify potential collaborations to pursue some of the most promising ‘hits’ identified during this project.

7.4 Conclusion

The methodologies and tools developed during the course of this project can facilitate the integration of existing biological and pharmacological data to allow for a systems approach to drug repositioning to be taken. Unlike the strategies described in Chapter 2, the work described in this thesis has developed an approach that considers all possible evidence about a set of drugs and their interactions. This holistic approach has resulted in a set of tools and integrated data sets that, support the queries and inferences outlined in the aims and objectives section at the start of this thesis (see Section 1.3).

With the term 'big data' being used ever more frequently, in both industry as well as academia, research focus is moving toward the way these vast amounts of data are handled. As such, it is expected that in the near future, there will a huge amount of accessible information to which the approaches implemented during this work can be applied. Furthermore, the approaches presented here are not limited to the field of drug repositioning. Searching for semantic subgraphs using the methodologies described has a vast range of possible applications. Although applications such as gene regulatory networks and social network analysis tend to analyse data with limited semantic types, many applications that utilise semantically complex data sets could benefit greatly from the research and approaches presented here.

Although the productivity of Research and Development (R&D) approaches to drug discovery have improved dramatically since the start of this project, computational drug repositioning is still of major importance to improving human health, through discovering new uses for existing drugs. The work presented here will also benefit greatly from the improved use of ontologies and controlled vocabularies pushed forward via projects such as Open PHACTS [130], ELIXIR[10] and eTRIKS[11].

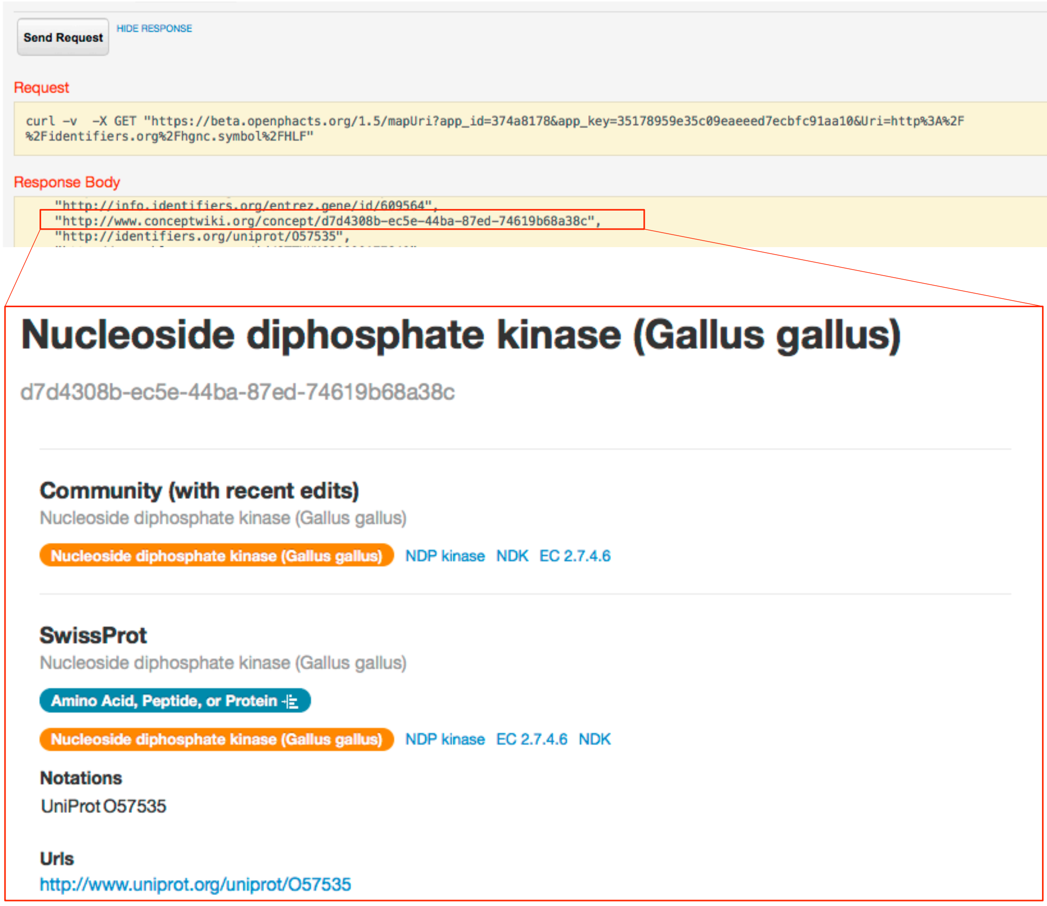
Finally, it is hoped that the work presented in this project will prove useful to the wider community and in turn help to improve the lives of those who are suffering from disease.

[10] www.elixir-europe.org

[11] www.etriks.org

A

Chapter 3 Appendix



Request

```
curl -v  -X GET "https://beta.openphacts.org/1.5/mapUri?app_id=374a8178&app_key=35178959e35c09eaeeed7ecbfc91aa10&Uri=http%3A%2F
%2Fidentifiers.org%2Fhgnc.symbol%2FHLF"
```

Response Body

```
    "http://info.identifiers.org/entrez.gene/id/609564",
    "http://www.conceptwiki.org/concept/d7d4308b-ec5e-44ba-87ed-74619b68a38c",
    "http://identifiers.org/uniprot/O57535",
```

Nucleoside diphosphate kinase (Gallus gallus)

d7d4308b-ec5e-44ba-87ed-74619b68a38c

Community (with recent edits)

Nucleoside diphosphate kinase (Gallus gallus)

Nucleoside diphosphate kinase (Gallus gallus) NDP kinase NDK EC 2.7.4.6

SwissProt

Nucleoside diphosphate kinase (Gallus gallus)

Amino Acid, Peptide, or Protein

Nucleoside diphosphate kinase (Gallus gallus) NDP kinase EC 2.7.4.6 NDK

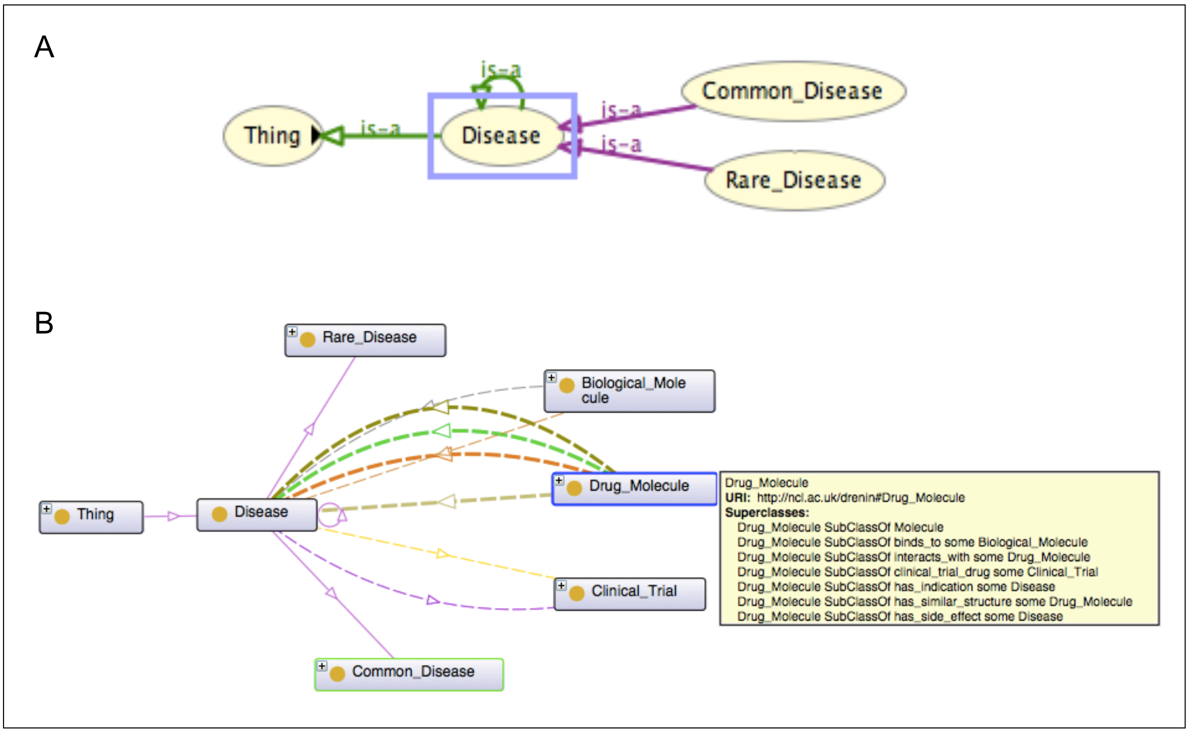
Notations

UniProt O57535

Urls

http://www.uniprot.org/uniprot/O57535

Figure A.1: **Gene URL mapping in Open PHACTS.** Open PHACTS RESTful API v1.5 `Map URL` call maps the hepatic leukaemia factor gene (HLF) from *Homo sapiens* to the nucleoside diphosphate kinase gene from *Gallus gallus* [1]



Drug_Molecule
URI: http://ncl.ac.uk/drenin#Drug_Molecule
Superclasses:
Drug_Molecule SubClassOf Molecule
Drug_Molecule SubClassOf binds_to some Biological_Molecule
Drug_Molecule SubClassOf interacts_with some Drug_Molecule
Drug_Molecule SubClassOf clinical_trial_drug some Clinical_Trial
Drug_Molecule SubClassOf has_indication some Disease
Drug_Molecule SubClassOf has_similar_structure some Drug_Molecule
Drug_Molecule SubClassOf has_side_effect some Disease

Figure A.2: **DReNInO disease representation.** (A) The `Disease` class in DReNInO has two children, `Rare_Disease` and `Common_Disease`. (B) Object properties involving `Disease` (dashed lines) capture the relations between other classes in DReNInO, such as `Biological_Molecule`, `Drug_Molecule` and `Clinical_Trial`. Also shown are the superclasses of `Drug_Molecule`. Images were created in Protege v4.3.

Table A.1: Concept classes in *Dat*.

ConceptClass	**Nodes**
Affymetrix_Probe	20,522
Biological_Process	19,046
Cellular_Component	2,731
Compound (DrugBank)	4,842
Compound (KEGG)	1,607
Disease	14,535
Enzyme_Classification	1,690
Enzyme	1,340
Gene	3,346
Kegg_Orthologs_Gene	2
Kegg_Orthologs_Protein	2
Molecular_Function	7,674
Pathway	436
Protein_Complex	196
Protein	22,665
Publication	45,059
Reaction	1,660
Target	3,500
Indication*	4,463

Table A.2: Relation types in *Dat*.

RelationType	Relations	Details
part_of_catalyzing_class	5,144	-
expressed_by	673	-
ubiquitinated_by	421	-
activated_by	4,430	-
has_function	50,922	-
regulated_by	4,333	-
adjacent_to	326	-
indirect_effect	778	-
part_of	5,332	-
produced_by	1,816	-
derives_from	210	-
has_participant	62,146	-
has_not_function	113	-
	4,813	-
repressed_by	4	-
dephosphorylated_by	410	-
share_intermediate	4,979	-
dissociated_from	51	-
interacts_with	37,166	Annotated with G-Sesame semantic similarity measures (no cutoff) (Du et al., 2009)
located_in	50,382	-
phosphorylated_by	1,537	-
binds_to	10,742	-
is_involved_in	15,400	-
published_in	109,061	-
is_a	48,726	-
has_similar_sequence	299,416	BLAST (E-value cutoff 1e-4)
inhibited_by	1,770	-
is_not_located_in	338	-
is_encoded_by	3,347	-
state_change_from	260	-
binds_to_encoding_mrna	22,514	-
protein_family	23,060	-
sim	12,256	2D-Tanimoto co-efficient calculated using (similarity cutoff 0.85) (O'Boyle et al, 2011)
is_part_of	573	-
member_is_part_of	1,957	-
consumed_by	1,845	-
participates_not	109	-
has_parent*	6,533	-
has_child*	2,018	-
may_treat*	3,744	-
may_prevent*	343	-
disgenet_involved_in*	16,098	-

Table A.3: Relation details in *Dat*

Relation Type	From ConceptClass	To ConceptClass
part_of_catalyzing_class	Enzyme	Enzyme Classification
	Protein	Enzyme Classification
expressed_by	Protein Complex	Protein Complex
	Protein	Protein
	Protein	Protein Complex
ubiquitinated_by	Protein Complex	Protein Complex
	Protein	Protein
	Protein	Protein Complex
activated_by	Protein	Compound (KEGG)
	Protein	Protein
	Protein Complex	Protein
	Protein	Protein Complex
has_function	Protein	Molecular Function
regulated_by	Biological Process	Biological Process
adjacent_to	Pathway	Pathway
indirect_effect	Protein	Protein
	Protein Complex	Protein
	Protein	Protein Complex
part_of	Cellular Component	Cellular Component
	Molecular Function	Molecular Function
	Biological Process	Biological Process
	Protein	Protein Complex
produced_by	Compound (KEGG)	Reaction
derives_from	Pathway	Pathway
has_participant	Protein	Biological Process
has_not_function	Protein	Molecular Function
catalyzed_by	Reaction	Enzyme
repressed_by	Protein	Protein
	Protein	Protein Complex
dephosphorylated_by	Protein	Protein
	Protein Complex	Protein
share_intermediate	Reaction	Pathway
	Reaction	Reaction
	Protein	Protein
	Protein Complex	Protein
	Protein	Protein Complex
dissociated_from	Protein	Protein
	Protein Complex	Protein
interacts_with	Compound (KEGG)	Protein
	Protein	Protein
located_in	Protein	Cellular Component
phosphorylated_by	Protein Complex	Protein Complex
	Protein	Protein
	Protein	Protein Complex

Continued on next page

Relation Type	**From ConceptClass**	**To ConceptClass**
binds_to	Protein Complex	Protein
	Compound (KEGG)	Protein
	Protein Complex	Protein Complex
	Compound (DrugBank)	Target
	Protein	Protein
	Protein	Protein Complex
is_involved_in	Protein	Disease
published_in	Protein	Publication
is_a	Enzyme	Protein
	Cellular Component	Cellular Component
	Molecular Function	Molecular Function
	Biological Process	Biological Process
	Protein	Target
	Target	Protein
has_similar_sequence	Target	Target
	Target	Protein
	Protein	Target
	Protein	Protein
inhibited_by	Protein	Compound (KEGG)
	Protein Complex	Protein Complex
	Protein	Protein
	Protein Complex	Protein
	Protein	Protein Complex
is_not_located_in	Protein	Cellular Component
is_encoded_by	KEGG Orth Protein	KEGG Orth Gene
	Protein	Gene
state_change_from	Protein	Protein
binds_to_encoding_mrna	Protein	Affymetrix Probe
protein_family	Enzyme Classification	Enzyme Classification
	Protein	Protein
sim	Compound (DrugBank)	Compound (DrugBank)
is_part_of	KEGG Orth Protein	Protein Complex
member_is_part_of	Reaction	Pathway
consumed_by	Compound (KEGG)	Reaction
participates_not	Protein	Biological Process
has_parent	Indication	Indication
has_child	Indication	Indication
may_treat	Compound (DrugBank)	Indication
may_prevent	Compound (DrugBank)	Indication
disgnenet_involved_in	Protein	Indication
	Target	Indication

B

Chapter 4 Appendix

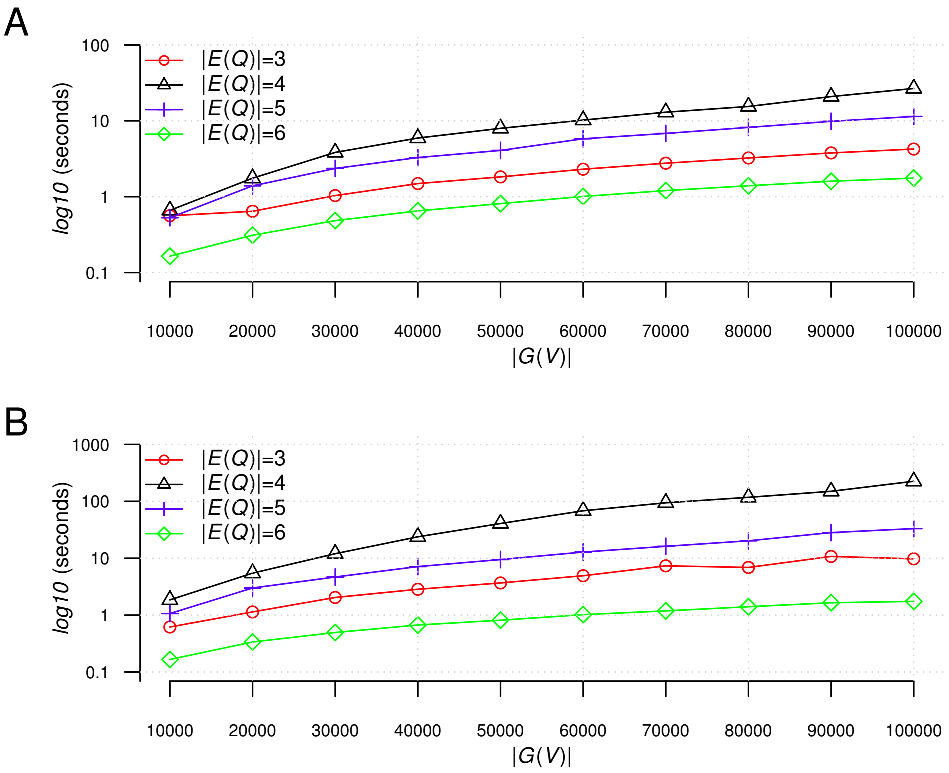


Figure B.1: **The effect of increasing the edge set of a semantic subgraph on the search time (in seconds) of a search.** Random semantic subgraphs were created with $|V(Q)|$ of 4. Edge sets ($|E(Q)|$) of the subgraphs ranged from 3-6. Random target graphs were created with node sets ranging from 1×10^4 to 1×10^5. The algorithm used one of two parameters: (A) all elements of the match must be greater than ST (top) or (B) all elements must cumulatively be greater than the ST (bottom).

C

Chapter 6 Appendix

Table C.1: Data sources used in the G-D integrated repositioning graph.

Source	**Version/Acc**	**NodeType**	**#Nodes**	**RelationType**	**#Rels**	**Attributes**
UniProtKB [214]	2015_08	Protein	20,203	-	-	UniProt UID, UniProt ID, Name
UniProtKB	2015_08	Gene	19,744	-	-	Entrez Gene Symbol, Entrez Gene ID
UniProtKB	2015_08	-	-	encoded_by	19,903	-
ORDO [135]	2/July '15	Rare_Disease	8,626	-	-	Name, MESH, OMIM, UMLS
ORDO	2/July '15	-	-	part_of	12,518	-
ORDO	2/July '15	-	-	has_parent	11,201	-
MeSH [120]	2015/Aug '15	Common_Disease	11,735∗	-	-	MeSH Header, MeSH, MeSHTree
MeSH	2015/Aug '15	-	-	is_a	23,829	-
DrugBank [113]	4.3/July '15	Small_Molecule	7,469	-	-	DBID, Name, Category, Group
DrugBank	4.3/July '15	-	-	binds_to	14,250	Action
ChEMBL [128]	20/Sep '15	-	-	binds_to	23,507	Activity type, Activity value
ChEMBL	20/Sep '15	-	-	-	-	Drug mechanism⋆
SIDER [82]	4/Aug '15	-	-	has_indication	4,488	-
NDFRT [243, 244]	Aug '15	-	-	has_indication	4,396	-
PREDICT [85]	-	-	-	has_indication	1,265	-
CTD curated [131, 245]	-	-	-	has_indication	18,540	-
SIDER	4/Aug '15	-	-	has_side_effect	67,934	-
Scored G-D	-	-	-	involved_in	309,885	Association score, Directionality◇

Data sources used in the creation of the repositioning dataset. ∗Made up of 5,370 descriptor records and 6,365 supplementary records. ⋆532 drug activity types (including agonist and antagonist) were taken from ChEMBL and mapped to drugs in the dataset. ◇3,459 G-D associations are annotated with the gene functionality resulting in a disease state, either loss-of-function (2,211) or gain-of-function (1,248).

Table C.2: GoF and LoF gene-disease associations.

Type	**#Ass.**	**#Unique**	**#Map**
GoF	16.3k	1,734	1,248
LoF	29k	3,059	2,211

GoF and LoF Gene-disease associations captured using linguamatics. *Note:* '# Ass.' = Associations from source, '#Map' = number mapped to MeSH,

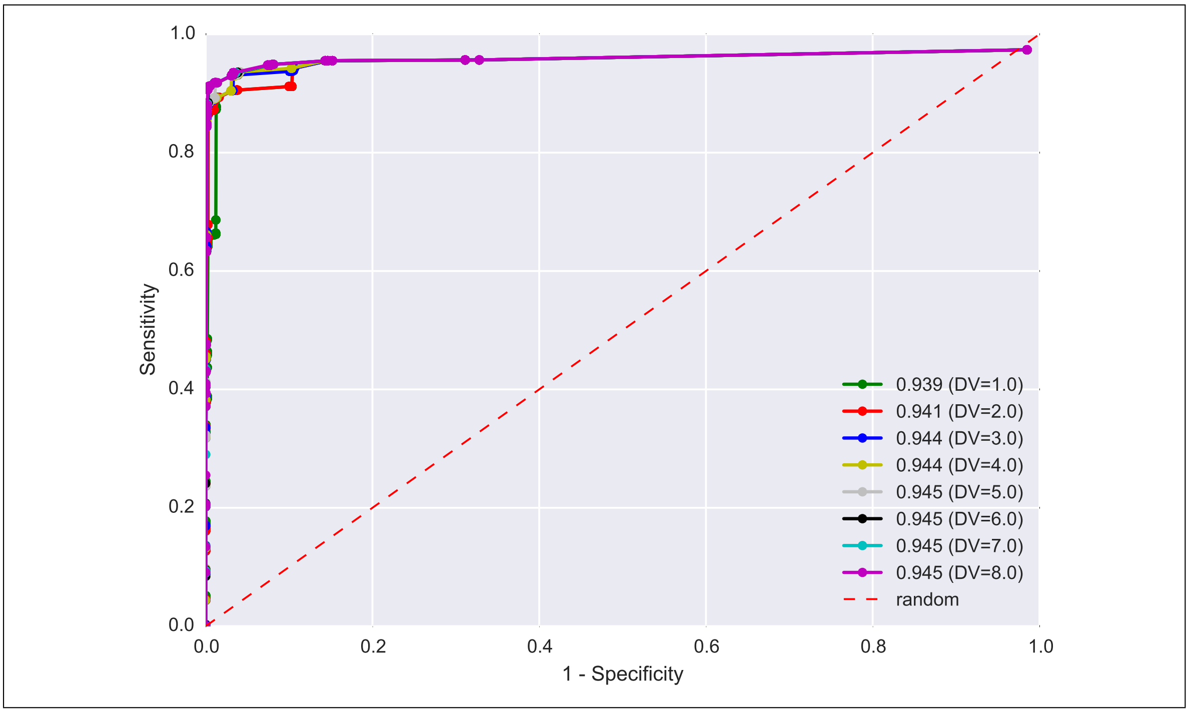


Figure C.1: **ROC curve when altering D value used to score associations with UniProt as the gold standard.** Using UniProt as the gold standard, all G-D associations were scored using D-Values from 1.0 - 8.0. It is seen that a D-Value (DV) of 5.0 (grey) gives the highest area under the curve AUC when validating using UniProt.

Table C.3: LLS for each dataset included in the gene-disease ranking.

	Test Source									
GS Src	**BEF**	**CTD**	**GOF**	**GWA**	**MGD**	**OMI**	**ORP**	**RGD**	**SEM**	**UNI**
BEF	-	9.26 (6)	9.97(5)	8.05 (9)	10.32 (4)	11.21 (2)	10.68 (3)	8.34 (8)	8.55 (7)	**11.21 (1)**
CTD	8.76 (9)	-	9.49 (7)	9.8 (5)	12.32 (3)	**14.12 (1)**	12.21 (4)	9.52 (6)	8.85 (8)	13.82 (2)
GOF	10.91 (7)	11.38 (6)	-	10.11 (9)	12.92 (4)	13.59 (3)	13.83 (2)	10.61 (8)	11.46 (5)	**14.53 (1)**
GWA	9.62 (7)	10.86 (4)	8.92 (9)	-	11.24 (3)	**12.06 (1)**	10.81 (5)	9.19 (8)	10.1 (6)	12.0 (2)
MGD	11.81 (8)	14.16 (4)	13.09 (5)	11.77 (9)	-	15.79 (2)	15.04 (3)	12.29 (7)	12.37 (6)	**15.91 (1)**
OMI	11.06 (9)	14.01 (4)	12.35 (5)	11.44 (6)	14.9 (2)	-	14.6 (3)	11.15 (8)	11.43 (7)	**16.78 (1)**
ORP	12.65 (9)	15.59 (4)	13.29 (7)	12.77 (8)	16.22 (3)	17.4 (2)	-	13.66 (6)	13.72 (5)	**17.5 (1)**
RGD	9.41 (9)	10.52 (5)	10.06 (6)	9.52 (8)	11.5 (3)	11.29 (4)	**11.73 (1)**	-	9.83 (7)	11.71 (2)
SEM	9.09 (8)	9.78 (6)	10.73 (5)	9.05 (9)	11.03 (3)	11.38 (2)	11.02 (4)	9.27 (7)	-	**11.5 (1)**
UNI	11.19 (8)	14.08 (4)	12.4 (5)	10.95 (9)	14.97 (2)	**16.57 (1)**	14.68 (3)	11.26 (7)	11.59 (6)	-
Mean	8.22 (10)	4.77 (5)	6 (6)	8 (9)	3 (3)	2.11 (2)	3.11 (4)	7.22 (8)	6.33 (7)	**1.33 (1)**

After applying the LLS method and alternating the gold standard, GS sources (left column), we see how every other source, the test sources (top row) perform in terms of identifying the ‘knowns’ captured in the GS. Performance is measured using the LLS score, which is shown. Furthermore, for each GS used, test sources are ranked in terms of performance (the higher the LLS score the better the performance of that test source). All ranks are shown in brackets and all scores are rounded to 2 decimal places.

Table C.4: Number of mappings returned for each MeSH therapeutic area after filtering.

Therapeutic Area		**# mapgs.**
C01	bacterial infections and mycoses	10,260
C02	virus diseases	16,566
C03	parasitic diseases	3,050
C04	neoplasms	86,438
C05	musculoskeletal diseases	20,531
C06	digestive system diseases	42,802
C07	stomatognathic diseases	7,139
C08	respiratory tract diseases	21,695
C09	otorhinolaryngologic diseases	3,005
C10	nervous system diseases	95,898
C11	eye diseases	12,776
C12	urologic and male genital diseases	18,899
C13	female genital diseases and pregnancy complications	23,405
C14	cardiovascular diseases	44,206
C15	hemic and lymphatic diseases	20,525
C16	congenital, hereditary, and neonatal diseases and abnormalities	51,524
C17	skin and connective tissue diseases	21,586
C18	nutritional and metabolic diseases	27,929
C19	endocrine system diseases	21,791
C20	immune system diseases	19,652
C21	disorders of environmental origin	2
C22	animal diseases	785
C23	pathological conditions, signs and symptoms	47,279
C24	occupational diseases	380
C25	chemically-induced disorders	7,410
C26	wounds and injuries	2,849
F01	behavior and behavior mechanisms	5,785
F02	psychological phenomena and processes	1,394
F03	mental disorders	27,574

Note: associations that include diseases that fall under multiple MeSH categories are duplicated in the counts (if a disease has multiple mesh tree terms from the same therapeutic area these are also counted multiple times). Only associations that survived the filtering steps are included. mapgs. = mappings.

Table C.5: F_1 scores

Sim	**# of inds.**	**TP**	**FP**	**FN**	**Precision (P)**	**Recall (R)**	**F-measure (F_1)**	**Mean Pruned**
1	295742	7494	288248	0	0.0253	1	0.0494	155528
0.7686	275934	6052	269882	1442	0.0219	0.8076	0.0427	133923
0.6333	239649	4204	235445	3290	0.0175	0.561	0.034	137806
0.5372	206227	3181	203046	4313	0.0154	0.4245	0.0298	141460
0.4628	180834	2618	178216	4876	0.0145	0.3493	0.0278	143843
0.4019	164765	2370	162395	5124	0.0144	0.3163	0.0275	145038
0.3504	156170	2248	153922	5246	0.0144	0.3	0.0275	145812
0.3059	152523	2209	150314	5285	0.0145	0.2948	0.0276	146241
0.2665	151287	2193	149094	5301	0.0145	0.2926	0.0276	146410
0.2314	150929	2188	148741	5306	0.0145	0.292	0.0276	146454
0.1996	150857	2188	148669	5306	0.0145	0.292	0.0276	146470

Note: F_1 score using each of the possible *Sim* scores whilst pruning potential side-effects from all mappings returned during the search. Predicted interactions were mapped to the known indications using a *Sim* of 1.0. *Note:* all values are corrected to 3 d.p. $F_1 = (2 \times \frac{P \times R}{P+R})$ TP= true positives, FP=false positives, FN=false negatives, '# of inds'=number of indications, Mean Pruned=mean ranking of pruned side effects.

Table C.6: Number of has_indication edges captured in *GenDat* involving the 1,188 approved small molecules.

Source	**# has_indication (%)**
CTD	14,761 (79.6)
SIDER4	3,641 (81.1)
PREDICT	1,210 (95.7)
NDFRT	2,586 (58.8)

Note: the percentage in brackets reflects the percentage of associations from source x that involves the drugs of interest. Sources cumulatively provide 18,889 unique has_indication edges.

Table C.7: Number of 18,889 known `has_indication` edges mapped to inferred associations using altering *Sim* values.

Sim	**# known has_indication (% of total)**
1.0	6,114 (34.2)
0.768	9,188 (51.38)
0.633	12,955 (72.65)
0.537	15,527 (87.10)
0.462	16,689 (93.60)
0.401	17,122 (95.74)
0.350	17,321 (96.88)
0.305	17,388 (97.50)
0.266	17,401 (97.58)
0.2314	17,407(97.6)

Note: of the 18,889, known `has_indication` edges, 1,006 involved 63 drugs of the 1,188 investigated for which the approach returned no mappings, leaving 17,883 that could potentially be validated.

Bibliography

[1] I. R. Edwards and J. K. Aronson, "Adverse drug reactions: definitions, diagnosis, and management.," *Lancet*, vol. 356, pp. 1255–1259, Oct. 2000.

[2] "disease, n.." OED Online. December 2015. Oxford University Press. [Accessed: 02-2016]. Available from: `http://www.oed.com/view/Entry/54151?rskey=uXLw6u&result=1`.

[3] J. Owens, "Determining druggability," *Nature Reviews Drug Discovery*, vol. 6, p. 187, Mar. 2007.

[4] S. K. Branch and I. Agranat, ""new drug" designations for new therapeutic entities: new active substance, new chemical entity, new biological entity, new molecular entity.," *Journal of medicinal chemistry*, vol. 57, pp. 8729–8765, Nov 2014.

[5] D. C. Swinney, "Phenotypic vs. target-based drug discovery for first-in-class medicines.," *Clinical pharmacology and therapeutics*, vol. 93, pp. 299–301, Apr 2013.

[6] D. B. T. Cox, R. J. Platt, and F. Zhang, "Therapeutic genome editing: prospects and challenges.," *Nature medicine*, vol. 21, pp. 121–131, Feb 2015.

[7] J. P. Hughes, S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery.," *British journal of pharmacology*, vol. 162, pp. 1239–1249, Mar. 2011.

[8] R. Capdeville, E. Buchdunger, J. Zimmermann, and A. Matter, "Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug.," *Nature reviews. Drug discovery*, vol. 1, pp. 493–502, July 2002.

[9] A. Mullard, "New drugs cost us[dollar]2.6 billion to develop," *Nat Rev Drug Discov*, vol. 13, pp. 877–877, 12 2014.

[10] A. Mullard, "2015 fda drug approvals.," *Nature reviews. Drug discovery*, vol. 15, pp. 73–76, Feb 2016.

[11] C. P. Adams and V. V. Brantner, "Estimating The Cost Of New Drug Development: Is It Really $802 Million?," *Health Affairs*, vol. 25, pp. 420–428, Mar. 2006.

[12] B. Booth and R. Zemmel, "Prospects for productivity.," *Nature reviews. Drug discovery*, vol. 3, pp. 451–456, May 2004.

[13] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht, "How to improve r&d productivity: the pharmaceutical industry's grand challenge.," *Nature reviews. Drug discovery*, vol. 9, pp. 203–214, Mar 2010.

[14] T. T. Ashburn and K. B. Thor, “Drug repositioning: identifying and developing new uses for existing drugs.,” *Nature reviews. Drug discovery*, vol. 3, pp. 673–83, Aug. 2004.

[15] K. B. Thor and M. A. Katofiasc, “Effects of duloxetine, a combined serotonin and norepinephrine reuptake inhibitor, on central neural control of lower urinary tract function in the chloralose-anesthetized female cat.,” *J Pharmacol Exp Ther*, vol. 274, no. 2, pp. 1014–24, 1995.

[16] M. R. Safarinejad, “Oral sildenafil in the treatment of erectile dysfunction in diabetic men: A randomized double-blind and placebo-controlled study,” *Journal of Diabetes and its Complications*, vol. 18, no. 4, pp. 205 – 210, 2004.

[17] L. Cordella, P. Foggia, C. Sansone, and M. Vento, “Performance evaluation of the VF graph matching algorithm,” *Proceeding ICIAP 9́9 Proceedings of the 10th International Conference on Image Analysis and Processing*, p. 1172, 1999.

[18] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, “A survey of current trends in computational drug repositioning,” *Briefings in Bioinformatics*, vol. 17, pp. 2–12, Jan. 2016.

[19] B. Booth and R. Zemmel, “Opinion/Outlook: prospects for productivity,” *Nature Reviews Drug Discovery*, vol. 3, pp. 451–456, May 2004.

[20] M. Dickson and J. P. Gagnon, “Key factors in the rising cost of new drug discovery and development,” vol. 3, pp. 417–29+, 2004.

[21] I. Kola and J. Landis, “Can the pharmaceutical industry reduce attrition rates?,” *Nature reviews. Drug discovery*, vol. 3, pp. 711–715, Aug 2004.

[22] M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal, “Computational drug repositioning: from data to therapeutics.,” *Clinical pharmacology and therapeutics*, vol. 93, pp. 335–341, Apr. 2013.

[23] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, “The price of innovation: new estimates of drug development costs,” *Journal of Health Economics*, vol. 22, pp. 151–185, Mar. 2003.

[24] D. W. Light and R. Warburton, “Demythologizing the high costs of pharmaceutical research,” *BioSocieties*, vol. 6, pp. 34–50, Feb. 2011.

[25] D. Cook, D. Brown, R. Alexander, R. March, P. Morgan, G. Satterthwaite, and M. N. Pangalos, “Lessons learned from the fate of AstraZeneca’s drug pipeline: a five-dimensional framework,” *Nature Reviews Drug Discovery*, vol. 13, pp. 419–431, May 2014.

[26] I. Melnikova, “Rare diseases and orphan drugs,” 2012.

[27] E. Tambuyzer, “Rare diseases, orphan drugs and their regulation: questions and misconceptions.,” *Nature reviews. Drug discovery*, vol. 9, pp. 921–929, Dec 2010.

[28] Y. Y. Li and S. J. Jones, "Drug repositioning for personalized medicine.," *Genome medicine*, vol. 4, p. 27, Mar 2012.

[29] N. J. Schork, "Personalized medicine: Time for one-person trials.," *Nature*, vol. 520, pp. 609–611, Apr 2015.

[30] O. Bahcall, "Precision medicine.," *Nature.*, vol. 526, p. 335, October 2015.

[31] J. Langedijk, A. K. Mantel-Teeuwisse, D. S. Slijkerman, and M.-H. D. B. Schutjens, "Drug repositioning and repurposing: terminology and definitions in literature.," *Drug discovery today*, vol. 20, pp. 1027–1034, Aug 2015.

[32] T. I. Oprea and J. P. Overington, "Computational and Practical Aspects of Drug Repositioning," *Drug Repurposing, Rescue, and Repositioning*, vol. 1, pp. 28–35, Feb. 2015.

[33] E. Fleming and P. Ma, "Drug life-cycle technologies," *Nat Rev Drug Discov*, vol. 1, pp. 751–752, Oct. 2002.

[34] E. L. Tobinick, "The value of drug repositioning in the current pharmaceutical market.," *Drug news & perspectives*, vol. 22, pp. 119–125, Mar 2009.

[35] P. Deotarse, A. Jain, M. Baile, N. Kolhe, and A. Kulkarni, "Drug repositioning: A review.," *International Journal of Pharma Research & Review.*, vol. 4, pp. 51–58, August 2015.

[36] S. Lee, K. Park, and D. Kim, "Building a drug-target network and its applications.," *Expert Opinion on Drug Discovery*, vol. 4, pp. 1–13, Oct. 2009.

[37] J. P. Overington, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?," *Nature reviews. Drug discovery*, vol. 5, pp. 993–996, Dec. 2006.

[38] V. J. Haupt, S. Daminelli, and M. Schroeder, "Drug promiscuity in pdb: Protein binding site similarity is key.," *PloS one*, vol. 8, p. e65894, Jun 2013.

[39] A. L. Hopkins, "Drug discovery: Predicting promiscuity," *Nature*, vol. 462, pp. 167–168, Nov. 2009.

[40] T. I. Oprea, J. E. Bauman, C. G. Bologa, T. Buranda, A. Chigaev, B. S. Edwards, J. W. Jarvik, H. D. Gresham, M. K. Haynes, B. Hjelle, R. Hromas, L. Hudson, D. A. Mackenzie, C. Y. Muller, J. C. Reed, P. C. Simons, Y. Smagley, J. Strouse, Z. Surviladze, T. Thompson, O. Ursu, A. Waller, A. Wandinger-Ness, S. S. Winter, Y. Wu, S. M. Young, R. S. Larson, C. Willman, and L. A. Sklar, "Drug repurposing from an academic perspective," *Drug Discovery Today: Therapeutic Strategies*, vol. 8, no. 3 - 4, pp. 61 – 69, 2011. Drug repurposing.

[41] A. Power, A. C. Berger, and G. S. Ginsburg, "Genomics-enabled drug repositioning and repurposing: insights from an iom roundtable activity.," *JAMA*, vol. 311, pp. 2063–2064, May 2014.

[42] A. Morales, C. Gingell, M. Collins, P. Wicker, and I. Osterloh, “Clinical safety of oral sildenafil citrate (viagra) in the treatment of erectile dysfunction.,” *Int J Impot Res*, vol. 10, pp. 69–73, Jun 1998.

[43] H. A. Ghofrani, I. H. Osterloh, and F. Grimminger, “Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond,” vol. 5, pp. 689–702+, 2006.

[44] W. H. Jost and P. Marsalek, “Duloxetine in the treatment of stress urinary incontinence.,” *Therapeutics and Clinical Risk Management*, vol. 1, no. 4, pp. 259 – 264, 2005.

[45] M. C. Michel and S. L. M. Peters, “Role of serotonin and noradrenaline in stress urinary incontinence.,” *BJU international*, vol. 94 Suppl 1, pp. 23–30, Jul 2004.

[46] B. Druker, *Imatinib as a Paradigm of Targeted Therapies*, vol. 91 of *Advances in Cancer Research*, pp. 1–30. Elsevier, 2004.

[47] M. Carroll, S. Ohno-Jones, S. Tamura, E. Buchdunger, J. Zimmermann, N. B. Lydon, D. G. Gilliland, and B. J. Druker, “Cgp 57148, a tyrosine kinase inhibitor, inhibits the growth of cells expressing bcr-abl, tel-abl, and tel-pdgfr fusion proteins.,” *Blood*, vol. 90, pp. 4947–4952, Dec 1997.

[48] M. C. Heinrich, D. J. Griffith, B. J. Druker, C. L. Wait, K. A. Ott, and A. J. Zigler, “Inhibition of c-kit receptor tyrosine kinase activity by sti 571, a selective tyrosine kinase inhibitor.,” *Blood*, vol. 96, pp. 925–932, Aug 2000.

[49] Z. Y. Pessetto, Y. Ma, J. J. Hirst, M. von Mehren, S. J. Weir, and A. K. Godwin, “Drug repurposing identifies a synergistic combination therapy with imatinib mesylate for gastrointestinal stromal tumor.,” *Molecular cancer therapeutics*, vol. 13, pp. 2276–2287, Oct 2014.

[50] O. S. Din and P. J. Woll, “Treatment of gastrointestinal stromal tumor: focus on imatinib mesylate.,” *Therapeutics and clinical risk management*, vol. 4, pp. 149–162, Feb 2008.

[51] Y. Y. Li, J. An, and S. J. M. Jones, “A computational approach to finding novel targets for existing drugs.,” *PLoS computational biology*, vol. 7, p. e1002139, Sep 2011.

[52] M. Johnson and G. Maggiore, “Concepts and application of molecular similarity.,” 1990.

[53] P. Sanseau and J. Koehler, “Editorial: Computational methods for drug repurposing,” *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 301–302, 2011.

[54] D. Rognan, “Chemogenomic approaches to rational drug design.,” *British journal of pharmacology*, vol. 152, pp. 38–52, Sept. 2007.

[55] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, "Predicting new molecular targets for known drugs.," *Nature*, vol. 462, pp. 175–181, Nov 2009.

[56] Y. Wang, S. Chen, N. Deng, and Y. Wang, "Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data.," *PloS one*, vol. 8, p. e78518, Nov 2013.

[57] F. Tan, R. Yang, X. Xu, X. Chen, Y. Wang, H. Ma, X. Liu, X. Wu, Y. Chen, L. Liu, and X. Jia, "Drug repositioning by applying 'expression profiles' generated by integrating chemical structure similarity and gene semantic similarity.," *Molecular bioSystems*, vol. 10, pp. 1126–1138, May 2014.

[58] J. T. Dudley, M. Sirota, M. Shenoy, R. K. Pai, S. Roedder, A. P. Chiang, A. A. Morgan, M. M. Sarwal, P. J. Pasricha, and A. J. Butte, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease.," *Science translational medicine*, vol. 3, p. 96ra76, Aug 2011.

[59] L. Huang, F. Li, J. Sheng, X. Xia, J. Ma, M. Zhan, and S. T. C. Wong, "Drug-comboranker: drug combination discovery based on target network analysis.," *Bioinformatics (Oxford, England)*, vol. 30, pp. i228–i236, Jun 2014.

[60] X.-M. Zhao, M. Iskar, G. Zeller, M. Kuhn, V. van Noort, and P. Bork, "Prediction of drug combinations by integrating molecular and pharmacological data.," *PLoS computational biology*, vol. 7, p. e1002323, Dec 2011.

[61] H. Huang, P. Zhang, X. A. Qu, P. Sanseau, and L. Yang, "Systematic prediction of drug combinations based on clinical side-effects.," *Scientific reports*, vol. 4, p. 7160, Nov 2014.

[62] Y. Sun, Y. Xiong, Q. Xu, and D. Wei, "A hadoop-based method to predict potential effective drug combination.," *BioMed research international*, vol. 2014, p. 196858, Jul 2014.

[63] Y. Liu, Q. Wei, G. Yu, W. Gai, Y. Li, and X. Chen, "Dcdb 2.0: a major update of the drug combination database.," *Database : the journal of biological databases and curation*, vol. 2014, p. bau124, Dec 2014.

[64] P. Imming, C. Sinning, and A. Meyer, "Drugs, their targets and the nature and number of drug targets," *Nature Reviews Drug Discovery*, vol. 5, pp. 821–834, Oct. 2006.

[65] G. Jin and S. T. C. Wong, "Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines.," *Drug discovery today*, vol. 19, pp. 637–644, May 2014.

[66] S. L. Kinnings, N. Liu, N. Buchmeier, P. J. Tonge, L. Xie, and P. E. Bourne, "Drug discovery using chemical systems biology: repositioning the safe medicine

comtan to treat multi-drug and extensively drug resistant tuberculosis.," *PLoS computational biology*, vol. 5, p. e1000423, Jul 2009.

[67] J. Lamb, E. D. Crawford, D. Peck, J. W. Modell, I. C. Blat, M. J. Wrobel, J. Lerner, J.-P. Brunet, A. Subramanian, K. N. Ross, M. Reich, H. Hieronymus, G. Wei, S. A. Armstrong, S. J. Haggarty, P. A. Clemons, R. Wei, S. A. Carr, E. S. Lander, and T. R. Golub, "The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease," *Science*, vol. 313, pp. 1929–1935, Sept. 2006.

[68] D. Vidović, A. Koleti, and S. C. Schürer, "Large-scale integration of small molecule-induced genome-wide transcriptional responses, kinome-wide binding affinities and cell-growth inhibition profiles reveal global trends characterizing systems-level drug action.," *Frontiers in genetics*, vol. 5, p. 342, Sep 2014.

[69] J. A. Parkkinen and S. Kaski, "Probabilistic drug connectivity mapping," *BMC bioinformatics*, vol. 15, no. 1, p. 113, 2014.

[70] J. Lamb, "The connectivity map: a new tool for biomedical research.," *Nature reviews. Cancer*, vol. 7, pp. 54–60, Jan 2007.

[71] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, and R. Edgar, "Ncbi geo: mining tens of millions of expression profiles–database and tools update.," *Nucleic acids research*, vol. 35, pp. D760–D765, Jan 2007.

[72] N. S. Jahchan, J. T. Dudley, P. K. Mazur, N. Flores, D. Yang, A. Palmerton, A.-F. Zmoos, D. Vaka, K. Q. T. Tran, M. Zhou, K. Krasinska, J. W. Riess, J. W. Neal, P. Khatri, K. S. Park, A. J. Butte, and J. Sage, "A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors.," *Cancer discovery*, vol. 3, pp. 1364–1377, Dec 2013.

[73] F. Iorio, T. Rittman, H. Ge, M. Menden, and J. Saez-Rodriguez, "Transcriptional data: a new gateway to drug repositioning?," *Drug discovery today*, vol. 18, pp. 350–357, Apr 2013.

[74] P. Sanseau, P. Agarwal, M. R. Barnes, T. Pastinen, J. B. Richards, L. R. Cardon, and V. Mooser, "Use of genome-wide association studies for drug repositioning.," *Nature biotechnology*, vol. 30, pp. 317–320, Apr. 2012.

[75] Y. Okada, D. Wu, G. Trynka, T. Raj, C. Terao, K. Ikari, Y. Kochi, K. Ohmura, A. Suzuki, S. Yoshida, R. R. Graham, A. Manoharan, W. Ortmann, T. Bhangale, J. C. Denny, R. J. Carroll, A. E. Eyler, J. D. Greenberg, J. M. Kremer, D. A. Pappas, L. Jiang, J. Yin, L. Ye, D.-F. Su, J. Yang, G. Xie, E. Keystone, H.-J. Westra, T. Esko, A. Metspalu, X. Zhou, N. Gupta, D. Mirel, E. A. Stahl, D. Diogo, J. Cui, K. Liao, M. H. Guo, K. Myouzen, T. Kawaguchi, M. J. H. Coenen, P. L. C. M. van Riel, M. A. F. J. van de Laar, H.-J. Guchelaar, T. W. J. Huizinga, P. Dieudé, X. Mariette, S. L. Bridges, A. Zhernakova, R. E. M. Toes,

P. P. Tak, C. Miceli-Richard, S.-Y. Bang, H.-S. Lee, J. Martin, M. A. Gonzalez-Gay, L. Rodriguez-Rodriguez, S. Rantapää-Dahlqvist, L. Arlestig, H. K. Choi, Y. Kamatani, P. Galan, M. Lathrop, RACI consortium, GARNET consortium, S. Eyre, J. Bowes, A. Barton, N. de Vries, L. W. Moreland, L. A. Criswell, E. W. Karlson, A. Taniguchi, R. Yamada, M. Kubo, J. S. Liu, S.-C. Bae, J. Worthington, L. Padyukov, L. Klareskog, P. K. Gregersen, S. Raychaudhuri, B. E. Stranger, P. L. De Jager, L. Franke, P. M. Visscher, M. A. Brown, H. Yamanaka, T. Mimori, A. Takahashi, H. Xu, T. W. Behrens, K. A. Siminovitch, S. Momohara, F. Matsuda, K. Yamamoto, and R. M. Plenge, "Genetics of rheumatoid arthritis contributes to biology and drug discovery.," *Nature*, vol. 506, pp. 376–381, Feb 2014.

[76] Y. Li and P. Agarwal, "A Pathway-Based View of Human Diseases and Disease Relationships," *PLoS ONE*, vol. 4, pp. e4346+, Feb. 2009.

[77] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases.," *Scientific reports*, vol. 5, p. 10888, Jun 2015.

[78] S. J. Hebbring, "The challenges, advantages and future of phenome-wide association studies.," *Immunology*, vol. 141, pp. 157–165, Feb 2014.

[79] H. Ye, Q. Liu, and J. Wei, "Construction of drug network based on side effects and its application for drug repositioning.," *PloS one*, vol. 9, p. e87864, Feb 2014.

[80] L. Yang and P. Agarwal, "Systematic Drug Repositioning Based on Clinical Side-Effects," *PLoS ONE*, vol. 6, pp. e28025+, Dec. 2011.

[81] M. Campillos, M. Kuhn, A.-C. C. Gavin, L. J. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity.," *Science (New York, N.Y.)*, vol. 321, pp. 263–266, July 2008.

[82] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs.," *Molecular systems biology*, vol. 6, p. 343, Jan 2010.

[83] R. Hoehndorf, A. Oellrich, D. Rebholz-Schuhmann, P. N. Schofield, and G. V. Gkoutos, "Linking pharmgkb to phenotype studies and animal models of disease for drug repurposing.," *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 388–399, 2012.

[84] T. M. Mitchell, *Machine Learning.* New York, NY, USA: McGraw-Hill, Inc., 1 ed., 1997.

[85] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine.," *Molecular systems biology*, vol. 7, June 2011.

[86] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, and D. Greco, "Drug repositioning: A machine-learning approach through data integration.," *J. Cheminformatics*, vol. 5, p. 30, 2013.

[87] M. P. Menden, F. Iorio, M. Garnett, U. McDermott, C. H. Benes, P. J. Ballester, and J. Saez-Rodriguez, "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties.," *PloS one*, vol. 8, p. e61318, Apr 2013.

[88] P. Zhang, F. Wang, and J. Hu, "Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity.," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2014, pp. 1258–1267, Nov 2014.

[89] F. Iorio, R. Bosotti, E. Scacheri, V. Belcastro, P. Mithbaokar, R. Ferriero, L. Murino, R. Tagliaferri, N. Brunetti-Pierri, A. Isacchi, and D. di Bernardo, "Discovery of drug mode of action and drug repositioning from transcriptional responses.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 14621–14626, Aug. 2010.

[90] C. Andronis, A. Sharma, V. Virvilis, S. Deftereos, and A. Persidis, "Literature mining, ontologies and information visualization for drug repurposing.," *Briefings in bioinformatics*, vol. 12, pp. 357–368, Jul 2011.

[91] L. B. Tari and J. H. Patel, "Systematic drug repurposing through text mining.," *Methods in molecular biology (Clifton, N.J.)*, vol. 1159, pp. 253–267, 2014.

[92] D. R. Swanson, "Fish oil, raynaud's syndrome, and undiscovered public knowledge.," *Perspectives in biology and medicine*, vol. 30, no. 1, pp. 7–18, 1986.

[93] P. Csermely, T. Korcsmáros, H. J. M. Kiss, G. London, and R. Nussinov, "Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review.," *Pharmacology & therapeutics*, vol. 138, pp. 333–408, Jun 2013.

[94] A. F. Shaughnessy, "Old drugs, new tricks.," *BMJ (Clinical research ed.)*, vol. 342, p. d741, Feb 2011.

[95] M. Schneider, "Defining Systems Biology: A Brief Overview of the Term and Field," in *In Silico Systems Biology* (M. V. Schneider, ed.), vol. 1021 of *Methods in Molecular Biology*, pp. 1–11, Humana Press, 2013.

[96] F. Riaz and K. M. Ali, "Applications of graph theory in computer science," pp. 142–145, IEEE, July 2011.

[97] X. M. Fernández-Suárez, D. J. Rigden, and M. Y. Galperin, "The 2014 Nucleic Acids Research Database Issue and an updated NAR online Molecular Biology Database Collection," *Nucleic Acids Research*, vol. 42, pp. D1–D6, Jan. 2014.

[98] N. T. Doncheva, T. Kacprowski, and M. Albrecht, "Recent approaches to the prioritization of candidate disease genes.," *Wiley interdisciplinary reviews. Systems biology and medicine*, vol. 4, pp. 429–442, Sep/Oct 2012.

[99] M. Y. Galperin, D. J. Rigden, and X. M. Fernández-Suárez, "The 2015 nucleic acids research database issue and molecular biology database collection.," *Nucleic acids research*, vol. 43, pp. D1–D5, Jan 2015.

[100] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, and J. Tegnér, "Data integration in the era of omics: current and future challenges.," *BMC systems biology*, vol. 8 Suppl 2, p. I1, Mar 2014.

[101] S. J. Cockell, J. Weile, P. Lord, C. Wipat, D. Andriychenko, M. Pocock, D. Wilkinson, M. Young, and A. Wipat, "An integrated dataset for in silico drug discovery.," *Journal of integrative bioinformatics*, vol. 7, no. 3, 2010.

[102] D. B. Searls, "Data integration: challenges for drug discovery," *Nature Reviews Drug Discovery*, vol. 4, pp. 45–58, Jan. 2005.

[103] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome research*, vol. 13, pp. 2498–2504, Nov 2003.

[104] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," 2009.

[105] J. Köhler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Rüegg, C. Rawlings, P. Verrier, and S. Philippi, "Graph-based analysis and visualization of experimental results with ONDEX," *Bioinformatics*, vol. 22, pp. 1383–1390, June 2006.

[106] C. T. Have and L. J. Jensen, "Are graph databases ready for bioinformatics?," *Bioinformatics (Oxford, England)*, vol. 29, pp. 3107–3108, Dec 2013.

[107] M. A. Rodriguez and P. Neubauer, "The graph traversal pattern," *CoRR*, vol. abs/1004.1001, 2010.

[108] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, and J. P. Overington, "Chembl: a large-scale bioactivity database for drug discovery.," *Nucleic acids research*, vol. 40, pp. D1100–D1107, Jan 2012.

[109] P. D'Eustachio, "Reactome knowledgebase of human biological pathways and processes.," *Methods in molecular biology (Clifton, N.J.)*, vol. 694, pp. 49–61, 2011.

[110] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, and A. M. Jenkinson, "The ebi rdf platform: linked open data for the life sciences.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 1338–1339, May 2014.

[111] T. R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing," *International Journal of Human-Computer Studies*, vol. 43, pp. 907–928, Nov. 1995.

[112] J. Taubert, K. Hassani-Pak, N. Castells-Brooke, and C. J. Rawlings, "Ondex web: web-based visualization and exploration of heterogeneous biological networks.," *Bioinformatics (Oxford, England)*, vol. 30, pp. 1034–1035, Apr 2014.

[113] D. S. Wishart, "DrugBank: a comprehensive resource for in silico drug discovery and exploration," *Nucleic Acids Research*, vol. 34, pp. D668–D672, Jan. 2006.

[114] UniProt Consortium, "Update on activities at the universal protein resource (UniProt) in 2013," *Nucleic acids research*, vol. 41, pp. D43–47, Jan. 2013. PMID: 23161681.

[115] M. Magrane and U. Consortium, "Uniprot knowledgebase: a hub of integrated protein data.," *Database : the journal of biological databases and curation*, vol. 2011, p. bar009, Mar 2011.

[116] L. J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, and C. von Mering, "String 8–a global view on proteins and their functional interactions in 630 organisms.," *Nucleic acids research*, vol. 37, pp. D412–D416, Jan 2009.

[117] Gene Ontology Consortium, "Creating the gene ontology resource: design and implementation.," *Genome research*, vol. 11, pp. 1425–1433, Aug 2001.

[118] J. Lomax, "Get ready to GO! A biologist's guide to the Gene Ontology," *Briefings in Bioinformatics*, vol. 6, pp. 298–304, Sept. 2005.

[119] D. Barrell, E. Dimmer, R. P. Huntley, D. Binns, C. O'Donovan, and R. Apweiler, "The goa database in 2009–an integrated gene ontology annotation resource.," *Nucleic acids research*, vol. 37, pp. D396–D403, Jan 2009.

[120] F. Rogers, "Medical subject headings.," *Bulletin of the Medical Library Association*, vol. 51, pp. 114–116, Jan. 1963.

[121] "Online mendelian inheritance in man: Omim." McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University Baltimore, MD. [Accessed: 09-2013]. Available from: `http://omim.org`.

[122] D. Welter, J. MacArthur, J. Morales, T. Burdett, P. Hall, H. Junkins, A. Klemm, P. Flicek, T. Manolio, L. Hindorff, and H. Parkinson, "The nhgri gwas catalog, a curated resource of snp-trait associations.," *Nucleic acids research*, vol. 42, pp. D1001–D1006, Jan 2014.

[123] "The nhgri-ebi catalog of published genome-wide association studies.." Burdett T (EBI), Hall PN (NHGRI), Hasting E (EBI) Hindorff LA (NHGRI), Junkins HA (NHGRI), Klemm AK (NHGRI), MacArthur J (EBI), Manolio TA (NHGRI), Morales J (EBI), Parkinson H (EBI) and Welter D (EBI). [Accessed: 09-2014]. Available from: `www.ebi.ac.uk/gwas`, version [v1.0].

[124] A. Bravo, J. Piaero, N. Queralt-Rosinach, M. Rautschka, and L. I. Furlong, "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research," *BMC Bioinformatics*, vol. 16, pp. 55+, Feb. 2015.

[125] T. C. Rindflesch and M. Fiszman, “The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text.,” *Journal of biomedical informatics*, vol. 36, pp. 462–477, Dec 2003.

[126] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, J. E. Richardson, and Mouse Genome Database Group, “The mouse genome database (mgd): facilitating mouse as a model for human biology and disease.,” *Nucleic acids research*, vol. 43, pp. D726–D736, Jan 2015.

[127] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, “Drugbank 4.0: shedding new light on drug metabolism.,” *Nucleic acids research*, vol. 42, pp. D1091–D1097, Jan 2014.

[128] A. P. Bento, A. Gaulton, A. Hersey, L. J. Bellis, J. Chambers, M. Davies, F. A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J. P. Overington, “The chembl bioactivity database: an update.,” *Nucleic acids research*, vol. 42, pp. D1083–D1090, Jan 2014.

[129] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, “The sider database of drugs and side effects,” *Nucleic Acids Research*, 2015.

[130] A. J. Williams, L. Harland, P. Groth, S. Pettifer, C. Chichester, E. L. Willighagen, C. T. Evelo, N. Blomberg, G. Ecker, C. Goble, and B. Mons, “Open PHACTS: semantic interoperability for drug discovery,” *Drug Discovery Today*, vol. 17, pp. 1188–1198, Nov. 2012.

[131] A. P. Davis, C. J. Grondin, K. Lennon-Hopkins, C. Saraceni-Richards, D. Sciaky, B. L. King, T. C. Wiegers, and C. J. Mattingly, “The comparative toxicogenomics database’s 10th year anniversary: update 2015.,” *Nucleic acids research*, vol. 43, pp. D914–D920, Jan 2015.

[132] “Unified medical language system ®,” May 5 2011. Release 2011AA. Bethesda (MD): National Library of Medicine (US); Available from: `http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.htm`.

[133] J. Chambers, M. Davies, A. Gaulton, A. Hersey, S. Velankar, R. Petryszak, J. Hastings, L. Bellis, S. McGlinchey, and J. P. Overington, “Unichem: a unified chemical structure cross-referencing and identifier tracking system.,” *Journal of cheminformatics*, vol. 5, p. 3, Jan 2013.

[134] M. Whirl-Carrillo, E. M. McDonagh, J. M. Hebert, L. Gong, K. Sangkuhl, C. F. Thorn, R. B. Altman, and T. E. Klein, “Pharmacogenomics knowledge for personalized medicine.,” *Clinical pharmacology and therapeutics*, vol. 92, pp. 414–417, Oct. 2012.

[135] “Orphadata: Free access data from orphanet.” ©INSERM 1997; [Accessed: 09-2014]. Available from: `http://www.orphadata.org`.

[136] S. C. Bull and A. J. Doig, “Properties of protein drug target classes.,” *PloS one*, vol. 10, no. 3, 2015.

[137] B. Gallagher, “Matching structure and semantics: A survey on graph-based pattern matching,” *AAAI FS*, vol. 6, pp. 45–53, 2006.

[138] J. Qian, A. Hintze, and C. Adami, “Colored Motifs Reveal Computational Building Blocks in the C. elegans Brain,” *PLoS ONE*, vol. 6, pp. e17013+, Mar. 2011.

[139] A.-L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell's functional organization.,” *Nature reviews. Genetics*, vol. 5, pp. 101–113, Feb 2004.

[140] R. Milo, “Network motifs: Simple building blocks of complex networks,” *Science*, vol. 298, pp. 824–827, Oct. 2002.

[141] E. Wong, B. Baur, S. Quader, and C.-H. Huang, “Biological network motif detection: principles and practice,” *Briefings in Bioinformatics*, vol. 13, pp. 202–215, June 2011.

[142] S. Mangan and U. Alon, “Structure and function of the feed-forward loop network motif,” *Proceedings of the National Academy of Sciences*, vol. 100, pp. 11980–11985, Oct. 2003.

[143] N. T. L. Tran, L. DeLuccia, A. F. McDonald, and C.-H. Huang, “Cross-disciplinary detection and analysis of network motifs.,” *Bioinformatics and biology insights*, vol. 9, pp. 49–60, Apr 2015.

[144] N. Betzler, R. van Bevern, M. R. Fellows, C. Komusiewicz, and R. Niedermeier, “Parameterized algorithmics for finding connected motifs in biological networks,” *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 8, pp. 1296–1308, Oct. 2011. PMID: 21282862.

[145] V. Lacroix, C. G. Fernandes, and M.-F. Sagot, “Motif search in graphs: application to metabolic networks,” *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 3, pp. 360–368, Dec. 2006. PMID: 17085845.

[146] J. Mullen, S. J. Cockell, H. Tipney, P. M. Woollard, and A. Wipat, “Mining integrated semantic networks for drug repositioning opportunities.,” *PeerJ*, vol. 4, p. e1558, Jan 2016.

[147] M. D. Rukhadze, G. S. Bezarashvili, N. S. Sidamonidze, and S. K. Tsagareli, “Investigation of binding process of chlorpromazine to bovine serum albumin by means of passive and active experiments,” *Biomedical chromatography: BMC*, vol. 15, pp. 365–373, Oct. 2001. PMID: 11559920.

[148] P. Mitchell, “Chlorpromazine Turns Forty,” *Aust NZ J Psychiatry*, vol. 27, pp. 370–373, Jan. 1993.

[149] D. Conte, P. Foggia, C. Sansone, and M. Vento, “Thirty Years Of Graph Matching In Pattern Recognition,” *International Journal of Pattern Recognition and Artificial Intelligence*, 2004.

[150] T. Washio and H. Motoda, “State of the art of graph-based data mining,” *ACM SIGKDD Explorations Newsletter*, vol. 5, p. 59, July 2003.

[151] J. R. Ullmann, “An algorithm for subgraph isomorphism,” *Journal of the ACM*, vol. 23, pp. 31–42, Jan. 1976.

[152] H. He and A. K. Singh, “Graphs-at-a-time,” p. 405, ACM Press, 2008.

[153] S. Zhang, S. Li, and J. Yang, “GADDI,” p. 192, ACM Press, 2009.

[154] S. Djoko, D. Cook, and L. Holder, “An empirical study of domain knowledge and its benefits to substructure discovery,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 9, pp. 575–586, Aug. 1997.

[155] L. Cordella, P. Foggia, C. Sansone, and M. Vento, “A (sub)graph isomorphism algorithm for matching large graphs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 1367–1372, Oct. 2004.

[156] R. Giugno and D. Shasha, “GraphGrep: A fast and universal method for querying graphs,” in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, vol. 2, pp. 112–115 vol.2, 2002.

[157] J. Ge and Y. Qiu, “Concept similarity matching based on semantic distance,” pp. 380–383, IEEE, Dec. 2008.

[158] N. F. Noy, “Semantic integration: a survey of ontology-based approaches,” *ACM SIGMOD Record*, vol. 33, p. 65, Dec. 2004.

[159] D. Seid and S. Mehrotra, “Semantically ranked graph pattern queries for link analysis,” pp. 296–299, IEEE, May 2007.

[160] E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[161] A. L. Hopkins, J. S. Mason, and J. P. Overington, “Can we rationally design promiscuous drugs?,” *Current opinion in structural biology*, vol. 16, pp. 127–136, Feb 2006.

[162] Q. Kuang, X. Xu, R. Li, Y. Dong, Y. Li, Z. Huang, Y. Li, and M. Li, “An eigenvalue transformation technique for predicting drug-target interaction.,” *Scientific reports*, vol. 5, p. 13867, Sep 2015.

[163] J. T. Dudley, T. Deshpande, and A. J. Butte, “Exploiting drug-disease relationships for computational drug repositioning,” *Briefings in Bioinformatics*, vol. 12, pp. 303–311, June 2011.

[164] S. Fakhraei, B. Huang, L. Raschid, and L. Getoor, “Network-based drug-target interaction prediction with probabilistic soft logic.,” *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, vol. 11, pp. 775–787, Sep/Oct 2014.

[165] H. Ding, I. Takigawa, H. Mamitsuka, and S. Zhu, “Similarity-based machine learning methods for predicting drug-target interactions: a brief review.,” *Briefings in bioinformatics*, vol. 15, pp. 734–747, Sep 2014.

[166] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, “Molecular docking: a powerful approach for structure-based drug discovery.,” *Current computer-aided drug design*, vol. 7, pp. 146–157, Jun 2011.

[167] J. Yang, Z. Li, X. Fan, and Y. Cheng, “Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization.,” *Journal of chemical information and modeling*, vol. 54, pp. 2562–2569, Sep 2014.

[168] J. Alvarsson, M. Eklund, O. Engkvist, O. Spjuth, L. Carlsson, J. E. S. Wikberg, and T. Noeske, “Ligand-based target prediction with signature fingerprints.,” *Journal of chemical information and modeling*, vol. 54, pp. 2647–2653, Oct 2014.

[169] G. Palma, M.-E. Vidal, and L. Raschid, “Drug-target interaction prediction using semantic similarity and edge partitioning,” in *The Semantic Web â AI ISWC 2014* (P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandefcifa, P. Groth, N. Noy, K. Janowicz, and C. Goble, eds.), vol. 8796 of *Lecture Notes in Computer Science*, pp. 131–146, Springer International Publishing, 2014.

[170] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, “Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.,” *Bioinformatics (Oxford, England)*, vol. 24, pp. i232–i240, Jul 2008.

[171] Y. Yamanishi, M. Kotera, M. Kanehisa, and S. Goto, “Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework,” *Bioinformatics*, vol. 26, pp. i246–i254, June 2010.

[172] S. Lee, K. Park, and D. Kim, “Building a drug-target network and its applications.,” *Expert opinion on drug discovery*, vol. 4, pp. 1177–1189, Nov 2009.

[173] R. Lappano and M. Maggiolini, “G protein-coupled receptors: novel targets for drug discovery in cancer,” *Nat Rev Drug Discov*, vol. 10, pp. 47–60, Jan. 2011.

[174] S. K. Bagal, A. D. Brown, P. J. Cox, K. Omoto, R. M. Owen, D. C. Pryde, B. Sidders, S. E. Skerratt, E. B. Stevens, R. I. Storer, and N. A. Swain, “Ion channels as therapeutic targets: A drug discovery perspective,” *Journal of Medicinal Chemistry*, vol. 56, no. 3, pp. 593–624, 2013. PMID: 23121096.

[175] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The Protein Kinase Complement of the Human Genome," *Science*, vol. 298, pp. 1912–1934, Dec. 2002.

[176] M. Drag and G. S. Salvesen, "Emerging principles in protease-based drug discovery.," *Nature reviews. Drug discovery*, vol. 9, pp. 690–701, Sept. 2010.

[177] G. J. Kaczorowski, O. B. McManus, B. T. Priest, and M. L. Garcia, "Ion channels as drug targets: the next gpcrs.," *The Journal of general physiology*, vol. 131, pp. 399–405, May 2008.

[178] D. Fabbro, S. Cowan-Jacob, H. Möbitz, and G. Martiny-Baron, "Targeting Cancer with Small-Molecular-Weight Kinase Inhibitors," in *Kinase Inhibitors* (B. Kuster, ed.), vol. 795 of *Methods in Molecular Biology*, pp. 1–34, Humana Press, 2012.

[179] S. E. Baranzini, "Revealing the genetic basis of multiple sclerosis: are we there yet?," *Current opinion in genetics & development*, vol. 21, pp. 317–324, Jun 2011.

[180] F. Martínez-Jiménez, G. Papadatos, L. Yang, I. M. Wallace, V. Kumar, U. Pieper, A. Sali, J. R. Brown, J. P. Overington, and M. A. Marti-Renom, "Target prediction for an open access set of compounds active against mycobacterium tuberculosis.," *PLoS computational biology*, vol. 9, p. e1003253, Oct 2013.

[181] G. Mugumbate, K. A. Abrahams, J. A. G. Cox, G. Papadatos, G. van Westen, J. Lelièvre, S. T. Calus, N. J. Loman, L. Ballell, D. Barros, J. P. Overington, and G. S. Besra, "Mycobacterial dihydrofolate reductase inhibitors identified using chemogenomic methods and in vitro validation.," *PloS one*, vol. 10, p. e0121492, Mar 2015.

[182] N. Dilmac, N. Hilliard, and G. H. Hockerman, "Molecular determinants of frequency dependence and ca2+ potentiation of verapamil block in the pore region of cav1.2.," *Molecular pharmacology*, vol. 66, pp. 1236–1247, Nov 2004.

[183] L. Asmal, S. J. Flegar, J. Wang, C. Rummel-Kluge, K. Komossa, and S. Leucht, "Quetiapine versus other atypical antipsychotics for schizophrenia.," *The Cochrane database of systematic reviews*, vol. 11, 2013.

[184] M. A. Gobbo and M. R. Louzã, "Influence of stimulant and non-stimulant drug treatment on driving performance in patients with attention deficit hyperactivity disorder: a systematic review.," *European neuropsychopharmacology : the journal of the European College of Neuropsychopharmacology*, vol. 24, pp. 1425–1443, Sep 2014.

[185] NIDA, "Propranolol for Treatment of Cocaine Addiction - 2," 2010. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2015 Apr 27]. Available from: `https://clinicaltrials.gov/ct2/show/NCT00000197` NLM Identifier: NCT00000197.

[186] M. U. of South Carolina, "Enhancing Disrupted Reconsolidation: Impact on Cocaine Craving," 2015. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2015 Apr 27]. Available from: `https://clinicaltrials.gov/ct2/show/NCT01822587` NLM Identifier: NCT01822587.

[187] U. of Missouri-Columbia, "Trial of Propranolol in Adults and Adolescents With ASD and Predictors of Response," 2015. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2015 Apr 27]. Available from: `https://clinicaltrials.gov/ct2/show/NCT02414451` NLM Identifier: NCT02414451.

[188] M. S. . D. Corp., "Study of Preladenant for the Treatment of Neuroleptic Induced Akathisia," 2014. In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2015 Apr 27]. Available from: `https://clinicaltrials.gov/ct2/show/NCT00693472` NLM Identifier: NCT00693472.

[189] J. T. Stewart, M. L. Mounts, and R. L. Clark, "Aggressive behavior in huntington's disease: treatment with propranolol.," *The Journal of clinical psychiatry*, vol. 48, pp. 106–108, Mar 1987.

[190] J. Wang, Z. Zhao, E. Lin, W. Zhao, X. Qian, D. Freire, A. E. Bilski, A. Cheng, P. Vempati, L. Ho, K. Ono, M. Yamada, and G. M. Pasinetti, "Unintended effects of cardiovascular drugs on the pathogenesis of alzheimer's disease.," *PloS one*, vol. 8, p. e65232, Jun 2013.

[191] M. Mantegazza, G. Curia, G. Biagini, D. S. Ragsdale, and M. Avoli, "Voltage-gated sodium channels as therapeutic targets in epilepsy and other neurological disorders.," *The Lancet. Neurology*, vol. 9, pp. 413–424, Apr 2010.

[192] A. Escayg and A. L. Goldin, "Sodium channel scn1a and epilepsy: mutations and mechanisms.," *Epilepsia*, vol. 51, pp. 1650–1658, Sep 2010.

[193] B. Holm, M. Sehested, and P. B. Jensen, "Improved targeting of brain tumors using dexrazoxane rescue of topoisomerase ii combined with supralethal doses of etoposide and teniposide.," *Clinical cancer research : an official journal of the American Association for Cancer Research*, vol. 4, pp. 1367–1373, Jun 1998.

[194] R. J. Xavier and J. D. Rioux, "Genome-wide association studies: a new window into immune-mediated diseases.," *Nature reviews. Immunology*, vol. 8, pp. 631–643, Aug 2008.

[195] J. F. Gusella, N. S. Wexler, P. M. Conneally, S. L. Naylor, M. A. Anderson, R. E. Tanzi, P. C. Watkins, K. Ottina, M. R. Wallace, and A. Y. Sakaguchi, "A polymorphic dna marker genetically linked to huntington's disease.," *Nature*, vol. 306, pp. 234–238, Nov 1983.

[196] M. G. Kann, "Advances in translational bioinformatics: computational approaches for the hunting of disease genes.," *Briefings in bioinformatics*, vol. 11, pp. 96–110, Jan. 2010.

[197] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, “A comprehensive review of genetic association studies.,” *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 4, pp. 45–61, Mar/Apr 2002.

[198] Y. Bromberg, “Chapter 15: disease gene prioritization.,” *PLoS computational biology*, vol. 9, p. e1002902, Apr 2013.

[199] R.-L. Liu and C.-C. Shih, “Identification of highly related references about gene-disease association.,” *BMC bioinformatics*, vol. 15, p. 286, Aug 2014.

[200] J. A. Blake, C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, and T. M. G. D. Group, “The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse,” *Nucleic Acids Research*, pp. gkt1225+, Nov. 2013.

[201] M. Shimoyama, J. De Pons, G. T. Hayman, S. J. F. Laulederkind, W. Liu, R. Nigam, V. Petri, J. R. Smith, M. Tutaj, S.-J. Wang, E. Worthey, M. Dwinell, and H. Jacob, “The rat genome database 2015: genomic, phenotypic and environmental variations and disease.,” *Nucleic acids research*, vol. 43, pp. D743–D750, Jan 2015.

[202] N. Rosenthal and S. Brown, “The mouse ascending: perspectives for human-disease models.,” *Nature cell biology*, vol. 9, pp. 993–999, Sep 2007.

[203] D. Smedley, A. Oellrich, S. Köhler, B. Ruef, Sanger Mouse Genetics Project, M. Westerfield, P. Robinson, S. Lewis, and C. Mungall, “Phenodigm: analyzing curated annotations to associate animal models with human diseases.,” *Database : the journal of biological databases and curation*, vol. 2013, p. bat025, May 2013.

[204] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky, “Emergent behavior of growing knowledge about molecular interactions.,” *Nature biotechnology*, vol. 23, pp. 1243–1247, Oct 2005.

[205] A. Bauer-Mehren, M. Rautschka, F. Sanz, and L. I. Furlong, “DisGeNET: a cytoscape plugin to visualize, integrate, search and analyze gene-disease networks,” *Bioinformatics*, vol. 26, pp. 2924–2926, Sept. 2010.

[206] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, and L. I. Furlong, “Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases.,” *PloS one*, vol. 6, p. e20284, Jun 2011.

[207] N. Rappaport, M. Twik, N. Nativ, G. Stelzer, I. Bahir, T. I. Stein, M. Safran, and D. Lancet, “Malacards: A comprehensive automatically-mined database of human diseases.,” *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. 47, pp. 1.24.1–1.24.19, Sep 2014.

[208] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, “Diseases: Text mining and data integration of disease-gene associations.,” *Methods (San Diego, Calif.)*, Dec 2014.

[209] F. Sams-Dodd, “Target-based drug discovery: is something wrong?,” *Drug discovery today*, vol. 10, pp. 139–147, Jan 2005.

[210] I. H. Gilbert, “Drug discovery for neglected diseases: molecular target-based and phenotypic approaches.,” *Journal of medicinal chemistry*, vol. 56, pp. 7719–7726, Oct 2013.

[211] S. Hoelder, P. A. Clarke, and P. Workman, “Discovery of small molecule cancer drugs: successes, challenges and opportunities.,” *Molecular oncology*, vol. 6, pp. 155–176, Apr 2012.

[212] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, “A Probabilistic Functional Network of Yeast Genes,” *Science*, vol. 306, pp. 1555–1558, Nov. 2004.

[213] J. Mullen, S. J. Cockell, P. Woollard, and A. Wipat, “An Integrated Data Driven Approach to Drug Repositioning Using Gene-Disease Associations,” *PLoS ONE*, vol. 11, pp. e0155811+, May 2016.

[214] UniProt Consortium, “Activities at the universal protein resource (uniprot).,” *Nucleic acids research*, vol. 42, pp. D191–D198, Jan 2014.

[215] “Orphanet: an online rare disease and orphan drug data base.,” 1997. ©INSERM 1997; [Accessed 19-July-2014]. Available from: `http://www.orpha.net`.

[216] “Semrep.” Bethesda (MD): National Library of Medicine (US); [Accessed: 02-2015]. Available from: `http://skr3.nlm.nih.gov`.

[217] I. Molineris, U. Ala, P. Provero, and F. Di Cunto, “Drug repositioning for orphan genetic diseases through conserved anticoexpressed gene clusters (cagcs).,” *BMC bioinformatics*, vol. 14, p. 288, Oct 2013.

[218] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, “Genenames.org: the hgnc resources in 2015.,” *Nucleic acids research*, vol. 43, pp. D1079–D1085, Jan 2015.

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

[220] H. Kilicoglu, G. Rosemblat, M. Fiszman, and T. C. Rindflesch, “Constructing a semantic predication gold standard from the biomedical literature.,” *BMC bioinformatics*, vol. 12, p. 486, Dec 2011.

[221] C. Leacock and M. Chodorow, “Combining local context and wordnet similarity for word sense identification,” in *MIT Press* (C. Fellfaum, ed.), (Cambridge, Massachusetts), pp. 265–283, 1998.

[222] B. T. McInnes, T. Pedersen, and S. V. S. Pakhomov, “Umls-interface and umls-similarity : open source software for measuring paths and semantic similarity.,” *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, vol. 2009, pp. 431–435, Nov 2009.

[223] J. Weile, K. James, J. Hallinan, S. J. Cockell, P. Lord, A. Wipat, and D. J. Wilkinson, “Bayesian integration of networks without gold standards,” *Bioinformatics*, vol. 28, pp. 1495–1500, June 2012.

[224] A. A. Fox, M. Pretorius, K.-Y. Liu, C. D. Collard, T. E. Perry, S. K. Shernan, P. L. De Jager, D. A. Hafler, D. S. Herman, S. R. DePalma, D. M. Roden, J. D. Muehlschlegel, B. S. Donahue, D. Darbar, J. G. Seidman, S. C. Body, and C. E. Seidman, “Genome-wide assessment for genetic variants associated with ventricular dysfunction after primary coronary artery bypass graft surgery.,” *PloS one*, vol. 6, p. e24593, Sep 2011.

[225] “Comparative toxicogenomics database (ctd).” Curated drug-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), MDI Biological Laboratory, Salisbury Cove, Maine, and NC State University, Raleigh, North Carolina. Available from: `http://ctdbase.org/`.

[226] T. P. O’Connor and R. G. Crystal, “Genetic medicines: treatment strategies for hereditary disorders.,” *Nature reviews. Genetics*, vol. 7, pp. 261–276, Apr 2006.

[227] “Pharmaadme.” [Accessed: 27-05-2015]. Available from: `http://www.PharmaADME.org`.

[228] “University of California, Irvine. Trial of Dasatinib (Sprycelĺ) in Subjects With Hormone-refractory Prostate Cancer.” In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2016 Mar 2]. Available from: `https://clinicaltrials.gov/ct2/show/NCT00570700` NLM Identifier: NCT00570700.

[229] “University Health Network, Toronto. Study Of Sunitinib In Patients With Recurrent Paraganglioma/Pheochromocytoma (SNIPP).” In: ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000- [cited 2016 Mar 2]. Available from: `https://clinicaltrials.gov/ct2/show/NCT00843037` NLM Identifier: NCT00843037.

[230] A. Ravaud, C. de la Fouchardière, J. Asselineau, J.-P. Delord, C. Do Cao, P. Niccoli, P. Rodien, M. Klein, and B. Catargi, “Efficacy of sunitinib in advanced medullary thyroid carcinoma: intermediate results of phase ii thysu.,” *The oncologist*, vol. 15, no. 2, pp. 212–3; author reply 214, 2010.

[231] M. Arock, C. Akin, O. Hermine, and P. Valent, “Current treatment options in patients with mastocytosis: status in 2015 and future perspectives.,” *European journal of haematology*, vol. 94, pp. 474–490, Jun 2015.

[232] W. Q. Qiu, M. Mwamburi, L. M. Besser, H. Zhu, H. Li, M. Wallack, L. Phillips, L. Qiao, A. E. Budson, R. Stern, and N. Kowall, “Angiotensin converting enzyme inhibitors and the reduced risk of alzheimer’s disease in the absence of apolipoprotein e4 allele.,” *Journal of Alzheimer’s disease : JAD*, vol. 37, no. 2, pp. 421–428, 2013.

[233] M. Sundqvist, A. Lundahl, M. B. Någård, U. Bredberg, and P. Gennemark, “Quantifying and communicating uncertainty in preclinical human dose-prediction.,” *CPT: pharmacometrics & systems pharmacology*, vol. 4, pp. 243–254, Apr 2015.

[234] A. J. Williams and S. Ekins, “A quality alert and call for improved curation of public chemistry databases,” *Drug Discovery Today*, vol. 16, pp. 747–750, Sept. 2011.

[235] R. Khare, J. Li, and Z. Lu, “Labeledin: cataloging labeled indications for human drugs.,” *Journal of biomedical informatics*, vol. 52, pp. 448–456, Dec 2014.

[236] R. Khare, J. D. Burger, J. S. Aberdeen, D. W. Tresner-Kirsch, T. J. Corrales, L. Hirchman, and Z. Lu, “Scaling drug indication curation through crowdsourcing.,” *Database : the journal of biological databases and curation*, vol. 2015, Mar 2015.

[237] T. K. Kelly, D. D. De Carvalho, and P. A. Jones, “Epigenetic modifications as therapeutic targets,” *Nat Biotech*, vol. 28, pp. 1069–1078, Oct. 2010.

[238] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, “The Protein Data Bank,” *Nucleic Acids Research*, vol. 28, pp. 235–242, Jan. 2000.

[239] R. Edgar, M. Domrachev, and A. E. Lash, “Gene expression omnibus: Ncbi gene expression and hybridization array data repository.,” *Nucleic acids research*, vol. 30, pp. 207–210, Jan 2002.

[240] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma, “Arrayexpress update–simplifying data submissions.,” *Nucleic acids research*, vol. 43, pp. D1113–D1116, Jan 2015.

[241] ENCODE Project Consortium, “An integrated encyclopedia of dna elements in the human genome.,” *Nature*, vol. 489, pp. 57–74, Sep 2012.

[242] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, C. J. Mungall, E. Arner, J. K. Baillie, N. Bertin, H. Bono, M. de Hoon, A. D. Diehl, E. Dimont, T. C. Freeman, K. Fujieda, W. Hide, R. Kaliyaperumal, T. Katayama, T. Lassmann, T. F. Meehan, K. Nishikata, H. Ono, M. Rehli, A. Sandelin, E. A. Schultes, P. A. C. ’t Hoen, Z. Tatum, M. Thompson, T. Toyoda, D. W. Wright, C. O. Daub, M. Itoh, P. Carninci, Y. Hayashizaki, A. R. R. Forrest, H. Kawaji, and FANTOM consortium, “Gateways to the fantom5 promoter level mammalian expression atlas.,” *Genome biology*, vol. 16, p. 22, Jan 2015.

[243] “Ndr-rt api,” July 2011. NLM. [Accessed: 08-2015]. Available from: `http://rxnav.nlm.nih.gov/NdfrtAPIs.html`.

[244] "Ndf-rt release notes," July 2011. NLM. Available from: http://evs.nci.nih.gov/ftp1/NDF-RT/ReadMe.txt.

[245] "Comparative toxicogenomics database (ctd)." Curated gene-disease data were retrieved from the Comparative Toxicogenomics Database (CTD), MDI Biological Laboratory, Salisbury Cove, Maine, and NC State University, Raleigh, North Carolina. Available from: http://ctdbase.org/.