Computational Approaches to Drug Repurposing Through Probabilistic Functional Integration of Disease-Gene networks and Graph Neural Networks

Aoesha Alsobhe

School of Computing

Newcastle University

This dissertation is submitted for the degree of

Doctor of Philosophy

May 2025

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.

Aoesha Alsobhe

May 2025

Publications

1. D. J. Skelton, **A. Alsobhe**, et al., ‘Drug repurposing prediction for COVID-19 using probabilistic networks and crowdsourced curation’, 2020, doi: 10.48550/ARXIV.2005.11088.

2. K. James, **A. Alsobhe**, S. J. Cockell, A. Wipat, and M. Pocock, ‘Integration of probabilistic functional networks without an external Gold Standard’, BMC Bioinformatics, vol. 23, no. 1, p. 302, Dec. 2022, doi: 10.1186/s12859-022-04834-4.

3. **A. Alsobhe**, J. Skelton, E. Anastasi, S.J. Cockell, A. Wipat and K. James,“Application of probabilistic integration to disease-gene networks: challenges and ways forward,” IEEE CIBCB 2022 short paper, This paper based on the work presented in Chapter one.

4. **A. Alsobhe**, S.J. Cockell, A. Wipat and K. James,“ Gene-gene associations as a gold standard for probabilistic integration of disease-gene networks” IEEE CIBCB 2023 short paper, This paper based on the work presented in Chapter two.

5. **A. Alsobhe**, P. Gater, K. James, S. Cockell, and A. Wipat, ‘A computational approach to drug repurposing through graph neural networks and biomedical knowledge graphs to predict drug-disease indications’, 2024, doi: 10.13140/RG.2.2.36572.40324, This paper based on the work presented in Chapter three.

6. K. James, **A. Alsobhe**, M. Pocock, A. Wipat, Simon J. Cockell, ‘A multi-faceted gold standard reduces data loss during probabilistic integration of disease-gene associations’, the paper based on the work presented in Chapter 2 is in preparation to be submitted to the Plos.

Acknowledgements

First and foremost, I extend my deepest gratitude to my esteemed supervisors: Prof. Anil Wipat, Dr. Simon Cockell, and Dr. Katherine James. Their unwavering guidance, unyielding support, and invaluable time have been the bedrock of my research journey. Weekly meetings, where they generously shared insights, provided constructive feedback, and showered me with encouragement, propelled this thesis forward. Their incredible support is the lifeline that breathes reality into this work, and I count myself fortunate to have them as my supervisors.

A tip of the hat to the phenomenal team led by Anil Wipat, whose constant support served as a compass during the twists and turns of my academic odyssey: Dr. James Skelton, Dr. Matthew Pocock, Dr. Elisa Anastasi, and Phoenix Gater.

Special appreciation is reserved for my cherished family, my pillars of strength. To my mother, father, sisters, and brothers, your unwavering encouragement propelled me forward.

A heartfelt thank you to my beloved children, Jana, Juwan, and Sarah, whose love, warm hugs, and daily kisses infuse life with meaning. Without them, the journey wouldn't be as vibrant.

Lastly, gratitude extends to the Saudi Arabian government for their generous funding and incredible financial support throughout my Ph.D. journey.

Abstract

Drug discovery is a time-consuming, costly, high-risk, and complex process. An alternative to traditional drug development is drug repurposing, which aims to find new uses for existing drugs. This approach significantly reduces time and cost, as much of the safety evaluation has already been completed. Computational approaches to drug repurposing help generate hypotheses about potential drug-disease indications, which can later be validated experimentally in the lab.

Network integration is a common computational technique in drug repurposing applications. These approaches combine multiple diverse data sources into a single heterogeneous biomedical integrated network. Such networks combine various types of biological data, including drugs, diseases, genes, and proteins, into a unified framework where biomedical entities are represented as nodes and their interactions as edges. Integrating diverse data sources is essential to gain a comprehensive picture of interconnected biological entities, which can then be mined to infer new hypotheses about drug repurposing opportunities.

The quality of these integrated networks is highly dependent on the experimental data they include. However, biomedical data is often noisy and incomplete, leading to a high rate of false results in existing networks. Therefore, there is an important need for methods to reduce noise during network integration. One proposed technique to produce accurate integrated networks is Probabilistic Functional Integrated Networks (PFINs), which assess data quality and generate confidence scores to filter out low-quality data before mining these networks for drug repurposing opportunities.

Disease-Gene Association (DGA) networks, where nodes represent diseases and genes and edges represent their associations, are the major building blocks for most biomedical integrated networks used in drug repurposing applications. Unfortunately, many available DGA networks contain a high rate of false results due to the quality of the biomedical data, which faces numerous challenges, including incorrect entries, missing values, inconsistencies, duplication, and various forms of bias. For instance, high-throughput experimental studies, which are commonly used to generate biological data, often produce incomplete and noisy data containing both false positives and false negatives. Although methods exist to score the confidence of DGAs, they are often unreliable. Many of these

scoring approaches rely on heuristic strategies that do not assess data quality prior to integration. For example, they often overlook the impact of duplicated data, which can artificially inflate confidence scores and distort the strength of associations. To address this gap, we investigated the applicability of PFINs to DGA networks by researching and developing novel strategies to build and evaluate DGA PFINs.

These accurate integrated DGA networks can be employed in various computational drug repurposing applications, including deep learning techniques. Deep learning has become the leading technique in most *in silico* applications for drug repurposing. Among deep learning methods, Graph Neural Networks (GNNs) have gained considerable attention due to their ability to learn complex relationships between drugs and related biological entities from heterogeneous biomedical integrated networks. Existing GNN applications in drug repurposing often overlook important aspects of data quality, such as noise and incompleteness. Given that the performance of GNNs is highly dependent on the quality of the integrated networks used for training, incorporating PFINs with GNNs could enhance their performance by reducing noise during network integration. To address these issues, we investigated the impact of incorporating the PFINs approach within GNNs on their performance. The constructed DGA PFIN was integrated with an existing network and used to train GNN models.

Another factor impacting the performance of GNNs, beyond data quality, is the lack of diverse data types in the integrated networks. Most existing GNN approaches are trained on networks with a limited number of node and edge types, often ignoring node features in the training process. We explored the impact of adding various types of nodes and edges to the integrated networks on GNN performance, as well as incorporating node features in the training process. The results showed that the performance of GNN models improved by incorporating these additional types of nodes and edges into the training networks. Furthermore, the proposed GNN models demonstrated significant enhancement by incorporating node features. Finally, the proposed GNN models were employed to predict drug-disease indications, and these predictions were validated and supported by the literature.

Table of Content

List of Figures

List of Tables

Acronyms

PFIN Probabilistic Functional Integrated Network

BN Bayesian network

DGA Network Disease-Gene Association Network

PPI Network Protein-Protein Interaction Network

DTI Drug-Target Interaction Network

DDI Drug-Drug Interaction Network

DrDI Drug-Disease Indication Network

DPD drug-protein-disease network

HDN Human Genetic Network

GNN Graph Neural Network

CNN Convolutional Neural Network

GCNN Graph convolutional Neural Network

GAT Graph Attention Network

GCMM Graph Convolution Network based on Multimodal Attention Mechanism

GraphSage Graph Sample and Aggregate

DNN Deep Neural Network

HTP High-throughput Studies

LTP Low-throughput Studies

HC High-confidence data

LC Low-confidence data

DSN Disease Similarity Network

GRN Gene Regulatory Network

MCL Markov Clustering Algorithm

LAP Label Propagation Algorithm

MCODE Molecular Complex Detection

MF Molecular Function

BP Biological Process

PCR Polymerase Chain Reaction

CC Cellular Component

AD Alzheimer's disease

CPCC Cophenetic Correlation Coefficient

CN Common Neighbors

JI Jaccard Index

AI Adar Index

PA Preferential Attachment

SNP Single Nucleotide Polymorphisms

NGS Next Generation Sequencing

LLS Log-Likelihood Score

WS Weighted Sum

BN Bayesian network

RNA-seq RNA sequencing

SVM Support Vector Machine

RF Random Forest

MoA Mode of Action

PheWAS Phenome-Wide Association Study

GWAS Genome-Wide Association Studies

WGS Whole-Genome Sequencing

HDN Human Genetic Network

DO Disease ontology

GO Gene Ontology

UMLS Unified Medical Language System

NCBI The National Centre for Biotechnology Information

PMID PubMed Identifier

EL Evidence Level

SSA Studies with Single Associations

SMA Studies with Multiple Associations

HEL High Evidence-Level Studies

LEL Low Evidence-Level Studies

MCS Multi-Curated Studies

SCS Single-Curated Studies

MG Monogenic Experimental Studies

IES Individual Experimental Studies

CDS Curated Data Sources

UniProt_CDS UniProt Scored CDS Network

OMIM_CDS OMIM Scored CDS Network

SSA_SMA Network SSA Scored SMA Network

HEL_LEL Network HEL Scored LEL Network

MCS_SCS Network MCS Scored SCS Network

MG_IES MG Scored IES Network

OMIM_IES OMIM Scored IES Network

BioGRID_IES BioGRID Scored Individual Experimental Studies Network

Reactome_IES Reactompathway Scored Individual Experimental Studies Network

IntAct_IES IntAct Scored Individual Experimental Studies Network

GGA PFINs Gene-Gene Association PFINs

DDA PFINs Disease-Disease Association PFINs

GGA_SSA_SMA Collapsed Gene-Gene Association SSA Scored SMA Network

GGA_HEL_LEL Collapsed Gene-Gene Association HEL Scored LEL Network

GGA_MCS_SCS Collapsed Gene-Gene Association MCS Scored SCS Network

GGA_MG_IES Collapsed Gene-Gene Association MG Scored IES Network

GGA_OMIM_IES Collapsed Gene-Gene Association OMIM Scored IES Network

GGA_BioGRID_IES Collapsed Gene-Gene Association BioGRID Scored IES Network

GGA_Reactome_IES Collapsed Gene-Gene Association Reactome Scored IES Network

GGA_IntAct_IES Collapsed Gene-Gene Association IntAct Scored IES Network

DDA_SSA_SMA Collapsed Disease-Disease Association SSA Scored SMA Network

DDA_HEL_LEL Collapsed Gene-Gene Association HEL Scored LEL Network

DDA_MCS_SCS Collapsed Disease-Disease Association MCS Scored SCS Network

DDA_MG_IES Collapsed Disease-Disease Association MG Scored IES Network

DDA_OMIM_IES Collapsed Disease-Disease Association OMIM Scored IES Network

DDA_BioGRID_IES Collapsed Disease-Disease Association BioGRID Scored IES Network

DDA_Reactome_IES Collapsed Disease-DiseaseAssociation Reactome Scored IES Network

DDA_IntAct_IES Collapsed Disease-Disease Association IntAct Scored IES Network

HBKG Heterogeneous Biomedical Knowledge Graph

AUROC Receiver Operator Characteristics Curve

AUPTC Area Under the Precision-Recall Curve

Chapter One

Introduction

1.1 Motivation

Over the past 30 years, new drugs have been released to treat conditions such as heart diseases, brain diseases, and many infections [1]. These drugs have helped improve and extend many patient lives worldwide. However, some new medications offer no distinct advantage in many cases over preexisting medications. Nearly one-third of new drugs are no better than older drugs, and some are worse [2]. Existing therapies, also known as "me-too" drugs, are ways for drug companies to establish market share for treatments for a particular disease [3]. Recently, older drugs have been gaining attention from drug companies, and have been used as treatments for new diseases. Many pharmaceutical companies are developing new drugs with the discovery of novel biological targets by applying drug repositioning; a process of identifying new therapeutic uses for existing drugs [3]. Drug repurposing can fulfil medical needs such as treatments for diseases that are rare and neglected since it has the potential to provide more effective treatment and cheaper alternative drugs in a short time than can be achieved using traditional drug discovery [4], [5], [6]. Since safety evaluations for repurposed drugs have already been completed during their initial development, the need for extensive time and cost for preclinical and early clinical trials is eliminated [6].

In silico drug repurposing methods, also known as computational methods, have accelerated the drug repurposing process by generating potential hypotheses about drug repurposing [7]. Computational approaches have gained considerable attention from researchers for two main reasons [8], [9]. First, with the revolution in high-throughput techniques, a massive amount of biological and biomedicine data has become available and is stored in numerous repositories and databases to allow access, sharing, and analysis, allowing for the inference of new hypotheses [10]. The second reason is computing power; rapid advancements in computational techniques and data sciences, including data mining [11], [12], machine learning techniques [13], and network integration [14], enable systematic analysis of the relationships between different biomedical entities, including those between drugs, diseases, genes, and proteins to identify potential novel indications for existing drugs [13], [15], [16], [17]. All of these factors have empowered computational approaches for drug repurposing.

Recently, computational approaches to drug repurposing have been emerging including network integration [14], machine learning [16], and deep learning [18]. However, the success of these techniques is highly dependent on data quality [19], [20], [21]. Studies have shown that data quality dramatically impacts the results of computational approaches [19], [21], [22], [23]. It has been proven how vital data quality is to the outcomes of machine learning techniques and how severely they are affected by low-quality data [19], [20], [21]. As a well-recognized aphorism in deep learning, the outcome is highly dependent on the nature of the input data; hence, the adage "garbage in, garbage out" holds significant relevance in the context of deep learning endeavours [24]. However, most applications of drug repurposing begin with the assumption that the data feeding the computational algorithms are highly accurate, consistent, complete, not duplicated, and not biased [12], [14], [17], [25], [26], [27]. However, this assumption is not always correct [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38].

Biomedical data suffers from several challenges related to data quality, including incorrect data, missing data, inconsistency, duplication, and bias [29], [31], [32], [33], [34], [35], [36], [37], [39]. Incorrect data may result from errors that occur during experiments that generate the data [36], [40]. For instance, biological data generated by high-throughput experimental studies, is incomplete and noisy [32], [33], [41], [42]. It has been estimated that high-throughput data (HTP) have high rates of false results [31], [32], [33]. False results can be either false positives, associations that are identified but do not exist, or false negatives, true associations that have not been identified yet. Duplicate data, in which identical entries are present multiple times within the dataset or across multiple datasets, is a common issue in many biological data sources [34], [35]. Duplicate data poses a significant issue, as it can skew analyses and lead to inaccurate results [35]. Missing data is also a common occurrence in biological datasets, stemming from factors such as incomplete recording or relying on a single dataset [23], [39], [43]. Missing data can reduce the effectiveness of computational approaches, and compromise the reliability of analyses [23]. Another issue of data quality is inconsistency in data which refers to contradictions within a dataset or between multiple datasets [34], [36], [44]. Additionally, different experimental techniques introduce unique biases, further complicating the accuracy and reliability of data analysis [37], [38]. Bias often comes from flaws in study design, measurement, or data collection methods, leading to inaccurate or misleading conclusions [37], [38]. Poor data quality compounds the challenge

of ensuring data reliability and integrity. Addressing these challenges is important for computational approaches to yield accurate analyses of biomedical data and reliable results.

Network integration is one of the most common computational approaches in drug repurposing [12], [14], [25], [26], [45], [46], [47], [48], [49], [50]. These networks combine multiple and diverse data types into a single unified network where nodes represent biological entities such as genes, diseases, drugs, and proteins, while edges represent biological relationships such as protein-protein interactions (PPIs) [51], disease-gene associations (DGAs) [52], drug-drug interactions (DDIs) [53], drug-disease indications (DrDIs) [54], and drug-target interactions (DTIs) [14]. Integrating multiple and diverse data sources is essential for inferring reliable instances of drug repurposing and understanding the complex interconnections between components of biological systems. These integrated networks can be analysed and mined using computational approaches, such as deep learning techniques, to generate new insights into drug repurposing. The success of network analysis and mining depends heavily on the quality of these networks. The accuracy of integrated networks is directly related to the quality of the data used to construct them. As mentioned previously, biological data often contain high rates of false results. Therefore, there is a critical need for approaches that can evaluate and ensure data quality during network integration.

One of the most powerful integration techniques to filter or at least reduce the noise during network integration is the use of Probabilistic Functional Integrated Networks (PFINs) [55], [56], [57], [58]. Using a high-quality gold standard dataset as a benchmark, this approach scores the datasets prior to integration to determine their level of confidence. Multiple pieces of evidence can then be combined into a network by integrating datasets according to their confidence levels. In PFINs, nodes represent biological entities such as genes, proteins, diseases or drugs, and the edges represent functional associations between nodes, indicating the confidence level of that edge. PFINs offer a robust solution to address various challenges associated with data quality, including duplicates, missing entries, inconsistency and inaccuracies. Firstly, PFINs ensure dataset independence, eliminating duplicate data and enhancing the reliability of integrated networks. Secondly, PFINs improve data completeness by integrating several datasets to make up for missing data [43]. If data is missing in one source, it may be found in another, allowing a more comprehensive coverage of data. Relying on a single data source may lead to gaps in coverage, particularly in biomedical research, in

which each data source may have its own unique focus. Thirdly, PFINs address data inconsistency by cross-validating data entries from several sources of evidence, increasing the confidence of entries that are supported and decreasing dependence on those that lack strong evidence [36], [44]. Fourthly, by comparing datasets to a high-quality gold standard, inaccurate data is found and eliminated from integration. Researchers have shown that PFINs can lower noise levels during integration [55]. Finally, PFINs can also reduce bias during network integration by integrating multiple and diverse sources of evidence [58]. Studies have shown that PFIN can reduce bias during integration [58], [59]. By leveraging the PFIN techniques, comprehensive and reliable biological networks can be developed, facilitating more accurate analyses and results.

DGA networks combine many DGA-related data sources into a single, integrated network, with genes and diseases acting as nodes and their associations as edges [60], [61], [62], [63]. By integrating various datasets into one network, researchers can uncover intricate relationships between diseases and genes that may not be apparent when analysing individual datasets separately [60], [61], [63]. This integration facilitates a holistic view of disease mechanisms and genetic factors, leading to insights that can inform drug discovery [27]. DGA networks are regarded as a foundational component for drug repurposing applications, often serving as a fundamental building block of most of the available heterogeneous biological integrated networks for drug repurposing [25], [27], [48], [64]. However, existing DGA networks exhibit varying levels of false positives, due to the inclusion of noisy data types such as HTP experimental data [31], [32] and text mining data. Therefore, there is a need for a technique to reduce noise during the process of network integration. PFINs have shown good results in reducing noise during network integration [55]. Despite the success of the PFIN approach in the applications of PPI networks, there exists a notable gap in its application to DGA networks. Exploring the application of the PFIN approach in the context of DGA networks could contribute to advancing drug repurposing applications by improving our understanding of the molecular basis of diseases.

In computational methods to drug repurposing, deep learning has been a prominent tool [18], [65], [66]. The potential of GNNs to exploit relational information included in complicated biomedical data represented as networks has drawn increasing interest among various deep learning approaches [25], [26], [64], [67], [68], [69], [70], [71]. GNNs are types of neural networks designed to operate on network-structured data where data is represented in the

form of a network [72], [73], [74], with nodes representing biological entities and edges representing biological relationships such as PPI networks [75], DrDI networks [54], DGA networks [63], or DTI networks [70]. The initial phase of GNNs often involves network integration techniques, where various biological networks such as PPI, DTI, and DGA networks are integrated into one integrated heterogeneous network [25], [48], [64], [67]. These integrated networks are employed for training GNNs to learn node representation to predict reliable drug targets or drug indications. Ensuring the accuracy of these integrated networks is important since they serve as inputs for GNNs. The accuracy of GNN outputs and predictions depends on the quality of these networks [21], [22]. The quality of integrated networks depends on the quality of individual datasets used in their construction. Given the prevalent issue of a high rate of false results in biological data, there is a necessity for an approach aimed at either removing or at least reducing noise during the process of network integration. An accurate biological integrated network can be constructed by evaluating the quality of each dataset before integration since datasets are different in their reliability [55]. However, existing GNN approaches present some limitations. Current GNNs operate on limited integrated networks that often contain missing data and have a restricted range of biological entities and relationships. Additionally, many of these networks suffer from a high rate of false results, particularly those constructed from high-throughput data. Training GNNs on extended biological networks, which include a broader variety of biological interactions, could improve their performance. Moreover, incorporating GNNs with PFINs could further improve GNN performance, as PFINs help reduce noise in the integrated networks used to train GNNs.

1.2 Project Aim and Objectives

This project aimed to research and develop computational techniques to predict novel uses or indications for existing drugs, exploring the use of the PFIN approach to enhance data quality, especially for disease gene networks, and finally incorporating a GNN based approach to predict new drug repurposing opportunities.

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

1. To research and develop the PFIN approach within the domain of DGA networks.

2. To research and develop network evaluation techniques to evaluate the performance of the resulting DGA PFIN.
3. To research and develop a GNN model and train the model on a recently developed biomedical knowledge graph for drug repurposing.
4. To research and develop GNN evaluation techniques to validate the developed model
5. To incorporate the DGA PFIN with the GNN model to enhance the performance of the GNN model.
6. To employ the validated model to predict novel DrDIs and subsequently validate these predictions by cross-referencing with existing biomedical literature and knowledge.

1.3 Contribution of the Work Presented

A. The first part of this work, investigating novel approaches to apply the PFIN approach to DGA networks, contributes:

    1. Identification of a novel set of strategies to define high-quality gold standards and individual datasets representing DGAs.

    2. Investigation using non-DGA gold standards such as PPI and pathway data to score DGA data.

    3. Developing a novel text-mining approach to define individual datasets representing DGAs.

    4. Construction of different DGA PFINs based on different strategies of individual datasets and gold standards definition.

    5. Identification of novel network analysis techniques to evaluate the performance of the constructed DGA PFINs.

B. The second part of this work, investigating a novel deep learning approach to predict drug-disease indications, contributes:

    1. Constructing a Heterogenous Biomedical Knowledge Graph (HBKG-integrated network) from a newly developed database; the NeDRex database.

2. Incorporating the DGA PFIN, developed in A, within the constructed biomedical knowledge graph.

3. Developing and training a novel GNN model on the constructed biomedical knowledge graph.

4. Validating the developed GNN model and employing it to predict novel drug-disease indications.

5. Setting and testing a set of hypotheses to enhance the developed GNN model.

1.4 Thesis structure

The remainder of this thesis is divided into the following chapters:

- Chapter 2 offers a comprehensive review of foundational concepts important to this thesis. Initially, it delves into graph theory and its applications in biological networks. Following this, the chapter explores network integration, elucidating Probabilistic Functional Integrated Networks and their applications in PPI networks. Next, it discusses DGA identification, including both experimental and computational methods, along with their applications in drug repurposing. Additionally, the chapter outlines GNNs and their applications in the domain of drug repurposing. Finally, it provides insights into drug repurposing strategies and various computational approaches employed in this field.

- Chapter 3 describes the biological datasets and computational methods utilised and developed in this project. The chapter describes the computational techniques used and developed to evaluate the outputs.

- Chapter 4 introduces the investigation of the applicability of PFINs to DGAs, exploring the identification of the main components required for constructing PFINs in the DGA context. The chapter introduces novel strategies for the identification of the gold standards and the individual datasets representing DGAs to build DGA PFINs. A detailed evaluation of the DGAs PFINs is then presented.

- Chapter 5 presents solutions to address the limitations outlined in Chapter 4, including challenges such as refining the identification of gold standards, improving the identification of individual datasets, and enhancing the evaluation techniques to assess the performance of the DGA PFINs. This chapter also introduces a novel text-mining approach to the identification of individual datasets representing DGAs.

- Chapter 6 introduces a novel approach to drug repurposing through the development of a GNN model trained on a recently constructed biomedical knowledge graph, rich with node and edge types as well as node and edge features, aimed at predicting potential drug repurposing opportunities. The chapter outlines the evaluation of this model and its application in predicting novel links between drugs and diseases. Furthermore, it presents a series of hypotheses and their testing to enhance the developed GNN model, including the incorporation of the PFINs, developed in Chapter 5, within the GNNs to enhance the GNN model performance.

- Chapter 7 discusses the implications of this work and suggests areas for future extension and analysis.

Chapter 2

Background

2.1 Introduction

The aim of this project was to develop computational approaches for drug repurposing by integrating PFINs within GNNs focusing on DGAs. In this chapter, four main topics that form the foundation of this project are reviewed. First, PFINs are explored, including how PFINs are constructed, their benefits in data quality issues, and their applications in PPIs. Second, DGAs, their computational and experimental techniques, and their benefits in drug repurposing strategies are discussed. Third, GNNs and their applications in drug repurposing are presented. Finally, drug repurposing strategies and their computational and experimental approaches are presented.

2.1.1 Probabilistic Functional Integrated Networks

Data integration is essential in fields such as systems biology [76], [77], e-commerce, retail [78], healthcare [79], and environmental science [80]. A holistic view of the mechanisms and patterns can be obtained by integrating data from multiple diverse datasets, an approach which can also facilitate hypothesis generation. The success of data integration approaches is highly dependent on the quality of the individual datasets used to generate the data to be integrated [81]. However, biomedical datasets, specifically DGA datasets, differ in terms of data quality and coverage [31], [33], [42], [82]. The accuracy and comprehensiveness of different biological datasets can be influenced by the experimental techniques being used. Therefore, any biological dataset may include false or incomplete data, a situation which emphasises the importance of addressing issues related to data quality and completeness. To address these challenges, various methods have been introduced to evaluate data quality and filter out noise [83], [84]. One common approach is the use of a scoring method that involves comparing the dataset with a high-quality gold standard dataset [50], [55], [56], [85], [86]. This comparison allows for the identification of potential noise, facilitating its reduction. Subsequently, the scored datasets, which potentially contain distinct information, are integrated to improve the overall comprehensiveness of the data.

The gold standard dataset contains associations that are considered to be biologically accurate and reliable. These associations can be derived from human expert-curated and peer reviewed databases such as the Biological General Repository for Interaction Datasets (BIOGRID) [87] for PPIs and Online Mendelian Inheritance in Man (OMIM) [88] for DGAs. Gold standards can be also constructed by integrating multiple, diverse curated data sources [89]. Confidence scores can be generated by comparing datasets against the gold standard data using statistical algorithms. Subsequently, these confidence scores from multiple scored datasets can be combined using methods such as weighted sums [55]. The result of this integration is a unified network known as a Probabilistic Functional Integrated Network (PFIN) [50], [55], [85]. In such networks, the edges are labelled with edge weights, which reflect the degree of confidence associated with each edge (For more details about PFINs see Section 2.3.3). Although PFIN approaches have been applied in PPIs, and have demonstrated success in many applications such as protein function predictions and PPI predictions [50], [55], [56], [57], [90], [91], [92], [93], [94], [95], their applications to DGAs remains limited.

2.1.2 Disease-Gene Associations

A disease-gene association (DGA) is the relationship between a gene and a particular disease. Defining DGAs involves identifying which genes are associated with the development of, progression of, or susceptibility to a disease [96]. Identifying DGAs is an important activity in biomedical research and industry, and enables a deeper understanding of disease mechanisms which facilitate drug repurposing applications [27], [96], [97], [98]. Current experimental methods to identify DGA include linkage studies [99], genome-wide association studies (GWAS) [100], functional assays [101], RNA interference screens [102], and animal models [103]. These experimental methods are resource-intensive in terms of time and cost.

Computational approaches to identify DGAs, such as network integration [61], [97], [104], machine learning [105], [106], [107], and text mining [108], [109], have lower expenses than experimental methods (Section 2.4.1). Computational methods typically involve the integration and analysis of experimental data to predict novel DGAs [110]. Experimental data

are often stored in databases that are designed to store information about genes, diseases, and the associations between them [111]. Some groups of related diseases or disorders are the subject of dedicated databases that focus exclusively on cataloguing DGAs related to those specific conditions. For instance, The Cancer Genome Interpreter (CGI) [112] is dedicated to cancer-related DGAs, Orphanet specialises in orphan diseases, which are rare conditions affecting a small percentage of the population and often lack sufficient medical research and treatment options [113], and PSYGENET focuses on psychiatric disorders [114] (For more information on DGAs databases, see Section 2.5).

Network integration has emerged as a common computational approach for DGA identifications, aiming to combine multiple diverse and multiple datasets to draw meaningful inferences [27], [61], [108]. Integrated DGA networks can be unweighted networks, where associations are integrated without considering the strength of supporting evidence [49], [97], [115], or heuristic weighted networks that incorporate the number of experimental evidence supporting DGAs but overlooking the quality assessment of these evidence [116], which may lead to noisy integrated networks. Therefore, in this work the PFIN approach is investigated for the integration of DGA data (Section 4.3).

2.1.3 Drug Repurposing

One of the domains that has benefited most from the improvement in DGA identification is drug repurposing, the process of finding novel therapeutic applications for existing drugs that were initially formulated for different medical purposes [27], [98]. There are many examples of drugs that have been repurposed. For example, aspirin was originally used as an analgesic and antipyretic to relieve pain and reduce fever [117]. Later, it was found to have antiplatelet and anticoagulant properties [118] and is now widely used for its cardioprotective effects in preventing heart attacks and strokes [119]. Another example is thalidomide, which was originally prescribed as a sedative and anti-nausea medication [120]. Later, thalidomide was discovered to possess anti-inflammatory effects [121], [122] and is now used for managing multiple myeloma and leprosy [122]. Drug repurposing has gained popularity due to its potential benefits such as reduced development costs and faster timelines for bringing drugs

to market [6]. Computational approaches to drug repurposing have been widely applied due to their lower costs and time when compared to experimental approaches [14], [27], [50], [68] (Computational approaches to drug repurposing are discussed in detail in Section 2.7.1).

2.1.4 Graph Modelling

Graph representations of complex systems, such as social networks and biological networks, are commonly used in computer science and bioinformatics. Biomedical data can be effectively represented and integrated as networks, enabling analysis of their underlying structure [72]. In these networks, biomedical entities such as genes, proteins, drugs, and diseases are depicted as nodes, while the biomedical relationships between these entities, such as interactions or associations, are depicted as edges [72], [123]. There are different types of networks used for specific purposes. For instance, simple networks like PPI networks [51] have only one type of node. In contrast, some biological networks include two types of nodes and are illustrated as bipartite graphs, such as DGAs networks [115], DTI networks [124], and DrDI networks [54]. More complex structures, known as multipartite networks, involve multiple biological entities, like diseases, drugs, proteins, and genes, to yield more precise and accurate outcomes [125], [126]. Graph modelling in biological data is discussed in Section 2.2, and is used to produce the networks in chapters 4, 5, and 6.

2.1.5 Graph Neural Networks

A neural network is an artificial intelligence model that draws inspiration from the structure and function of the human brain. It includes layers of interconnected artificial neurons that process data and learn from it through training [66]. A GNN is a specialised type of neural network tailored to process data presented in the form of graphs [127]. The fundamental concept behind GNNs involves learning node representations, or node embeddings, by gathering information from neighbouring nodes within the graph. This aggregation process typically occurs through multiple layers of neural networks, enabling the model to capture both local and global patterns in a graph. GNNs models learn meaningful representations for each entity through multiple convolutions, and capture complex patterns within the graph. The combination of graph modelling and deep learning techniques has empowered GNN approaches to biological data [26], [107], [127], [128], [129], [130]. In the field of drug

repurposing, GNNs are increasingly adopted to capitalise on the intricate relationships among biological entities like drugs, proteins, genes, diseases, and molecular pathways [25], [26], [68], [69], [70], [71], [130], [131], and studies have demonstrated that GNNs outperform conventional neural networks [132], [133]. GNNs and their use for biological network data are described in detail in Section 2.6.1

2.2 Graph Modelling

Biomedical data can be represented as a graph [134]. A graph is a powerful way to represent and analyse complex relationships and structures in various biomedical applications [74]. In mathematics and computer science, a graph is a data structure that includes a collection of nodes and a collection of edges that connect pairs of nodes. Nodes represent entities of interest while edges represent various kinds of associations between nodes [123]. For example, in a biological network, nodes might represent biological entities such as drugs, genes, proteins, or diseases, and edges might represent interactions or associations between these entities [74]. In this thesis, the terms "network" and "graph" are used interchangeably to refer to the same concept.

A simple graph G can be represented as:

$$G = (V, E) \tag{2.1}$$

Where G is the graph, $V = \{v_1, .., v_n\}$ is the set of vertices and $E \subseteq \{(v_i, v_j) | v_i, v_j \in V\}$ is the set of edges.

2.2.1 Application of Graph Modelling to Biological Networks

Graph modelling is a powerful approach in the field of biological networks, where it is used to represent and study complex relationships within biological systems [74]. Biological networks capture interactions between various biological entities, such as genes, proteins, diseases, drugs, and pathways. For example, in a unipartite undirected network such as a PPI network, nodes belong to a single set (*proteins*), and edges indicate physical interactions [135]. PPI networks can be weighted, assigning a numerical value to each edge to provide

additional information about the confidence of interactions between proteins [136]. Disease Similarity Networks (DSNs) are another example of a unipartite network in which nodes belong to a single set (diseases) and edges indicate similarity between diseases [62]. DSNs can be weighted where edge weight represents the similarity score between diseases [137]. The similarity scores can be calculated based on the number of shared drugs, shared genes, or comorbidity scores, which quantifies the presence and severity of multiple medical conditions co-occurring in an individual [138].

Another type of biological network is an undirected bipartite network such as a DGA network [63], in which nodes represent two types of entities-*diseases* and *genes*- and the edges represent the associations between these entities. These networks can help in understanding the genetic basis of diseases, identifying potential disease genes [63], [139], [140]. DGA networks can be unweighted networks, in which the edges between diseases and genes are binary, meaning that they represent the presence or absence of an association, regardless of the confidence score of the association [140]. In weighted DGA networks numerical values are assigned to the edges to reflect the confidence scores of the associations between diseases and genes [141].

Biological networks can also be directed. For example, Gene Regulatory Networks (GRNs) model the interactions between genes, where genes are represented as nodes, and directed edges represent the regulatory relationships between them [142]. If Gene A regulates the expression of Gene B, there is a directed edge from Gene A to Gene B. This directionality indicates that Gene A influences the expression of Gene B. Another example of a directed biological network is a metabolic network, which represents the biochemical reactions in a cell with nodes representing metabolites and enzymes. Directed edges indicate the direction of metabolic reactions, reflecting the relationships between enzymes and the metabolites they catalyse. These networks can have edges between nodes of the same type, such as enzymes connecting to enzymes or metabolites connecting to metabolites [143].

A more complex type of biological network is a multipartite network, a heterogeneous network containing multiple types of nodes and edges [25]. For example, a network may have *drugs, proteins, genes, and diseases*, as nodes, with various types of relationships between

them including *drug-target-protein, gene-encoding-protein, disease-gene associations, and drug-disease indications, protein-protein interactions, drug-drug similarity, disease-disease similarity.* Figure 2.1 shows diverse categories of biological networks.

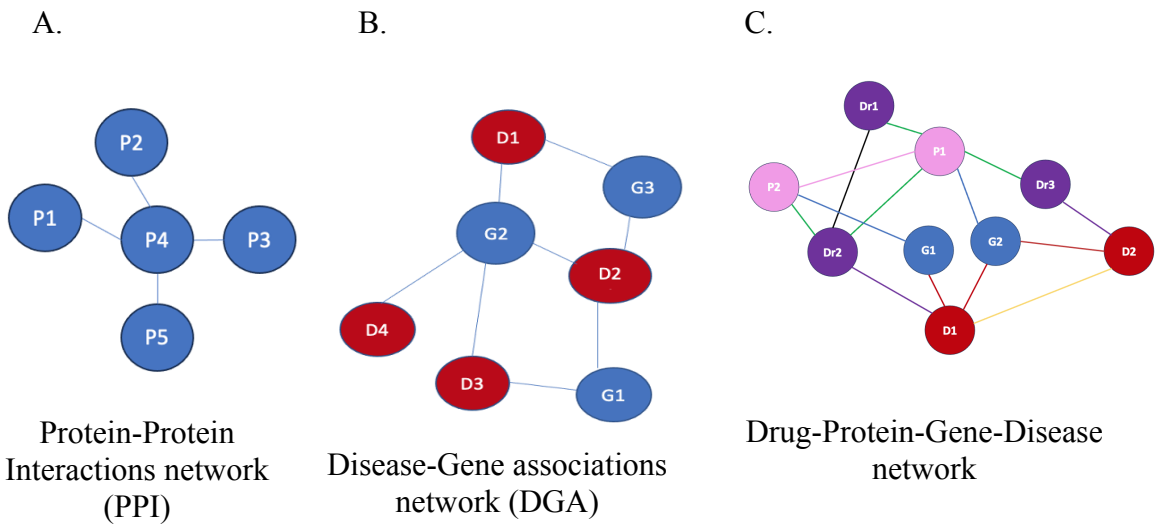


Figure 2.1 A variety of biological networks. A. A PPI network with uniform nodes and edges. B. A DGA network showing red disease nodes and blue gene nodes. C. A multipartite network with diseases (red), genes (blue), drugs (purple), and proteins (pink). Edges represent drug-target (green), gene-protein (blue), disease-gene (red), drug-disease (purple), protein-protein (pink), drug similarity (black), and disease similarity (yellow).

2.3 Network Integration

The vast amount of biological data has increased the use of data integration approaches [76], [77], [144] particularly for understanding cellular processes and molecular interactions [144]. A number of data integration methods have been introduced to enhance the reliability of new findings, and one of the most common is network integration [134], [145]. Network integration involves combining multiple diverse data sources into a unified network to gain a comprehensive understanding of the connections between biological components [14], [61], [104]. There are many data sources for DGAs [111]. However, a single data source cannot completely cover all DGAs since each data source provides information about DGAs relevant to its motivation and purpose. For example, several databases are dedicated to specific diseases, such as the Cystic Fibrosis Mutation Database[1], the Breast Cancer Information Core (BIC)[2] [146], and the Alzheimer's Disease Biomarkers Comprehensive Database (ABCD)[3]

[1] http://www.genet.sickkids.on.ca/
[2] https://research.nhgri.nih.gov/bic/
[3] http://www.bioinfoindia.org/abcd/

[147]. In addition, experimental studies in DGAs concentrate on a particular disease or a group of related diseases to delve deeply into genetic factors, disease mechanisms, and potential treatments. Nevertheless, the components of living systems exhibit extensive interconnections, and diseases can have direct and indirect impacts on each other [148], [149]. For instance, many individuals concurrently experience multiple diseases such as a person with both diabetes and heart disease. Diabetes and heart disease share common pathophysiological mechanisms. For example, individuals with diabetes often experience insulin resistance which can contribute to the development and progression of heart diseases [150]. Additionally, both diabetes and heart disease have shared risk factors, including high blood pressure, which can exacerbate each other [151]. Recognizing the relationships between diseases is instrumental in managing comorbidity more effectively [152]. Some diseases share common genetic and molecular pathways [148]. Investigating these shared elements can enhance our insight into disease mechanisms and potentially lead to the design of therapies that target multiple diseases. Identifying common pathways and genes allows researchers to uncover shared biological processes underlying different diseases [153]. For instance, mutations in the BRCA1 and BRCA2 genes are linked to elevated risk of developing breast and ovarian cancers [154]. Researchers have found that these genes are involved in repairing damaged DNA and maintaining genomic stability, processes important for preventing cancer development [155]. This interconnectedness is a fundamental principle in systems biology and personalised medicine. The integration of multi-omic data in DGAs offers the capacity to amalgamate disparate lines of evidence pertaining to DGAs, thereby enhancing the confidence of these associations. For instance, a DGA derived from a combination of Genome-Wide Associations Studies (GWAS), RNA sequencing, clinical medical records, and data from animal models provides a multifaceted, heterogeneous evidentiary basis, thereby reinforcing the validity of the respective DGA. Integrating DGAs from various data sources can reduce noise by reinforcing high confidence associations from multiple sources while diminishing low confidence associations present in just one source. The integration of these data sources can improve low-confidence associations that may not be apparent when looking at each data source individually [90]. Integrated network analyses have demonstrated enhanced accuracy across various applications. For instance, identification of DGAs, prediction of DTI, and prediction of protein functions [27], [85], [90].

Various methods exist for integrating DGAs, one of which involves building a network from datasets, depicting genes and diseases using nodes and using the edges of the network to capture the relationships between these nodes [12], [27], [45], [156]. However, in this approach, the strength or reliability of the evidence supporting these associations is not considered, and the final integrated network is unweighted (Figure 2.2.A).

An alternative network integration method based on heuristic methods can also be utilised, in which the weight of edges can show how many lines of evidence support each association [116], [136]. Associations backed by multiple evidence sources are considered to be stronger than those with just one source, and this weight indicates the confidence in the edge (Figure 2.2.B). Since weighted integrated networks can be more informative than unweighted networks [157], [158], this approach to network integration can aid in improving network performance. Various heuristic methods can be used to generate edge weights in DGA networks such as frequency, co-occurrence, and similarity measures [60], [116], [159], [160]. Frequency methods can be used to generate the edge weight based on the frequency of observed DGAs in the literature or in data sources [116], [160]. DGAs that are frequently mentioned are assigned higher edge weights. Co-occurrence methods can be used, which apply text mining techniques to scientific literature to measure the frequency of co-occurrence of DGA entities within the text corpus can also be used [159]. Similarity measures can also be used to assess the confidence of DGAs incorporating other data types such as GO-based similarity or PPIs [60].

However, these heuristic approaches for constructing weighted integrated networks do not take into account the quality of data prior to integration. For instance, DisGeNET generates DGA confidence scores based on the frequency of DGAs in curated data sources, animal models sources and literature data sources, ignoring duplicate data between these data sources which upweighted the scores (discussed in more detail in Section 2.4.1.2). Given the prevalence of false associations in biological data [31], [33], especially from high-throughput techniques, including poor-quality data in an integration can lead to a network with a lot of noise [81]. An accurate biological integrated network can be built by considering the quality of each dataset before integration, since datasets differ in their reliability. Consequently, more sophisticated integration techniques are needed to eliminate, or at least significantly decrease,

false associations during the network integration.

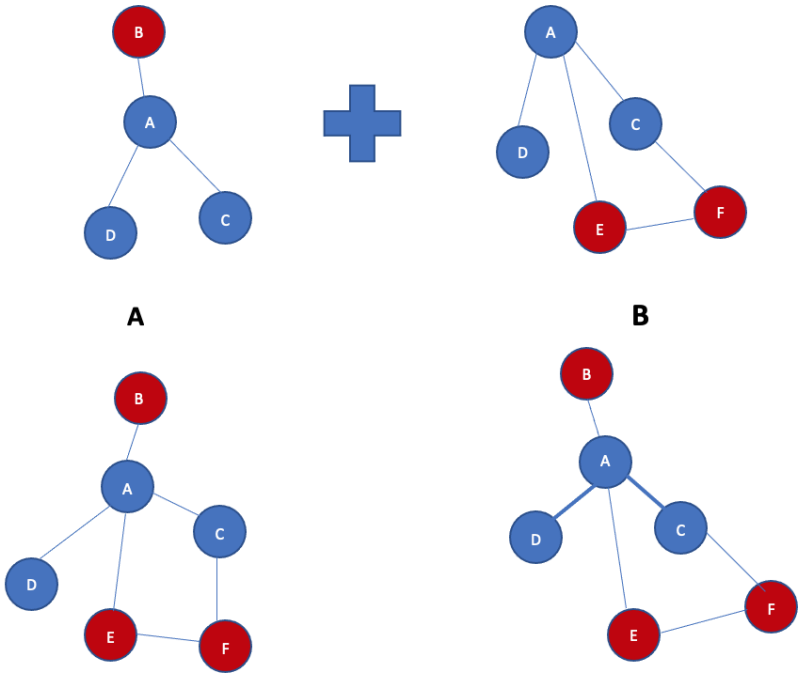


Figure 2.2 Network integration techniques. Two datasets share three common nodes (blue circles). ***A*** *shows an unweighted network formed by merging associations equally.* ***B*** *shows a weighted network where edge thickness reflects the number of supporting evidence lines. Adapted from* [86].

2.3.1 Data Quality

Biological data, especially data generated from high throughput experimental studies, is incomplete and noisy [21], [29], [31], [32], [33], [42]. It has been estimated that HTP data has high rates of false results [31], [82], [161]. False results can be either false positives or false negatives. False positive associations refer to associations identified but which do not exist, while false negative associations refer to true associations that have not been identified. Colhoun *et al*. estimated that the proportion of false-positive results in studies investigating the link between a genetic variation and a disease could be as high as 95% [162]. Ioannidis *et al*. argued that many published research findings are not replicable due to limitations in hypothesis testing and reliance on *p*-values, which can lead to high rates of false positives [163]. Ioannidis *et al*. indicated that the proportion of false positives could exceed 50% [163]. However, Jager *et al*. critiqued Ioannidis's study for relying on assumptions rather than empirical data, estimating a false positive rate of around 14% in biomedical studies, which is

significantly lower than Ioannidis's estimate [164].

False results in DGA experiments can arise due to biological or technical factors [163], [165]. Technical factors and biological factors represent distinct aspects influencing experimental outcomes in scientific research. Technical factors refer to aspects of an experiment that are related to the methodology, equipment, and procedures used. Technical factors primarily influence the reliability and reproducibility of experimental procedures. These factors can lead to variations, errors, or artefacts in the data. Biological factors pertain to the inherent characteristics of living organisms, biological systems, or samples under investigation. Biological factors introduce complexity and variability into experimental systems which may lead to differences in outcomes between individuals, tissues, or experimental conditions [165].

Technical factors play a considerable role in possibly generating false results in DGA experimental studies. For example, in various DGA studies, DNA amplification techniques, such as the Polymerase Chain Reaction (PCR), are used to make multiple copies of a particular DNA region. PCR is a highly sensitive and commonly used method for DNA amplification, but it can be susceptible to various types of errors. Some common sources of PCR errors include: primer design errors [166], suboptimal reaction conditions [167], or DNA contamination [168]. With respect to biological factors, false negatives can result from biological masking. The presence of one biological factor can mask the effects of another, leading to overlooked true associations if these relationships are not considered during the analysis [165], [169].

Due to issues with low-quality data [31], and the fact that the effectiveness of computational methods greatly depends on data quality [21], there is a considerable need for approaches that assess data quality, thereby reducing noise before performing data integration. Several techniques have been applied to assess the quality of biological data [55], [82], [83], [84], [93], [140], [161], [170], [171]. One of these techniques is cross-species comparison. This technique is used to investigate the molecular basis of diseases across different species, typically focusing on model organisms that share genetic and physiological similarities with

humans [170]. This approach can help to reveal false results and enhance the understanding of disease genetics. For example, if a particular gene is implicated in a disease in humans, researchers can examine whether the homolog exists in other species and if its disruption also leads to a similar disease phenotype. Consistent results across different species may provide high confidence evidence of the gene's function in the disease and validate the original association [171].

The use of additional data types can also be beneficial for reducing noise in various types of biological data analyses [84], [172]. For instance, annotating genetic variants or genes with functional information helps prioritise those relevant to disease, reducing false associations [173]. For example, Quick *et al*. conducted a study aimed at enhancing the effectiveness of gene-based association tests in Genome-Wide Association Studies (GWAS) by integrating diverse heterogeneous annotations, including comprehensive coding and tissue-specific regulatory annotations, with GWAS summary statistics [174]. Using additional annotations can filter out irrelevant genes or variants and reduce the false positives rates.

The topological structure of a network can also be used to detect false associations [75], [140], [175]. The presence or absence of multiple paths between DGAs in DGA networks can be used to confirm the existence or absence of these DGAs [176]. For example , Lei *et al*. [173] developed a network propagation algorithm called InLPCH, which effectively reduced false positives when predicting disease-related genes, and enhanced prediction accuracy by considering multiple paths within the network. Figure 2.3 illustrates multiple paths in a DGA network.

Network models, a graph representation for biological associations, can also be a beneficial framework for filtering false associations in biological data [177], [178]. When data fits a network model well, it might provide additional evidence for the validity of the DGAs [161], [179]. A well-defined network model captures the underlying biological relationships between nodes such as DGAs.

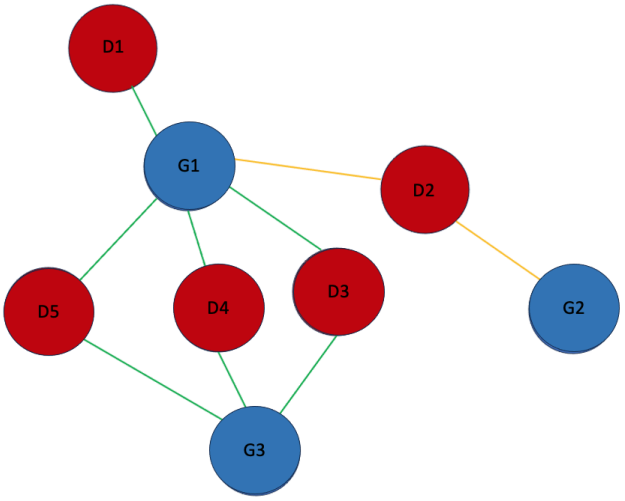


Figure 2.3. Multiple paths in a DGA network. Blue circles represent diseases, red circles represent genes. Green lines show multiple paths linking gene G3 to disease D1; yellow lines show a single path between G2 and D1.

Another commonly used approach in PPI data for assessing the quality of a dataset is scoring against high-confidence data, typically generated using well-established experimental techniques [50], [55], [57], [86]. This method is used to estimate the amount of noise present in a dataset. High-confidence data are typically derived from sources that are known to be reliable. These could be data generated through LTP experimental techniques (LTP studies) [57] or from well-established curated databases [55], [57]. This technique has been widely employed within PPI networks and has demonstrated remarkable effectiveness in reducing noise [50], [55], [56], [57], [91], [92], [94]. It has been applied to protein functional prediction applications [57] and drug repurposing applications [50]. Nevertheless, its application to bipartite networks, such as predicting DGAs, remains limited.

2.3.2 Gold Standard Data

Some existing methods employed to reduce noise in biological networks may inadvertently remove true positive findings, leading to the presence of false negatives in the network [83], [180], [181]. Constructing a precise biological network involves filtering out false results while retaining important data. However, integrating multiple diverse datasets can present challenges due to variations in their reliability and accuracy, making it challenging to estimate the quality of each dataset before integration.

A widely employed approach to assess the quality of datasets involves comparing them with a benchmark known as the *gold standard* [55], [56], [57], [182]. Gold standard data is a high-quality, well-curated dataset that serves as a benchmark or reference for assessing the accuracy and reliability of datasets or new results coming from computational methods. This dataset is assembled by experts, often through comprehensive literature reviews, experimental validation, and rigorous quality control. It also provides a trustworthy foundation for evaluating the performance of various computational approaches. This benchmark represents datasets of the utmost quality and confidence [13], [14], [15], [17], [21], [225].

An appropriate gold standard should comprise accurate, high-quality interactions or associations. For instance, gold standards for PPIs include those related to shared metabolic pathways, shared biological processes, shared molecular functions, or shared memberships in protein complexes. For example, Gene Ontology (GO)[4] [183] terms, which are curated by human experts, or interactions in the Human Protein Reference Database HPRD[5] [184] have been directly used as gold standards for PPI data [95]. The Kyoto Encyclopaedia of Genes and Genomes (KEGG)[6] [185] and the Reactome pathway database can also serve as a PPI gold standard datasets [228]. The BIOGRID[7] [87] database and the IntAct[8] database [187] are further examples of resources that can be used as PPI gold standards.

Small-scale interactions from LTP experiments known for their high quality, can serve as gold standards. For instance, a dataset comprising physical interactions from low-throughput studies in BIOGRID [87] can be considered a high confidence dataset and can be used to assess the quality of the datasets [57].

[4] http://geneontology.org/

[5] https://www.hprd.org/

[6] https://www.genome.jp/kegg/

[7] https://thebiogrid.org/

[8] https://www.ebi.ac.uk/intact/home

The overlap between datasets can be used to identify gold standards and provide a reliable way to assess data quality. Vicente et al. explored using intersections of datasets from KEGG, GO, and PPIs to establish a reliable benchmark for evaluating data quality. They proposed four intersection gold standard (IGS) datasets: KEGG+GO, KEGG+PPIs, GO+PPIs, and KEGG+GO+PPIs. Despite being smaller, these IGS datasets effectively captured the shared relationships across all datasets and provided a practical and unbiased solution for assessing data performance across diverse sources.

A good example of a gold standard dataset for disease-gene associations is the Online Mendelian Inheritance in Man (OMIM)[9] database [88]; manually curated by human experts, and an authoritative resource that catalogues genetic disorders and the associated genes. Other databases and resources such as the Genetic Association Database (GAD)[10] [188], Clinical Variation Database (ClinVar)[11] [189], and the Genome-Wide Association Studies Catalog[12] [190] may also serve as valuable curated sources for creating gold standard datasets for disease-gene associations.

2.3.3 Probabilistic Functional Integrated Networks

One of the most effective methods of integration to reduce the data noise during the integration of multiple data sources integration is the use of Probabilistic Functional Integrated Networks (PFINs) [50], [55], [59], [85], [86], [191], [192]. In functional networks, nodes correspond to biological entities such as genes, proteins, or diseases, and the edges to functional associations between nodes. A PFIN has edge weights which indicate the level of confidence in the combined evidence for that edge. This confidence score is present when there is evidence of a functional association between the nodes from at least one data source [55]. The edge weights are produced by statistical comparison against a gold standard dataset

[9] https://www.omim.org/

[10] https://maayanlab.cloud/Harmonizome/resource/Genetic+Association+Database

[11] https://www.clinicalgenome.org/data-sharing/clinvar/

[12] https://www.ebi.ac.uk/gwas/

[55], [59], [85], [86], [90]. The resulting networks can be used to infer novel associations between nodes.

An approach using Bayesian statistics was developed by Lee and his colleagues [55]. This method calculates a log-likelihood score (LLS) for each dataset, and can be used to compare datasets:

$$LLS = ln(\frac{P(L|E)/\sim P(L|E)}{P(L)/\sim P(L)}) \quad (2.2)$$

where *P(L|E)* and ~*P*(*L*|*E*) represent the frequencies of edges or linkages *(L)* in the dataset *(E)* observed and not observed in the gold standard, respectively, and *P*(*L*) and ~*P*(*L*) represent the prior expectation of linkages or edges observed and not observed in the gold standard, respectively. Only datasets with an LLS score greater than zero are included in the integration. The P(L) is calculated by dividing the total number of linkages in the gold standard over the number of all possible linkages in the gold standard, calculated using all types of nodes captured in the gold standard, whilst ¬*P(L)* was determined by dividing the total number of linkages not in the gold standard over all possible linkages in the gold standard. The *P(L|E)* was calculated by dividing the number of true positives (the number of linkages observed in both the dataset *E* and the gold standard) over the total number of linkages in the dataset E, whilst ¬*P(L|E)* was calculated by dividing the total number of false positives (the number of linkages observed in the dataset *E* but not in the gold standard) over the total number of linkages in the dataset *E*.

The scored datasets can be integrated in several ways. Simple integration includes summing each confidence score to produce a network, in which an edge weight is the integrated sum of all the evidence for that edge. This integration method requires the datasets to be independent, which is difficult in biological data. Independent datasets mean that each dataset provides unique and separate evidence of interactions or associations, without being influenced by or duplicating data from other datasets. Achieving independence among datasets is challenging in biological data due to various interconnected factors and dependencies [55], [193], [194], [195]. To overcome this issue, Lee and his colleagues [55]

introduced a weighted sum during integration to successively down-weight evidence scores in order of magnitude:

$$WS = \sum_{i=1}^{n} \frac{L_i}{D^{i-1}} \tag{2.3}$$

Datasets are integrated in the order of their LLS scores from the highest to lowest, where L_1 is the highest confidence score for the edge and L_n is the lowest confidence score in a set of n datasets. Division of the score by the D parameter means that, while the highest score is integrated unchanged, subsequent weights are progressively decreased. Therefore, a D value of *1.0* would produce a simple summed network (Figure 2.4), and higher values successively down-weight the confidence scores. Figure 2.5 shows the overview of Lee's two steps method for building a Probabilistic Functional Integrated Network [55].

The PFIN approach has been used to reduce noise during data integration in PPI networks. Fraser *et al*. [56] proposed a framework rooted in functional genomics to construct probabilistic gene interaction networks. In this framework, diverse large datasets, such as physical interaction data, microarray co-expression data, and genetic interaction studies, were integrated using statistical approaches. Statistical methods were proposed for quality control to address the noise and errors inherent in functional genomics datasets. Each dataset was scored against benchmark tests to distinguish accurate from inaccurate data. These benchmarks involved assessing how well known interactions are recovered, evaluating co-localization in subcellular compartments, or analysing coexpression patterns across experiments. The aim was to quantify the error rate in experiments using benchmarks, enabling the weighting of interactions based on their performance. Different weighting schemes, including Bayesian statistics or probabilistic approaches were employed. The resulting gene networks were not binary but consisted of probabilistic weights for gene-gene linkages. These weights reflected both the confidence of the interactions and the measured error in linkage determination.

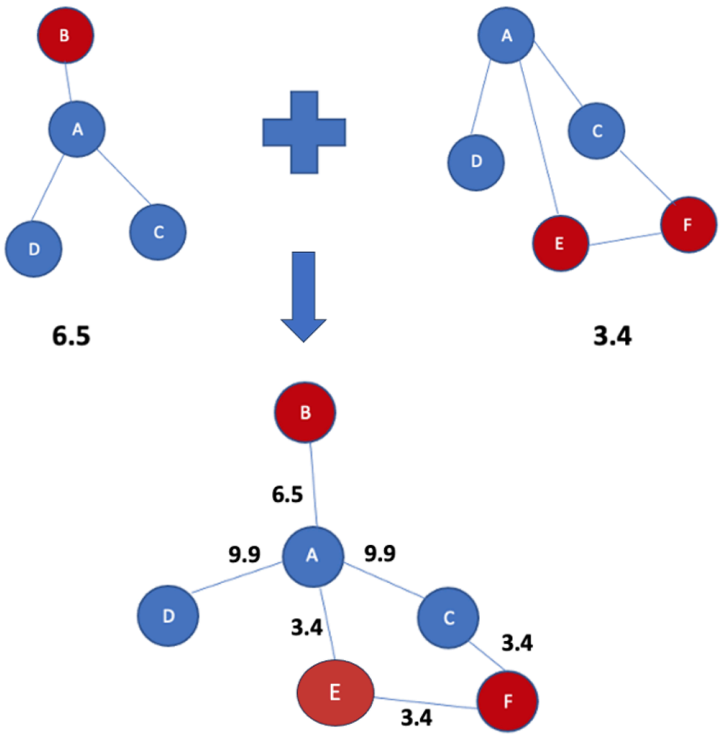


Figure 2.4 Probabilistic Functional Integrated Networks (PFINs). Edge weights reflect summed confidence scores from two datasets (6.5 and 3.4). Shared edges are weighted 9.9, while unique edges retain their original weight. This illustrates a naïve integration method. Adapted from [86].

James *et al*. [85] also developed an integration technique that scores and integrates both high-throughput and low-throughput data from a single source database consistently without the need for an external gold standard dataset. It was found that this integration technique is easier and faster than using an external gold standard. Alexeyenko *et al*. introduced FunCoup, a computational framework designed to reconstruct global networks of functional interactions by integrating diverse proteomics and genomics data from multiple organisms [94]. The naïve Bayesian network was applied. Training datasets included positive and negative examples of PPIs. The positive gold standard was constructed by compiling four functional interaction classes from various databases: IntAct, HPRD, BIND, KEGG, and UniProt. Bayes' rule was used by incorporating the background evidence probability instead of relying on a negative dataset. The naïve Bayesian network was trained and the resulting classifier was used to determine functional interactions.

Xia et al. developed IntNetDB, a computational tool for building human PPI networks, using

a Naïve Bayes probabilistic model to integrate 27 datasets, including genomic, proteomic, and functional data [95]. The HPRD served as the positive gold standard, while a dataset from Rhodes et al. provided the negative gold standard. By incorporating phenotypic distances and genetic interactions, the study aimed to improve PPI network accuracy and coverage, resulting in 180,010 predicted PPIs among 9,901 human proteins.

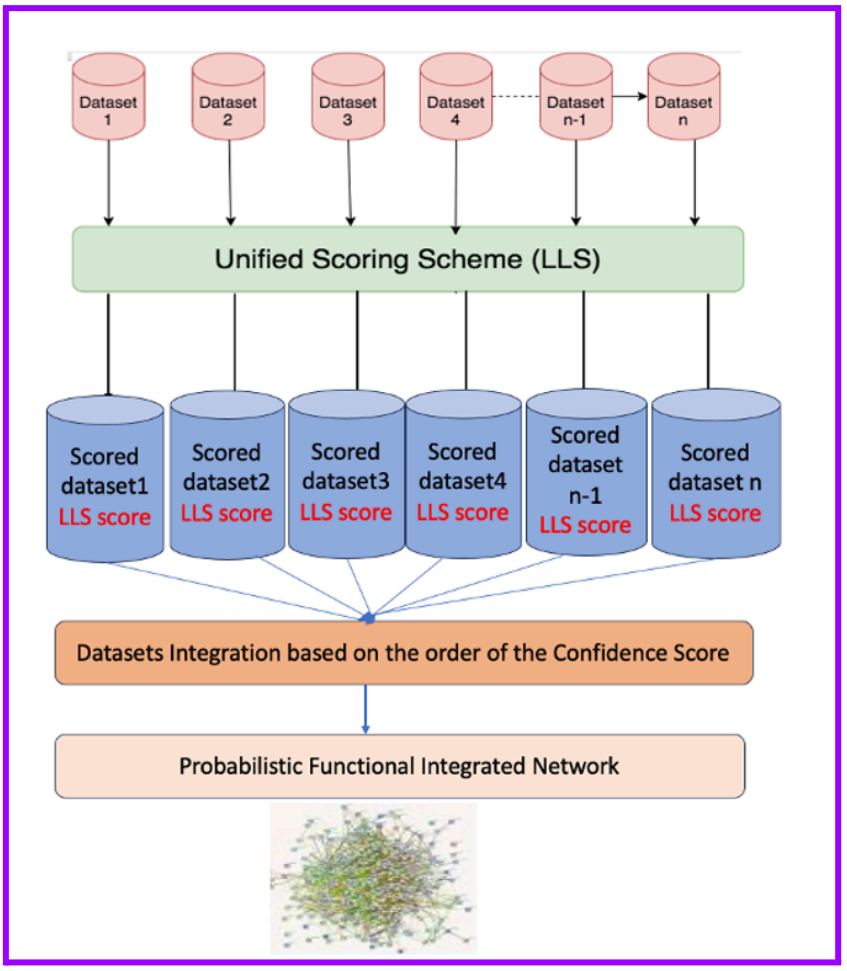


Figure 2.5 Overview of Lee's integration method. Datasets are first scored against gold standards using LLS, then integrated via weighted sum based on confidence scores.

Wang *et al*. developed a Bayesian network (BN) classifier for predicting functional modules in complex human diseases, particularly hepatocellular carcinoma (HCC) [92]. A customised BN classifier was designed to integrate diverse information from different sources and predict PPIs. The BN classifier was able to group molecules into functional modules or pathways based on their primary function and biological features. The application of this model to HCC led to the identification of functional modules associated with the development and

progression of HCC.

The discussed approaches aim to construct accurate probabilistic functional networks using different methodologies. Both PFIN and FunCoup use Bayesian networks to integrate diverse datasets and predict functional interactions, focusing on noise reduction and improved accuracy through confidence scoring and thresholding. While PFIN and FunCoup target global networks across organisms, IntNetDB focuses on human PPI networks, and Wang et al.'s Bayesian classifier targets functional modules in complex human diseases. These methods enhance network accuracy and coverage, but probabilistic methods for DGA networks remain underexplored, with most existing approaches relying on heuristic methods. This work explores applying PFINs to DGA networks.

2.4 Disease-Gene Associations

The identification of DGAs leads to numerous benefits across multiple domains of biomedical research and healthcare. DGA identification can help to improve the understanding of disease. DGAs provide insights into the genetic basis of diseases, aiding in understanding the molecular mechanisms underlying disease development and progression [196]. DGAs can be beneficial in targeted therapies knowing which genes are associated with specific diseases enables the development of targeted therapies that address the root causes of diseases, potentially leading to more effective treatments [197]. DGAs can also be useful in personalised medicine, allowing the tailoring of treatments to individuals based on their genetic profiles. Personalised medicine enhances treatment efficacy and minimises the occurrence of adverse reactions [198]. Genetic testing based on DGAs enables early disease detection and risk assessment. For example, genetic testing for newborn babies can help to detect conditions such as sickle cell anaemia in their early stages. This early identification allows for timely intervention, preventive measures, or appropriate medical treatments. Genetic testing facilitates preventive measures to reduce disease progression [199]. Identifying DGAs can provide valuable guidance for directing drug discovery applications, identifying potential targets for new drugs [200]. DGA identification also aids in repurposing existing drugs for new indications, speeding up the development timeline [201]. DGAs have

led to the identification of biomarkers, which are used for disease diagnosis, prognosis, and treatment response monitoring [202].

DGAs can be constructed using either experimental or computational approaches, or both. Experimental approaches include the observation of gene expression patterns, interactions, and functional roles within disease contexts. The power of experimental techniques lies in their capacity to provide direct evidence of gene involvement in diseases. Such experiments can reveal the exact molecular pathways through which genes contribute to disease processes. Experimental approaches often demand substantial time, resources, and financial investment. Dealing with complex genetic interactions and multiple factors such as environmental conditions concurrently can present challenges requiring careful experimental planning. Experimental data might be influenced by biases, noise, and variations introduced during the experimentation process.

In contrast, computational methods involve employing bioinformatics and data analysis tools to handle massive amounts of biological data and predict gene-disease associations. The strength of computational techniques lies in their scalability, as they can evaluate vast datasets comprising thousands of genes and diseases simultaneously. Computational approaches allow for the uncovering of hidden associations by integrating a variety of data sources in an efficient timeframe and at a reduced cost. However, the accuracy of computational predictions heavily depends heavily on the quality and completeness of the input data.

2.4.1 Computational Approaches to Disease-Gene Associations

Investigating the relationship between genes and diseases is expensive due to the time-consuming nature of experimental validation, such as linkage studies [99], genome-wide association studies (GWAS) [203], and RNA interference screens [102]. Computational approaches to identifying or predicting disease-gene associations can reduce costs and time [110], [204].

Recently, several computational approaches to the prediction of disease-gene associations

have been developed to understand the mechanisms of diseases [110], [204], [205], [206], [207]. These methods can be divided into machine learning-based techniques [105], [108], [203], [208], [209], text mining-based approaches, and network-based approaches [62], [63], [140], [182], [210], [211].

2.4.1.1 Machine Learning Methods

Machine learning methods are commonly used to identify disease-gene associations. Genomic features that could have a function in disease mechanisms can be identified by machine learning models [204]. These techniques extract the characteristics of disease-related genes to build predictive models. Machine learning models use various biological features to learn patterns associated with disease-gene associations. These biological features include genomic features such as gene expression data, genetic variants [106], pathway information, and Gene Ontology annotations such as information about the biological processes, molecular functions, and cellular components associated with genes [105], clinical information about patients such as symptoms [212], medical history, and demographics, and phenotypic data such the observable characteristics of individuals with a particular disease.

For example, Krishnan *et al*. proposed a machine-learning approach based on a human brain-specific gene network. A weighted support vector machine (SVM) classifier was developed to predict the probability of association between each gene with Autism spectrum disorder [208]. Using a recently developed human brain-specific functional interaction network, they predicted the candidate genes of Autism spectrum disorder, across the genome and systematically characterised the developmental and functional features of autism's molecular phenotype.

Deep learning, a subfield of machine learning, has shown promise in predicting disease-gene associations by using the power of neural networks to learn complex patterns from large-scale biological data. For example, Chen and co-workers proposed a CNN for identifying disease-genes associations [209]. First, the symptomatic information of the disease was retrieved, and the sequence information of the disease gene product protein was obtained. Then the DGA information was mapped into a grayscale image. Finally, a CNN

model for predicting disease-related genes was constructed.

Recently, GNNs have been applied extensively to predicting DGAs, and it has been demonstrated that the GNN-based approaches outperformed traditional methods [107]. GNNs are a powerful tool for using the biological networks such as DGAs networks, and/ or PPI networks to predict disease-gene associations. For example, Azadifar *et al*. introduced a novel semi-supervised learning approach based on graph convolutional networks, utilising a newly devised feature vector construction for each gene [107]. The method started with the construction of three distinct feature vectors for each gene, utilising information sourced from the GO database, including the molecular function, cellular component, and biological process terms. Subsequently, a graph convolutional network was trained on these vectors, utilising PPI network data to identify potential candidate genes for diseases. Hidden layer representations encoding both local graph structures and node features were uncovered by the model. The study concluded that the proposed semi-supervised learning method effectively discriminated and ranked potential disease genes. This was achieved through the utilisation of a graph convolutional network and an innovative approach that generated three distinct feature vectors for genes.

2.4.1.2 Network Integration Methods

Relying on a single data source for these data-driven approaches to predicting DGAs is insufficient, as most disease-gene association datasets have their own motivation and purpose. Currently, several network integration-based techniques have been developed to identify disease-gene associations [140], [182], [210], [213]. These approaches tend to combine multiple diverse types of data sources to create more comprehensive and informative networks to predict disease-gene associations than can be achieved using only a single data source. The primary goal of these integrated networks is to integrate complementary information from various sources, such as gene expression profiles, PPI networks, pathways, and functional annotations, to gain insights into the complex interactions underlying disease mechanisms.

Each data source can be represented as a separate network in which nodes represent

biological entities such as genes, proteins, or diseases,and edges represent disease-gene associations. The integration process involves combining individual networks into a unified network. This can be achieved simply by integrating DGAs from different data sources without considering the strength or reliability of the individual associations. This approach results in an unweighted network where the presence or absence of an edge between disease and gene signifies their association, without any indication of the degree or quality of that association.

For example, Zeng *et al*. integrated heterogeneous networks of PPIs, gene-phenotype associations, and phenotype–phenotype similarity [214]. These networks were used to prioritise novel gene-disease associations. Franke *et al*. integrated three different gene networks based on GO annotations, PPIs and gene expression with information on pathways added from the KEGG and Reactome databases to prioritise novel gene-disease associations [182] . Zou *et al*. [211] created a phenome–interactome network by identifying human protein complexes that comprise proteins known to be involved in many types of diseases. Zhang *et al*. [215] constructed an expanded Human Disease Network by integrating disease gene information with PPI information, and then the network's topological features and functional properties were analysed.

While this method is straightforward and easy to implement, ignoring the strength or quality of edges discards valuable information about the level of evidence or reliability of associations in each data source. Therefore, an unweighted network lacks the ability to distinguish between strong and weak associations.

Another integration approach that involves edge weighting for gene-disease associations is more sophisticated and takes into account the strength of evidence supporting each association. This approach assigns confidence scores to edges based on heuristic methods that consider factors such as the numbers of lines of evidence supporting associations, enabling the creation of a weighted network which can provide a more accurate representation of the underlying disease-gene associations.

For example, Piñero *et al* integrated data sources related to human gene-disease associations

and variant-disease associations (VDAs) from diverse repositories, encompassing Mendelian, complex, and environmental diseases [141]. The team introduced two distinct scoring systems to prioritise both gene-disease and variant-disease associations based on the confidence of their supporting evidence. These scores range from 0 to 1, considering factors such as the number and type of sources (level of curation, model organisms), and the number of publications supporting the association. The presence of duplicate studies among these sources has not been considered in this calculation. The omission of duplication considerations could potentially lead to an inflation of the score, as the same evidence might be counted multiple times, influencing the perceived strength of the DGA. The impact of the duplicate data on the confidence scores was investigated in chapter 4 (Section 4.3.5)

Strande *et al*. also devised a comprehensive heuristic approach for establishing confidence levels for DGAs [160]. This approach evaluates the relevance and reliability of genetic and experimental evidence that supports or contradicts a potential DGA. The framework included a semi-quantitative system that quantifies the potency of evidence for DGAs. This system aligns with qualitative categories, encompassing "Definitive," "Strong," "Moderate," "Limited," "No Reported Evidence," and "Conflicting Evidence." Within ClinGen's structure, these classifications, determined via the framework, are subject to review by relevant disease experts. These experts assess and potentially adjust the classifications based on their clinical expertise.

The classification "*Disputed*" is applied to genes lacking proven causal links to human monogenic diseases but possessing experimental data, such as model system findings, indicating potential disease-related functions. The "*Limited*" category requires at least one variant claimed as disease-causing to exhibit compelling genetic proof supporting the connection to a human disease, regardless of gene-level experimental data. The "Moderate" class involves additional clinical and supportive experimental evidence, which may stem from multiple studies or a single robust study. The "*Strong*" classification necessitates replication of the DGA in independent publications, along with substantial genetic and experimental data. A DGA is deemed "*Definitive*" when, beyond accumulating convincing genetic and experimental support, the relationship has been replicated and sufficient time has passed (typically more than three years) to allow for the emergence of any conflicting

evidence since the initial publication.

Although heuristic integration methods for generating confidence scores are straightforward and easy to apply, these approaches often do not consider the data quality of the datasets before integration. Therefore, employing more advanced methods rooted in statistical approaches could offer enhanced accuracy when generating confidence scores. For example, the PFIN approach has been used to generate confidence scores for PPIs, and has also produced promising results in reducing noise during integration [55]. However, this probabilistic approach has remained limited to disease-gene confidence scores. In this work, the applicability of the PFIN approach on disease-gene networks was investigated.

Available networks for DGAs have various levels of false positives due to the use of noisy data types; for example, results from HTP experimental studies are affected by noise and biases, resulting in many false results [31], [32], [33], [82]. Computational methods in systems biology typically rely on the quality of experimental data, and it has been shown that experimental data quality considerably impacts the results of these computational methods [19], [20], [21], [28]. Data quality is one of the most common challenges for computational approaches to the prediction of DGAs (Section 2.3.1 for data quality challenges).

Current network integration methods used to predict DGAs either involve using unweighted networks [63], [140], [211] or using heuristic methods to measure the confidence scores of DGAs (Section 2.3) [141]. However, unweighted networks may be less informative than weighted networks, which provide a richer representation of the relationships between diseases and genes, capturing how confident we are in the reliability of those connections. In a weighted network, the use of edge weights provides a measure of the confidence of DGAs. To mitigate noise, various strategies can be employed. Thresholding involves setting a minimum weight value and removing edges below this threshold to filter out less significant associations. Alternatively, keeping only the top_*k* edges with the highest weights

concentrates on the most substantial associations, effectively reducing noise. Weight-trimming techniques, such as scaling or normalisation, can also help attenuate the impact of outliers. Clustering based on edge weights groups similar associations, aiding in noise reduction while preserving underlying patterns. However heuristic methods lack the means to assess the data quality, and face several limitations, including the presence of duplicate data, resulting in duplicated evidence of associations. For example, the team of DisGenNET [141] developed a heuristic method to generate the confidence scores of DGAs taking into account the number and type of sources (level of curation, model organisms), and the number of publications supporting the association. However, the quality of the evidence is not taken into account, and the duplicate evidence between data sources was not removed before the scoring, which leads to bias within the scoring, upscoring associations with duplicate evidence [116].

Unlike current network integration-based methods for predicting DGAs, the PFIN approach has shown the potential to reduce noise during integration, since PFINs take the quality of individual datasets into account by confidence scoring before integration (Section 2.3.3) [55], [56], [59], [86], [90]. The PFIN approach has been applied to various applications such as protein function prediction and new PPI prediction, and it has shown encouraging results in PPI networks [50], [55], [57].

While the PFIN approach already involves considerable efforts to reduce noise during data integration in unipartite PPI networks [50], [55], [57], this approach has remained limited to bipartite DGA networks (for more details about PFINs, refer Section 2.3.3). Several unique challenges occur when applying this approach to such bipartite DGA networks. First, the lack of identification of accurate and good coverage (containing diverse types of diseases) gold standard data for DGAs. A gold standard is a benchmark dataset which can be used to evaluate the quality of datasets and also can be used to validate outputs from studies (Section

2.3.2). One of the main challenges in applying PFINs to DGA networks is the lack of reliable and comprehensive gold standard data for DGA data. Gold standard data is an important component in building PFINs because it serves as the benchmark for scoring and assessing dataset quality (Section 2.3.2). Accurate and comprehensive gold standards should encompass diverse types of diseases and reflect high-confidence DGAs. Studies in PPI networks have shown that gold standards significantly impact the performance of PFINs [192]. Therefore, the selection of the gold standard is one of the most important steps in building PFINs. However, current DGA datasets often lack either the high-quality data or the comprehensive coverage needed in terms of diseases, leading to potential biases and limitations in the performance of PFINs. Addressing this issue is critical to ensure the validity of PFINs, facilitating more accurate predictions and insights into disease mechanisms.

Second, the lack of identification of appropriate individual datasets to represent DGAs poses a challenge. Many existing DGA datasets are either not of high quality or lack comprehensive coverage of diseases. Selecting the appropriate individual datasets is essential for building PFINs, as these datasets are scored and then integrated into the network. Studies in PPI networks have shown that individual datasets significantly impact PFINs' performance. Therefore, individual datasets need to be high-quality, up-to-date, and provide good coverage of diseases.

Third is the lack of appropriate evaluation techniques for the robustness of such bipartite-weighted DGA networks. Most available robustness network analysis techniques, such as link prediction and network clustering analysis (Section 4.3.3.1), were originally developed for unipartite or unweighted networks such as PPI networks. While some studies have attempted to adapt these techniques to bipartite-weighted networks, they still face several limitations (Sections 4.3.3.1). Therefore, there is a need to search and develop appropriate and robust evaluation techniques specifically tailored for bipartite-weighted DGA

networks. Therefore, the aim of this work was to address these limitations by investigating the applicability of this approach to DGA networks.

2.4.1.3 Text Mining Methods

Text mining methods play a significant role in disease-gene identification by extracting valuable information from the vast amount of biomedical literature available in the form of scientific articles, clinical reports, and other textual sources. These methods leverage natural language processing (NLP) techniques to automatically process and analyse text, enabling the discovery of potential associations between diseases and genes [216].

Frankild *et al*. introduced a text mining method targeting disease-gene associations [108]. They developed a confidence scoring mechanism based on co-occurrence to assess the relationship between a gene and a disease. This scoring quantifies the frequency with which genes and diseases are mentioned together in the text corpus. The scoring approach is designed to reflect the varying strengths of co-occurrences. Co-occurrences within the same sentence are considered more robust evidence than co-occurrences spanning multiple sentences within a paragraph. Similarly, co-occurrences within a paragraph outweigh those occurring across paragraphs within a paper.

2.5 Biomedical Databases

Numerous data sources are available, providing the potential to facilitate the applications of computational methods to drug repurposing focusing on DGAs. These data repositories can be classified into databases manually curated by biocurator experts; animal models which focus on animal-related data; inferred databases which are derived from statistical analyses and predictions; integrated databases which automatically integrate information from multiple diverse sources; and literature data which involve text mining from the biomedical literature. Examples of each category are outlined in Table 2.1.

Table 2.1. A survey of some available biomedical databases including their types, names, descriptions and focus.

Method of Data Generation	Biological Focus of Data	Databases Name	Description of the Data Source	Link and Reference
Manually curated data	Protein	Universal Protein Resource (UniProt)	• Focuses on protein sequences including annotation data, functional details, and protein-coding genes for human, yeast, and bacteria. • Combines expert-reviewed UniProtKB/Swiss-Prot with unreviewed UniProtKB/TrEMBL sequences. • Involves expert biocuration teams reviewing experimental and predicted data	https://www.uniprot.org/ [217]
		Human Protein Atlas (HPA)	• Includes protein expression across various tissues, organs, and cell types. • Provides protein expression changes linked to diseases • Providing Immunohistochemistry (IHC) data in to locate proteins in tissue samples • Provides visual representations of protein localization within tissues • Includes both physical PPIs and genetic interactions.	https://www.proteinatlas.org/ [218]
		Biological General Repository for Interaction Datasets (BioGRID)	• Gathers experimental data from scientific literature, high-throughput and low-throughput experiments, and curated databases. • Employs manual curation, with curators meticulously reviewing literature and	https://thebiogrid.org/ [87]

			experimental data to ensure reliability. ● Includes supplementary information, including experimental techniques, and original research articles. ● Regular updates incorporate new interaction data, improving the overall quality	
		Reactome	● Provides a comprehensive overview of biological pathways, reactions, and processes across diverse organisms. ● Concentrates on elucidating molecular processes within cells, driving various biological functions such as metabolism, signal ● Contains a detailed catalogue of pathways, representing organised sequences of molecular reactions and interactions leading to specific cellular processes. ● Provides manual curation by experts, ensuring the reliability of the data.	https://reactome.org/ [186]
	Gene			
		Gene Ontology (GO)	● Describes gene functions and products ● Uses a structured vocabulary to categorise functions into molecular, biological, and cellular components. ● GO terms are organised hierarchically, showing relationships between different gene functions. ● Annotations are based on experimental evidence or computational predictions.	http://geneontology.org/ [183]

		National Centre for Biotechnology Information (NCBI)	• Offers a wide range of resources in bioinformatics, biomedical, and molecular biology fields including PubMed, providing access to a vast scientific literature collection, and GenBank, facilitating exploration of DNA sequences. • Provides resources for genomics, proteins, and genes, enhancing understanding in these areas.	https://www.ncbi.nlm.nih.gov/ [11]
	Disease	Online Mendelian Inheritance in Man (OMIM)	• A comprehensive database of human genes and genetic phenotypes • Includes manually curated DGAs. • Focuses on the molecular relationship between genetic variation and phenotypic expression. • Updated manually by the biocurator.	https://www.ncbi.nlm.nih.gov/omim [88]
		Orphanet	• Includes manual curated rare diseases, DGAs and orphan drugs • Provides high-quality data on rare diseases including a classification of rare diseases, DGAs and mappings with other terminologies, and a proportion of associations taken from OMIM	https://www.orpha.net/consor/cgi-bin/index.php [113]
		Clinical Genome Resource(ClinGen)	• Dedicated to evaluating the clinical importance of genes and variants. • Manually curated to assess DGAs • Provides clinical validity classifications based on evidence from literature and other sources. • Utilises a gene curation process to evaluate the confidence of DGAs based	https://clinicalgenome.org/ [219]

			on publicly available evidence, including genetic, experimental, and contradictory evidence from the scientific literature. ● The curation process involves assigning a clinical validity classification using methods and metrics developed by the ClinGen (For more details about evidence level metric developed by ClinGen, see Section 2.4.2.2)	
		Genomic England	● Publicly available knowledge base for creating, storing, and querying virtual gene panels related to human disorders ● Utilises a crowdsourcing tool to integrate reviews from the global clinical and scientific community. ● Expert biocurator team evaluates gene lists to assess evidence supporting or refuting DGAs ● Evaluation process includes a traffic-light system indicating confidence levels: Green: Highest confidence (supported by at least three sources); Amber: Moderate confidence (supported by two sources); Red: Lower confidence (supported by one source).	https://www.genomicsengland.co.uk/
		Cancer Genome Interpreter (CGI)	● Foucs on interpreting and analysing cancer genomic data. ● Integrates information from various sources for identifying potential driver mutations and therapeutic options. ● Provides evidence-based interpretations from	https://www.cancergenomeinterpreter.org/home [112]

			scientific literature, clinical trials, and databases	
		Psychiatric disorders Gene association NETwork (PsyGeNET)	• Explores psychiatric diseases and related genes. • Extracts DGAs from scientific literature using the text mining tool BeFree. • The text mining process is followed with a curation process conducted by expert biocurator teams in the field. • The curation team involves 22 experts from diverse backgrounds, including psychiatry, neuroscience, medicine, psychology, and biology.	http://www.psygenet.org/web/PsyGeNET/menu;jsessionid=1dgwnbgzb7ip91af3l9kd2bn4q [114]
	Drug	DrugCentral	• Provides information about drugs such as authorised medications and experimental compounds, drug nomenclature, molecular configurations, modes of operation, drug applications, drug contraindications, and recommended dosages. • Contains characteristics of drugs, including their interactions with receptors, enzymes, and molecular targets. • Explores facets of drug metabolism, covering how substances are absorbed, distributed, metabolised, and excreted by the body.	https://drugcentral.org/ [220]
		DrugBank	• Provides detailed information about drugs, chemical structure, indications, contraindications, dosages, and administration routes. • Includes information about the molecular mechanisms by which drugs exert their effects, how drugs target	https://go.drugbank.com/ [221]

			specific proteins, enzymes, receptors, or other biological components, and potential drug-drug interactions. ● Provides information about possible side effects, adverse reactions, and toxicity associated with drugs. ● Identifies the molecular targets that drugs interact with, offering insights into their therapeutic mechanisms ● Includes information about ongoing or completed clinical trials for some drugs.	
Inferred data	Disease	Genome-Wide Association Studies (GWAS Catalog)	● Includes information about DGA and DVA identified in GWAS. ● Provides a reference and link to each GWAS study for further exploration and verification. ● Gathers information from a variety of GWAS studies conducted worldwide. ● Focusses on genetic associations with various traits, phenotypes, and diseases. ● Regularly updated to incorporate new findings and enhance the overall quality of the data.	https://www.ebi.ac.uk/gwas/ [222]
		Genome-Wide Association Studies Database (GWASdb)	● Provides comprehensive data curation and knowledge integration for GWAS ● Focuses on GWAS identified significant trait/disease-associated SNPs. ● Integrates data from various sources related to GWAS	https://maayanlab.cloud/Harmonizome/resource/GWASdb [190]

			studies ● Regularly updated to incorporate new GWAS findings.	
		Clinical Variation Database(ClinVar)	● Foucses on the clinical significance of genetic variations and their associations with health and diseases. ● Involves a curation process to assess the clinical importance of genes and variants based on evidence from literature and other sources. ● Assigns clinical validity classifications using methods and metrics developed by the ClinGen. These classifications are based on publicly available evidence, including genetic, experimental, and contradictory evidence from scientific literature. ● Collaborates ClinGen, to enhance the quality(see Section 2.4.2.2 for quality assessment process developed by ClinGen)	https://www.ncbi.nlm.nih.gov/clinvar/ [189]
		Human Phenotype Ontology (HPO)	● A structured vocabulary used for describing traits and clinical features in human diseases. ● Supports phenotype descriptions in genetic and medical conditions. ● Aids in accurate phenotype annotations, identifying DGAs, and genomics. ● Organised hierarchically with terms having identifiers, synonyms, and definitions. ● While it is designed for human phenotypes, it can annotate traits in other	https://hpo.jax.org/app/ [223]

			species, enabling comparative studies.	
Integrated data	Protein	Integrated Interaction Database (IID)	• Integrates known and predicted interactions between eukaryotic proteins. • Covers 30 tissues in model organisms and humans, providing tissue-specific PPIs. • Incorporates experimentally detected PPIs, orthologous PPIs, and high-confidence computationally predicted PPIs. • Users can input proteins, match them to orthologs, and find interactions specific to various tissues. • Includes data for model organisms such as yeast, worm, fly, rat, mouse, and human.	http://dcv.uhnres.utoronto.ca/iid/Download/ [224]
	Disease	DisGeNET	• Integrates information on human DGAs and VDAs from different sources including curated, animal models, inferred, and literature data sources. • The scope includes Mendelian, complex, and environmental diseases. • Employs two scoring systems to prioritise associations, one is for DGAs, and the other is for VDAs (for more details about the confidence scores , see Section 2.4.2.2)	https://www.disgenet.org/ [116]
	Drug	NeDRex	• An interactive network medicine platform designed for disease module identification and drug repurposing. • Comprises three main components: a knowledge base (NeDRexDB); a	https://nedrex.net/tutorial/intro.html [48]

			Cytoscape app (NeDRexApp); and an API (NeDRexAPI). ● Integrates information from 21 existing data sources, including NCBI, DrugCentral, CTD, DrugBank, SIDER, HPO, Mondo, OMIM, DisGeNET, HPA, BioGRID, Reactome, UNIPROT, GO, BioOntology, UniChem, Repotrial, and ClinVar. ● Includes relationships between drugs, molecules, disorders, genes, proteins, and variants. These associations include information about indications, targets, side effects, contraindications, molecular similarity, phenotypes, gene associations with disorders, gene expression, protein interactions, pathways, signatures, and more.	

2.6 Graph Neural Networks(GNNs)

The fundamental idea behind GNNs is to learn and propagate information through a graph by iteratively updating node representations based on information from neighbouring nodes. This process is often referred to as message passing. GNNs encode both the node's own features and the features of its neighbours, allowing them to capture complex patterns and dependencies in graph-structured data. The key components of a GNNs include (Figure 2.6) [129]:

Node Embeddings: Each node in the graph is associated with an initial embedding or feature vector that represents its characteristics or attributes (Figure 2.6.A).

Message Passing: GNNs iteratively aggregate information from neighbouring nodes and update node embeddings using learned transformation functions. These functions enable nodes to exchange information and consider their local context (Figure 2.6.B).

Aggregation: Information from neighbouring nodes is aggregated to form a new representation for each node. Various aggregation methods, such as summation, mean, or attention mechanisms, can be employed (Figure 2.6.C).

Update Function: The aggregated information is combined with the node's current embedding using an update function. This function is designed to capture the way in which the node's representation should evolve based on its neighbours' information.

Pooling and Output Layers: GNNs can include pooling layers to down sample graphs and output layers for specific tasks, such as node or graph classification. Figure 2.7 shows two-layer graph neural network structure.

There are many types of GNNs, such as Graph Convolutional Networks (GCNs) which were introduced by Thomas Kipf and Max Welling in 2017 in Amsterdam [225], GraphSage which was introduced by Leskovec and two of his Stanford graduate students in 2017 [226], and Graph Attention Networks (GATs) which were introduced by researchers from Cambridge University in 2018 [227].

While GCNs, GATs, and GraphSAGE all operate on graph-structured data, they differ in terms of their aggregation mechanisms, attention mechanisms, weight sharing, information propagation, scalability, and use cases. The choice of which model to use depends on the specific task, graph structure, and computational considerations. Since this work deals with big data and more than a million edges, and due to the limited resources of memory and computations, GraphSAGE was employed to predict links between drugs and diseases.

A.

B.

C.

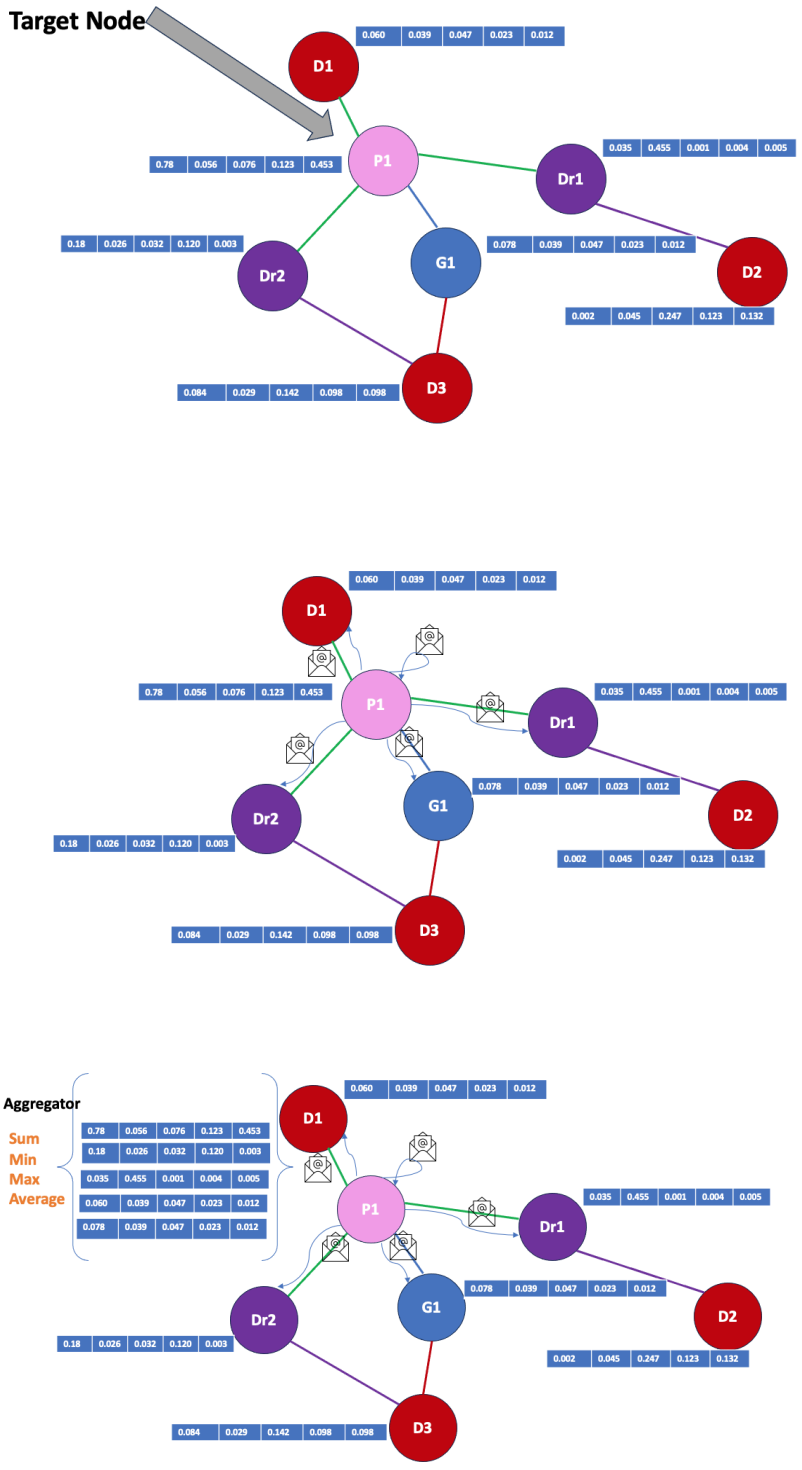


Figure 2.6. Neighbourhood exploration in graph convolutional networks. Each node is assigned a feature vector, including the target node P1. Through message passing, P1 aggregates features from its neighbours (Dr1, Dr2, D1, G1). Aggregator functions (e.g., mean, max) then combine these features to generate node embeddings.

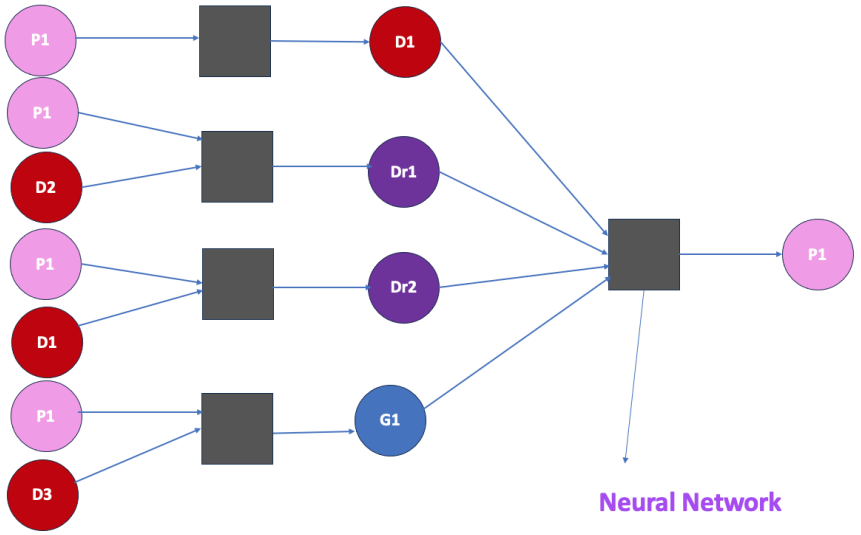


Figure 2.7. Two-layer graph neural network structure. The node embedding of the target node P1 is generated by aggregating information from the node's neighbours in two depth layers. The figure was adapted from https://web.stanford.edu/class/cs224w/slides/08-GNN.pdf

GraphSAGE operates by sampling a fixed-size neighbourhood for each node and then aggregating the sampled neighbours' features. This approach employs aggregation functions such as mean, min, max, or sum aggregation to combine neighbours' features. GraphSAGE does not learn separate weights for different nodes, and the aggregation scheme is fixed based on the chosen aggregation function. This method focuses on local neighbourhoods and is designed for scalable inductive learning on large graphs. GraphSAGE is scalable to large graphs due to its neighbourhood sampling strategy, making it suitable for applications where the graph is large. By sampling neighbourhoods and using aggregation functions, GraphSAGE optimally utilises computational resources and memory, making it feasible for training on large-scale graphs without requiring excessive resources. (Figure 2.8).

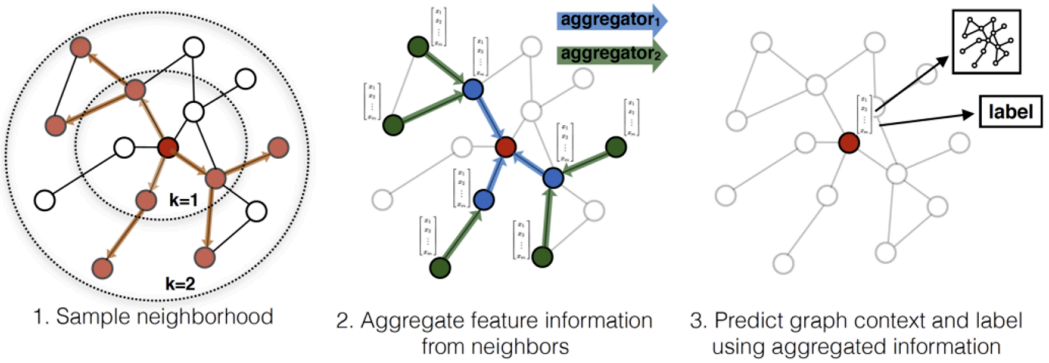


Figure 2.8. Sampling and aggregation in GraphSAGE. A sample of neighbouring nodes contributes to the embedding of the central node. The figure is from [226].

2.6.1 Graph Neural Network Applications in Drug Repurposing

Recently, there has been growing interest in utilising GNNs in drug repurposing applications, which aim to find new uses for existing drugs. This increasing attention is due to the ability of GNNs to extract meaningful information from complex biomedical knowledge graphs. Biomedical information represented as graphs offers a clear and efficient way to highlight data structure. The integration of deep neural networks with graph-based data provides a promising approach to addressing the challenge of drug repurposing.

For example, Sadeghi *et al*. constructed a complex heterogeneous graph called the drug-protein-disease (DPD) network representing the connections between drugs, proteins, diseases, and drug side effects [26]. They assigned labels to each drug-protein interaction edge in the DPD network, indicating the diseases associated with both the drug and protein. They then designed a multi-label ranking approach that incorporated a graph neural network architecture, which operated on this complex graph-structured data. This approach used the interaction patterns and features of drug and protein nodes. Their method outperformed existing techniques on the same dataset.

Doshi *et al*. developed a drug repurposing approach using a graph neural network model [130]. This model aimed to predict potential treatments for newly identified diseases. They employed a multi-layered heterogeneous network containing approximately 1.4 million

edges, which capture complex interactions among around 42,000 nodes representing drugs, diseases, genes, and human anatomies. The model was constructed with an encoder-decoder architecture to generate scores for drug-disease pairs that are under assessment. Moreover, the application of their proposed model to a coronavirus disease (COVID-19) dataset is highlighted. The results show that many drugs from the proposed model predicted list are already being studied for their potential effectiveness against the disease.

Zhang *et al.* introduced a new comprehensive model named Graph Convolution Network based on Multimodal Attention Mechanism (GCMM) [131]. This model integrated multiple types of data, including known drug-disease relationships, drug chemical similarity, drug therapeutic similarity, disease semantic similarity, and disease target-based similarity, into a heterogeneous network. The GCMM employed a Graph Convolution Network encoder to learn representations of drugs and diseases from various perspectives. The GCMM's performance surpassed that of four recently proposed deep-learning models across the majority of evaluation criteria. The GCMM demonstrated its reliability in predicting drug-disease relationships and displayed improvements in relevant metrics. The study included a case study focused on AD. Notably, four out of five medications identified by GCMM as having the highest potential correlation with AD are substantiated by literature or experimental research.

While GNNs hold promise in the field of drug representation, they are not without limitations. Despite their potential, several challenges and drawbacks persist in their applications to drug repurposing. Certain approaches tend to prioritise addressing the problem of missing data over data quality concerns. They assume that available data is of high quality, despite the fact that the efficacy of these methods heavily depends on data quality. Ignoring data quality can undermine the success of these techniques. Moreover, present GNN-based methods often operate within constrained knowledge graphs, characterised by limited nodes and edge types, along with restricted features. This limited scope can restrict the model's ability to capture the full complexity of drug-disease relationships. Importantly, the majority of GNN models overlook the inclusion of node features, often due to their absence in many existing databases. Nonetheless, integrating node features could potentially enhance the performance of these models. Furthermore, many existing graph learning techniques

concentrate primarily on node features and often disregard the valuable information stored in edge features. In drug repurposing, edge features contain valuable information about graph structures that are not adequately leveraged by existing methods.

This work focused on overcoming these constraints by incorporating PFINs technique within GNNs, aiming to tackle data quality concerns. Additionally, a GNN framework was deployed on a recently developed extended heterogeneous biomedical knowledge graph called NeDRex. This graph encompasses various node and edge types, comprising 11 distinct types of nodes and 19 different types of edges, thus addressing the limitations associated with limited knowledge graphs. Furthermore, node features were integrated into the GNN model to improve its performance.

2.7 Drug Repurposing

Traditional drug discovery is a complex, lengthy, high-risk and costly process. On average, the process of bringing a new drug to the market typically requires 10 to 15 years, according to industry group PhRMA[13] [228], [229], [230]. It is also estimated that the overall mean expenditure involved in developing a novel pharmaceutical compound is $298 million to $2.3 billion, according to the latest annual report in 2023 from Deloitte[14]. Annually, over 90% of pharmaceutical compounds do not progress through the development process, primarily due to high toxicity or inefficacy [231]. The incorrect identification of the drug target is one of the most common factors of inefficacy, which leads to a high rate of attrition [232]. Attrition refers to the elimination of potential drug candidates for reasons such as ineffectiveness or safety concerns which lead to setbacks in the drug development pipeline [231], [232]. Despite rapid advances in technologies and genomics over time, the rate of newly approved drugs remains consistent with that of previous years; analyses of clinical trial data from 2010 to 2017 show that the success rate of drug development remains at 10%–15% [231], [232]. Only 5% of oncology drugs that undergo Phase I, Phase II, and Phase III clinical trials are

[13] https://phrma.org/policy-issues/Research-and-Development-Policy-Framework

[14] https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/life-sciences-health-care/deloitte-uk-seize-digital-momentum-rd-roi-2022.pdf

ultimately approved. Phase I includes testing the safety and the dosage of a drug in a small group of healthy individuals, while Phase II involves testing the efficacy of the drug in a larger population with the targeted disease. Phase III consists of large-scale clinical trials to confirm effectiveness and safety, providing data for regulatory approval and Phase IV involves post-marketing surveillance to monitor the performance and safety of the drug under real-world conditions [233]. There are a large number of rare diseases (diseases that affect < 1 in 2,000 people), making the traditional approach to drug development less cost-effective for these diseases[234].

Drug repositioning (or repurposing) encompasses various processes and approaches that aim to uncover alternative applications or find new uses for medications that are already available in the market [235]. This strategy is considered efficient, economical, and less risky than developing new drugs. The costs and risks of traditional drug discovery can be reduced. The repurposed drugs have often already been proven safe in humans through successful completion of Phase I, Phase II, and Phase III clinical trials, so the risk of failure due to safety reasons is low [9], [236]. The approximate cost of developing a new drug through a drug repurposing strategy is, on average, $300 million [228], [230]. This approach is beneficial in the case of rare diseases, which affect a small population, for whom there is limited interest in new drug development due to the risk and low economic benefits of investment in this area [234]. The drug discovery timeline is reduced as a significant portion of the preclinical testing and safety evaluation will have already been conducted (Figure 2.9) [229]. Due to these benefits, approximately 30% of recently approved drugs in the US have undergone the process of repositioning [237].

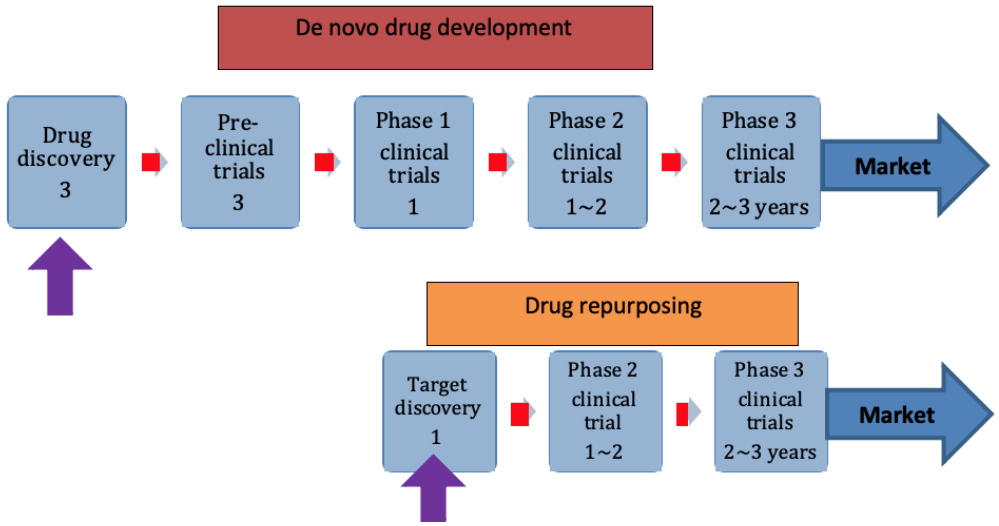


Figure 2.9. Comparison of traditional drug development and drug repurposing. The purple arrow indicates initiation time, and numbers show average duration (in years). Repurposing skips preclinical and phase 1 trials, as these are already completed. Adapted from [9], [238].

2.7.1 Methods for Computational Drug Repurposing

Computational approaches incur lower costs than experimental methods [7], [9], [236]. Their progress has been facilitated by the availability of biological data from diverse sources, such as genomics, proteomics, chemo-proteomics, and phenomics [134], [144]. These data sources have provided the foundation for constructing heterogeneous biological networks that can be analysed using network mining repurposing algorithms [12], [239], [240], leading to the discovery of new insights and connections [14], [45], [64]. The combination of these factors has contributed significantly to the advancement of computational drug repositioning approaches, which typically involve techniques such as data mining, machine learning, deep learning, and network analysis [45], [50], [64], [68], [241], [242]. *In silico* drug repositioning can be divided into drug-centric (finding new uses for existing drugs based on drug similarity [14], [70], [124]) and disease-centric (identifying effective drugs for a specific disease based on disease similarity [62], [98]).

Currently, a majority of research in *in silico* drug repositioning focuses on exploring the

interconnections among various biomedical entities such as drugs, diseases, genes, and proteins[12], [25], [26], [27], [64]. The drug repositioning domain stands to benefit from the implementation of advanced computational techniques that enable the detection of these intricate relationships, thereby facilitating the identification of novel therapeutic possibilities for existing drugs [64].

2.7.1.1 Machine learning-based approaches

Recently, numerous machine learning methods have been developed for drug repurposing applications, such as logistic regression (LR), SVM, random forests , neural networks , and deep learning . Machine learning models use biological data to predict novel relationships between drugs and diseases. Firstly, feature vectors are extracted from drugs' properties, such as drug fingerprint, chemical structures, and side effects, while phenotype data characterise diseases. Secondly, machine learning models are trained on these drug and disease feature vectors. Finally, associations between drugs and diseases can be predicted based on these models.

In their study, Gottileb *et al*. created a model called PREDICT, which relies on the similarity between drugs, diseases, and integrated similarity values. By employing logistic regression, the researchers utilised these features to predict drugs that are similar in their effects for similar diseases [243]. Liu *et al*. developed SPACE, a predictive model built on logistic regression, to determine a drug's therapeutic chemical class. This model integrates various data sources to achieve accurate predictions regarding a drug's classification for therapeutic purposes [156]. Napolitano *et al*. employed an SVM approach to predict therapeutic drug classes by considering molecular targets, drug chemical structures, and gene expression similarities. These features were combined to generate a unified drug similarity matrix. The similarity matrix was then used as a kernel for SVM classification [17]. In their study, Wang *et al.* utilised an SVM model that integrated molecular activities, drug chemical structures, and side effects. By combining these three types of data, they created a kernel function for the SVM classifier, resulting in a method that is more efficient than other approaches [244].

In most recent published studies, deep learning has emerged as a groundbreaking field within

machine learning, proving to be immensely influential in drug repurposing and development. This technique has seen significant advancements and is now at the forefront of innovative approaches in this domain [18], [132]. Previous studies showed that deep neural network (DNN)-based drug repurposing methods exhibit superior performance when compared to conventional machine learning techniques such as SVM or RF. DNN-based methods have been reported as more effective for predicting drug targets, showcasing their potential for enhanced accuracy and efficiency in this context [18], [132], [133].

Aliper *et al*. employed a deep learning approach to analyse gene expression profile data, aiming to predict the therapeutic types of drugs. Their findings revealed that DNNs outperformed SVM in this task. Other studies have indicated that multitask learning using deep learning-based approaches yields better results than traditional machine learning methods such as SVM or RF [15]. Previous works have also demonstrated the effectiveness of deep learning-based methods in accurately assessing drug toxicity [245].

Because of its ability to grasp intricate relationships between features and outcomes from vast datasets, deep learning, as a subdivision of machine learning, holds immense promise in the field of drug discovery [246]. Among the diverse range of deep-learning models, GNNs have been receiving growing attention in the drug discovery domain [246].This surge in interest is attributed to the fact that GNNs can present molecules and interactions in a clear and succinct manner [246]. The strength of GNNs is derived from their capacity to effectively model biological interactions using graph representations, harnessing the power of both graph modelling and deep learning techniques.

GNNs are gaining attention in drug discovery because they constantly exhibit better performance than traditional neural networks such as deep convolutional neural networks (DCNN). For example, Torng *et al.* developed a framework that leveraged graph convolutional neural networks (GCNN) to predict interactions between proteins and ligands across 102 protein targets. Their model outperformed other deep learning approaches, including 3DCNN and NN [71]. Lim and colleagues introduced a novel method for conducting structure-based virtual screening using GNNs with a focus on the 3D protein-ligand binding pose. Their approach surpassed the performance of CNNs in this

context [70]. Jiang *et al.* developed a predictive model for drug-target interactions, which involved creating two graphs representing the molecular structure and protein information. The researchers then employed two GNNs to extract relevant information from these graphs and make predictions of the affinity of binding between the ligand and the target protein [69].

2.7.1.2 Network-based approaches

Networks are a knowledge-based approach based on integrating diverse data sources to represent a wide range of interactions. These interactions include but are not limited to DDIs, DTIs, drug-disease, disease-disease, DGAs, PPIs, and transcriptional and signalling networks to infer unknowns or hidden drug-disease relationships [53]. For example, Cheng *et al.* incorporated network-based inference by assessing the similarities among drug-based, target-based, and network-based interactions. They utilised these similarities as a basis for predicting drug-target interactions [247]. Wu and colleagues employed a drug-disease heterogeneous network model to create a specialised network for identifying interconnected modules of drugs and diseases. This approach allowed them to extract valuable information concerning potential drug-disease indications [248]. Haeberle *et al*. developed a drug repurposing technique focused on cancer drugs. Their method involved exploring the potential off-target effects of these drugs on cancer cell signalling pathways [249].

In particular, semantic networks, a type of network-based approach, have a powerful capability to predict complicated relationships between biological entities from semantic meta-knowledge graphs [12], [27], [50]. A massive amount of semantic information in the biomedical literature enables the construction of semantic networks by integrating multiple data sources. Semantic network-based methods are used in many biomedical applications, including DGA identification, DTI prediction, and drug repurposing. In drug repurposing applications, semantic network-approaches include three steps (Figure 2.10): firstly, biological entity relationships are extracted from prior information. Secondly, semantics networks are constructed by adding this prior information. Finally, network mining algorithms are developed to uncover novel drug-disease indications.

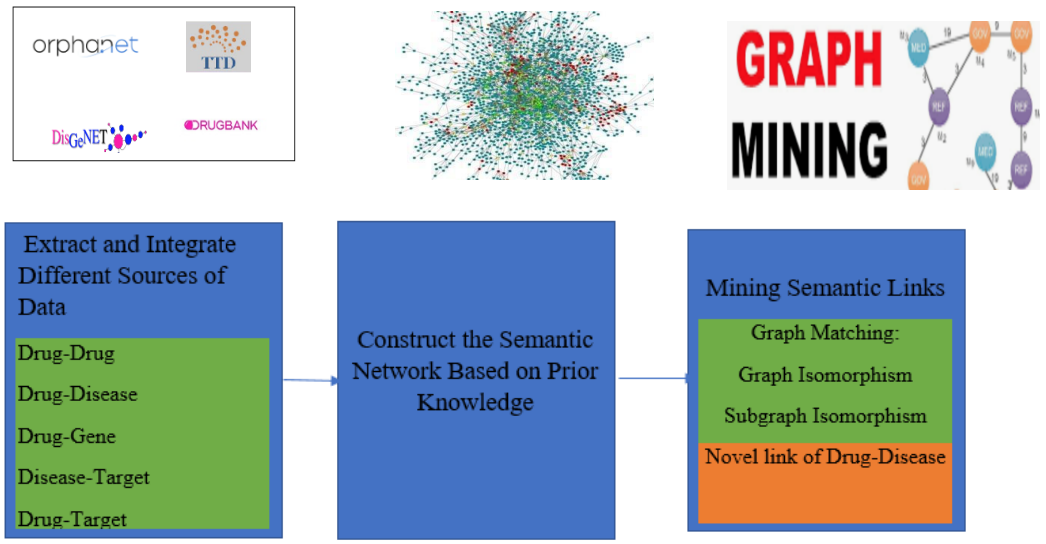


Figure 2.10. Workflow of semantic network inference.

Mullen *et al.* [12], [27] developed a data-driven approach for drug repurposing based on semantic network-based techniques. A Bayesian statistical method was used to generate DGA confidence scores according to prior knowledge. The confidence scores were used to rank the DGAs. The ranked DGAs were integrated with other biological relationships, such as *PPIs*, *protein-encoded-by genes*, *drug has indications*, and *drug-has-target,* to build a semantic integrated network for drug repurposing. A semantic network mining algorithm, including semantic subgraph matching, was developed to uncover novel drug-disease indications. The author concluded that nitrendipine, a potent calcium channel blocker, could be used to treat hypokalemic periodic paralysis. Skelton *et al*. employed a blend of network algorithms and manual curation to explore integrated semantic knowledge graphs. A probabilistic functional integrated network for PPIs was constructed to identify drug repurposing opportunities for COVID-19. Their results reported eight potential repurposing opportunities identified [50].

Chen *et al.* [250] developed a semantic network with different biological entities such as drugs, chemical compounds, targets, proteins, side effects, and diseases. A statistical model was applied to predict drug-target relationships. The topological structure and the semantics of the drug-target subgraph were considered. Similar drug-drug pairs across various disease categories could indicate potential repositioning opportunities. The suggested model inferred certain DTIs and potential repurposed drugs. An example includes the prediction that barbiturates, typically used to treat migraines, could also effectively alleviate insomnia,

supported by literature evidence. However, this approach is made challenging by the need to construct a semantic network by integrating heterogeneous data. Previous studies show that the accuracy of predicting biological entity relationships has improved when using semantics-based networks. However, constructing a semantic network by integrating multiple diverse data sources is challenging.

2.7.1.3 Text mining-based approaches

Several text-mining techniques have been developed to uncover novel drug-disease indications in the literature. For example, Cheng *et al.* [109] implemented a web-based text mining system for mining associations between different biomedical entities such as diseases, genes, proteins, and drugs. Diverse text mining techniques have been applied to a massive collection of biomedical databases to extract informative abstracts, paragraphs, or sentences. Li *et al.* [251] developed a model to extract clinical pharmacogenomics (PGx) gene-drug-disease associations using data from clinical trials. Relevant text in clinical trial records, which was obtained from ClinicalTrials.gov was determined and used as a dictionary to identify PGx entities. The co-occurrence of PGx concepts in each clinical trial was considered to identify gene-drug-disease associations. Then each clinical trial was indexed using its identified gene-drug-disease relationships. Therefore, given a PGx gene, the model can identify related diseases and drugs within the corresponding clinical trials. Likewise, given a PGx gene-drug or gene-disease pair, the model can return to clinical trials in which the PGx pair is or has been studied.

2.8 Summary and Conclusions

Network integration is a key method in computational drug repurposing, combining diverse data sources into heterogeneous networks. The quality of these networks, crucial for reliable inferences, depends heavily on the accuracy of the data. DGA networks are essential for uncovering disease mechanisms; however, existing DGA networks suffer from high false positive rates, particularly from high-throughput data, and the current scoring methods are often unreliable. This thesis addressed these issues by developing PFIN techniques to improve DGA network accuracy, reduce noise, and enhance reliability. These improved

networks were then integrated into broader biomedical networks for deep learning with GNNs, enabling more accurate predictions of drug-disease links and supporting drug repurposing efforts.

Chapter 3

Methods

This project aimed to research and develop computational approaches to highlighting drug repurposing opportunities based on the use of PFINs and GNNs focusing on DGAs. To achieve this aim, several existing data sources, tools, and methods were employed and integrated to build and assess the developed PFINs and GNNs. This chapter outlines the data sources, tools and computational methods applied in this project. The data sources that were used in this project will be introduced in Section 3.1 (see Chapters 4-6). The computational methods and evaluation techniques employed to build and evaluate the PFINs will be discussed in Sections 3.2.1 to 3.3.4 (see Chapters 4-5). The computational methods and the evaluation techniques developed to build and evaluate the biomedical literature mining approaches to extracting DGA experimental techniques will be introduced in Sections 3.5.1 to 3.5.3.1 (see Chapter 5). The computational methods and the evaluation techniques developed to build and evaluate the GNNs will be introduced in Sections 3.6.1 to 3.6.4 (see Chapter 6).

3.1 Data Sources

DisGeNET[15] (v7.0, 2020, SQL download) [141] was used as a source for DGAs for several reasons [141]. First, DisGeNET is a comprehensive DGA database that integrates associations from several sources that cover different biomedical aspects of diseases. Second, DisGeNET is one of the largest publicly available collections of genes and variants associated with human diseases. Third, DisGeNET integrates data from multiple, diverse data types, including expert-curated repositories, GWAS catalogues, animal models, and the scientific literature [141]. Fourth, DisGeNET is annotated using controlled vocabularies. The vocabulary used for diseases is the CUIs from the UMLS Metathesaurus[16]. The vocabulary used for genes is the official gene symbol, from the NCBI[17] [11]. This standardisation of identifiers eliminates the need for identifier mapping, as unified identifiers are already in

[15] https://www.disgenet.org/

[16] https://www.nlm.nih.gov/research/umls/index.html

[17] https://www.ncbi.nlm.nih.gov/

place. Finally, several original metrics are provided by the DisGeNET team to assist the prioritisation of genotype-phenotype relationships, such as DGA score, evidence level, and evidence index. Only manually curated DGAs were used. Curated data sources are meticulously reviewed and verified by human experts, ensuring that the data presented is accurate and reliable since these data sources undergo rigorous quality control processes to eliminate errors, duplicate, inconsistencies, and outdated data, reducing the noise in the data source. The quality control process results in high-quality data that can be trusted for various applications (Section 2.5).

OMIM[18] (downloaded March 4th, 2022) [88] was chosen as an external source for the DGA gold standard for many reasons: OMIM data is meticulously curated by human experts who evaluate, and synthesise data from scientific literature, research studies, and clinical reports [88]. OMIM surpasses other data sources in reliability because of its rigorous curation process, which involves meticulous evaluation and synthesis of scientific literature by expert geneticists. This process ensures high accuracy and comprehensive coverage of genetic and phenotypic relationships. This manual curation ensures that the data presented in OMIM are accurate and reliable. OMIM is also regularly updated to incorporate new DGAs. This incorporation ensures that the database reflects the latest and most up-to-date DGAs. OMIM also provides information on both the clinical manifestations of genetic disorders and on the underlying molecular basis which bridges the gap between clinical practice and molecular research. The OMIM database categorises disorders based on genetic associations, indicated by numbers in parentheses. Categories include: 1) disorders associated with a gene but with an unknown defect, 2) disorders mapped through linkage without identified mutations, 3) disorders with a known molecular basis and identified mutations, and 4) disorders caused by contiguous gene deletions or duplications. Only associations where the molecular basis of the disorder is known (category 3) and disorders with known susceptibility are recommended for use, the probability of using robust and accurate genetic data. Only DGAs for which the molecular basis of the disorder is known were included in this chapter (Section 2.5).

[18] https://www.omim.org/

Monogenic DGAs[19] [252] were also used as an external source for the DGA gold standard. A dataset of rare monogenic diseases, causative genes, and DGAs was used as a source for monogenic experimental studies. This dataset contains details on monogenic, rare diseases with known genetic causes. The dataset was downloaded from the figshare database[20] in a spreadsheet and nanopublication. In monogenic studies, a single gene mutation is responsible for the development of a disease. The clear link between a single gene and a monogenic disorder provides a solid basis for validation, which makes monogenic DGAs reliable benchmarks with a high degree of accuracy for genetic studies. A set of 4166 rare monogenic disorders, 3163 causative genes and 4292 associations was used as a source for monogenic experimental studies[21] [252].

BioGRID[22] [87] *(Homo sapiens* version 4.4.213, August 2022) database was also chosen as a source for the Gene-Gene Association (GGA) Gold Standard since it is manually curated and contains expert-reviewed interactions, ensuring high-quality data. BioGRID contains both physical and genetic PPIs from 28 experimental types including high throughput and low throughput. BioGRID is among the most comprehensive databases for *H. sapiens*, encompassing 22 different data types with source metadata linking back to the original publications. Curators furnish standardised, unique gene identifiers and provide NCBI IDs for all genes and proteins, eliminating the need for ID mapping. The database is regularly updated to incorporate new findings and maintain its relevance in PPIs. Only physical and low throughput interactions were used (for more details, see Section 2.5).

Reactome[23] [253] (version 81, 2022) database was also used as a source for the GGA Gold Standard. It covers a variety of biological pathways, such as metabolic, signalling, and regulatory pathways. The database provides information about each step in these pathways, which help to understand cellular processes. This database is curated by human experts to maintain the accuracy and reliability of the data. The curation process involves the review of scientific literature, which helps maintain the quality of the data. Reactome offers pathway information for multiple species, including humans, model organisms, and other commonly studied organisms. Reactome is regularly updated to include the latest research findings. It is

[19] https://www.nature.com/articles/s41597-021-00905-y
[20]https://springernature.figshare.com/articles/dataset/Metadata_record_for_A_resource_to_explore_the_discovery_of_rare_diseases_and_their_causative_genes/14140661
[21] https://doi.org/10.6084/m9.figshare.14140661
[22] https://thebiogrid.org/
[23] https://reactome.org/download-data

an open-source, manually curated, and peer-reviewed pathway database. Only human PPIs were used (see Section 2.5 for more details).

IntAct[24] [187] database was also used as a source for the GGA Gold Standard since it is manually curated by experts to ensure data quality. It is a resource for molecular interaction data, such as protein-protein, protein-DNA, and protein-RNA interactions, which are collected from literature curation and direct user submissions. It maintains quality control measures to ensure data reliability and works with other molecular interaction databases to integrate and ensure consistency. IntAct includes a wide variety of species, such as humans, rats, and mice. The platform offers easy-to-use search and analysis tools and ontology and annotation systems are employed for organisation. It is integrated with pathway databases and is updated regularly. *Homo sapiens* PPI interactions were used (For more details, see Sections 2.5).

EFO[25] [254], (v3.54.0, copyright 2014), Experimental Factor Ontology, was used as a resource for identifying the experimental techniques utilised in DGA studies. EFO was used to develop a dictionary comprising terms related to experimental techniques, facilitating the text-mining process for extracting such information from DGA literature. EFO is a resource that systematically describes experimental variables across various biological domains, encompassing several biological ontologies such as anatomical structures, disease classifications, and chemical compounds. By standardising the representation of experimental factors, EFO facilitates data integration, meta-analysis, and cross-study comparisons, enabling researchers to derive new insights from aggregated data. EFO enables the discovery of relationships between experimental factors and biological outcomes across diverse datasets and links experimental variables to biological entities, such as genes, proteins, and phenotypes. EFO was downloaded in Web Ontology Language (*OWL*) format. The ontology was explored using the ontology editing and visualisation tool *Protégé*[26]. The ontology parser, *Owlready2*[27], was used to parse the ontology terms from the *OWL* file.

[24] https://www.ebi.ac.uk/intact/home
[25] https://www.ebi.ac.uk/efo/
[26] https://protege.stanford.edu/
[27] https://owlready2.readthedocs.io/en/latest/

EDAM[28] [255], (version 1.25) Ontology of bioscientific data analysis and data management, was also used as a source for DGA experimental techniques. EDAM is a structured vocabulary designed to describe concepts and operations related to bioinformatics data analysis and management. EDAM covers a wide range of bioinformatics topics, including data types, formats, operations, algorithms, software tools, databases, and workflows used in the analysis, management, and interpretation of biological data. EDAM offers a structured set of concepts, complete with preferred terms, synonyms, and definitions. EDAM was also used to construct the dictionary containing DGA experimental terms. *EDAM* was downloaded in *OWL* format. The *Owlready* parser was employed to extract ontology terms from the *OWL* file.

PubMed[29] [256] was used to extract the abstract from articles in DGA literature. PubMed is an online database maintained by the NCBI. It provides access to a vast collection of citations and abstracts from biomedical literature, including articles from scientific journals, conference proceedings, and more. To retrieve DGA-related abstracts, the *Entrez Programming Utilities (E-utilities)* API was utilised for programmatic access to PubMed. The search was specifically tailored to articles associated with the DisGeNET database by leveraging PubMed IDs directly sourced from DisGeNET's curated data repositories. This method ensured that only high-quality, manually curated studies were included, while filtering out data from animal models and text mining sources. By using these PubMed IDs, the search criteria were precisely restricted to abstracts linked to curated data sources within DisGeNET, with a focus on DGAs.

The *abstracts* were retrieved in *XML* format using Python's *Bio.Entrez module* from the *Biopython* library. During the retrieval process, metadata such as publication date, journal name, and article title were also captured alongside the *abstracts*. After retrieval, the *XML* file was parsed to extract the *abstracts*, which were then stored in a structured text file format. This file was used as input for subsequent text mining procedures aimed at identifying experimental techniques described in the DGA studies. By matching terms from a predefined dictionary of DGA experimental techniques with the text in the *abstracts*, these procedures sought to systematically extract and analyse experimental methods mentioned in the studies.

[28] https://edamontology.org/page…
[29] https://pubmed.ncbi.nlm.nih.gov/

PMC[30] [256] was used to retrieve the full articles in DGA literature to extract the method section from the DGA experimental studies. PMC stands for PubMed Central, which is a free digital repository that archives and provides access to full-text scholarly articles in the biomedical and life sciences field. To retrieve these articles, PMIDs sourced from the DisGeNET database were utilised. The PMIDs were used to identify and access the corresponding full-text articles in PMC. *The E-utilities* API was employed to search for and download the full-text articles. The *requests* library in Python facilitated interaction with the API to fetch the XML files of the articles. Upon obtaining the XML files, the *xml.etree.ElementTree* library was used to parse the XML structure, and *BeautifulSoup* function was employed to extract the "*Methods*" sections from the articles. These sections are important for understanding the experimental techniques used in the DGA studies. The extracted method sections were then stored in a structured text file format. This file served as input for subsequent text mining procedures aimed at identifying experimental techniques described in the DGA studies. By matching terms from a predefined dictionary of DGA experimental techniques with the text in the methods sections, these procedures systematically extracted and analysed the experimental methods mentioned in the studies. The methods section provides more detailed information about the methodologies compared to the abstracts.

NeDRexDB[31] [48] was used to construct an integrated heterogeneous network for drug repurposing. NedRexDB integrated data from various biomedical databases. Combining multiple databases allows for the creation of diverse networks that represent different types of biomedical entities (e.g., diseases, genes, drugs) and the associations between them. These networks can be used in drug repurposing applications (For more details, see Section 2.5).

RepoDB[32] [257] is a standard database for drug repurposing. RepoDB contains approved and failed drug-indication pairs. The approved indications were drawn from DrugCentral and failed indications were drawn from the American Association of Clinical Trials Database. RepoDB was used as a benchmark for the computational repurposing approaches employed in this thesis.

[30] https://www.ncbi.nlm.nih.gov/pmc/
[31] https://nedrex.net/tutorial/intro.html
[32] https://repodb.net/

3.2 Network Integration

3.2.1 Confidence Scoring

The confidence score was calculated by scoring the datasets against the Gold Standard using the Bayesian statistics approach developed by Lee and colleagues (see Section 2.3.3) [55], which calculates a log-likelihood score (LLS) for each dataset (3.1):

$$lls^{L}(E) = \ln\left(\frac{P(L|E)/\neg P(L|E)}{P(L)/\neg P(L)}\right) \qquad \textbf{3.1}$$

where, $P(L|E)$ and $\neg P(L|E)$ represent the frequencies of associations L in dataset E observed and not observed in the gold standard, respectively, and $P(L)$ and $\neg P(L)$ represent the prior expectation of associations observed and not observed in the gold standard, respectively. Datasets that did not have a positive score or had no score were discarded. Datasets scoring infinity were given a score of $\ulcorner h \urcorner + 1$, where $\ulcorner h \urcorner$ is the highest *LLS* score.

3.2.2 Network Integration

The networks were produced by integrating the confidence scores using the weighted sum (see Section 2.3.3) introduced by Lee and colleagues [55]. This method integrates the datasets in the order of their confidence scores, giving a higher weighting to datasets with higher confidence while allowing for dependencies between them (3.2):

$$WS = \sum_{i=1}^{n} \frac{L_i}{D^{(i-1)}} \qquad \textbf{3.2}$$

Where L_1 is the highest confidence score, and L_n *is* the lowest confidence score of a set of n datasets. Division of the score by the D parameter means that, while the highest score is integrated unchanged, subsequent weights are progressively down-weighted. With a D value of one, the integration is a simple sum of the scores. With a higher D value, lower-ranking scores contribute little or nothing to the integration, down-weighting low-quality data and improving network performance. The D values were chosen based on the one that provided the highest network performance regarding link prediction and clustering analysis (see

Section 4.3.2). In the resulting network, highly weighted edges have high confidence (see Sections 4.3.2.2 and 5.3.1.1).

3.3 Network Visualisation and Evaluation

3.3.1 Visualisation and Topological Analysis

Networks were visualised in Cytoscape[33] Version 3.8.2 [258]. The network Analyser plugin[34] version 3.8.2 for Cytoscape was used to calculate the topological statistics of networks.

3.3.2 Network Clustering

The networks were clustered using the weighted Markov (MCL)[35] [259]. The MCL algorithm was chosen for its efficiency, robustness, and flexibility through the inflation parameter in clustering large networks. The MCL algorithm divides the network into non-overlapping clusters based on topological structures and edge weights. The inflation parameter influences the rate at which clusters grow and merge during the clustering process. A higher inflation value typically results in more granular clusters, whereas a lower inflation value leads to larger, more inclusive clusters. The inflation value was chosen based on the network performance that provides the highest average of cluster connectedness in each case (Section 4.3.3.2 and Section 5.3.1.2). The clusters were evaluated for their connectedness and cohesion using various similarity methods described in the next section.

3.3.3 Network Clusters Evaluation

After clustering the networks, the clusters were analysed for their connectedness and coherence using several evaluation techniques. This section outlines the methodologies employed to assess the clusters generated by the MCL clustering algorithm, including the DGA clusters, disease clusters, and gene clusters.

[33] https://cytoscape.org/

[34] https://manual.cytoscape.org/en/3.7.2/Network_Analyzer.html#:~:text=NetworkAnalyzer%20computes%20a%20comprehensive%20set,as%20the%20characteristic%20path%20length.

[35] https://micans.org/mcl/

3.3.3.1 Analysis of DGA Clusters

The DGA clusters were analysed to evaluate the cohesion of the network clusters by examining diseases and their associated genes within each cluster. The hypothesis was that diseases closely associated with their related genes, indicated by a high confidence score (high edge weight), would be clustered together in a single cluster. To test this hypothesis, the average cohesion of the clusters was computed. The related genes of a disease, *D(G)*, were defined as the average of its associated genes [260]:

$$D(G) = \frac{1}{n}\sum_{i=1}^{m} g_i \qquad \textbf{3.3}$$

Where *n* is the total number of disease genes in the whole network, *m* is the total number of disease genes in a cluster, and g is the disease genes

The related genes of a cluster *C(G)* are defined as the average of related genes for the cluster:

$$C(G) = \frac{1}{n}\sum_{i=1}^{n} D_i(G) \qquad \textbf{3.4}$$

Where *n* is the total number of diseases in the cluster, and *D(G)* is the average of related genes for each disease.

Finally, the average of related genes for the whole network is defined as:

$$N(G) = \frac{1}{n}\sum_{i=1}^{n} C_i(G) \qquad \textbf{3.5}$$

Where *n* is the total number of clusters in the network, and *C(G)* is the average of related genes for each cluster (see Section 4.3.3.2).

3.3.3.2 Analysis of Disease Clusters

The disease clusters were analysed using four similarity measures: disease semantic similarity, genetic similarity, gene semantic similarity and disease treatment similarity. All

these methods were adapted from [260]. Disease cluster analysis was used to assess network performances (see Section 5.3.2.3).

A. Correlation analysis of disease clusters with disease genetic network

Correlation analysis was conducted between the disease clusters and the disease genetic network using the method introduced in [260]. A weighted human genetic network (HDN) was constructed using GWASCAT[36] [222] and GWASDB[37] [190]. In this network, diseases are represented as nodes and an edge is added between two diseases if they share common genes. The weight of each edge indicates the number of common genes between the connected diseases. Using the genetic information in HDN, the pairwise shared genes were calculated within each disease cluster of the disease-disease similarity network obtained from the weighted MCL clustering algorithm. The shared genes of a cluster, denoted as *C(G)*, are defined as the average of pairwise shared genes, as shown in Equation 3.6:

$$C(G) = \frac{1}{m} \sum_{d1 \neq d2,\ d1,d2 \epsilon D} g(d1, d2) \qquad \textbf{3.6}$$

Where d_1, d_2 are pairwise diseases in the cluster, $g(d1,d2)$ is shared genes between $d1$ and $d2$, D is the disease node set in each cluster and m is the number of total pairwise diseases in each cluster.

B. Correlation analysis of disease clusters with disease treatment network

To compute the shared drugs of a cluster of the disease-disease similarity network, first, a DrDI network based on DrugCentral[38] (downloaded database dump 11/01/2022, Postgres v14.4) [220] and CTD[39] (downloaded May 30 2022) [261] was constructed. In the DDN network, diseases are represented as nodes and an edge is added between two diseases if they share common drugs. The weight of each edge indicates the number of common drugs between the connected diseases. Pairwise shared drugs, in a cluster of disease-disease similarity network, were calculated based on the DrDIN and share drugs of a cluster *C(D)* are defined as the average of pairwise shared drugs in Equation 3.7:

[36] https://www.ebi.ac.uk/gwas/
[37] https://maayanlab.cloud/Harmonizome/resource/GWASdb
[38] https://drugcentral.org/
[39] https://ctdbase.org/

$$C(D) = \frac{1}{m} \sum_{d1 \neq d2,\ d1,d2 \epsilon D} g(d1, d2) \quad \textbf{3.7}$$

where $d1$,$d2$ are pairwise diseases in a cluster, $d(d1,d2)$ are shared drugs between $d1$ and $d2$, D is the disease node set in each cluster and m is the number of total pairwise diseases in each cluster.

C. Correlation analysis of disease clusters with disease semantic network

Disease ontology[40] (DO) [262] was used for computing the disease semantic similarity of pairwise disease in a cluster of the disease-disease similarity network. The semantic similarity between pairs of diseases, $sim(d1, d2)$, was computed using the *DOSE* (Disease Ontology Semantic and Enrichment analysis, version 3.28.2)[41] [263], Bioconductor R package version 3.18, which computes semantic similarity among DO terms. Semantic similarity of a cluster $C(SIM)$ is computed as the average of pairwise disease semantic similarity as shown in Equation 3.8

$$C(SIM) = \frac{1}{m} \sum_{d1 \neq d2,\ d1,d2 \epsilon D} sim(d1, d2) \quad \textbf{3.8}$$

where $d1$, $d2$ are pairwise diseases in a cluster, $sim(d1,d2)$ is the semantic similarity between $d1$ and $d2$, D is the disease node set in each cluster and m is the number of total pairwise diseases in each cluster.

D. Correlation analysis of disease clusters with gene semantic network

To compute the gene semantic similarity within disease clusters, the set of genes associated with the diseases contained in each cluster of the disease-disease similarity network was identified (see Section 5.3.2.3). Gene Ontology (GO)[42] [264] was then utilised to compute the semantic similarity between pairs of genes within the related genes set for each cluster. The semantic similarity between pairs of genes, $sim(g1, g2)$, was computed using the R Bioconductor package(version 3.18)-GOSemSim (version 2.28.1) [265], for semantic

[40] https://disease-ontology.org/
[41] https://bioconductor.org/packages/release/bioc/html/DOSE.html
[42] https://geneontology.org/

similarity computation among GO terms. Semantic similarity of a cluster $C(SIM)$ is computed as the average of pairwise gene semantic similarity in Equation 3.9:

$$C(SIM) = \frac{1}{m} \sum_{g1 \neq g2, g1, g2 \epsilon G} sim(g1, g2) \quad \textbf{3.9}$$

where $g1$,$g2$ are pairwise common genes in a cluster, $sim(g1,g2)$ is the semantic similarity between *g1* and *g2*, G is the disease node set in each cluster and *m* is the number of total pairwise genes in each cluster (see Section 5.3.2.3).

E. Comparing Disease Clusters with Random Networks

The disease clusters were compared with random networks in terms of shared genes, shared drugs, and disease semantic and gene semantic similarities. A random network was constructed to have the same network structure as the original, with the preservation of node numbers, edge numbers, and network degree distribution in the original network [266], [267]. The nodes, the edges and the edge weights were randomly shuffled. The configuration model was used to create the random network that has the same structure as the original network. The node degrees were preserved by assigning and randomly pairing stubs to form edges between two sets of nodes: diseases and genes. Edge weights were maintained by shuffling them among the newly formed edges. One hundred random networks were generated and average shared genes, shared drugs, disease semantic, and gene semantic similarities within each cluster were computed for each network.

3.3.3.3 Analysis of Gene Clusters

The gene clusters were evaluated using the functional enrichment analysis [268]. The functional enrichment analysis aims to identify GO terms associated with MF, BP, and CC contexts, which were significantly overrepresented by the genes within a cluster of the gene-gene interaction network (see Section 5.3.2.2). This analysis utilised the R package *BioStats*[43] version 3.18 [269]. The statistical significance of a GO term within a cluster was determined using a hypergeometric test, which evaluated its overrepresentation compared to what would be expected by chance. A functionally enriched cluster indicates that the observed number of genes annotated with a particular function (GO term) exceeds the

[43] https://bioconductor.org/packages/release/bioc/html/GOstats.html

expected number based on the reference list of all human genes. The expected value for a function is calculated as the number of genes with that specific function within the given cluster, relative to the entire list that exists in the curated DGA data sources in DisGeNET [141].

The gene clusters were evaluated using functional homogeneity, heterogeneity and specificity measures introduced by Kaalia and colleagues [135]. The p-value was determined by considering the odds ratio and the threshold of 0.0001 gave the maximum modularity with the highest odds ratio of the GO terms. The enriched functions were then ranked based on their significance levels. Subsequently, the enriched functions across all clusters were compiled into a set denoted as *F*.

The *functional homogeneity* (*H)* of a cluster is defined as the homogeneity of maximally enriched function in the cluster [135]. Functional homogeneity measures the functional coherence of the clusters while functional heterogeneity indicates how exclusive the clusters are for the function across all gene clusters [135]. The functional specificity value measures how exclusively the cluster is enriched by the specific biological function [135]. To measure the *functional homogeneity* of a cluster concerning a specific function, the proportion of genes annotated by that function within the total gene count of the cluster was calculated. The homogeneity of a function $f \in F$ within a cluster was determined by the fraction of clusters where the function $f \in F$ was found to be enriched. The functional homogeneity is computed in equation 3.10:

$$homogeneity = \frac{fn}{N} \quad \textbf{3.10}$$

where f_n is the number of genes annotated by the function *f* and *N* is the total number of genes in the cluster. The functional heterogeneity of a cluster is defined as the proportion of the cluster where the function $\boldsymbol{f \in F}$ is enriched. That is,

$$heterogeneity = \frac{Kf}{K} \quad \textbf{3.11}$$

Where k_f is the number of clusters enriched with the function *f* and *K* is the total number of clusters. Finally, the *functional specificity* for an enriched function is defined as follows:

$$Specificity = homogeneity + \frac{1}{heterogeneity} \quad \textbf{3.12}$$

3.3.4 Link Prediction

The link prediction technique was employed to assess the performance of the networks (Section 4.3.3.1 and 5.3.1.2). The Jaccard Index (JI) was specifically utilised for predicting new links between diseases and genes [270]. JI considers the degree of nodes in a network and represents a normalised version of the Common Neighbours algorithm [271]. The JI was selected because it yielded better performance in predicting network links compared to the other four neighbourhood algorithms. Additionally, JI takes into account the degree of nodes within a network and provides a normalised version of the common neighbours algorithm [271]. The normalisation ensures that similarity scores are based on shared neighbours rather than solely on node degree. Since the networks are weighted and bipartite, the modified version of the JI, which is tailored for a weighted bipartite network, was employed using the following formula [270]:

$$JI(x,y) = \sum_{z \in s(x) \cap s(y)} \frac{max\, i \in (T(x) \cap T(y))\{\omega(x,i)+\omega(z,i)\}+\omega(y,z)}{\sum_{a \in S(x)} max\, j \in (\Gamma(x) \cap \Gamma(a))\{\omega(x,j)+\omega(a,j)\}+ \sum_{b \in S(y)} \omega(b,y)} \quad \textbf{3.13}$$

Where *s(x)* represents *x*'s second neighbours where *x* is a gene node. This leads to a set of genes. *s(y)* represents y's direct neighbours where *y* is a disease node. This leads to a set of genes. *Γ (x)* is the set of neighbours of a node *x*. *Γ (y)* is the set of neighbours of a node *y* in the network, and *(x, y)* is the link weight between nodes *x* and *y*. Since this network is bipartite, nodes *x* and *z* are at the same part of the network, and there cannot be any direct

link between *x* and *z*. Therefore, the 2-length path from *x* to *z* through their common neighbour *i*, which yields the maximum sum of weights, is used instead of ω *(x, z)*.

Ten-fold cross-validation was employed to assess the network performance [272], [273]. This method was chosen because the link prediction task can be computationally intensive, particularly for bipartite weighted networks with large sizes. The network was randomly partitioned into ten subsets, each containing approximately an equal number of DGAs. In each iteration, the JI was applied to nine subsets while using the remaining subset as the test data to validate the accuracy of the predicted links. This process was repeated ten times, and the sensitivity and specificity calculated from each iteration were averaged to produce a final score [272]. The Receiver-operating characteristic curve was used to evaluate the performance network link prediction [274]. The standard error (SE) of the AUC was estimated using the *aucSE()* function from version 1.0.0 of the *auctestr* package in R, which leverages its equivalence to the *Wilcoxon* statistic.

3.4 Confidence Score for Disease-Gene Associations Developed by DisGeNET

DisGeNET developed a confidence score for DGAs (see Sections 2.3 and 2.4.1.2). The DisGeNET score (*S*) for DGAs is determined by a formula that includes contributions from various sources[141]:

$$S=C+M+I+L \quad \textbf{3.14}$$

Where:

$$C = \begin{cases} 0.6 & if\ N_{sources_i} > 2 \\ 0.5 & if\ N_{sources_i} = 2 \\ 0.3 & if\ N_{sources_i} = 1 \\ 0 & otherwise \end{cases}$$

Where Nsourcesi represents the number of curated sources supporting a DGA, i ∈ CGI, ClinGen, Genomic England, CTD, PsyGeNET, Orphanet, and UniProt.

$$M = \begin{cases} 0.2 & if\ N_{sources_j} > 0 \\ 0 & otherwise \end{cases}$$

Where *M* accounts for the support from animal models, and j ∈ RGD, MGD, and CTD.

$$I \quad = \begin{cases} 0.1 & if\ N_{sources_k} > 0 \\ 0 & otherwise \end{cases}$$

where *I* represent the number of inferred sources supporting a DGA and k ∈ HPO, ClinVar, GWASCat, and GWASDB.

$$L \quad = \begin{cases} 0.1 & if\ N_{pubs} > 9 \\ N_{pubs} * 0.01 & if\ N_{pubs} < 9 \end{cases}$$

where *L* represents the number of publications supporting aDGA and Npubs ∈ LHGDN and BeFree.

3.5 Text Mining Techniques

This section outlines the text mining techniques used to extract the experimental techniques employed in the DGA experimental studies from biomedical literature obtained from PubMed and PMC (Section 5.3.3.2).

3.5.1 Parsing Abstract and Method Sections from DGA Articles

The *XML* files of DGA articles were retrieved from PubMed and PMC using the Entrez Programming Utilities (*EUtils*)[44] provided by NCBI. The *efetch* function from the *Bio.Entrez* module, part of the *Biopython* package, was utilised to download the articles in *XML* format. The process involved setting the *Entrez.email* parameter to identify the user to NCBI, and defining a function to fetch *XML* data by calling *Entrez.efetch* with a batch of PMIDs. The list of PMIDs was divided into smaller batches to efficiently manage large datasets. For each batch, the *efetch* function was called to retrieve the *XML* data, which was then processed using *BeautifulSoup* to parse the *XML*. The *Beautiful Soup*[45] Python version 14.12.0 was used to parse *XML* files of DGA articles retrieved from PubMed and PMC. From these files, both

[44] https://www.ncbi.nlm.nih.gov/books/NBK25500/
[45] https://beautiful-soup-4.readthedocs.io/en/latest/

the abstract and method sections of each article were extracted. The method section provides more details about the experimental techniques employed in the study to explore DGAs compared to the Abstract section.

3.5.2 Parsing EFO and EDAM Ontologies

Experimental terms from the EFO and EDAM ontologies were extracted to construct a DGA experiment dictionary. The *Protege*[46] version 5.6.1 ontology browser was used to identify the relevant branches related to DGA experimental techniques within the ontologies containing these experimental terms. Experimental terms from the EFO and EDAM *owl* files were parsed using *Owlready2*[47] `version 0.46` ontology parser. These ontologies include equivalent terms with different names. For instance, *'polymerase chain reaction*' appears under different labels such as '*PCR*' in *EFO* and '*PCR* experiment' in *EDAM*. Synonym matching between the *EDAM* and *EFO* ontologies was carried out by first extracting lists of terms and their associated synonyms from both ontologies. Synonyms were identified in the metadata of the ontologies using properties like *oboInOwl:hasExactSynonym* in both EFO and EDAM. Once the synonym lists were compiled, they were compared directly between the two ontologies. If a synonym in *EDAM* exactly matched a synonym or primary term in *EFO*, these terms were mapped as equivalents. For example, the term '*PCR*' in *EFO* and *'PCR experiment'* in *EDAM*, where *'polymerase chain reaction'* is listed as a synonym, were mapped as equivalents due to the shared synonym. To ensure accuracy and relevance, a manual review of a sample of these mappings was performed, which is critical for constructing a reliable dictionary.

3.5.3 Clustering DGA Experimental Studies Based on Their Experimental Techniques

Experimental techniques employed in DGA studies were extracted from DGA articles using terms combined from both the EFO and the EDAM. A binary similarity matrix was constructed, where rows represent studies and columns represent experimental terms. A value of one indicates the presence of a term in a study, while zero indicates its absence. Equivalent terms with different expressions were unified into standardised synonyms using the synonym list (Section 3.1). The hierarchical clustering dendrogram [275] was then applied to group these studies based on their experimental terms. The hierarchical clustering was performed

[46] https://protege.stanford.edu/
[47] https://owlready2.readthedocs.io/en/latest/

using the *Scipy* package (version 1.11.4). Hamming distance was chosen to measure the distance between two vectors of experimental terms due to its suitability for binary clustering, aligning with the binary representation of our data [276]. The distance between clusters was computed using Ward's method due to its highest cophenetic correlation coefficient (CPCC) [277]. The distance threshold was selected based on maximising the Silhouette Score [278], effectively minimising the distance within clusters while maximising the distance between them.

3.5.3.1 Hierarchical Clustering Dendrogram Validation

The CPCC was used to choose the optimal linkage method for the hierarchical clustering [277], [279]. The CPCC measures the quality of a hierarchical cluster by assessing how well the dendrogram preserves the original pairwise distances between data points. A high CPCC indicates that the dendrogram accurately represents these distances, ensuring a more reliable clustering outcome. Conversely, a low CPCC suggests that the dendrogram poorly represents the original distances, leading to less reliable and potentially misleading clustering results. The Silhouette score was used to choose the optimal distance threshold [278]. Silhouette score is used to calculate the goodness of a clustering technique. The silhouette score evaluates clustering quality by measuring both cohesion (how close points are within a cluster) and separation (how distinct clusters are from each other). It ranges from -1 to 1, with higher scores indicating well-defined and well-separated clusters. A high silhouette score signifies accurate and meaningful clustering.

3.6 Graph Neural Network

This section outlines the techniques used in Chapter 6 to build and evaluate the GNN models for link prediction in drug repurposing applications (Section 6.3.3).

3.6.1 Encoder

The encoder was used to generate node embeddings. The encoder included two layers of GraphSAGE (Section 2.6) [280]. The GraphSAGE model was chosen because it is effective in handling large graphs and achieving good generalisation performance for inductive learning on graph-structured data. Since there are two of these layers, final embeddings are generated using the 2-hop neighbourhood as shown in equation 3.15:

$$h_v^k = \sigma(W^k\, MEAN(\{h_v^{k-1}\} \cup \{h_u^{k-1}, \forall\, u \in N(v)\})) \quad \textbf{3.15}$$

Where h_v^k is the embedding of node *v* at layer *k*, W^k is the weights learnable matrix of a layer, *N(v)* the neighbour sampler function, h_u^k *is* the embedding of the neighbour nodes of *v,* and $\sigma(x)$ the nonlinearity function, *k* is the number of layers, the aggregator function is the *MEAN* operator, the aggregated neighbourhood vector *AGG*= $\{h_u^{k-1}, \forall\, u \in N(v)\}$ and the updated embedding is $\sigma(W^k\, MEAN(\{h_v^{k-1}\} \cup AGG$).

3.6.2 Decoder

The decoder was used to calculate the probability of an edge existing between a disease and a drug. To compute it, the sigmoid function was applied to the dot product of the embeddings, represented in Equation 3.16:

$$PS = \sigma(X.Y^T) \quad \textbf{3.16}$$

Where *PS* is the probability score of the predicted edge between a disorder and a drug, $\boldsymbol{X}$ is the disorder embedding vector, $\boldsymbol{Y^T}$ is the transpose of the drug embedding vector and $\sigma(x)$ the sigmoid function.

3.6.3 Loss Function

During the training phase, the models' parameters were adjusted using the Binary Cross Entropy with Logit Loss (BCELogitLoss) as the loss function. The BCELogitLoss was chosen because the drug repurposing task was framed as a link prediction problem between drugs and diseases. Since this problem is a binary classification task, where it is predicted whether a link exists between a drug and a disease, the Logit loss function is well-suited. The loss function is computed as it can be seen in equation 3.17:

$$\begin{aligned} &l(x,y) = L = \{l_1, \ldots, l_N\}^T, \\ &l_n = -[y_n \bullet log\sigma(x_n) + (1 - y_n) \bullet log(1 - \sigma(x_n))] \end{aligned} \quad \textbf{3.17}$$

x_n represents the prediction made by the model, while y_n indicates the true label, which determines whether the edge exists or not.

3.6.4 Evaluation Metrics

The evaluation metrics for assessing the GNN model were the area under the receiver operating characteristic (ROC) curve (AUCROC) [274], the area under the precision-recall (PR) curve (AUCPR) [281], and the accuracy (ACC). AUCROC evaluates the trade-off between the true positive rate (TPR) (3.18), also known as *Sensitivity*, and the false positive rate (FPR) (3.19), calculated as $1-$*Specificity*, across different decision thresholds. TPR, or sensitivity, measures the proportion of true positives out of all actual positives, while FPR measures the proportion of false positives out of all actual negatives. AUCPR quantifies the trade-off between precision and recall across various decision thresholds. Precision (3.20), also known as positive predictive value, assesses the proportion of true positives out of all predicted positives. Accuracy measures the proportion of correctly classified instances among all instances (3.21). The formulas for these metrics are as follows:

$$TPR = Recall = Sensitivity = \frac{TP}{TP+TN} \quad \textbf{3.18}$$

$$FPR = 1 - Specificity = \frac{FP}{TN+FP} \quad \textbf{3.19}$$

$$Precision = \frac{TP}{TP+FP} \quad \textbf{3.20}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad \textbf{3.21}$$

Where TP, TN, FP, FN means true positive, true negative, false positive and false negative respectively.

3.7 Identifier Standardisation

In DisGeNET, genes are annotated with the official gene symbols from the NCBI, while diseases are characterised using Concept Unique Identifiers (CUIs) sourced from the Unified Medical Language System® (UMLS) Metathesaurus® (version UMLS 2019AA). Given that various data sources of gene-disease associations may employ different disease vocabularies, the initial step before constructing disease-gene PFINs involves establishing a standardised representation for both genes and diseases. Determining an optimal disease representation is

not a straightforward endeavour due to the existence of numerous competing disease classifications and ontologies. These classifications serve diverse purposes and often lack integration, resulting in inconsistencies across systems.

For disease identifier mapping, OMIM disease identifiers were mapped to UMLS. Different identifier mappers were employed for identifier mapping, including DO cross-referencing[48] [262], Mondo disease ontology equivalent[49] [282], and the UMLS Metathesaurus [283]. The UMLS disease identifiers were mapped to DO identifiers using DO cross-referencing. Similarly, to ensure consistency in gene representation, gene identifiers, in the monogenic data source, were mapped to NCBI gene identifiers. Ensembl gene identifiers were mapped to NCBI gene identifiers using the cross references from HGNC[50]. Additionally, gene identifiers, from the Reactome database, were mapped to NCBI gene identifiers using the UniProt Mapping service[51].

[48] https://disease-ontology.org/about/cross_references

[49] https://mondo.monarchinitiative.org/#:~:text=the%20mondo%2Dwith%2Dequivalent%20edition,%2C%20HP%2C%20RO%2C%20NCBITaxon.

[50] https://www.genenames.org/

[51] https://www.uniprot.org/id-mapping

Chapter 4

Investigating the Applicability of Probabilistic Functional Integrated Networks to Disease-Gene Networks

4.1 Introduction

Existing DGA networks contain high level of noise, and current methods for DGA network integration fail to produce accurate DGA networks [104], [211], [284], [285], [286], [287], [288] (Section 2.4.1.2). PFINs offer a powerful technique for reducing noise during network integration (Section 2.3.3) [55], [56], [59], [86], [90], [289], [290]. The main aim of this study was to investigate the applicability of the PFIN approach to DGA networks. While PFINs have been widely applied to PPI networks, several challenges arise when applying this approach to bipartite DGA networks (for a detailed discussion of these challenges, see Section 2.4.1.2). Therefore, to test the applicability of the PFIN approach for predicting associations between genes and diseases, several tasks needed to be investigated:

1. To research and develop strategies to identify appropriate gold standard data for DGAs.
2. To research and develop strategies to identify appropriate individual datasets to represent DGAs to be scored against the Gold Standard data and then integrated into a network.
3. To research and develop appropriate robust evaluation techniques for assessing the performance of resulting PFIN.

To build PFINs in DGAs, two main components need to be identified: gold standard data and individual datasets. In Section 4.3.1 of this chapter, various strategies for identifying gold standard data and individual datasets for DGAs are presented, as outlined in Objective one (Section 1.2). Next, the focus is placed on network integration in Section 4.3.2, including the processes of dataset scoring and integration. Finally, the evaluation of the networks constructed in Section 4.3.2 is described in Section 4.3.3, addressing Objective 2 (Section 1.2). These sections collectively aim to establish a comprehensive framework for developing and assessing robust DGA PFINs.

4.2 Source Data

DisGeNET A large number of data sources has been developed to store DGAs (Section 2.5). In this chapter, DisGeNET 2021-v7.0 was chosen as the data source (Section 3.1) [141]. In the work reported in this chapter, only curated data sources from DisGeNET were used [116]. Figure 4.1 shows the numbers of unique genes, diseases, DGAs, and evidence in each curated data source. "Evidence" refers to the PubMed ID[52] that identifies the individual experimental study which generated the corresponding DGA. The CGI does not list PubMed IDs.

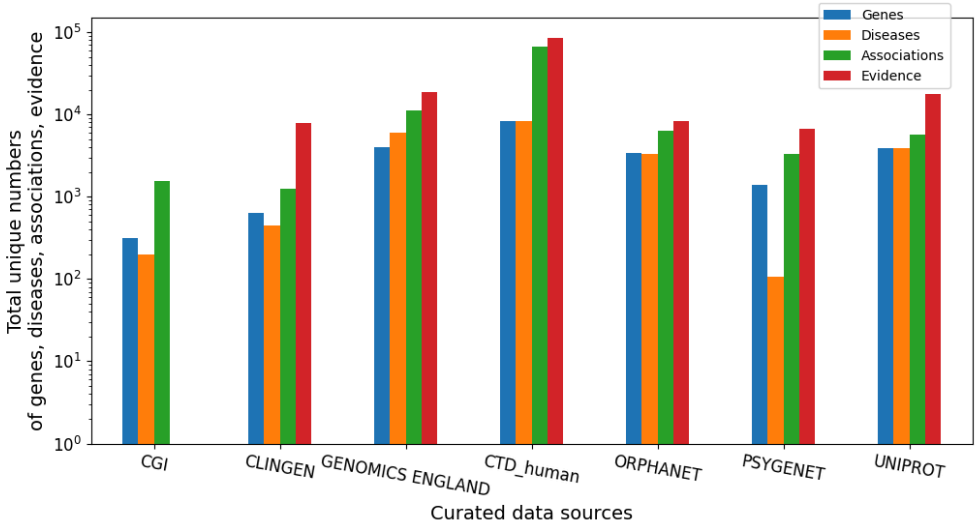


Figure 4.1. The total numbers of unique genes, diseases, associations, and evidence in each curated data source. CTD_human has the highest number of genes, diseases, and associations, whereas the lowest number of diseases is from PSYGENNET.

OMIM The Online Mendelian Inheritance in Man (OMIM) database[53] [88] was used as an external gold standard to score DGAs from DisGeNET (Section 3.1).

Monogenic Diseases: Monogenic experimental studies [252] (Section 3.1) were also used as an external gold standard to score DGAs from DisGeNET, given the well-established relationship between monogenic diseases and their associated genes [291].

[52] https://pubmed.ncbi.nlm.nih.gov/
[53] https://www.omim.org/

4.3 Results and Discussion

4.3.1 Identification of Gold Standard Data and Individual Datasets for Disease-Gene Associations

In order to build a DGA PFIN, two main components are required: individual datasets which represent independent evidence of DGAs, and gold standard data which contains high confidence set of DGAs (Section 2.3.2 for more details about gold standard data) [292], [293], [294], [295]. It has been shown that the changes in the gold standard data and individual datasets used considerably impact the quality and the performance of the PFINs [192]. Therefore, choosing the gold standard data and the individual datasets is an important step in the construction of the PFINs. The gold standard must be of high confidence and reliability, accurately reflecting true DGAs, and up to date.

Most of the available DGA data sources were designed for the investigation of a specific disease or group of related diseases. For example, the ORPHANET[54] database was specifically created for rare diseases, the CGI database is intended to assist in identifying genetic alterations in tumours that are responsible for the disease, and the PSYGENET[55] database is a resource for psychiatric diseases and their associated genes. These data sources may have a low overlap with a gold standard, as they may focus on a set of diseases which is not in the gold standard. Therefore, it is necessary to search for and develop appropriate strategies to define suitable gold standards and individual datasets which represent separate evidence of DGAs. Due to the lack of identification of accurate, high-coverage gold standard data for DGAs, different approaches based on different strategies were developed to generate the gold standard data and the individual datasets and compared to investigate which of these proposed approaches be were suitable for identifying the gold standard data and the individual datasets.

The gold standard data were chosen both internally and externally (Section 2.3.2). In the external gold standard-based approach, the gold standard was chosen from a separate high-quality data source different from the individual datasets to be scored. In the internal

[54] https://www.orpha.net/consor/cgi-bin/index.php

[55] http://www.psygenet.org/web/PsyGeNET/menu;jsessionid=16qc59g1ueuxo1sryh9w9dmkih

gold standard-based approach, the gold standard and the individual datasets to be scored come from a single data source. In this chapter, the two approaches were investigated to determine the most suitable approach to building the bipartite DGA PFIN. In the internal gold standard-based approach, the gold standard and the individual datasets to be scored were both generated from DisGeNET based on individual experimental studies or data sources. Using the external gold standard-based approach, the gold standard was chosen from a database separate from DisGeNET, specifically the OMIM database and the monogenic experimental studies. Two distinct approaches were used to identify both the gold standard data and the individual datasets for comparison. One approach was based on curated data sources in DisGeNET, while the other was based on individual experimental studies in DisGeNET. Therefore, the DisGeNET database was subdivided into two levels: by data source; and by individual study.

In this thesis, the term "*gold standard*" refers to high-quality and reliable data that serve as a benchmark for evaluating the quality of datasets. The term "*individual datasets*" refers to the distinct datasets that represent a set of DGAs, which are scored against the gold standard data and subsequently integrated into a network to produce PFINs. The term "*data source*" refers to a single curated data source within the DisGeNET database. The term "*individual experimental study*" refers to a single DGA experimental study documented in DisGeNET, identified by its unique PubMed ID. In the individual experimental study-based approach, a single DGA experimental study is used to represent the individual dataset, whereas in the data source-based approach, a single curated data source is employed to represent the individual datasets.

4.3.1.1 Data Source-Based Approach to Identifying the Gold Standards and the Datasets

Two approaches were used to identify the gold standard and the individual datasets based on the data source. One approach utilised an internal gold standard from DisGeNET, the UNIPROT database. The other approach relied on an external gold standard distinct from DisGeNET, the OMIM database. In the internal approach, DisGeNET was divided into several subsets according to the original data sources from which DisGeNET acquired the data, including CTD_human, PSYGENET, UNIPROT, ORPHANET, GENOMICS_ENGLAND, CLINGEN, and CGI. Among these curated data sources, one was

chosen as the gold standard, while the remaining curated data sources were employed as separate individual datasets; each data source represented one dataset. Figure 4.2 shows the overview of the Data Source-Based approach to identifying the gold standards and the individual datasets.

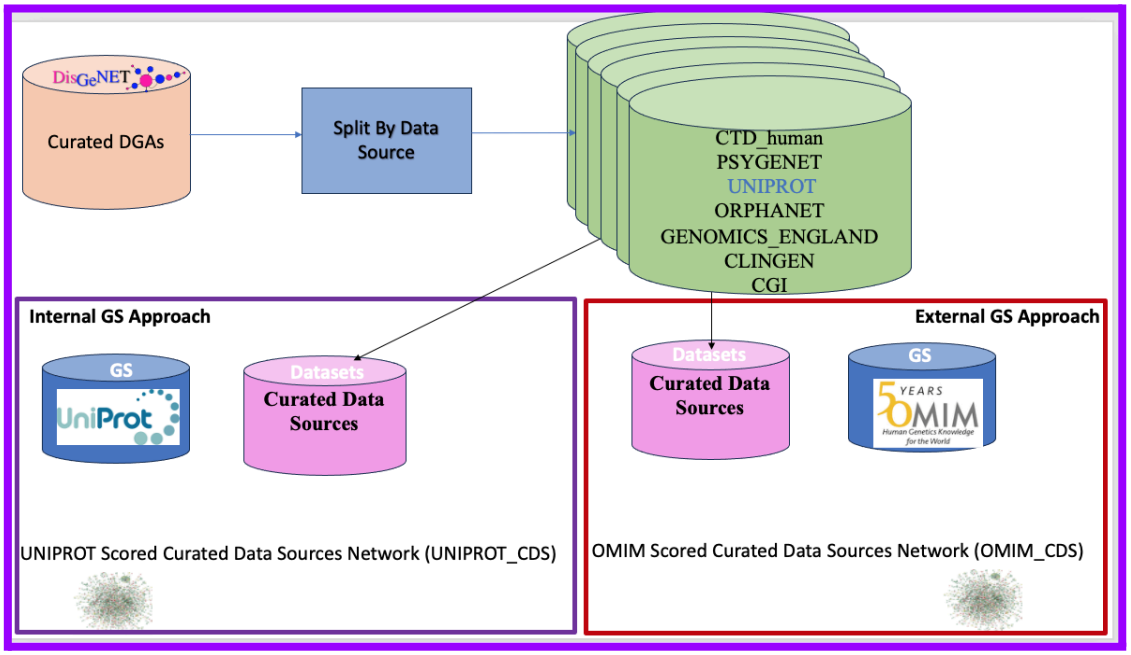


Figure 4.2. Data source-based gold standard approach. DisGeNET was split by source, using UNIPROT (internal) or OMIM (external) as gold standards. The remaining curated sources from DisGeNET served as individual datasets. Blue indicates gold standards; pink shows individual datasets.

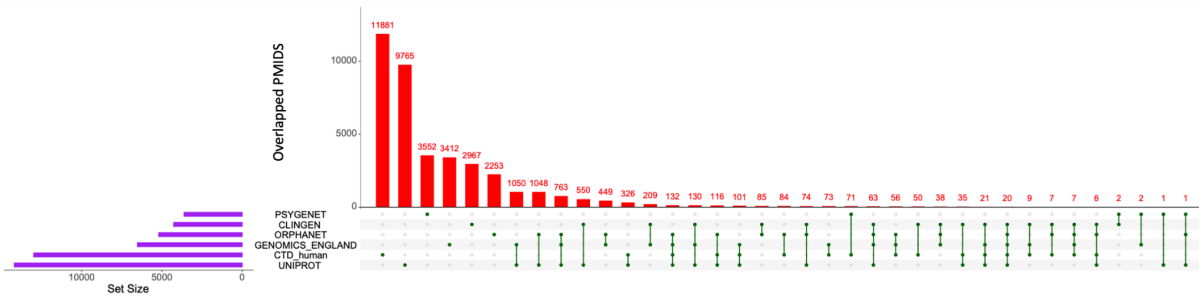
As DisGeNET comprises multiple manually-curated data sources, there is a possibility of these sources curating the same studies. In the the PFIN approach, the gold standard and individual datasets must be independent to prevent any biases that can arise from duplicate data in different data sources. Duplicating data between the gold standard and the individual datasets will lead to an overlap between the individual datasets and the gold standard, introducing bias into the resulting scores of the individual datasets. Moreover, duplicating data within individual datasets may bias the integrated scores by integrating the same evidence multiple times. Such redundancy can skew the network and give excessive weight to edges supported by duplicated evidence. Consequently, the initial step in constructing a PFIN involves eliminating redundancy among data sources, ensuring that each DGA within an experimental study is uniquely represented within a single data source. Duplicate studies

were identified by matching PMIDs[56]. A PMID (PubMed Identifier) is a unique number assigned to each entry in the PubMed database of life sciences and biomedical literature, used to reference specific studies. If duplicate DGAs were found within the same PMID, one instance of these DGAs was kept in a dataset at random, and the other duplicates were discarded. For example, the DGA "*Leigh syndrome, LRPPRC*" from study PMID 12529507 is present in UNIPROT, CTD_human, ORPHANET, CLINVAR, and GENOMICS_ENGLAND. Since this DGA is duplicated across these data sources, it was removed from all but one, with the retained instance chosen at random.

To identify the optimal strategy for addressing redundancy, an investigation into the duplication of studies across the seven data sources was conducted. Situations were identified in which multiple data sources contained identical studies. For example, UNIPROT and GENOMICS_ENGLAND had 1,050 overlapping studies (PMIDs), while UNIPROT and PSYGENET had 1,048 overlapping studies. Figure 4.3. A shows the overlap among the data sources in the experimental studies. However, instances were identified in which one data source contained more DGAs from a specific study than another. If study A was present in two data sources, each data source held a distinct subset of DGAs from that study. For example, PMD 17994018 was present in GENOMICS_ENGLAND and CTD_human. Each data source held a distinct subset of DGAs. GENOMICS_ENGLAND included *{(Aortic Aneurysm Familial Thoracic 6, ACTA2), (Multisystemic smooth muscle dysfunction syndrome, ACTA2), (Moyamoya disease 5, ACTA2)}* and CTD_human included *{(Aortic Aneurysm Thoracoabdominal, ACTA2), (Aortic Aneurysm Thoracic, ACTA2)}*. This discrepancy could be attributed to differing curation strategies among different curated data sources (Section 2.5). In such cases, the strategy for removing redundancy must operate at the *Data Source-Study-DGA* level, rather than solely at the *Data Source-Study* level. For instance, UNIPROT and GENOMIC ENGLAND contained some identical studies, but each data source incorporated distinct subsets of DGAs originating from these studies. Figure 4.3.B shows the extent of duplicate data among data sources concerning DGAs from specific studies.

[56] https://pubmed.ncbi.nlm.nih.gov/

A.



B.

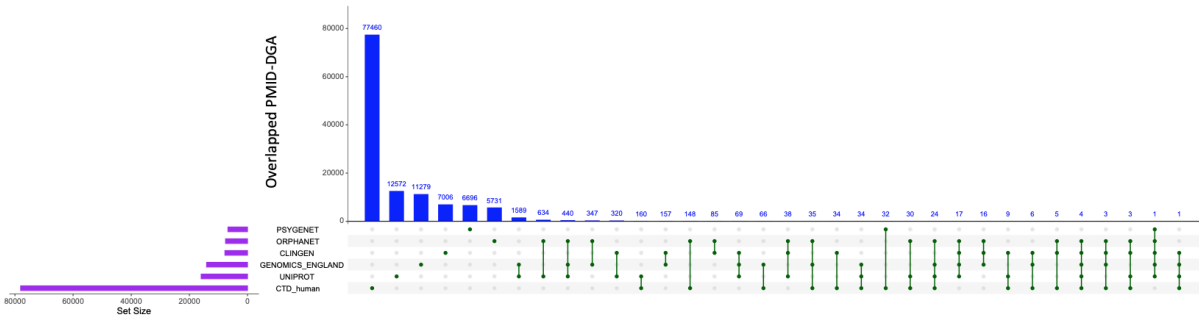


Figure 4.3. A. Overlap between curated data sources in terms of experimental studies. B. Overlap between data sources in terms of DGAs generated by particular experimental studies.

It was found that 16,193 studies had multiple association types for the same DGAs, either within a single data source or across different data sources. This situation often arose due to the use of various experimental methods within a single study to investigate the same DGAs. A single study might employ two different experimental methods to examine the same DGA, with each experiment providing distinct lines of evidence for that DGA. In such cases, the strategy for removing redundancy must work at the *Data Source-Study-Association Type-DGA* level, rather than solely at the *Data Source-Study-DGA* level. For example, the DGA *(Noonan Syndrome, RAF1)* from study PMID 17603483 is present in two data sources: UNIPROT and GENOMICS_ENGLAND. In UNIPROT, the association type is listed as '*Genetic Variation*,' while in GENOMICS_ENGLAND, it is categorised as '*Biomarker*'.

However, this is not the case in DisGeNET. The association types are assigned by DisGeNET using the DisGeNET association type ontology, and by going back to the original data sources it was found that no information was provided regarding association types.

Additionally, by investigating the DisGeNET association type ontology, it was discovered that the annotations to all the association types are children of the same parent, which is "*biomarker"*. Therefore, the association type does not provide a separate line of evidence for DGAs. Removing redundancy was done at the *Data Source-Study-DGA* level rather than solely at the *Data Source-Study-Association type-DGA* level. To eliminate redundant data, a random selection process was employed among data sources. Each DGA generated by a particular study is retained in only one data source. For instance, if three data sources contain the same DGA from the same study, the DGA is randomly removed from two of those data sources, ensuring that it exists in only one source.

UniProt Database: An investigation was conducted on each of the curated data sources within DisGeNET to assess their suitability as a gold standard. Each data source was sequentially chosen as the gold standard, and the remaining curated data sources were scored against this gold standard using the Bayesian statistics approach developed by Lee *et al*., which calculates a log-likelihood score for each dataset (Equation 3.1) [55]. These scored data sources were then ranked based on their confidence scores. When each curated data source was considered as the gold standard, UNIPROT consistently ranked first in most instances (Table 4.1). Consequently, the decision was made to use the UNIPROT data source as the gold standard, with the remaining curated data sources (CTD_human, PSYGENET, ORPHANET, GENOMICS_ENGLAND, CLINGEN, and CGI) designated as individual datasets.

OMIM Database: The OMIM database was also used as an external gold standard for scoring the curated data sources within DisGeNET . OMIM is not included in DisGeNET. Utilising an external gold standard led to several challenges during scoring and integration. Firstly, there is a risk of data redundancy because different databases might contain identical studies. It is necessary to eliminate duplicate data between the gold standard (OMIM database) and the individual datasets (the curated data sources from DisGeNET) before scoring. This step is essential, due to the distinct curation strategies employed by each data source, which may result in the presence of duplicate data between the OMIM database and the original data sources from which DisGeNET is derived. There might also be differences in identifier

formats, requiring mapping between the gold standard and the individual datasets before scoring.

Table 4.1. LLS score for each test dataset when altering the gold standard. After applying the LLS method and alternating the gold standard, gold standard (left column), we see how every other dataset, the test dataset (top row), performs in terms of identifying the 'knowns' captured in the gold standard. Performance is measured using the LLS score, which is shown. The highest LLS scores are represented in bold. The number in parentheses represents the rank.

	Individual datasets						
Gold Standard	**UNIPROT**	**GENOMICS_ ENGLAND**	**CLINGEN**	**CTD_human**	**ORPHANET**	**PSYGENET**	**CGI**
UNIPROT	-	**14.50 (1)**	13.78 (2)	12.40 (4)	13.35 (3)	7.58 (6)	9.97 (5)
GENOMICS_ ENGLAND	**14.02 (1)**	-	12.26 (3)	11.49 (4)	12.58 (2)	6.82 (6)	10.24 (5)
CLINGEN	**18.35 (1)**	15.13 (3)	-	14.70 (4)	15.20 (2)	lost	13.15 (5)
CTD_human	**11.41 (1)**	10.46 (4)	10.47 (3)	-	10.54 (2)	lost	9.59 (5)
ORPHANET	**14.70 (1)**	13.92 (2)	13.04 (3)	2.65 (5)	-	lost	11.65 (4)
PSYGENET	**9.239 (1)**	7.45 (2)	lost	lost	lost	-	lost
CGI	12.87 (4)	13.65 (2)	**14.89 (1)**	12.87 (4)	13.53 (3)	lost	-

Since OMIM uses OMIM identifiers for diseases and DisGeNET uses UMLS for disease identifiers, the process of identifier mapping for diseases between UMLS and OMIM was executed using the Metathesaurus (Section 3.7). However, it is important to note that approximately 2,476 diseases, 1,740 genes, and 2,772 associations in the OMIM database could not be mapped to DisGeNET. Various methods for identifier mapping were explored (Section 3.7). The Disease Ontology (DO), which cross-maps to UMLS and provides extensive cross-referencing, was used. However, it produced only limited mapping, with only 94 terms from OMIM to UMLS (1.21% mapped using DO cross-referencing). In contrast, the Mondo ontology had 6,324 terms between OMIM and UMLS, achieving a 65% mapping rate. The UMLS Metathesaurus contained 95,421 terms, with 66% mapped from UMLS to OMIM. These results suggested that the Mondo ontology and the UMLS Metathesaurus offered the most effective identifier mapping between OMIM and UMLS. The gold standard

might have poor overlap with the datasets due to differing focuses, since both the gold standard and the datasets originated from distinct sources.

DisGeNET was also subdivided into individual experimental studies to identify the gold standard and the individual datasets. The next section reports the investigation into the utilisation of the individual experimental study-based approach to define the gold standard and the individual datasets.

4.3.1.2 Individual Experimental Study-Based Approach for Identifying the Gold Standards and the Datasets

In this section, different strategies were investigated to define the gold standard and the individual datasets based on individual experimental studies. A PMID was used to identify these studies. Each PMID indicates a single study designed to investigate either a single disease and its related genes or a group of related diseases and their genes. In the individual experimental study-based approach, the manually curated DGAs from DisGeNET were divided by PMID into individual experimental studies to identify DGAs produced by different experimental studies. The gold standards were identified both externally and internally. In the internal approach, the individual experimental studies from DisGeNET were subdivided by *experimental scale level*, *experimental confidence level*, and *experimental curation level* to identify the gold standards and the individual datasets. In the external approach, all the individual experimental studies were used as the individual datasets, with OMIM and Monogenic experimental studies used as the gold standards. Figure 4.4 shows an overview of the individual experimental study-based approach for identifying the gold standards and the individual datasets.

Experimental Scale Level: Data generated by LTP experimental techniques can be considered to be of high confidence, and can be used to assess data quality [57], [90]. Data generated by HTP experimental techniques are noisy and contain a high proportion of false results [296]. Therefore, LTP studies can be used as gold standard data to estimate the quality of HTP studies. For example, in PPI PFINs, small-scale LTP interactions, considered to be high quality, were used as a gold standard to assess the quality of the HTP studies [57], [90]. For instance, the LTP interactions from BioGRID were used to score the quality of the HTP

studies in PPIs [50], [57], [90]. It has been shown that using LTP studies to score HTP studies produces improved performance compared to the use of an external gold standard [57]. This method, which tends to generate both gold standard and datasets from one single curated data source, produced more significant results than using an external gold standard [57]. The scale of the experimental studies was used previously to identify the gold standard and the individual PPI datasets [57], [90]. The experimental scale has been used to distinguish between LTP and HTP datasets, and a 100 interaction cut-off was used to split the experimental studies from the BioGRID database into LTP studies and HTP studies [50], [57], [86], [90]. However, applying this approach to DGA data could pose challenges due to the lack of information on the experimental techniques utilised in disease-gene studies. Many existing data sources for DGAs do not provide details about the experimental techniques employed in generating these associations. For instance, neither DisGeNET nor its original data sources offer details about the techniques employed in generating DGA data. The absence of this information poses a challenge in determining whether the studies utilised HTP or LTP techniques.

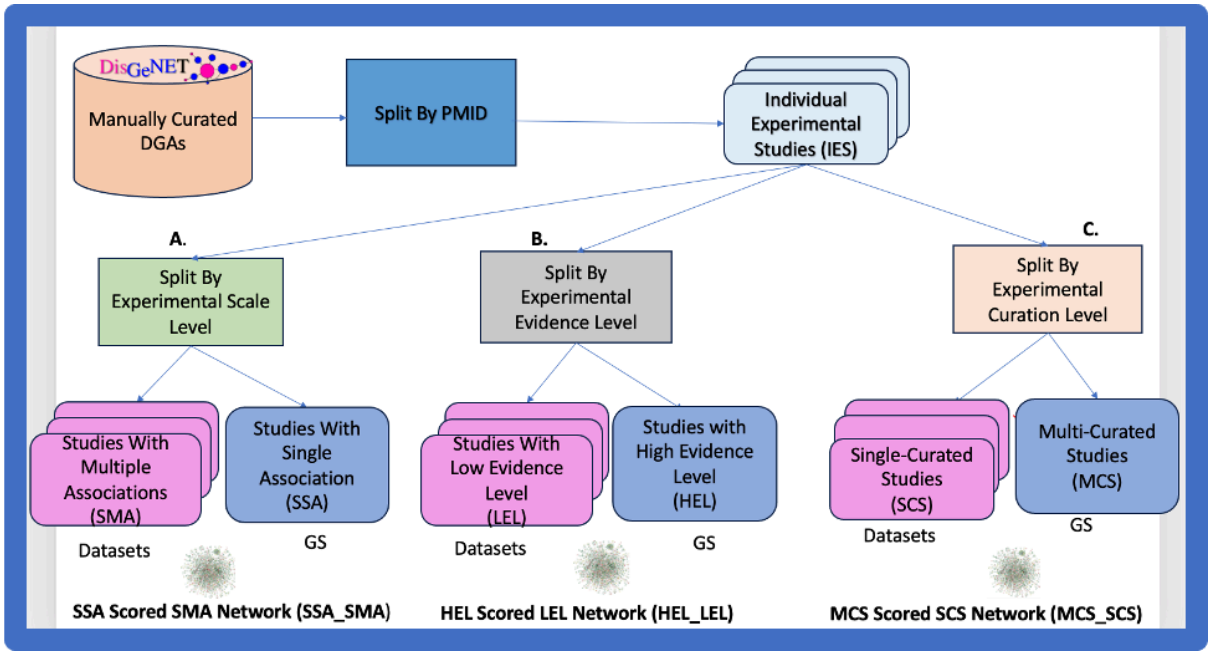


Figure 4.4. Internal gold standard approach based on individual experimental studies. DisGeNET was split by PMID and classified by experimental scale, confidence, and curation level. In the SSA_SMA method, large-scale studies served as datasets and small-scale studies as the gold standard. In HEL_LEL, studies with over 50% strong or definitive DGAs were used as the gold standard; others served as datasets. In MCS_SCS, multi-curator studies formed the gold standard, while single-curator studies were used as datasets.

In the research reported in this section, it was investigated whether individual experimental studies in DisGeNET that resulted in fewer DGAs are more accurate than those that resulted in multiple DGAs. The experimental studies range in size from a single DGA to 1,040 DGAs. Approximately 55% of the experimental studies contained a single association, while 45% of the studies contained multiple associations (Figure 4.5). A cut-off value of one association was chosen to differentiate between large-scale and small-scale studies, following a systematic evaluation of various factors within the resulting PFIN. This comprehensive assessment included five considerations: the distribution of confidence scores; edge weights; data loss rate during scoring; the percentage of total unique genes, diseases, and associations integrated; and the connectedness of the resulting clusters in the network. First, setting the threshold at one provided the highest number of individual datasets, unique diseases, unique genes, and unique DGAs considered for scoring and then integrated into the final network (Figure 4.6.A). However, as the threshold increased, the percentage of individual datasets, unique genes, unique diseases, and unique DGAs within individual datasets decreased. A lower threshold increases coverage in individual datasets but decreases coverage in the gold standards (Figure 4.6.B). While a threshold of one provided the highest number of individual datasets, genes, diseases, and DGAs considered for scoring, it resulted in a low rate of these elements, from the original size of these elements considered for scoring, after scoring (Figure 4.6.C), with a high rate of data loss compared to other thresholds (Figure 4.6.D). For example, a threshold of 100 maintained a high rate of datasets, diseases, genes, and DGAs from the original size, with the lowest data loss rate. This is because a threshold of 100 had fewer datasets (less than 20%) and DGAs (less than 22%) considered for scoring, with low coverage of diseases, genes, and DGAs. Higher thresholds led to a lower number of datasets before scoring, resulting in less data loss due to fewer datasets being considered, and a smaller integrated network (Figure 4.6.E). Additionally, a threshold of one demonstrated a wide range of confidence scores, from eight to 18, and provided a higher average LLS score than other thresholds (Figure 4.6.E). It also showed variations in edge weights, ranging from eight to 63, and produced the highest average edge weight (Figure 4.6.F). This threshold resulted in a larger network size. As the threshold increased, network size decreased, due to fewer datasets being considered for scoring (Figure 4.6.G). Finally, setting the threshold at one provided the highest average clustering cohesiveness, which refers to the degree of similarity among the diseases and their related genes within a cluster, indicating how tightly

grouped the diseases and genes were and how well they shared common biological characteristics. The average connectedness decreased with higher thresholds, making a threshold of one optimal for cluster cohesiveness. Overall, setting the threshold at one produced variations in both confidence score and edge weight distributions. This threshold minimised data loss while ensuring a substantial representation of distinct genes, diseases, and DGAs. The chosen threshold also resulted in the highest average level of network cluster connectedness, as illustrated in Figure 4.6.H. Consequently, studies with single associations were grouped and used as the gold standard to score those with multiple associations (SSA_SMA). Division of DisGeNET at the one association cut-off produced 17,855 Studies with Multiple Associations (SMA) datasets and 21,719 Studies with Single Associations (SSA) datasets.

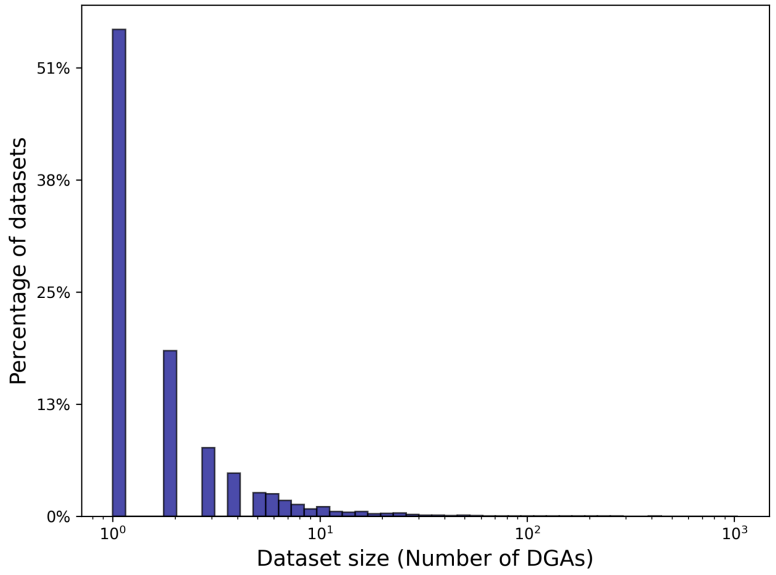
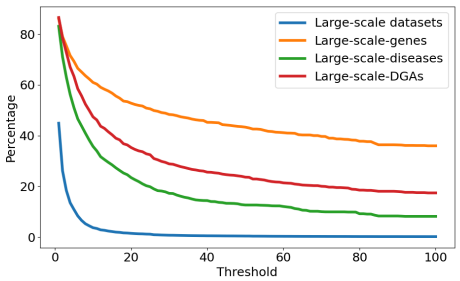


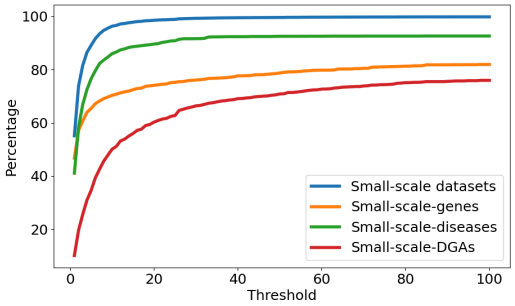
Figure 4.5. Distribution of dataset sizes by number of DGAs. The left histogram shows datasets with a single DGA; the right shows those with multiple DGAs. Y-axis indicates percentage of all datasets.

***Experimental Confidence Level**:* The Evidence Level (EL) is a quantifiable measure introduced by ClinGen to measure confidence in a DGA. EL was categorised qualitatively into five classes: *"Definitive," "Strong," "Moderate," "Limited," and "Disputed"* (For more details about the EL, refer Section 2.4.1.2) [160]. Genomics England introduced an evidence criterion called the traffic light system, which uses slightly different categories: Green (go) means "High Evidence," Amber (pause) means "Moderate Evidence," and Red (stop) means "Low Evidence" (see Section 2.4.1.2). DisGeNET incorporated the metrics from both

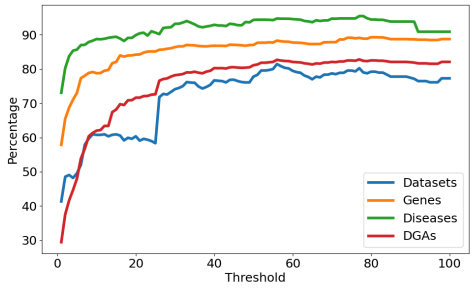
A.


Large-scale datasets
Large-scale-genes
Large-scale-diseases
Large-scale-DGAs
Percentage
Threshold

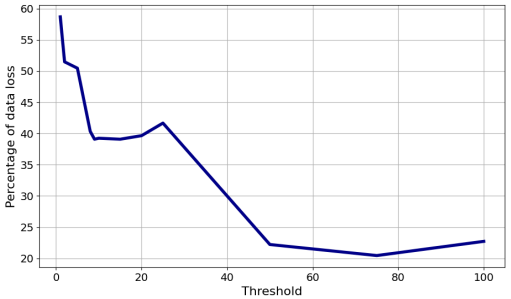
B.


Small-scale datasets
Small-scale-genes
Small-scale-diseases
Small-scale-DGAs
Percentage
Threshold

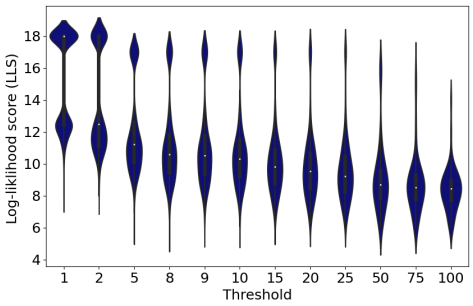
C.


Datasets
Genes
Diseases
DGAs
Percentage
Threshold

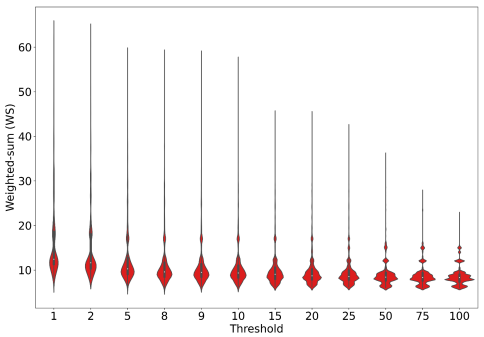
D.


Percentage of data loss
Threshold

E.


Log-likelihood score (LLS)
Threshold

F


Weighted-sum (WS)
Threshold

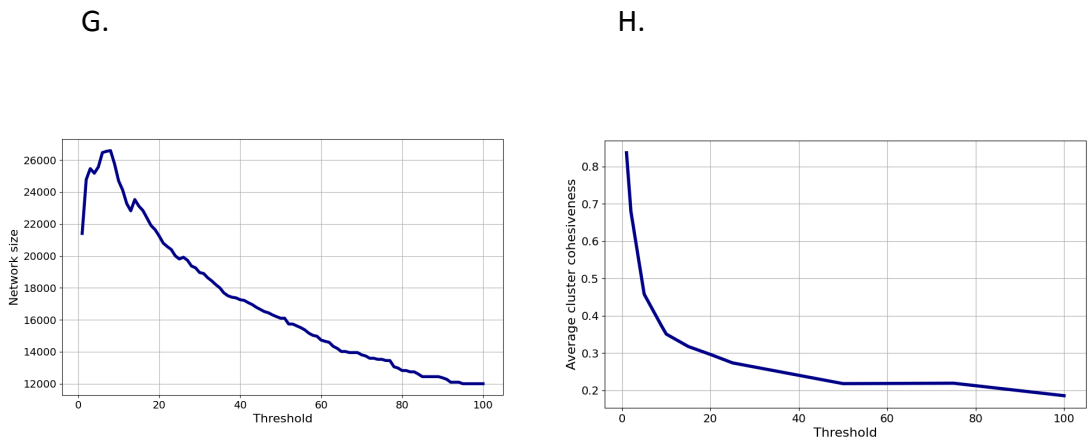


Figure 4.6 Systematic analysis for selecting the gold standard threshold. Panels A–C show percentage changes in datasets, genes, diseases, and DGAs with increasing thresholds. Panels D and E display broader distributions of confidence and edge weights at threshold one. Panel F shows that using one association maximizes network size. Panel G presents data loss, while Panel H shows highest cluster cohesion at threshold one.

reported by Genomics England PanelApp with the categories established by ClinGen: DGAs identified as High Evidence by Genomics England PanelApp were categorised as strong in DisGeNET. DGAs marked as Moderate Evidence are similarly categorised as moderate, and associations labelled as Low Evidence are categorised as limited. It was hypothesised that experimental studies containing a high rate of high evidence level (*definitive* or *strong*) DGAs are high-confidence studies, while experimental studies containing a low rate of high evidence level of DGAs are low-confidence studies. Therefore, the high evidence level studies (HEL) were grouped and used as the gold standard to assess the quality of low evidence level studies (LEL) which were treated as individual datasets.

DisGeNET was split by PMID into individual studies, and each study contains DGAs marked as definitive, strong, moderate, limited, or disputed. Each study was treated as a single dataset. However, most of these studies contained low rates of definitive, and strong DGAs. Only 22% of these studies contain 100% of high evidence level DGAs, whereas 70% of these studies contain around 20% of high evidence level DGAs, as shown in Figure 4.7. A cut-off of 50% was chosen to distinguish between high-confidence and low-confidence studies based on the evidence-level distribution among the datasets. Therefore, datasets having less than 50% of their DGAs marked with definitive or strong were treated as low-confidence datasets and used as datasets to be scored. Datasets having at least 50% of DGAs marked with definitive or strong were treated as high-confidence data and used as the gold standard to assess the quality of the datasets (Figure 4.7.B). Therefore, high evidence-level studies scored

low evidence-level studies (HEL_LEL). The DisGeNET database was split by evidence level to 12,064 HEL datasets and 27,510 LEL datasets. The HEL datasets contain 50% or more of DGAs classified as definitive and strong. In contrast, the LEL datasets contain less than 50% of DGAs classified as definitive and strong.

As the EL is derived exclusively from ClinGen and the Genomics England PanelApp, only DGAs found within these databases, along with those overlapping with ClinGen and Genomics England PanelApp, receive an EL label. Consequently, in calculating the EL rate for a study, only DGAs with an assigned EL were taken into account. The absence of the EL for DGAs from the remaining curated data sources in DisGeNET introduced a limitation in defining the Gold Standard and datasets using this approach. In total, 25,157 studies have DGAs not labelled with an evidence level, resulting in 70,395 unique DGAs not labelled with an evidence level. In contrast, 14,418 studies reported DGAs with an evidence level, accounting for 6,970 unique DGAs.

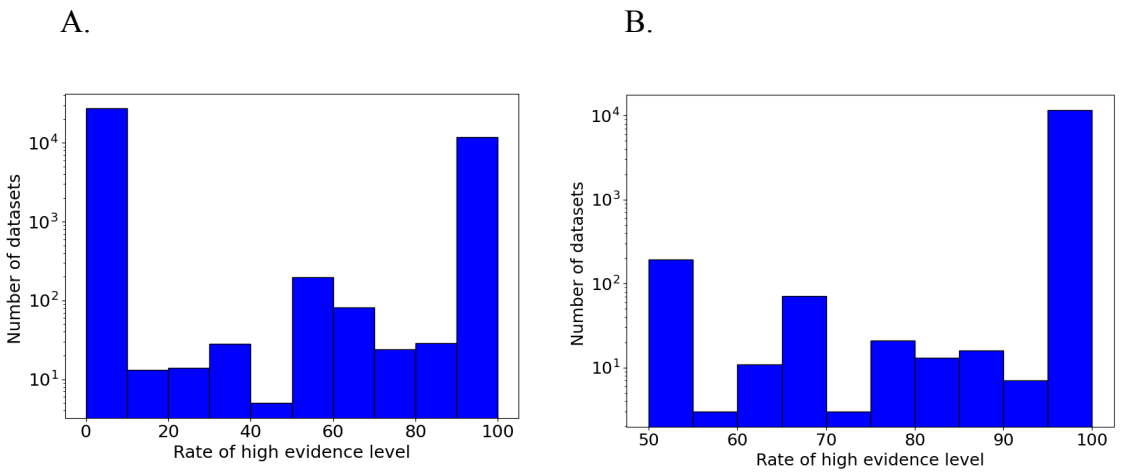


Figure 4.7. Distribution of high evidence-level DGAs across datasets. In A, 22% of datasets contain 100% high-evidence DGAs, while 70% contain only 20%. B shows the distribution of High Evidence Rates ranging from 50% to 100%.

Experimental Curation Level:. Curated data sources contain DGAs that have been manually curated by expert biocurators. The intersection between these curated data sources, particularly those involving experimental studies, can provide valuable insights. This overlap can aid in assessing the confidence level of an experimental study. It was hypothesised that experimental studies curated by multiple biocurators may carry higher confidence levels compared to those curated by only one biocurator. This higher confidence arises from the

consensus among multiple curation experts regarding these studies. As a result, experimental studies curated by multiple biocurators were considered high-confidence data, and grouped to form the gold standard. Conversely, experimental studies curated by a single biocurator were regarded as low-confidence datasets and were utilised as datasets. The overlap among datasets has been previously used to identify gold standards and served as a reliable benchmark for evaluating data quality (For more details, see Section 2.3.2) [297]. To determine multi-curated experimental studies among the curated data sources, the intersection between these curated sources concerning experimental studies was computed. The Genomic England Panel and UNIPROT exhibited the greatest degree of overlap in their studies. The second highest level of overlap was observed between UNIPROT and ORPHANET. The smallest degree of overlap occurred between Genomic England and PSYGENET. This low overlap may be attributed to distinct focuses between UNIPROT and PSYGENET, as PSYGENET specialises in psychiatric disorders. Figure 4.3.A shows the overlapping studies among the curated data sources. DisGeNET was split by PMID into 5,744 datasets that were curated by multiple experts and 33,830 datasets curated by a single expert. The curation level of the experimental studies has previously been used to generate the confidence scores of DGAs [141]. However, it was not properly used as the overlap between curated data sources was not removed before computing the confidence score (Section 4.3.5).

A gold standard was additionally identified externally based on the individual experimental study approach, in which each study represented a distinct dataset. Monogenic experimental studies as well as DGAs available in the OMIM database were used as the gold standard. Figure 4.8 shows the external approach based on individual experimental studies to identify the gold standard and the individual datasets.

OMIM database: DGAs sourced from OMIM were employed to score the individual experimental studies in DisGeNET (OMIM_IES). DisGeNET was subdivided by individual experimental studies, resulting in 39,574 datasets. The overlapped studies between DisGeNET and OMIM were excluded from DisGeNET, amounting to a total of 2,065 overlapped studies. Consequently, 37,509 individual studies were considered for scoring.

***Monogenic Experimental Studies**:* The work reported in this section utilised monogenic experimental studies as the gold standard data for scoring individual experimental studies from DisGeNET (MG_IES). Figure 4.8 shows the external approach to identifying the gold standard and the individual datasets. The utilisation of an external gold standard dataset introduced several challenges, one of which was the difference in identifier formats, necessitating identifier mapping before scoring. In the monogenic dataset, the DGAs were annotated with OMIM identifiers for diseases and Ensembl identifiers for genes. Consequently, the mapping of diseases between UMLS Concept Unique Identifiers (CUIs) and OMIM was required, utilising the Mondo disease ontology, and the mapping of genes between Ensembl and NCBI was carried out using information from HGNC (Section 3.7). In total, 3,358 diseases out of 4,166, 2714 genes out of 3,163, and 3,636 associations out of 4,292 were successfully mapped. It is important to note that only diseases and genes that could be mapped to UMLS disease IDs and NCBI gene IDs, respectively, were considered.

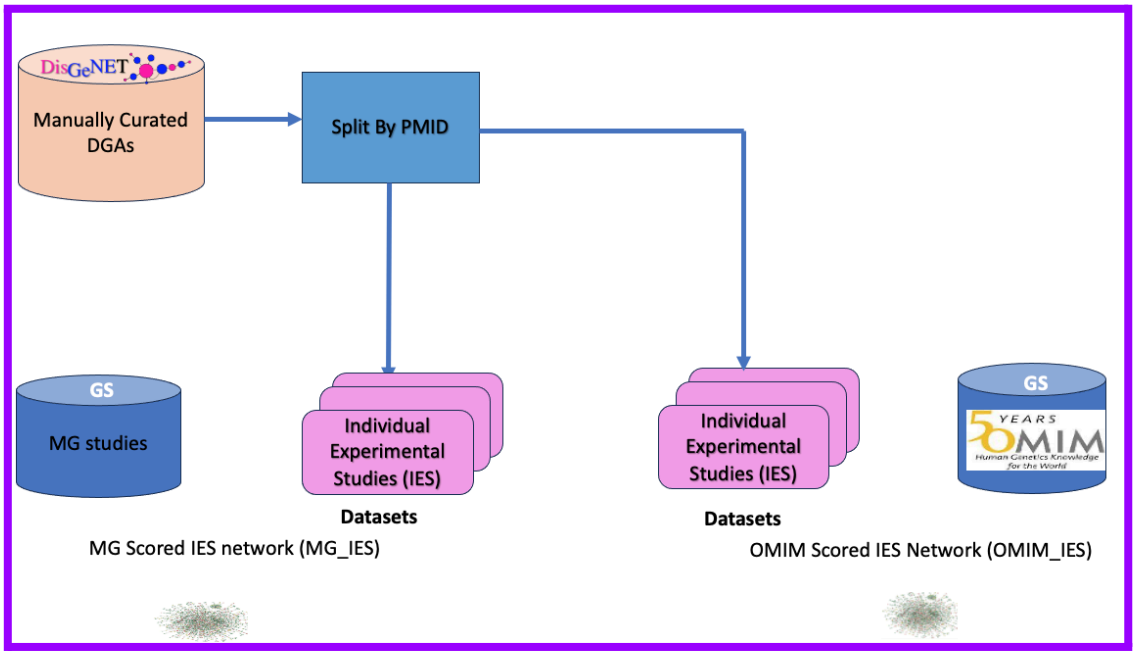


Figure 4.8 Overview of the external individual experimental study-based approach. DisGeNET studies were scored against two external gold standards—monogenic DGAs and OMIM. Gold standards and datasets were selected from separate sources.

DisGeNET was subdivided by individual studies, resulting in 39,574 datasets. The overlapped studies between DisGeNET and monogenic disease studies were excluded from

DisGeNET, amounting to a total of 2,678 overlapped studies. Consequently, 36,896 individual studies were selected to be scored against monogenic studies.

4.3.2 Network Integration

In the previous section, two approaches were developed to identify gold standards and individual datasets. Building on these approaches, seven DGA PFINs were constructed. The first approach, based on data sources, used curated data sources to represent individual datasets, including the UniProt-scored Curated Data Sources network (UniProt_CDS) and the OMIM-scored Curated Data Sources network (OMIM_CDS). The second approach, based on individual experimental studies, treated individual experimental studies as individual datasets, including the Studies with Single Association and Studies with Multiple Associations network (SSA_SMA), High-Evidence-Level studies scored Low-Evidence-Level studies network (HEL_LEL), Multi-Curated Studies scored Single-Curated Studies network (MCS_SCS), Monogenic studies scored Individual Experimental Studies network (MG_IES), and OMIM scored Individual Experimental Studies (OMIM_IES). To ensure consistency, a standardised network naming format was employed: 'a gold standard name_ an individual dataset name'. In each network, the identified individual datasets were scored against the corresponding gold standard using Lee's method (see Section 3.2 for details). Figure 4.9 presents an overview of the construction of DGA PFINs based on Lee's approach.

4.3.2.1 Confidence Scoring of the Datasets Using Log-Likelihood Score

The confidence scores were calculated by scoring the individual datasets against their respective gold standards using the Bayesian statistics approach developed by Lee *et al*., which calculates a LLS for each dataset (Equation 3.1) [55]. Datasets were then ranked based on the order of their confidence scores. Datasets that received zero or negative scores were excluded from the integration. The LLS distribution is shown in Figure 4.10.

In the data source-based approach, curated data sources within DisGeNET were first scored against UniProt to produce the UniProt_CDS network and then scored against OMIM to produce the OMIM_CDS network. The PSYGENET database was omitted due to receiving a negative score against the gold standard in both UniProt and OMIM. This result suggests that

both UniProt and OMIM have low overlap with the PSYGENET database, likely due to differences in their focus. The resulting confidence scores of the datasets were similar in both gold standards, resulting in limited variability in the distribution of the confidence scores.

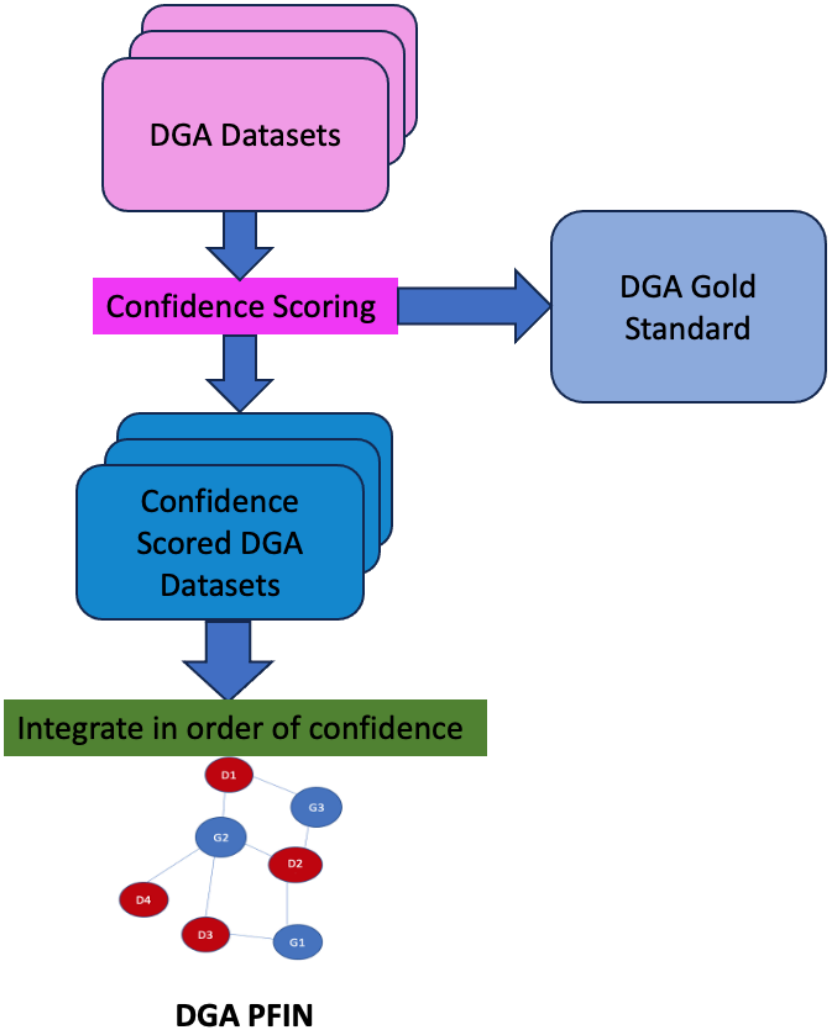


Figure 4.9. DGA PFIN is produced using the method developed by Lee and colleagues [55] for integrating the datasets in order of confidence rank.

The confidence scores of the UniProt_CDS and OMIM_CDS are presented in Table 4.2. Adopting OMIM as the gold standard led to an improvement in confidence scores compared to UniProt. For instance, the confidence score of GENOMICS_ENGLAND increased from 14.50 to 16.03. Despite the change in the LLS scores, the ranks of the datasets remained the same in both gold standards. This change in scores is primarily due to the difference in the size of the gold standards, with OMIM being smaller than UniProt. The smaller size of OMIM may improve the overlap between the datasets and the gold standard, resulting in

higher confidence scores. However, these scores are not directly comparable due to the different sizes of the gold standards. The stability in rankings indicates that while the absolute confidence scores can vary with the size of the gold standard, the relative performance of the datasets is consistent. Therefore, the rankings provide a more reliable measure of dataset quality than the scores alone.

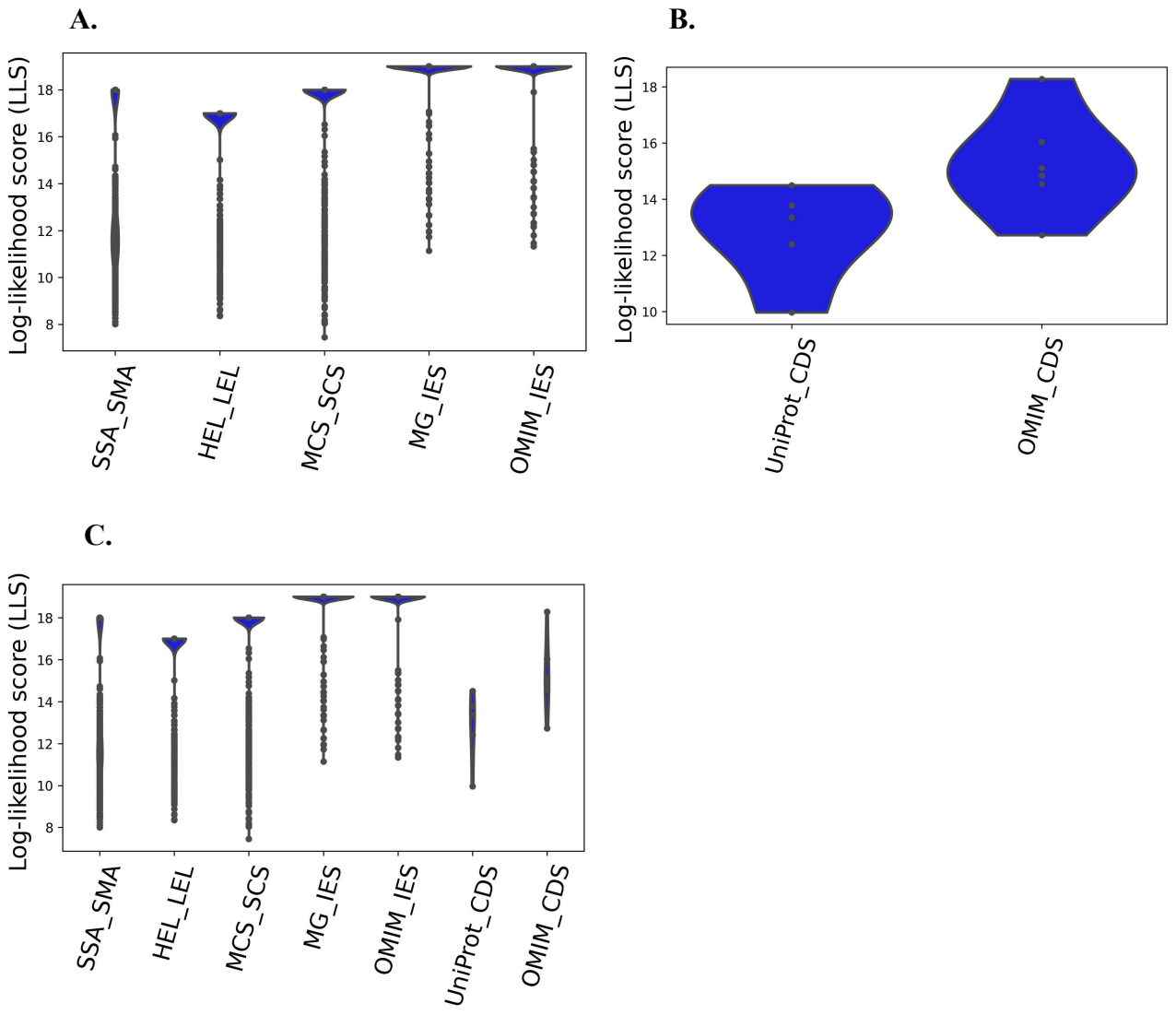


Figure 4.10. Log-likelihood score (LLS) distributions for experimental study- and data source-based approaches using internal and external gold standards. SSA_SMA, HEL_LEL, and MCS_SCS represent internal comparisons; MG_IES, OMIM_IES, UniProt_CDS, and OMIM_CDS represent external comparisons.

In the individual experimental study-based approach, after scoring various datasets against their respective gold standards, distinct patterns of data loss and scoring emerged. Datasets may be lost during scoring if they have a low overlap with the gold standard data, thus receiving NaN, zero, or negative scores. Datasets may also score infinity if they have a perfect overlap with the gold standard. The individual experimental study-based approach

had a higher rate of data loss than the data source-based approach. In this approach, each DGA experimental study**,** which typically focuses on a single disease or a group of related diseases and their associated genes, represents an individual dataset. When scoring these individual datasets against the gold standard, the dataset may be lost if the disease of interest in the DGA experimental study is not included in the gold standard, leading to exclusion due to a lack of overlap. In contrast, in the data source-based approach**,** each curated data source from DisGeNET represents an individual dataset. These curated data sources aggregate multiple DGA experimental studies, thereby increasing the range of diseases covered. This broader coverage may increase the likelihood that the diseases are included in the gold standard, reducing data loss and improving the alignment between datasets and the gold standard. For instance, when scoring the SMA datasets against the SSA gold standard, 10,545 datasets were discarded and a final set of 7,310 datasets were integrated. Similarly, scoring

Table 4.2. Summary of the LLS scores and the integration order for the data source-based approach including UNIPROT-scored curated data sources (UniProt_CDS) and OMIM_scored Curated Data sources(OMIM_CDS).

Datasets	**UniProt_CDS**		**OMIM_CDS**	
	Confidence Score	**Rank**	**Confidence Score**	**Ranks**
UniProt	-	-	18.28	1
GENOMICS ENGLAND	14.50	1	16.03	2
CLINGEN	13.78	2	15.11	3
ORPHANET	13.35	3	14.84	4
CTD_human	12.40	4	14.54	5
CGI	9.97	5	12.72	6

the LEL datasets against HEL resulted in 20,435 datasets being discarded. Therefore, a final set of 7,075 datasets were integrated. The scoring of the SCS datasets against the MCS gold

standard resulted in 18,469 datasets being discarded. Consequently, a final set of 15,361 datasets was integrated. Scoring the individual experimental studies from DisGeNET against the monogenic experimental studies resulted in a total of 9,539 datasets being integrated, while 27357 datasets were discarded. Finally, scoring the experimental studies from DisGeNET against the DGAs from the OMIM database led to a total of 11,378 datasets being integrated, while 26,131 datasets were discarded. The individual experimental study-based approach exhibited high data loss rates, with 81.53% for SSA_SMA, 82.12% for HEL_LEL, 61.18% for MCS_SCS, 70.00% MG_IES, and 66.03% for OMIM_IES. The increased data loss observed in the individual experimental study-based approach can be attributed to the fact that each DGA experimental study tends to focus on a specific disease or related groups of diseases. Consequently, the focus of the datasets may not align with the coverage of the gold standard, resulting in low overlap and increased data loss. Among individual experimental study-based networks, internal gold standards, including SSA_SMA and HEL_LEL, showed more data loss since a subset of the data is excluded for use as gold standards compared to external gold standards, including MG_IES and OMIM_IES. Table 4.3 shows the level of individual study loss during scoring in the individual experimental study-based approaches. In the SSA_SMA, HEL_LEL, and MCS_SCS networks, a subset of the data was excluded for use as a gold standard.

Table 4.3 Level of individual study loss during scoring in the individual experimental study-based approaches. Datasets are lost if they have negative scores (poor overlap with the gold standard) or no score (no overlap with the gold standard). In SSA_SMA, HEL_LEL, and MCS_SCS networks, a subset of the data is excluded for use as a gold standard.

Integrated Network Name	Number of Datasets with Negative LLS Scores	Number of Datasets Without LLS Scores	Size of Gold Standard Subset	Number of Datasets Lost%
SSA_SMA	5003	5542	21719	32264(81.53%)
HEL_LEL	5162	15273	12064	32499(82.12%)
MCS_SCS	6098	12371	5744	24213(61.18%)
MG_IES	338	27019	-	27357 (70.00%)
OMIM_IES	418	25713	-	26131(66.03%)

The data source-based approach resulted in a small amount of data loss, with only one dataset lost. PSYGENET was lost when scored against the external gold standard, UniProt, and also lost against the external gold standard, OMIM. This low rate of data loss may be due to the

fact that each data source is considered as a dataset, encompassing multiple individual experimental studies. This broader scope of data sources may enhance the dataset's coverage in terms of diseases, which can improve its alignment with the gold standard by covering a wider range of diseases. However, the purpose of the gold standard is to assess the confidence of datasets, and a range of scores is expected, including some that may not align closely.

Datasets may score infinity if they overlap highly with the gold standard. Dealing with datasets that receive infinite scores can be challenging, since these datasets share the highest score, potentially resulting in low variability in confidence score distribution. Infinity scores therefore share the same ranking, which may impact the integrated score of the DGAs, since integration depends on the dataset order. Since the datasets are integrated based on their ranks, the variability in the weighted integrated scores may be decreased. The individual study-based approach resulted in higher infinity scores compared to the data source-based approach, due to the high overlap with the gold standard. The individual study-based showed higher overlap in terms of individual genes, diseases, and DGAs (Table 4.4).

The individual experimental study-based approach resulted in a larger range of confidence score distribution (Figure 4.10) with the lowest confidence scores being 7.99, 8.36, 7.46, 11.14, and 11.33 for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES respectively while the highest scores were 18.00, 17.00, 18.00, 19.00, and 19.00 for SSA_SMA, HEL_LEL, MCS_SCS, and MG_IES, respectively, in comparison to the data source-based approach, in which the lowest scores were 9.97 and 12.72 for UniProt_CDS and OMIM_CDS, respectively, while the highest scores were 14.50 and 18.28 for UniProt_CDS, and OMIM_CDS, respectively. The individual experimental study-based approach had lower variability in LLS distribution due to the high rate of infinity scores with ~53.45%, ~ 74.28%, ~54.59 %, ~74.41 %, and ~70.00% for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, respectively.

The individual experimental study-based approach, in which the individual experimental studies were treated as individual datasets, resulted in small datasets compared to the data source-based approach, in which data sources were treated as individual datasets (Figure 4.1 for data source-based approach datasets size and Figure 4.5 for individual experimental

study-based approach datasets size). The size of the datasets may influence the confidence scores. Small datasets may exhibit high overlap with gold standard data, resulting in infinity scores, or conversely, low overlap with gold standard data, leading to dataset loss. Datasets of small size may be too limited to be treated as a single dataset, since reliable confidence

Table 4.4. Gold Standard overlap. The overlap in terms of diseases, genes, and associations of the individual study-based approach including SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES, and the data source-based approach including OMIM_CDS, and UniProt_CDS.

		Individual Experimental Study-based Approaches					Data Source-based Approaches	
		Internal gold standard based approaches			External gold standard based approaches		Internal gold standard	External gold standard
	DisGeNET	SSA_SMA	HEL_LEL	MCS_SCS	MG_IES	OMIM_IES	UniProt_CDS	OMIM_CDS
Genes (Datasets)	9703	7669	8256	8654	8890	8994	9173	9303
Diseases (Datasets)	11181	7950	7866	8722	9663	9789	10942	10956
Associations (Datasets)	84038	66022	66234	68839	76655	76865	82279	82894
Genes (gold standard)	-	5540	3394	3436	2714	2490	3778	2490
Diseases (gold standard)	-	6530	5642	6636	3358	3075	2954	3075
Associations (gold standard)	-	16486	13632	13355	3449	3219	4355	3219
Overlap (Gene)%	-	54.13	30.94	35.00	28.21	27.07	32.00	25.62
Overlap (Diseases)%	-	51.52	39.59	57.00	25.27	26.21	33.70	25.00
Overlap (Associations) %	-	8.51	3.78	7.72	3.45	3.59	5.20	3.43

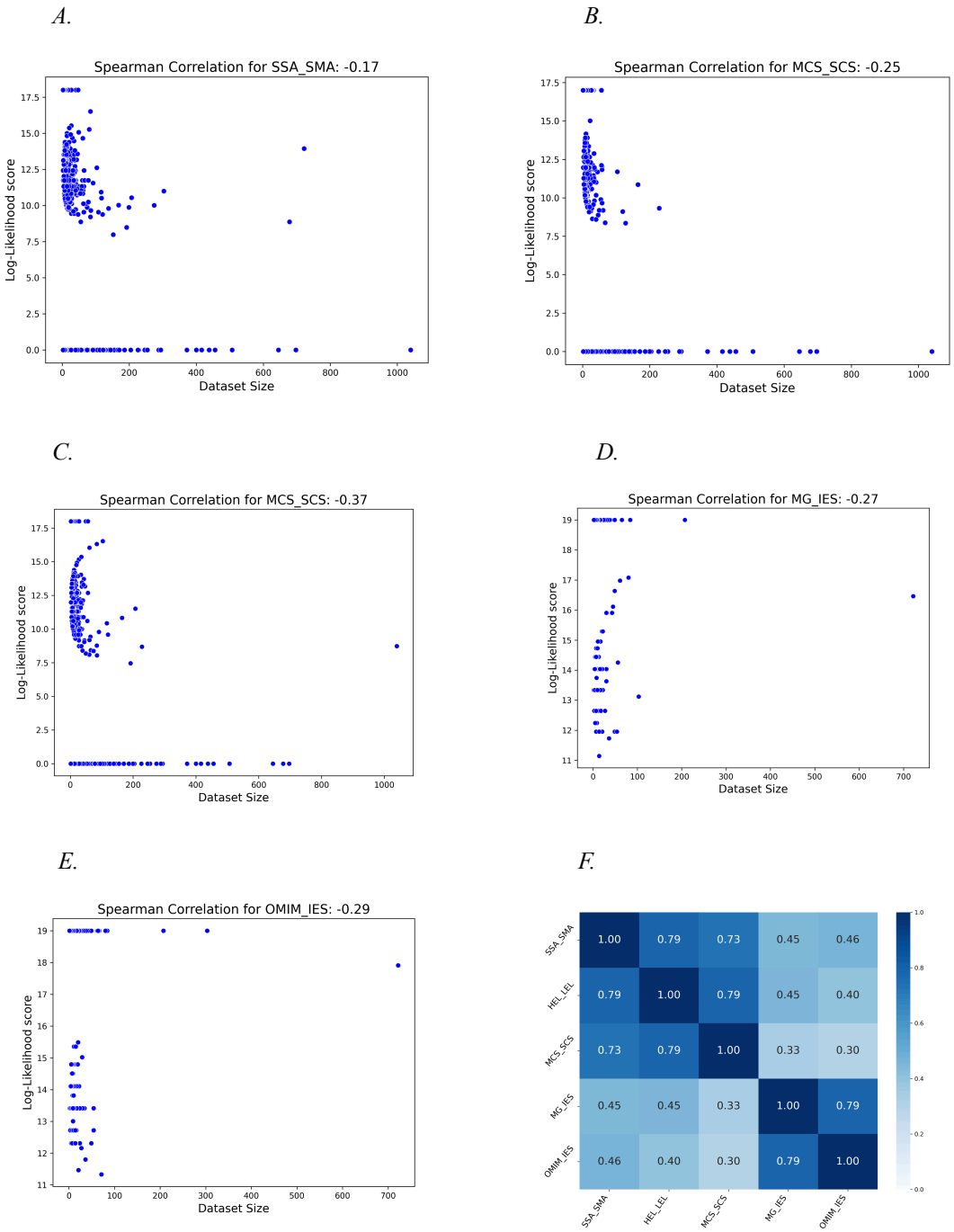


Figure 4.11 Correlation between dataset size and log-likelihood score (LLS). No strong correlation was observed. Panels A–E show correlations across SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES networks. Zero values on the y-axis indicate negative or NaN LLS scores. Panel F shows LLS correlations across datasets common to the individual experimental study-based networks.

scoring against the gold standard would not be possible due to the scale difference. The dataset size would be too small for accurate confidence assessment compared to the gold standard's size , given that no gold standard offers complete coverage of diseases and genes.

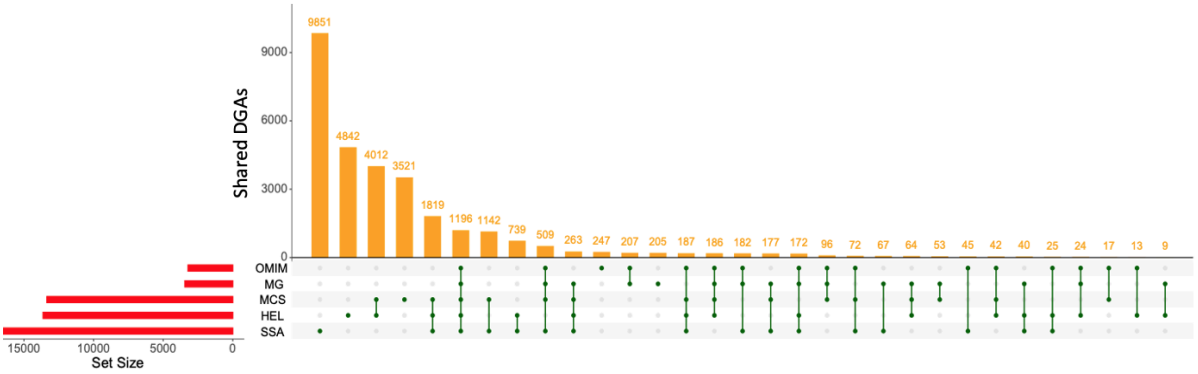
The results revealed that larger datasets, specifically those with more than approximately 250 entries, frequently receive negative or NaN scores, which are represented as 0 on the y-axis (Figure 4.11). This observation may suggest a potential correlation between dataset size and LLS, where larger datasets tend to be lost. Figure 4.11 shows the correlation between the dataset size and the LLS for individual study-based networks. The correlation analysis of the LLS for the shared scored datasets (Figure 4.11 F), based on the individual experimental study-based approach, indicates that each gold standard measured different aspects of DGAs. For instance, the LLS of MG_IES showed a high correlation with the LLS of OMIM_IES (0.79), while the LLS of MCS_SCS was highly correlated with the LLS of HEL_LES (0.79) and SSA_SMA (0.73). Similarly, HEL_LEL was highly correlated with SSA_SMA (0.79). This correlation suggests that external gold standards, such as OMIM and monogenic DGAs (the source of the gold standard is separated from the source of individual datasets which is DisGeNET), are more similar to each other than internal gold standards, such as MCS, SSA, and HEL (both the gold standard and the individual datasets are derived from the same source; DisGeNET). Experimental studies curated by multiple biocurators tend to contain a higher proportion of high evidence levels (strong and definitive) and also tend to have single DGAs (Figure 4.11.F). The MCS showed a high correlation with the HEL (0.79) and the SCS (0.73). These three gold standards demonstrated strong correlations with one another.

Despite using different features of DGAs to generate these gold standards, such as experimental studies containing a high proportion of high evidence level studies, experimental studies containing single DGA, and experimental studies curated by multiple experts, all these gold standards have a similar subset of DGAs. The high correlation of the LLS between the MCS_SCS and HEL_LEL networks is due to the high overlapped DGAs between the gold standards used to score the networks, specifically the MCS gold standard and the HEL gold standard (Figure 4.12.A) and the highest overlapped scored datasets (Figure 4.12.B) between these two gold standards. Similarly, the high correlation of the LLS between OMIM_IES and MG_IES was due to the highly overlapped scored datasets (Figure 4.12. B).

4.3.2.2 Integration of the Scored Datasets Based on the Confidence Scores Using Weighted Sum

Seven integrated networks were constructed based on the two approaches developed to generate the datasets and the gold standard: the data source-based approach and the individual experimental study-based approach. These networks are SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES, UniProt_CDS, and OMIM_CDS. This integration was performed using the weighted sum integration method developed by Lee and colleagues [55](Sections 2.3.3 and 3.2, Equation 3.2) [55]. Figure 4.13 shows the weighted sum distribution for the seven PFINs. The selection of the *D* value for each network was based on identifying the value that produces the highest performance in network link prediction. Therefore the *D* values of 1.1, 1.4, 1.4, 1.5, 1.5, 1.4, and 1.3 were chosen for the SSA_SMA, the HEL_LEL,

A.



B.

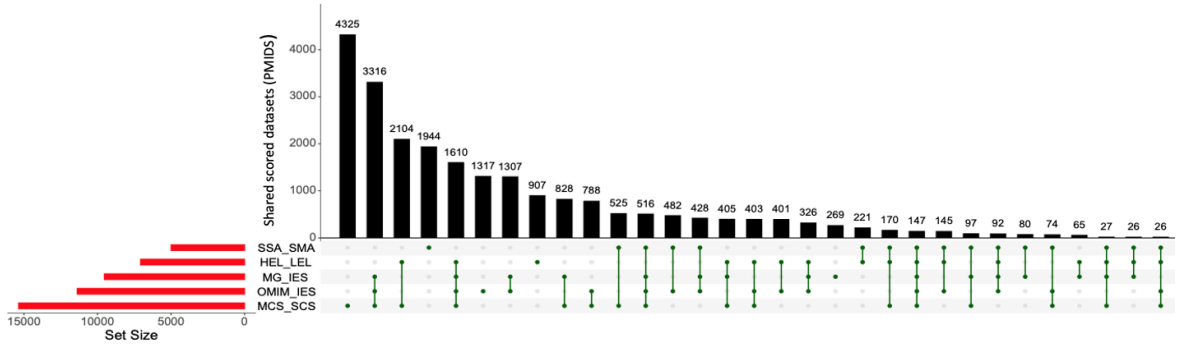


*Figure 4.12 Overlap in the experimental study-based approach. **A** shows 1,196 DGAs shared across five gold standards; **B** shows 147 datasets shared across their corresponding scored networks.*

the MCS_SCS, the MG_IES, the OMIM_IES, the UniProt_CDS, and the OMIM_CDS, respectively (Figure 4.13.D).

The differences between the gold standards among the seven integrated networks in both approaches can be observed in the topological structures of these integrated networks (Table

4.5), which reflects the edges included and excluded based on these gold standards. These differences are also evident in the edge weights (Figure 4.13), which reflect both the confidence scores (Figure 4.10) and the number of scored datasets supporting these edges (DGAs). The edge weights differed among the seven integrated networks, reflecting the differences in the confidence score distributions and the number of datasets included in the integration of each network (Figure 4.13). The individual experimental study-based approach, SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, had higher edge weight distribution than the data source-based approach, UniProt_CDS, and OMIM_CDS. The individual experimental study-based approach also had a larger range and a higher variation of edge weight distribution compared to the data source-based approach. The edge weights of the individual experimental study-based approach were highly correlated (Figure 4.14). For example, the edge weight of the OMIM_IES was highly correlated with the MG_IES (Spearman correlation coefficient = 0.76, $p < 0.0001$). Similarly, the edge weight of the MCS_SCS was highly correlated with the HEL_LEL (Spearman correlation coefficient = 0.74, $p < 0.0001$). This finding can confirm that the EL metric developed by ClinGen based on the frequency of the DGA in the literature matches with the curation strategies used in other curated data sources. There was also a high correlation (Spearman correlation coefficient = 0.67, $p < 0.0002$) between the SSA_SMA and the MCS_SCS. Similarly, the edge weights of the data source-based approach were highly correlated, for example UniProt_CDS is highly correlated with OMIM_CDS (Spearman correlation coefficient = 0.74, $p < 0.0002$).

The topological structures among the seven networks are markedly different (Table 4.5). The data source-based approach produced larger networks than the individual experimental study-based approach, with the highest total number of DGAs for OMIM_CDS (80,867). Additionally, the data source-based approach had a higher number of disease and gene nodes than the individual experimental study-based approach, with OMIM_CDS having the highest number of disease nodes (11,108) and gene nodes (9,417). The smaller size of networks in the individual experimental study-based approach is attributable to the high rate of data loss during scoring.

In the individual study-based approach, removing duplicate evidence was not required, as each DGA experimental study is treated as an individual, distinct dataset. Each piece of DGA

evidence in this approach is independent, aligning with the requirements of the PFIN framework. Conversely, in the data source-based approach, where each data source represents

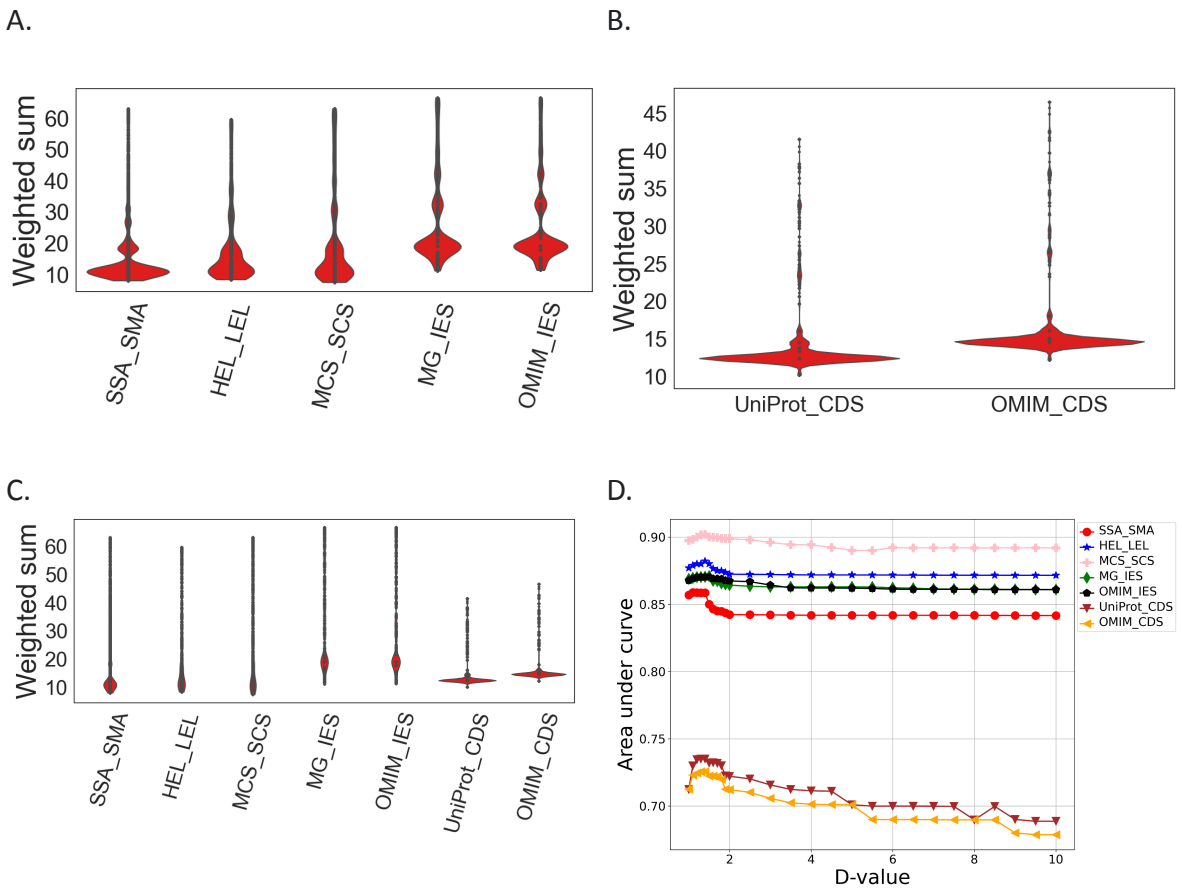


Figure 4.13 Weighted sum distribution. ***A–B*** *show distributions for experimental study-based networks (SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES) and data source-based networks (UniProt_CDS, OMIM_CDS).* ***C*** *combines both.* ***D*** *shows integration using confidence-ranked datasets weighted by a D value, optimized for link prediction. Selected D values: 1.1 (SSA_SMA), 1.4 (HEL_LEL, MCS_SCS, UniProt_CDS), 1.5 (MG_IES, OMIM_IES), and 1.3 (OMIM_CDS).*

a separate individual dataset, eliminating duplicate evidence is important. This need arises because different data sources may curate the same DGAs based on overlapping experimental studies, leading to duplicate evidence.

To assess the impact of duplicate evidence on the PFINs in the data source-based approach, the LLS scores of the datasets and the weighted sums of the DGAs were compared both with and without duplicate evidence. The findings indicated that the LLS scores of the datasets and the weighted sums of DGAs were higher with duplicate evidence. For instance, in the UniProt_CDS network, both the LLS scores and weighted sums with duplicate evidence were elevated compared to those without duplicates (Figure 4.15.A and 4.15.B). Similarly, in the

OMIM_CDS network, the scores and weighted sums were higher when duplicates were included. Although the impact of duplicates was small, the presence of duplicates still influenced the LLS scores and weighted sums, which could affect the PFIN performance.

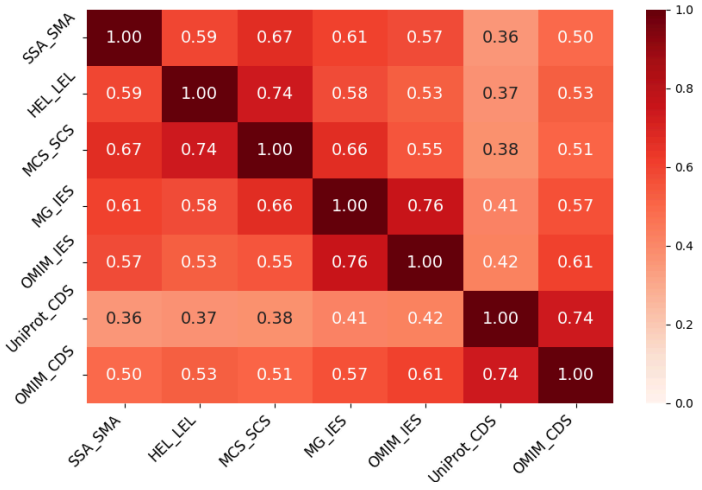


Figure 4.14 Correlation of weighted sums for common DGAs across seven networks: SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES, UniProt_CDS, and OMIM_CDS.

Overall, the data source-based approach yielded fewer connected components in the integrated networks, with 424 for UniProt_CDS and 435 for OMIM_CDS. In contrast, the individual experimental study-based approach exhibited higher numbers of connected components: 476 for SSA_SMA, 674 for HEL_LEL, 1,218 for MCS_SCS, 1,254 for MG_IES, and 1,038 for OMIM_IES. The high number of connected components in the individual study-based approach indicates that these networks are less connected and more isolated, likely due to the smaller number of DGAs and the small number of diseases and genes included in these networks. Conversely, the data source-based approach tends to be more connected, owing to the larger numbers of DGAs, diseases, and genes incorporated into these networks. The data source-based networks, including UniProt_CDS and OMIM_CDS, showed higher average disease and gene degrees compared to individual experimental study-based networks (SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES).

Table 4.5. Network statistics. Topological characteristics for the seven networks: individual study-based networks including SSA_SMA, HEL_LEL, MG scored network, MC scored network, OMIM scored network, and UniProt scored network. Statistics were calculated using the Cytoscape NetworkAnalyser plugin and clustering was carried out using the Markov Clustering Algorithm (MCL).

		Individual Experimental Study-based Networks					Data Source-based Networks	
Topological Structure	DisGeNET (Curated DGAs)	SSA_SMA	HEL_LEL	MCS_SCS	MG_IES	OMIM_IES	UniProt_CDS	OMIM_CDS
Number of disease nodes	11181	6382	3050	5341	4911	5816	10866	**11108**
Number of gene nodes	9703	5020	2069	3430	2511	2909	8865	**9417**
Number of DGAs	84038	24775	7220	12495	7792	11080	79116	**80867**
Average number of neighbours	8.386	4.674	3.664	3.719	2.780	3.120	**8.530**	8.231
Average gene degree	9	4.431	3.490	3.643	3.103	3.809	**8.922**	8.588
Average disease degree	8	3.152	2.367	2.339	1.587	1.905	7.273	**7.280**
Network diameter	14	18	22	20	26	22	14	14
Network radius	7	1	1	1	1	1	1	1
Connected components	417	476	674	1218	1254	1038	424	435
Characteristic path length	4.766	5.497	7.328	7.215	8.410	7.696	4.834	4.833
Clusters	1864	1345	593	957	915	1156	2067	2167

Specifically, UniProt_CDS and OMIM_CDS had a disease degree of 7.27, 7.28 and a gene degree of 8.92, 8.58, respectively, whereas the experimental study-based networks had disease degrees ranging from 1.59 to 3.15 and gene degrees ranging from 3.10 to 4.43. The node degree distribution follows a power law (Figure 4.17), indicating that most diseases are associated with few genes, while a few disorders are hubs linked to many genes (Figure 4.16). The largest connected components varied in size across the networks, with SSA_SMA having 10,303 nodes and 24,077 edges, HEL_LEL having 4,620 nodes and 8,600 edges, MCS_SCS having 5,704 nodes and 10,607 edges, MG_IES having 3,987 nodes and 5,541 edges, and OMIM_IES having 5,968 nodes and 9,311 edges.

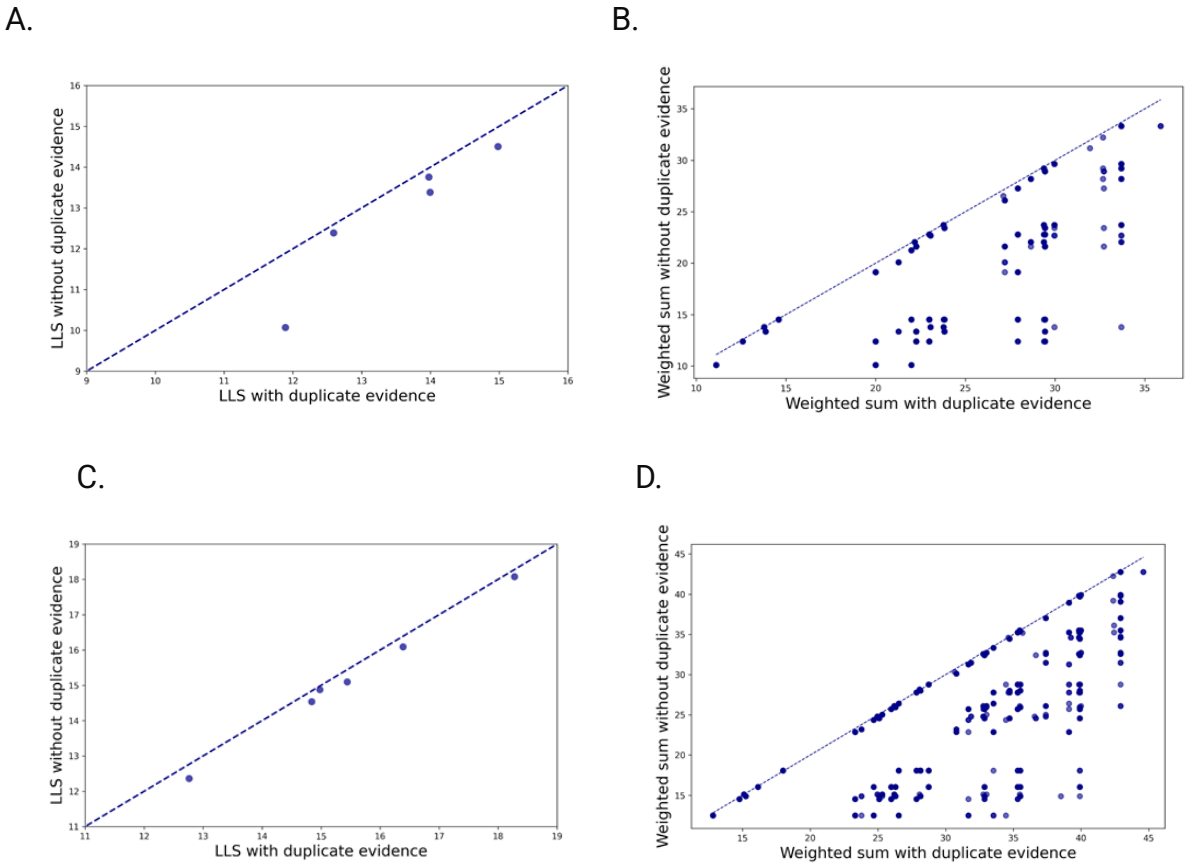


Figure 4.15. Scatter plots showing the impact of duplicate evidence on LLS and weighted sum values in UniProt_CDS and OMIM_CDS networks. In both networks, Panels A and C show LLS comparisons with and without duplicates, while Panels B and D show weighted sum comparisons. Duplicate evidence leads to minor but consistent increases, with most points near the y = x line.

The analysis of overlap between the seven networks in terms of genes, diseases, and DGAs revealed differences attributable to the variances in the gold standards (Figure 4.18). The highest overlap in diseases was observed between OMIM_CDS and UniProt_CDS, with 2,963 diseases shared exclusively between these two networks. The second highest overlap was seen between OMIM_CDS, UniProt_CDS, and SSA_SMA (Figure 4.18.A). Across all seven networks, a total of 1,177 diseases were common. There were 1,862 genes common to all networks, with the highest overlap between UniProt_CDS and OMIM_CDS (Figure 4.18.B). Additionally, the analysis revealed significant overlap in DGAs, particularly between UniProt_CDS and OMIM_CDS. However, only 1,917 DGAs were shared among all networks, reflecting the influence of differences in the gold standards (Figure 4.18.C).

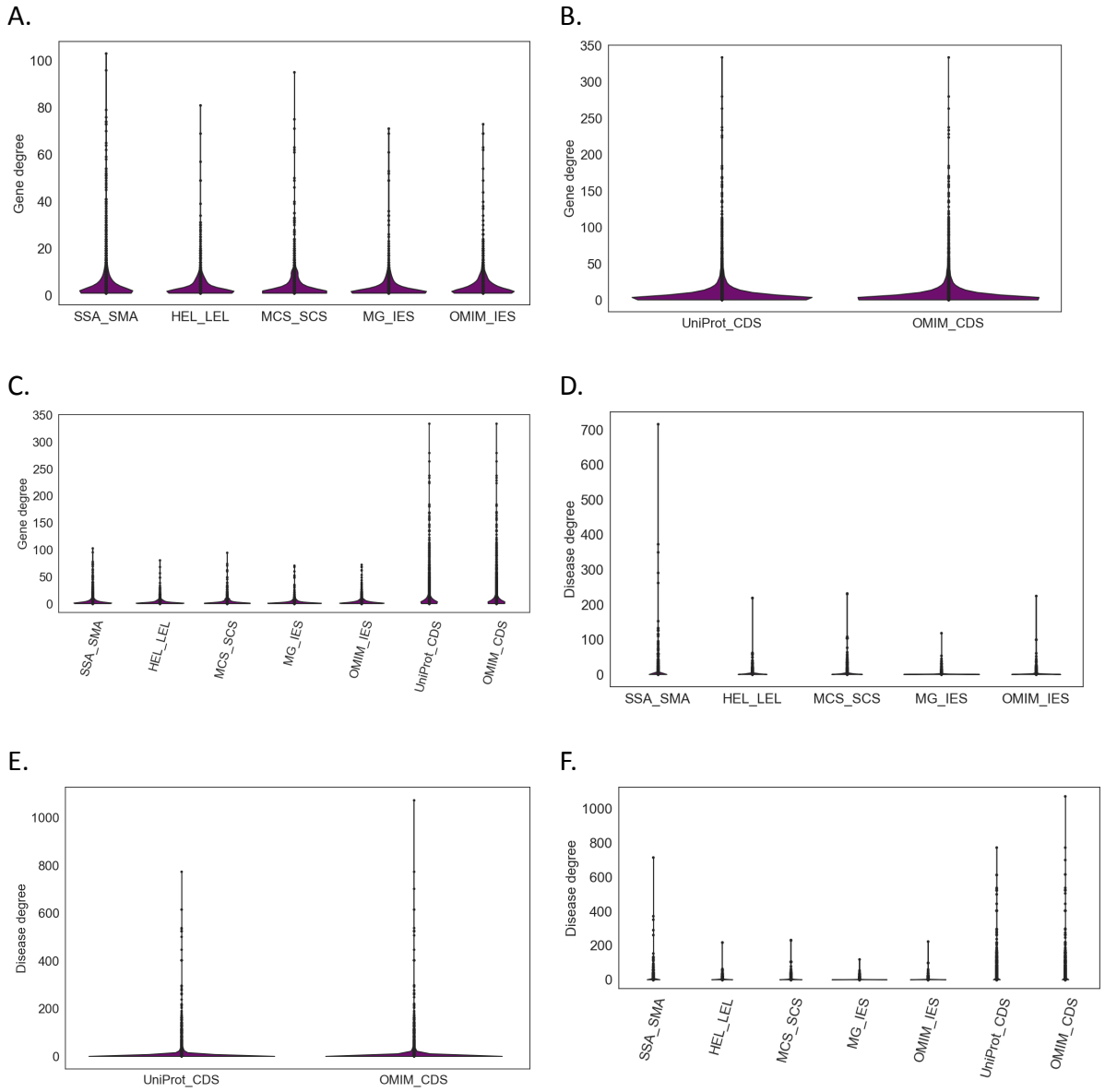
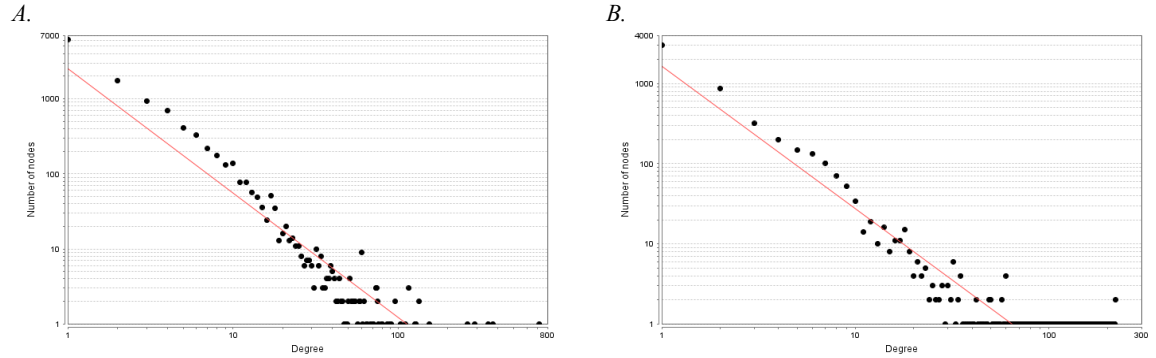


Figure 4.16 Node degree distributions. (A–C) Gene degrees for individual study-based networks, data source-based networks, and all networks. (D, F) Disease degrees for the same sets of networks.



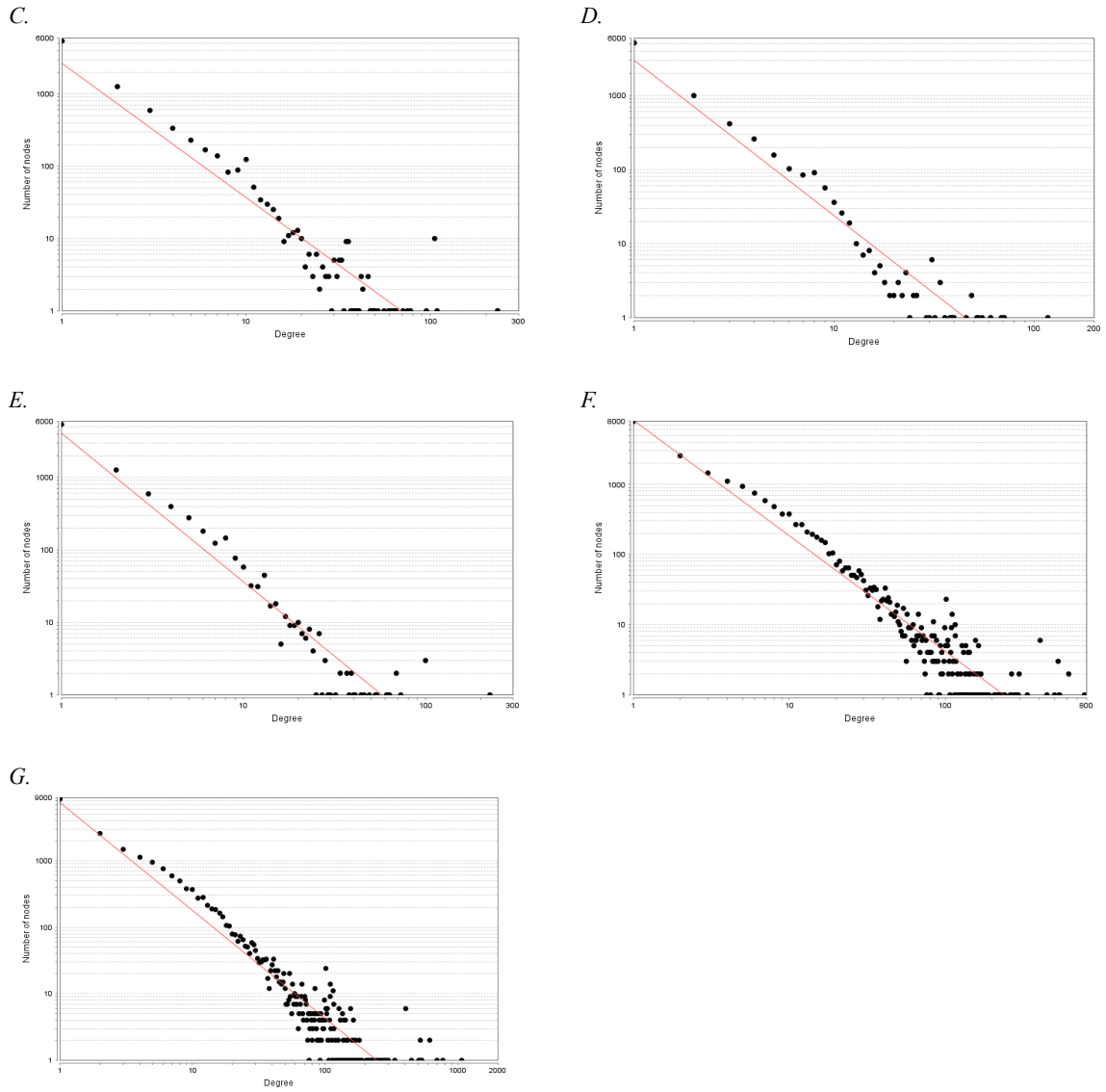


Figure 4.17 Node degree distributions fit a power law model. Panels A–G show results for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES, UniProt_CDS, and OMIM_CDS, with correlations ≥ 0.996 and R^2 values between 0.834 and 0.898, indicating strong fits across all networks.

4.3.3 Network Evaluation

The networks were evaluated using two techniques:link prediction and clustering analysis.

4.3.3.1 Link Prediction

The ability of the networks to predict hidden DGAs was evaluated using the JI algorithm; specifically the weighted bipartite version of JI (Section 3.3.4, Equation 3.13) [270]. The JI, a common neighbour-based algorithm, was selected for its ability to include both network topological structure and edge weights when predicting links between diseases and genes (

Section 3.3.4). Among common neighbour-based algorithms, JI was chosen for its normalised version, which ensures predictions of DGAs are not influenced by high node degree. Additionally, JI demonstrated the highest network performance compared to other existing common neighbour algorithms. Ten-fold cross-validation was employed to assess the network link prediction performance (Section 3.3.4) [273]. This method was chosen because the link prediction task can be computationally intensive, particularly for bipartite weighted networks of large size.

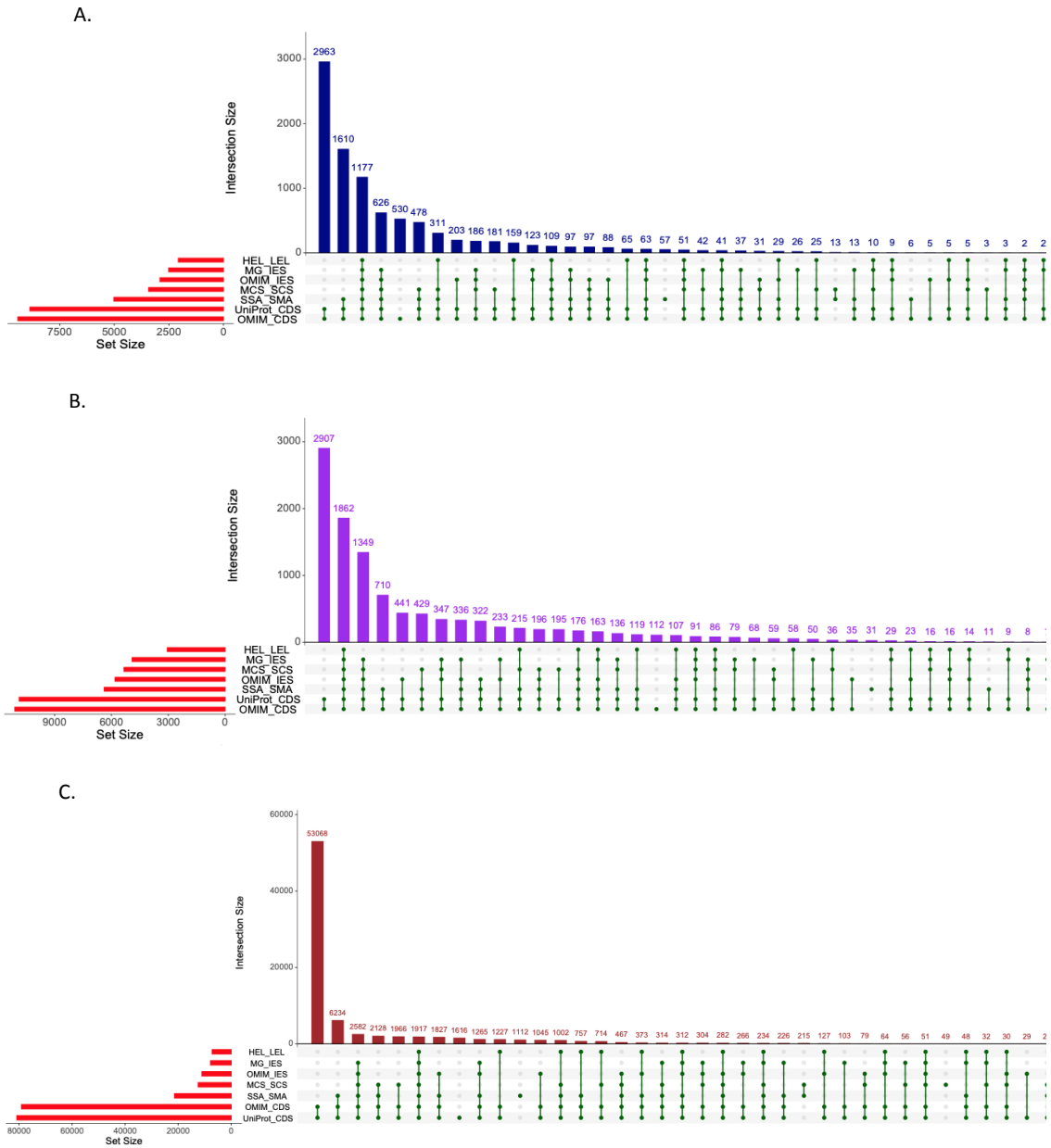


Figure 4.18 Overlaps among integrated networks. (A) Diseases, (B) Genes, (C) DGAs.

The individual study-based networks produced higher AUC values than the data source-based networks. Specifically, SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES achieved AUCs of 0.861, 0.882, 0.902, 0.871, and 0.870, respectively, whereas UniProt_CDS and OMIM_CDS attained AUCs of 0.74 and 0.73, respectively (Figure 4.19). The SE was used to assess the AUCs, and the differences in AUC were found to be statistically significant across all networks.

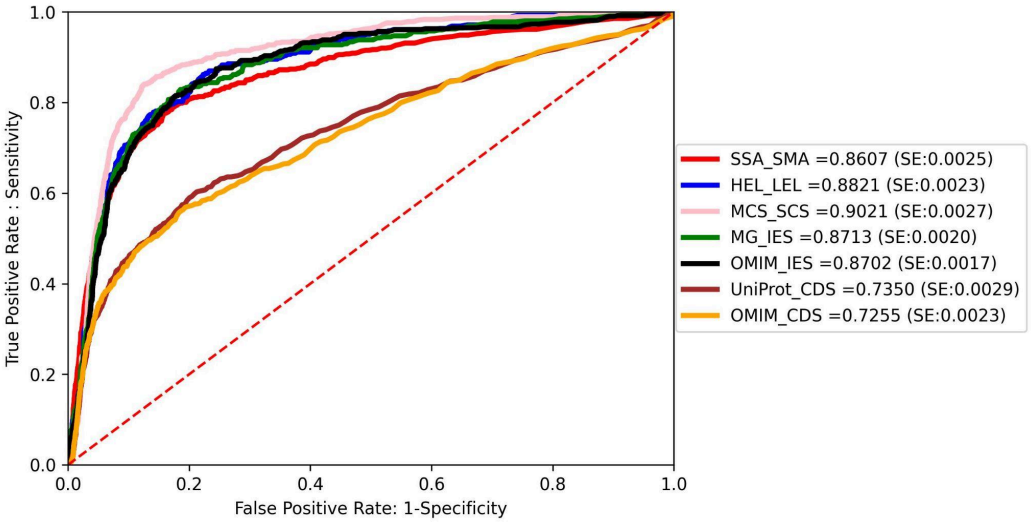


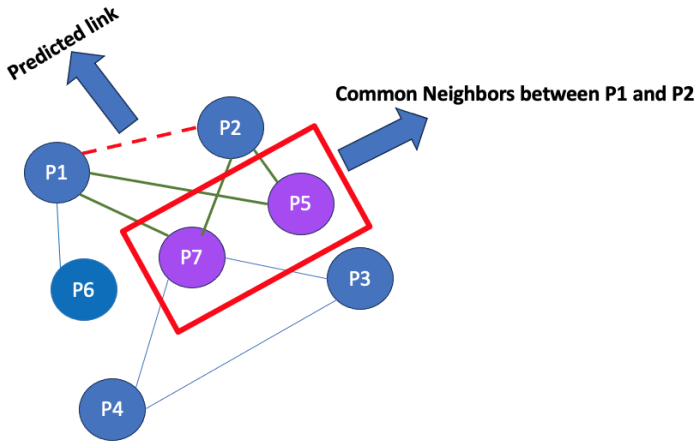
Figure 4.19: ROC curves for link prediction in individual study-based networks (SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES) and data source-based networks (UniProt_CDS, OMIM_CDS).

Link prediction methods, such as neighbourhood-based methods, were initially developed for unweighted and unipartite networks such as PPI networks, however, these methods encounter significant challenges when applied to bipartite networks such as DGA networks, particularly when these networks are weighted. In bipartite networks, nodes of two distinct types (e.g., diseases and genes) are connected, requiring neighbourhood-based methods to use indirect similarity measures that alternate between these node types. This complexity is further exacerbated in weighted networks, where varying interaction strengths complicate similarity calculations. However, Liu et al. [298] and Kart et al. [270] adapted these methods for bipartite networks, with Liu et al. focusing on drug-target interactions and Kart et al. modifying algorithms for weighted networks. Both approaches shift the focus from direct neighbours to second-degree neighbours, facilitating the identification of common neighbours within the same node category. Figure 4.20 illustrates the differences in link prediction

between unipartite and bipartite networks. Figure 4.20.B illustrates the modified version of neighbour-based methods for bipartite networks.

Network link prediction for bipartite and weighted networks also presents some limitations. For example, randomly splitting the network into ten subnetworks for cross validation may result in two disconnected subnetworks, which can hinder the prediction of links. For instance, as illustrated in Figure 4.21, the removal of a link from *D1* to *G2* leads to the formation of two disconnected subnetworks, making the prediction of a link between *D2* and *G1* impossible.

A.



B.

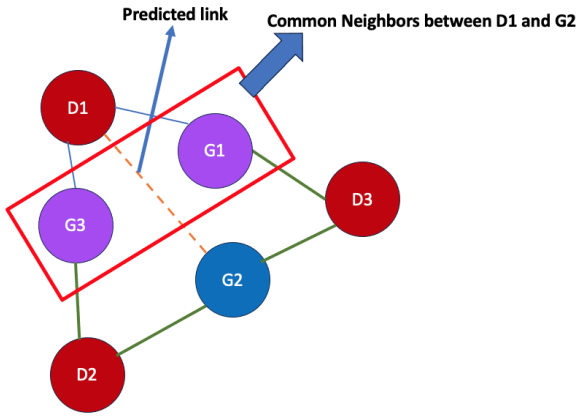


Figure 4.20. Link prediction using common neighbours in unipartite and bipartite networks. (A) In the unipartite network, predicted links (e.g., between P1 and P2) are based on shared neighbours (P5 and P7), shown with green edges. (B) In the bipartite network, link prediction between D1 and G2 uses paths of length two to identify common neighbours (G1 and G3), visualised with green lines.

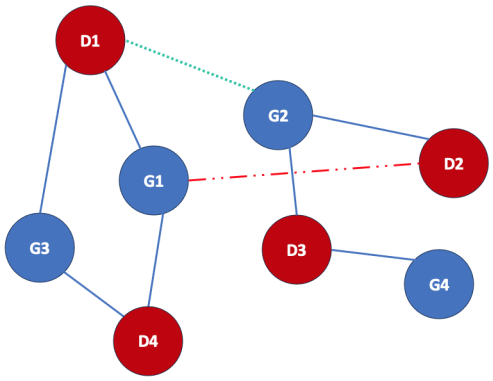


Figure 4.21. Limitation of common neighbours in bipartite DGA networks. Removing the D1–G2 link (green) leads to the formation of two subnetworks, potentially hiding a D2–G1 link (red).

4.3.3.2 Clustering

The second technique used for evaluation was network clustering analysis. The weighted MCL clustering algorithm was used to cluster the networks (Section 3.3.2) [299]. The MCL was chosen for several reasons (Section 3.3.2). The inflation values were determined based on those that produced the highest average of network cluster cohesiveness, with values set at 1.5 for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, and 1.8 for UniProt_CDS and OMIM_CDS (Figure 4.22). The network clusters were analysed to investigate the clustering of diseases and their related genes. To assess the cohesiveness of the network clusters, diseases and their related genes within clusters were examined . It was hypothesised that if a disease is associated with a gene with a high confidence score (edge weight), the disease and that gene should cluster together. To test this hypothesis, the average cohesiveness of the clusters was calculated (Section 3.3.3.1, Equation 3.5).

The individual study-based networks generated fewer clusters than the data source-based approach due to their smaller size, a result of the higher data loss associated with individual study-based approaches. For example, SSA_SMA contained 1,345 clusters ranging from 468 to three nodes, the HEL_LEL network had 593 clusters ranging from 113 nodes to three nodes, MCS_SCS had 957 clusters ranging from 150 nodes to three nodes. MG_IES had 915 clusters ranging from 75 nodes to three nodes, and OMIM_IES had 1,156 clusters ranging from 136 nodes to three nodes. In contrast, the data source-based networks resulted in a larger number of clusters: UniProt_CDS network had 2,067 clusters ranging from 312 nodes to three nodes, while OMIM_CDS had 2,167 clusters ranging from 586 nodes to three nodes.

The average cluster size was similar among all the networks. For instance , that for SSA_SMA, HEL_LEL, OMIM_CDS, and UniProt_CDS was four nodes, whereas for MCS_SCS and MG_IES, it averaged five nodes. Figure 4.23 shows the cluster size distribution.

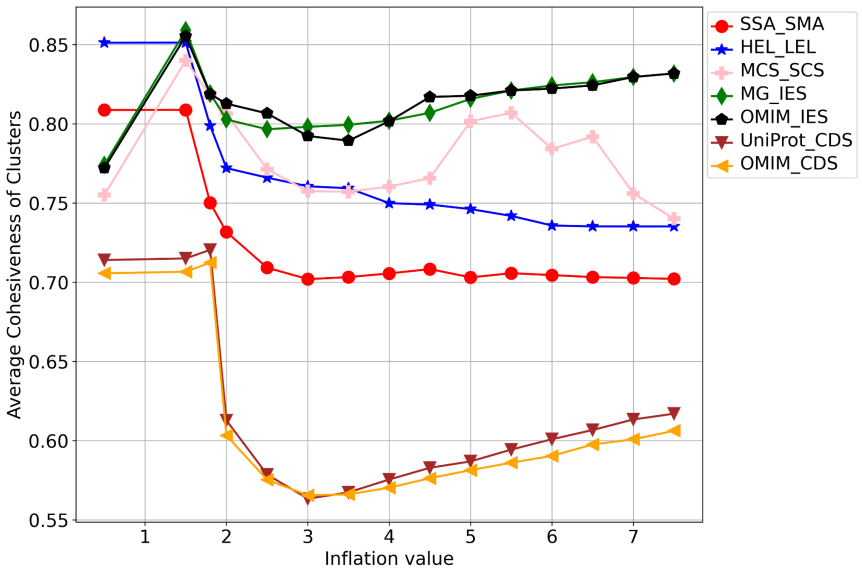


Figure 4.22. Selection of *inflation values were chosen for the MCL clustering based on maximum network clustering cohesiveness.*

The individual study-based networks produced a higher average cluster cohesiveness compared to the data source-based approach, with averages of 0.81, 0.85, 0.84, 0.86, and 0.87 for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, respectively, in contrast to UniProt_CDS with 0.72 and OMIM_CDS with 0.71. Figure 4.24 illustrates the distribution of average cluster cohesiveness for each network.

It was hypothesised that if a disease is associated with a gene with a high confidence score (indicated by a high edge weight), the disease and its related gene should cluster together. Consequently, the average cohesiveness of the clusters should increase when applying edge weight thresholding, aiming to minimise the number of edges with lower weights. To test this hypothesis, an edge weight threshold was applied to the networks and the average cluster cohesiveness was calculated to observe any changes. The selection of the edge weight threshold for each network was based on the network's average edge weight distribution, cluster size distribution, and the number of clusters. The threshold was chosen to ensure that

improvements in the average connectedness of clusters resulted from filtering out low-quality DGAs, rather than from a reduction in the network's size, which could falsely enhance connectedness. Changes in network size due to the edge weight threshold may create misleading interpretations of average connectedness based on size reduction rather than the removal of low-confidence data. Consequently, edge weight thresholds of 12, 12, 13, 13, 13, 17, and 15 were chosen for the SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES, UniProt_CDS, and OMIM_CDS respectively. Figure A.1 illustrates the network size and the

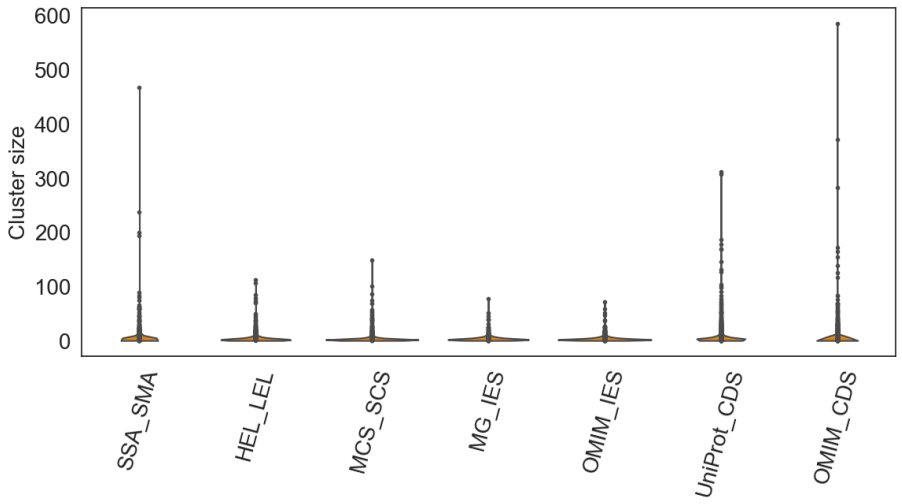


Figure 4.23. Distribution of cluster size for all networks.

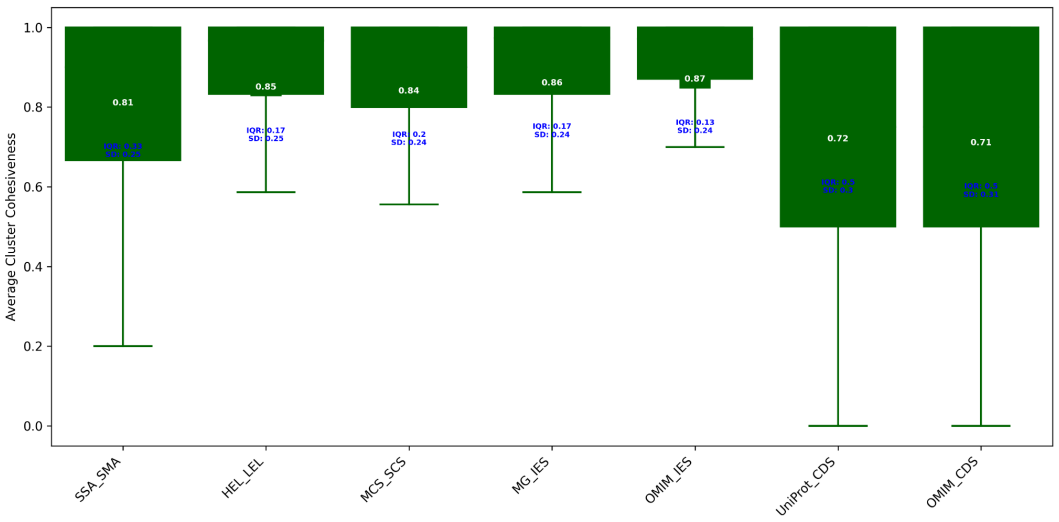


Figure 4.24. Average cluster cohesiveness per network using MCL clustering. White text shows the mean; SD and IQR indicate standard deviation and interquartile range, respectively.

number of clusters at different edge weight thresholds. Figure A.2 shows the cluster size distribution across different edge weight thresholds.

The average cluster cohesiveness for both the individual experimental study-based networks and data source-based networks increased upon applying edge weight thresholds. Specifically, SSA_SMA improved by 0.04, HEL_LEL by 0.01, MCS_SCS by 0.07, MG_IES by 0.01, and OMIM_IES by 0.007 (Figure 4.25). The data source-based networks, including OMIM_CDS and UniProt_CDS, also exhibited improvements in their averages.

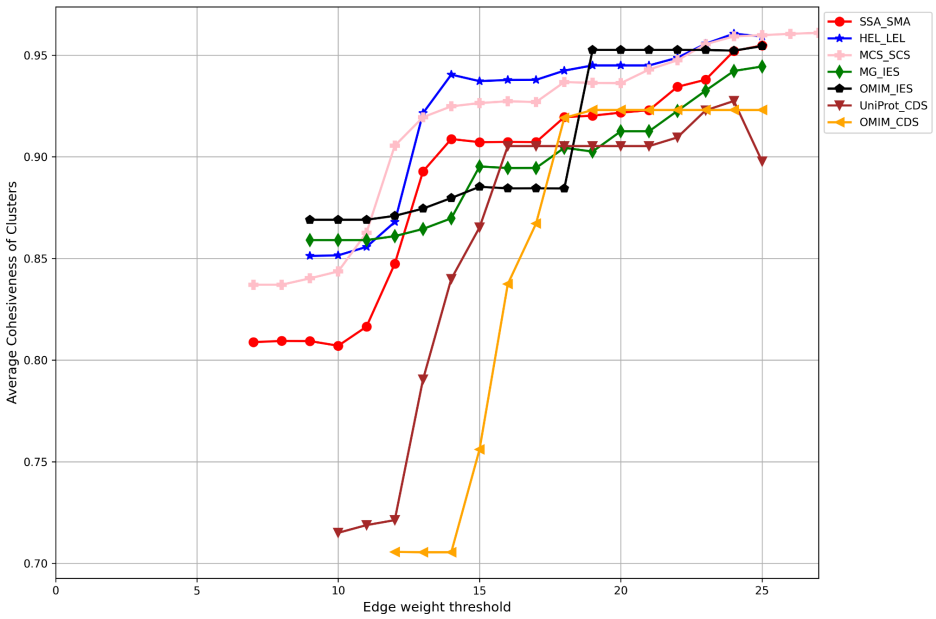
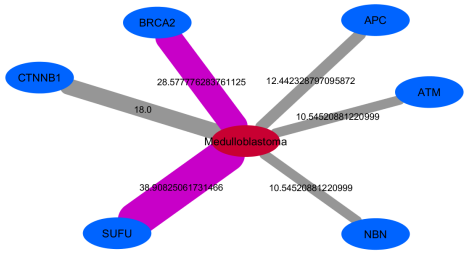


Figure 4.25. Network cluster cohesiveness. Average cohesiveness increased by 0.01 in SSA_SMA, 0.02 in HEL_LEL, and 0.03 in MG_IES and MCS_SCS. No improvement was observed in OMIM_CDS and UniProt_CDS networks.

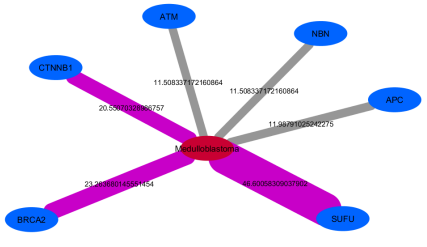
In both network analysis techniques, the individual experimental study-based networks consistently outperformed the data source-based networks. The individual experimental study-based networks demonstrated superior performance, with higher AUC values than the data source-based approach. These networks also exhibited greater average network cluster cohesiveness. Despite the data source-based approach presenting a lower rate of data loss compared to the individual experimental study-based approach, the latter still produced better performance.

The differences between the gold standards were reflected in the topological structures of the networks and the confidence scores between diseases and genes. For example, medulloblastoma, the second most common brain tumour in children, is connected to 50 genes in DisGeNET. However, the node degree and the neighbourhood groups of this disease vary among the networks due to differences in the gold standard. For instance, in individual experimental study-based networks such as SSA_SMA and MCS_SCS, this disease is associated with six genes—ATM, APC, NBN, SUFU, BRCA2, and CTNNB1—with the highest confidence observed with SUFU, while the lowest is with NBN and ATM. Similarly, in MG_IES, the disease is associated with five genes, with SUFU exhibiting the highest confidence and NBN and ATM the lowest. Notably, CTNNB1, present in SSA_SMA and MCS_SCS, is lost in MG_IES. In OMIM_IES, the disease is associated with four genes—ATM, NBN, SUFU, and BRCA2—with SUFU displaying the highest confidence and NBN and ATM the lowest. Additionally, genes such as CTNNB1 and APC, found in other networks, are lost in OMIM_IES. Conversely, in data source-based networks including UniProt-CDS and OMIM_CDS, the disease is associated with all the genes associated with it in DisGeNET, with a disease degree of 50 due to the relatively small data loss rate in this approach compared to the individual experimental study approach. However, the highest confidence is connected with this disease in both OMIM_CDS and UniProt_CDS, involving the same six genes present in the individual study-based approach: ATM, APC, NBN, SUFU, BRCA2, and CTNNB1. Figure 4.26 shows the topological structure visualisations of all networks. The networks were visualised using Cytoscape, as detailed in Section 3.3.1.

A.



B.



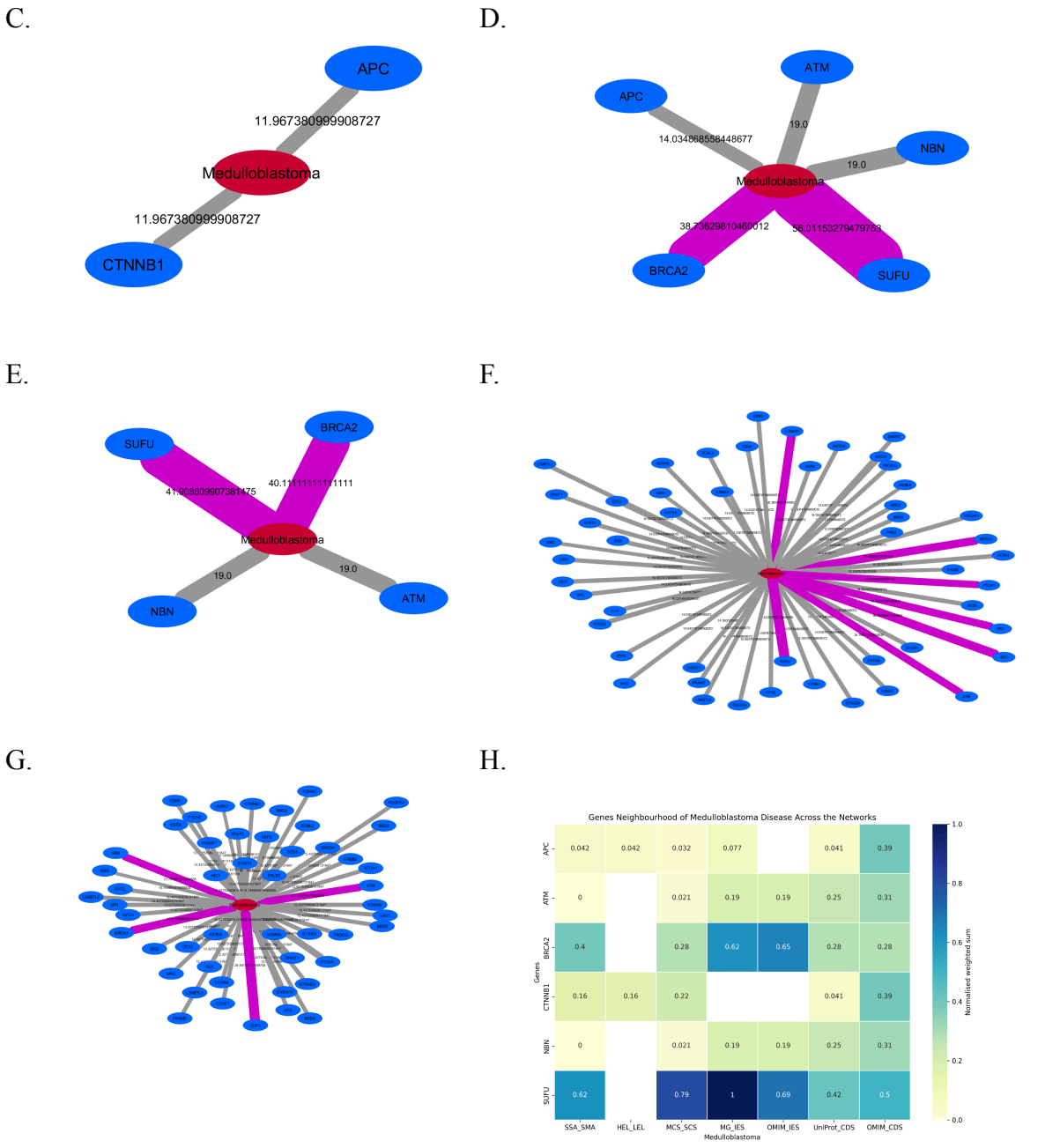


Figure 4.26. Medulloblastoma gene neighbourhoods across networks. (A–E) Individual study-based networks: SSA_SMA, MCS_SCS, HEL_LEL, MG_IES, OMIM_IES. (F–G) Data source-based networks: UniProt_CDS, OMIM_CDS. Red = disease node, blue = gene nodes, purple edges = highest confidence. (H) Heatmap of Medulloblastoma gene neighbourhoods across all networks; white indicates missing genes.

The heatmap illustrates the variation in the weighted sum of Medulloblastoma and its related genes across different networks. Genes such as SUFU and BRCA2 demonstrated higher normalised weighted sums in networks like MG_IES and OMIM_IES, indicating strong associations with Medulloblastoma. In contrast, genes such as ATM and APC showed lower normalised sums, reflecting weaker associations. Additionally, CTNNB1 exhibited a

relatively low weighted sum, particularly in the SSA_SMA and HEL_LEL networks, and was absent in the MG_IES and OMIM_IES networks. While still involved in Medulloblastoma, its confidence is lower compared to SUFU and BRCA2. Differences in the weighted sums are due to the distinct gold standards used to score the networks. These gold standards measure different aspects of the disease. For example, BRCA2 attained higher confidence scores in the OMIM_IES and MG_IES networks compared to MCS_SCS and SSA_SMA, while SUFU maintained consistent confidence scores across all networks, demonstrating a stable association with Medulloblastoma, regardless of the gold standards employed.

Clustering analysis of networks unveiled distinct topological structures in different networks. For instance, in the individual experimental study-based approach, It was observed that SSA_SMA and MCS_SCS, as well as OMIM_IES, grouped medulloblastoma together with Joubert Syndrome 32 and Localised Primitive Neuroectodermal Tumour. Similarly, within MG_IES, Medulloblastoma is clustered with Joubert Syndrome 32, Localised Primitive Neuroectodermal Tumour, Meningioma, and Oculo-dento-digital syndrome. Several studies have demonstrated the biological and genetic connections that explain the clustering of medulloblastoma with the disorders: Joubert Syndrome, Localised Primitive Neuroectodermal Tumour, Meningioma. For instance, Hatten *et al*. showed that Medulloblastoma and Joubert Syndrome share common genetic disruptions in the Sonic Hedgehog (SHH) signalling pathway, which is important for cerebellar development and ciliary function [300]. These cilia-related defects are a hallmark of both Joubert Syndrome and SHH-subtype Medulloblastoma [300]. Similarly, Youn *et al*. reported that mutations in cilia-related genes lead to overlapping developmental and tumorigenic pathways, providing a molecular basis for the observed clustering of these two conditions in genomic analyses [301].

Louis *et al*. demonstrated that medulloblastoma and Primitive Neuroectodermal Tumour share a common embryonic origin and frequently exhibit MYC amplifications and WNT pathway mutations [302]. Their genetic similarities explain the co-occurrence of Medulloblastoma and Primitive Neuroectodermal Tumour in several multi-omic datasets. Furthermore, Taylor *et al*. emphasised that these tumours often show similar histopathological characteristics, which further justifies their close association [303].

Taylor et al. showed that meningioma and medulloblastoma share genetic links through mutations in PTCH1, a gene implicated in both Nevoid Basal Cell Carcinoma Syndrome and sporadic medulloblastoma [303]. Additionally, individuals with Neurofibromatosis type 2 (NF2), a condition predisposing patients to both meningiomas and medulloblastomas, often exhibit the co-occurrence of these tumors. A study examining NF2 patients with multiple cranial meningiomas found that such tumors frequently cluster due to shared genetic factors, including chromosomal instability in regions like 22q, which may contribute to their progression [304]. These findings further support the observation of tumor co-occurrence and clustering in genomic studies of NF2.

In the case of Oculo-Dento-Digital Syndrome (ODDD), Sinyuk *et al*. highlighted that mutations in GJA1, which encodes connexin 43, affect gap junction communication, a process that can be disrupted in cancers such as Medulloblastoma [305]. This dysregulation of cellular communication likely underpins the clustering of Medulloblastoma with ODDD in the current analysis, as shared disruptions in signalling pathways contribute to both syndromes.

Finally, Peixoto *et al.* explained that primary cilia, which play an important role in regulating signalling pathways, are implicated in both ciliopathies and tumorigenesis [306]. They noted that ciliary dysfunction leads to the loss of primary cilia, a feature often seen in many tumours, including Medulloblastoma. This insight provides a genetic and molecular rationale for why Medulloblastoma, Joubert Syndrome, and other disorders with ciliary dysfunction tend to group together. These studies collectively support the observed clustering and offer a deeper understanding of the shared pathways involved.

In the data source-based approach including UniProt_CDS and OMIM_CDS, the clustering analysis revealed that Medulloblastoma is grouped with related conditions, indicating a disease group clustering in which a cluster of related diseases are grouped together such as child Medulloblastoma and adult Medulloblastoma. Figure 4.27 illustrates the clusters of Medulloblastoma diseases across all networks.

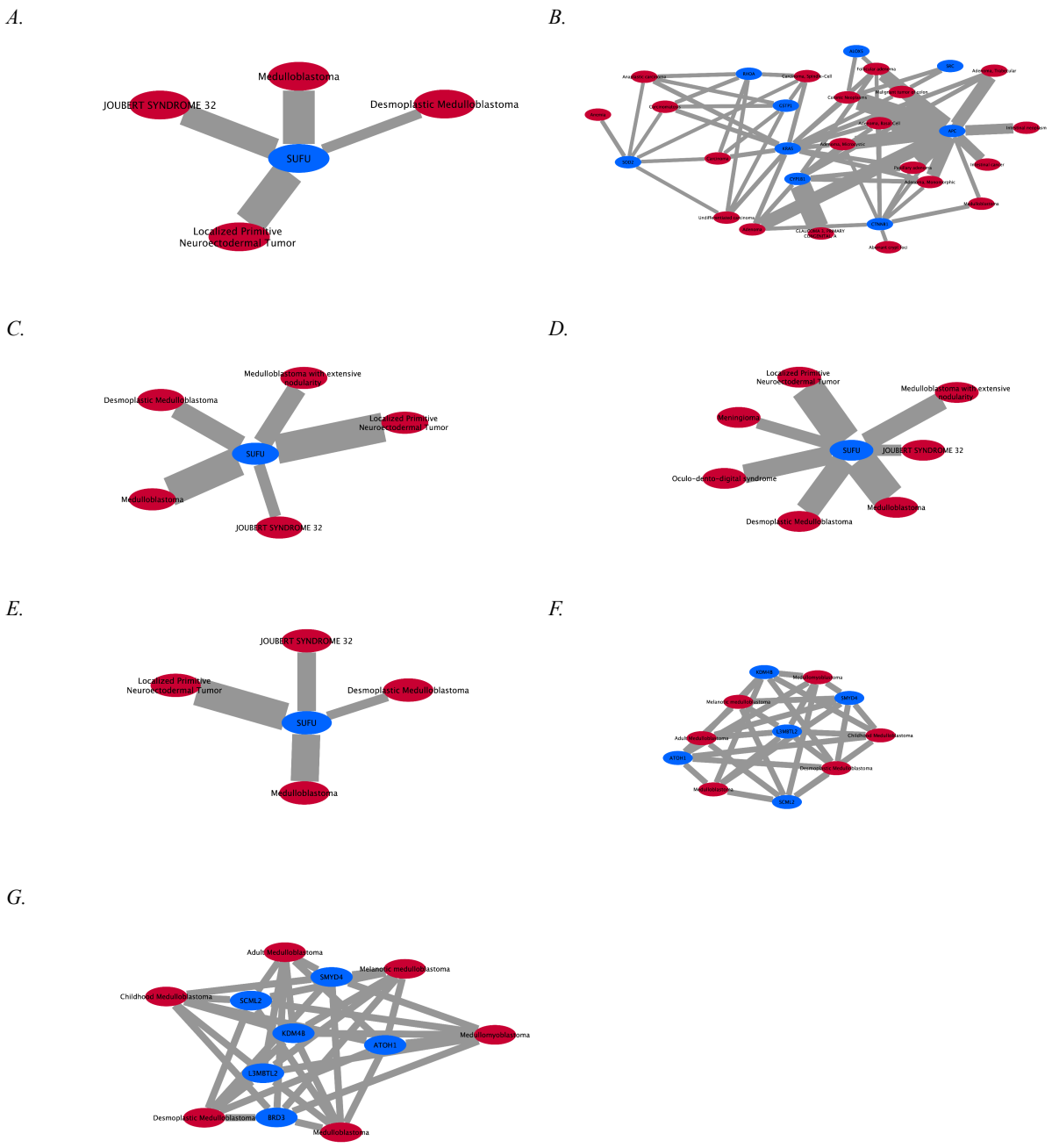


Figure 4.27: Medulloblastoma clusters across all networks. (A–E) Individual study-based networks; (F–G) Data source-based networks. Red = disease nodes, blue = gene nodes, edge width = confidence level.

4.3.4. Investigating External DGA Gold Standards Outside DisGeNET

Alternative sources of DGA data, including GAD [188], CoMAGC [307], and PolySearch [109], were investigated as gold standards. For instance, PolySearch was used as a gold standard, however we found that it focuses on ten specific diseases and their associations, covering 522 DGAs. However, when mapped to DisGeNET using UMLS, only 378 DGAs

were matched. Similarly, we explored CoMAGC as a gold standard, but its focus on cancer disorders may result in discrepancies between datasets and the gold standard, with only 3 unique diseases identified and a lack of disease identifiers, making it challenging to align with other datasets. We also considered the Genetic Associations Database (GAD) as a gold standard, known for its comprehensive coverage of DGAs, encompassing 5330 DGAs with 1652 genes and 923 diseases. Unfortunately, this database is not up-to-date, having been retired with frozen data since 2014. However, the gold standard data need to be up-to-date to maintain relevance.

4.3.5 Existing Methods for DGA Confidence Score Calculation: DisGeNET Score for Disease-Gene Associations

While probabilistic methods like PFINs are commonly used to generate confidence scores in PPI network integration [50], [56], [57], heuristic methods are more prevalent for integrating DGAs and generating confidence scores (Section 2.3) [141], [160]. These heuristic approaches rely on factors such as the frequency, co-occurrence, or similarity measures of DGAs. However, unlike PFINs, which integrate independent evidence of PPIs, heuristic methods have limitations, such as the inflation of confidence scores due to duplicate evidence of DGAs (Section 2.3). For example, DisGeNET employed a heuristic approach to develop its DGA confidence score (Section 2.4.1.2).

A DGA score was developed by DisGeNET [141] to rank the DGAs according to their level of evidence. The DGA score ranges from 0 to 1 and considers the number and type of sources (curation degree, model organisms) and the number of publications supporting the association. The DisGeNET score (S) for DGAs is determined by a formula that includes contributions from various sources (Equation 3.14). However, the presence of duplicate studies among these sources was not considered in this calculation. The omission of duplicate considerations could potentially lead to an inflation of the score, as the same evidence might be counted multiple times, influencing the perceived strength of the DGA. The DisGeNET DGAScore (S) was recalculated to investigate whether the redundancy between data sources was considered during the scoring process. The score was recalculated by applying the same algorithm developed by DisGeNET, allowing redundancy among data sources. It was found that the recalculated score within the duplicate data was identical to the DGAScore assigned

by DisGeNET, which means that removing overlapped DGAs were not considered in the scoring process. To investigate the effect of data duplication on the DGAScore, the redundancy among the data sources was removed (Figure 3.4 for duplicate data between curated data sources), and the same algorithm developed by DisGeNET was applied to recalculate the score. It was found that the recalculated score avoiding duplicate data differed from the score assigned by DisGeNET. Even though the difference is slight, duplicate data can lead to biases within the network, upweighting edges with duplicate evidence. For example, the (Leigh Syndrome, LRPPRC) DGA from the study PMID 12529507 is present in UNIPROT, CTD_human, ORPHANET, CLINVAR, and GENOMICS_ENGLAND. According to the heuristic method developed by DisGeNET for calculating confidence scores, the (Leigh Syndrome, LRPPRC) DGA is reported to be supported by five pieces of evidence. However, this is not accurate, as the same evidence (same DGA generated by same experimental study PMID 12529507) is represented multiple times across different curated data sources. Figure 4.28.A shows the correlation between the DGAScore assigned by DisGeNET and the recalculated score without data duplication. Addationaly, the density plot showed that low DGAScores increased when duplicate data was removed, while the density of low DGAScores decreased with the presence of duplicate data. This indicates that duplicate data can upweight the confidence scores, even if the effect was slight(Figure 4.28.B and C). When comparing the scores assigned by DisGeNET with the edge weights of the data source-based networks, we observed a high correlation between the DGAScore and OMIM_CDS (0.84) as well as UniProt_CDS (0.65) (Figure 4.28.D). Similarly, for the individual experimental study-based approach, the DGAScore exhibited a high correlation with MG_IES (0.56) and OMIM_IES (0.62)(Figure 4.28.D).

The difference between DGAScores with and without redundancy was slight due to the thresholds applied to determine the number of data sources supporting each DGA (see Section 3.4). A threshold of 9 was used for literature data, a threshold of 2 for curated data sources, and a threshold of 0 was applied for both modelling and inferred data sources. These thresholds could lead to no difference between DGAscores with duplicated evidence and those without; for example, if the number of animal models supporting a DGA is greater than zero, it would yield a contribution of 0.2, regardless of whether there is duplicate evidence or not. In contrast, in literature data sources, if the threshold is greater than 9, it would contribute 0.1, which may result in a difference between scores with duplicate evidence and

those without. Although the impact of redundancy on the scores is slight, it nonetheless affects the overall confidence in the DGA associations.

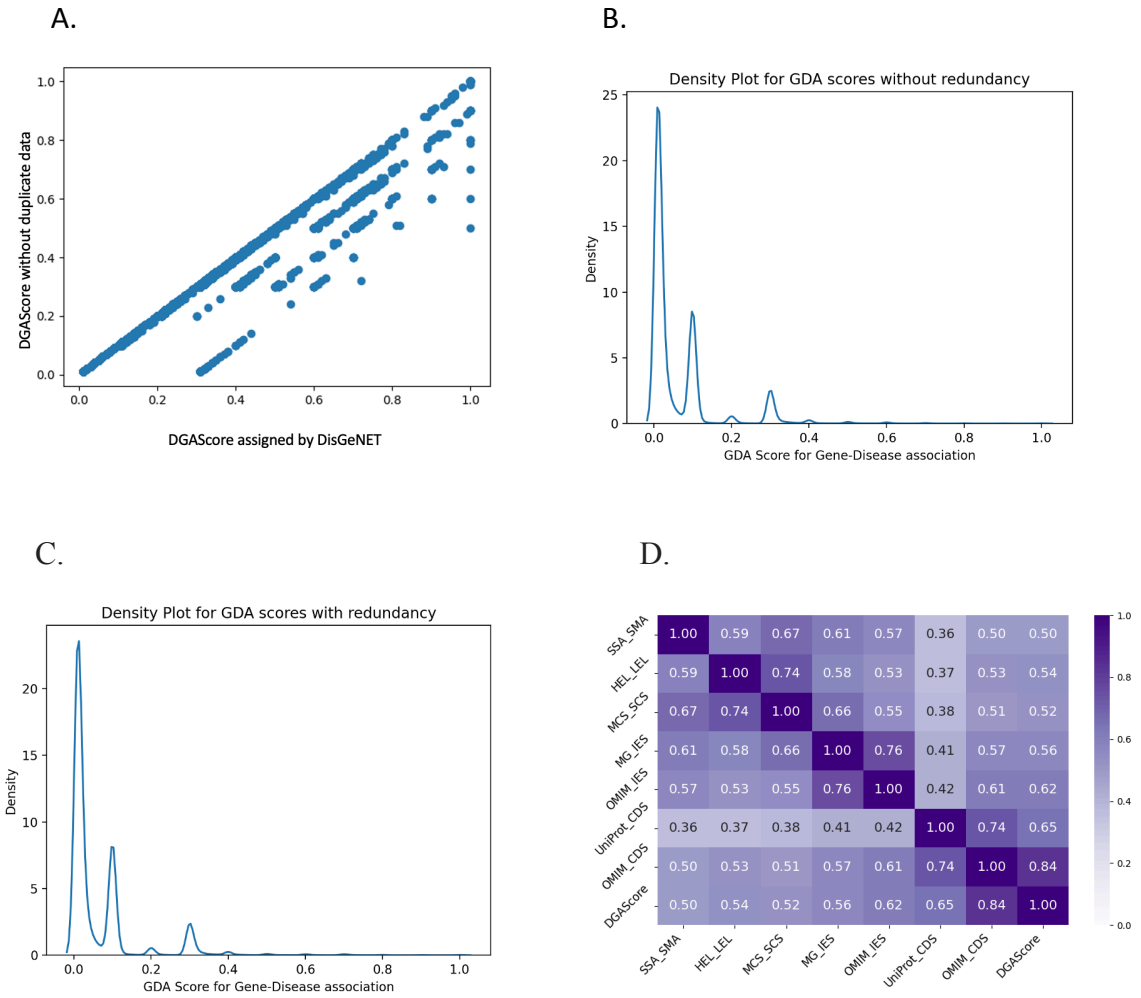


Figure 4.28. Impact of duplicate data on DGA score distribution. (A) Correlation between DisGeNET scores with and without redundancy, showing slight differences due to upweighting from duplicate evidence. (B–C) Density plots show higher low-score density without redundancy (B) and slightly lower peak with redundancy (C), indicating distortion from duplicates. (D) Correlation between weighted sums across seven networks and DisGeNET DGA scores.

4.4 Conclusion

In this work, novel strategies for constructing and evaluating DGA PFINs were researched and developed. Two innovative approaches were investigated to identify the most suitable methods for determining the DGA gold standard data and selecting the individual DGA datasets. Unlike existing DGA networks, which are either unweighted or rely on heuristic

methods based on frequency, co-occurrence, or similarity of DGAs, to generate DGA confidence scores, DGA PFINs provide a more accurate and unbiased approach to DGA network integration. By scoring DGA datasets against high-confidence gold standards prior to integration, PFINs effectively reduce noise and ensure that only high-quality datasets are incorporated. The PFIN approach guarantees that datasets remain independent and free from duplicate data which make the PFIN approach a more robust method for DGA network construction. In contrast, heuristic methods such as the DisGeNET confidence score are vulnerable to inflated confidence scores due to duplicate data, where repeated evidence is disproportionately represented.

The applicability of the PFIN approach to DGA networks was explored through the development of two key strategies: the data source-based approach and the individual experimental study-based approach. In the data source-based approach, curated data sources from DisGeNET were evaluated against two primary gold standards: an internal gold standard within DisGeNET (the UniProt_CDS network) and an external gold standard (the OMIM_CDS network). The individual experimental study-based approach employed two methods. The first method involved dividing individual experimental studies into gold standard data and individual datasets based on experimental scale level (SSA_SMA network), experimental evidence level (HEL_LEL network), and experiment curation level (MCS_SCS network). The second method scored all individual experimental studies from DisGeNET against external gold standards derived from OMIM (OMIM_IES network) and monogenic data (MG_IES network). As a result, seven integrated networks were developed for comparison.

The data source-based networks experienced less data loss compared to the individual experimental study-based networks. This was likely because treating each single curated data source as an individual dataset improves disease coverage. Since each data source includes multiple experimental studies on DGAs, it broadens the range of diseases represented within the dataset and increases the likelihood of their inclusion in the gold standard. However, certain data sources, such as PSYGENET and GCI, were lost when scored against the gold standard, as they focus on specific diseases. To address this issue, future work could prioritise the use of general-purpose data sources, which cover a broader range of diseases, as individual datasets. Nevertheless, specific-purpose data sources, such as those focused on

Alzheimer's, cancer, or psychiatric disorders, can still be employed to construct PFINs for targeted applications.

Duplicate data between the gold standard and individual datasets can introduce bias into the resulting scores. Identifying the optimal strategy for handling redundancy and removing duplicate data remains a challenge. In this study, duplicate data were removed randomly, retaining duplicates in only one data source. However, this method may not be the most effective for eliminating duplicates. Future strategies may involve removing duplicates based on dataset rank, where duplicates are removed from lower-ranked datasets. Alternatively, duplicates could be eliminated from larger datasets, preserving smaller ones to prevent dataset size from skewing the results.

The use of external gold standards, such as OMIM, Monogenic, GAD, CoMAGC, and PolySearch, required disease identifier mapping, which led to the loss of some DGAs, as not all disease IDs could be mapped to DisGeNET. In this work, the Mondo ontology and Metathesaurus were used for disease identifier mapping. However, some DGAs from OMIM and Monogenic datasets remained unmapped. Additionally, some disease IDs were mapped to multiple identifiers, necessitating manual curation to determine unique IDs. In databases like PolySearch, only disease names are provided, making it challenging to use names as unique identifiers. To resolve these issues, future efforts will include manual processing of the mapping process or the use of machine learning techniques to enhance the accuracy of disease ID mapping.

Overall, the results indicated that the individual study-based approach outperformed the data source-based approach, achieving higher AUCs in link prediction and greater average cluster cohesion in cluster analysis. In the next chapter, the focus will be on addressing the limitations encountered with the individual experimental study-based approach, while setting aside the data source-based approach for future work. Specifically, alternative gold standards, such as PPIs and pathways, will be explored. Scoring DGA datasets against non-DGA gold standards may allow us to investigate molecular-level mechanisms underlying disease associations.

For individual dataset identification, treating each experimental study as a separate dataset leads to a high rate of data loss, as most DGA experimental studies concentrate on specific

disease or group of related diseases. If a disease is not included in the gold standard, its dataset is discarded. At present, achieving a gold standard that encompasses all diseases is not feasible. The high rate of data loss should be based on dataset quality, as PFINs tend to include high-quality evidence, rather than on datasets focused on well-studied diseases. In the next chapter, a text mining approach, that systematically groups individual experimental studies into larger datasets based on the similarity of the experiment techniques used to generate DGAs, will be introduced. This method may expand the focus of datasets by treating multiple experimental studies as a single dataset, rather than single individual studies.

Chapter 5
Constructing Disease-Gene Association PFINs with Gene-Gene Association Gold Standards

5.1 Introduction

In the previous chapter, the applicability of the PFIN approach for building DGA networks was explored using DGA data. Multiple and diverse strategies were developed to define the two main components of the PFIN: the gold standard data and the individual datasets. Two primary methods were introduced for identifying gold standards and individual datasets: the individual experimental study-based approach and the data source-based approach. The findings revealed that the individual experimental study-based approach outperformed the data source-based approach. Consequently, the focus of the work described in this chapter was to address the limitations inherent in the individual experimental study-based approach to identify the individual datasets and the gold standard. Certain limitations were identified across three distinct dimensions: the definition of gold standard data, the definition of individual datasets (Section 4.3.1), and the evaluation of network performance (Section 4.3.3.1). Limitations in the gold standards and individual datasets definition manifested as significant data loss and a prevalence of infinite scores, leading to reduced variability in confidence scores (Section 4.3.2.1). As a result, the purpose of the work described in this chapter was to tackle these constraints. In the context of the PPI networks, it is effective to treat individual studies as separate datasets, contributing distinct pieces of evidence [57]. However, when dealing with DGA data, this approach may not be the most optimal one, given that the majority of studies primarily focus on a single disease or a group of related diseases. This bias results in the potential exclusion of a dataset if it does not correspond to the diseases included in the gold standard. Some data loss is anticipated as PFINs aim to prioritise higher-quality evidence and facilitate network thresholding. However, this reduction should be guided by data quality rather than the focus on well-studied diseases. However, creating a gold standard that includes every possible disease is not currently possible [308]. Furthermore, there is a lack of suitable network analysis methods for assessing the effectiveness of these integrated probabilistic networks. Many of the existing analysis techniques were initially developed for unipartite and unweighted networks [309]. Although some efforts have been made to modify these approaches for unipartite weighted

networks [270], [271], they continue to exhibit certain constraints (Section 4.3.3.1) [271]. In this chapter, it was investigated how to tackle the constraints present in the identification of the gold standard data, the individual datasets, and the techniques used to evaluate network performance. One of the main constraints in identifying the gold standard and individual datasets was the high rate of data loss. In response, it was explored whether using a new type of gold standard could mitigate the data loss issues. Therefore, several objectives needed to be achieved:

1- Identification of the gold standard data through alternative data types, non-DGA gold standards including shared pathways or highly reliable PPIs. In this scenario, DGA datasets were evaluated against their molecular foundations, such as PPIs and pathway analysis. Individual DGA datasets were assessed based on their alignment with established cellular associations, and evaluating all DGAs on their concordance with established biological knowledge. As it was observed a significant data loss when employing DGA gold standard data, in this chapter, it was explored whether the use of non-DGA gold standards data can help mitigate this problem.

2- Applying text mining techniques to DGA literature, particularly focusing on the experimental methodologies used to identify DGAs. In this scenario, DGA experimental studies were grouped based on their experimental techniques. This approach allowed studies utilising similar techniques to be grouped together, thus treating them as individual datasets representing distinct evidence of DGAs. As it was observed a significant data loss when treating an individual experimental study as an individual dataset, in this chapter, it was investigated whether treating multiple individual experimental studies as an individual dataset can help mitigate this problem.

3- In order to overcome the limitations in network evaluation techniques, the scored DGA individual datasets were collapsed and integrated into Gene-Gene Association (GGA) PFINs and Disease-Disease Association (DDA) PFINs to enable evaluation of the resulting PFINs using unipartite networks evaluation techniques.

5.2 Source data

DisGeNET [141] 2021-v7.0 was used as the data source for the individual datasets definition. In this context, curated DGA experimental studies from DisGeNET were used to represent the individual datasets for building DGA PFINs. Each distinct DGA experimental study corresponds to a separate individual dataset (Sections 2.5 and 3.1).

BioGRID [87] version 4.4.213, August 2022, for *Homo sapiens* was chosen as the data source for the non-DGA gold standard data (Section 3.1).

Reactome pathway [253]**,** version 81, 2022, was also used as a non-DGA gold standard (Section 3.1).

IntAct [187], version 1.0.4, was also used as a source of non-DGA gold standard (Section 3.1).

The **EFO** ontology [254] was used to build a dictionary containing experimental terms to extract experimental techniques employed in biomedical literature. EFO is a structured, hierarchical ontology developed to annotate experimental variables in biomedical research. EFO is designed to work in conjunction with other ontologies, such as the Gene Ontology [264], and the Human Phenotype Ontology [223] (Section 3.1).

EDAM [255] ontology was also used to build a dictionary containing experimental terms to extract experimental techniques employed in biomedical literature (Section 3.1).

5.3 Results and Discussion

5.3.1 Addressing Limitations in Gold Standard Data Definition

In the previous chapter, two methods were presented for defining individual datasets and the gold standard data: the individual experimental study-based approach, which utilised individual experimental studies as datasets, and the data source-based approach, which utilised curated data sources as datasets. The findings from the previous chapter demonstrated that constructing PFINs through the individual study-based approach, in which DGA

experimental studies are used to represent individual datasets, so that each DGA experimental study is treated as a single individual dataset, yielded superior performances in network analysis techniques compared to the data source-based approach. Consequently, in this chapter, the individual study-based approach was used for defining the datasets. The individual experimental study-based approach was proven effective in defining distinct datasets for constructing PPI PFINs, yielding promising outcomes [57], [90], [192] (Section 2.3.3). Nonetheless, this approach presented two major issues when applied to DGA data: a high rate of data loss and a high rate of infinity scores. The high rate of data loss and infinity scores stems from the fact that the majority of DGA studies focus on a single disease or related disease groups. Consequently, datasets, represented by individual DGA experimental studies, were excluded if the diseases they focused on were not included in the gold standard. Datasets containing diseases included in the gold standard exhibited high overlap, resulting in infinity scores. Data loss is a significant concern because this high rate of data loss should primarily occur due to the quality of the dataset rather than the absence of overlap between the dataset's focus and the gold standard. Additionally, the high rate of infinity scores poses another issue, as datasets with those scores are assigned the highest scores, leading to a lack of variability in the confidence scores distribution. This uniformity in confidence scores impacts the weighted sum, as datasets are integrated based on their ranking from highest to lowest scores. Creating a gold standard that includes every disease is not currently possible.

A potential approach could be to use gold standard data of a different type, non-DGA gold standard, such as shared pathway or high-confidence PPIs [308]. In this case datasets would be scored on how well they reflect known cellular associations [310]. Therefore, in the research presented here, new types of gold standard data, including PPI gold standard (BioGRID and IntAct), and shared pathway (Reactome pathway) were investigated for use as gold standards. These gold standards are external and separate data sources from DisGeNET. The use of non-DGA gold standards were evaluated by comparing them with the use of DGA gold standards defined in the previous chapter (Section 4.3.1.2). Therefore, for individual datasets identification, manually-curated data from DisGeNET was split by the PMID to identify the individual experimental studies generated by DGAs. The individual experimental studies from DisGeNET were scored against the three non-DGA gold standards.

Interactions from BioGRID were split by throughput and experimental system type to identify the low throughput individual studies containing physical interactions only. Dividing BioGRID by interaction types and throughput techniques produced 775,325 total unique physical interactions and 107,076 total unique low throughput interactions. The overlap between physical and low throughput interactions amounted to 106,411 unique interactions and 16,427 unique genes. Only interactions of the physical and low throughput types were considered, self-loop interactions were removed, and only human data was used. Using low throughput interaction studies is often favoured because they are typically considered to be of higher quality and more reliable compared to high throughput studies, involving analysing a smaller number of interactions. Similarly, physical interactions are often considered more reliable than functional interactions, providing direct evidence of molecular contacts between proteins. Therefore, low throughput and physical interactions were used to build the gold standards. Table 5.1 shows the statistics of the interactions in BioGRID.

Interaction pathways from Reactome were also investigated for use as a source for the non-DGA gold standard data. Only Human data was used which contains 4968 total unique genes and 20334 total unique interactions. The UniProt gene identifiers were mapped to NCBI identifiers. After identifier mapping, 4856 total unique genes and 19599 total unique interactions were considered. Retrieve/ID mapping tool from UniProt was used for identifier mapping (Section 3.7). Self-loop interactions were removed.

Interactions from IntAct were also used as non-DGA gold standard data. Only human interactions were used. The total number of unique genes was 1429 and only genes mapped to NCBI were considered, in total 1413 genes. ID mapping was performed to map UniprotKB identifier to NCBI identifier using Retrieve/ID mapping tool from UniProt (Section 3.7). The total number of unique interactions was 8059 and only 7994 mapped interactions were considered. These three non-DGA gold standards were used to score the individual experimental studies from DisGeNET (Figure 5.1).

5.3.1.1 Datasets Scoring and Integration

The division of DisGeNET by PMID produced 39,574 individual experimental studies (IES). Out of these studies, 34,024 consist of only one gene associated with either a single disease or multiple diseases, while 918 studies lack any shared diseases among all possible pairs of

genes. Since the non-DGA gold standard evaluation necessitates a minimum of two genes and at least one common disease between them, a total of 34,942 datasets could not be evaluated using the non-DGA gold standards. The confidence scores were calculated by scoring the datasets against the gold standard data using the Bayesian statistics approach developed by Lee and co-workers, which calculates a log- likelihood score for each dataset (Equation 3.1). Datasets are lost if they have poor overlap. Datasets may score infinity if they have perfect overlap with the gold standard. The IES were assessed against the three non-DGA gold standards: BioGRID, Reactome pathways, and IntAct, respectively. When scored against BioGRID, 1,216 datasets were scored, with 559 of them yielding infinite scores and 38,358 (97%) datasets were lost, resulting in a total of 23148 DGAs. Scoring IES against shared pathway interactions resulted in 990 datasets, out of which 542 received infinite scores, and 38584 (97%) were lost, producing 18954 DGAs. In the case of IntAct, 62 datasets were scored, and 32.27% of them received infinite scores, and 39512 (99.84%) were lost, leading to a total of 4,787 DGAs.

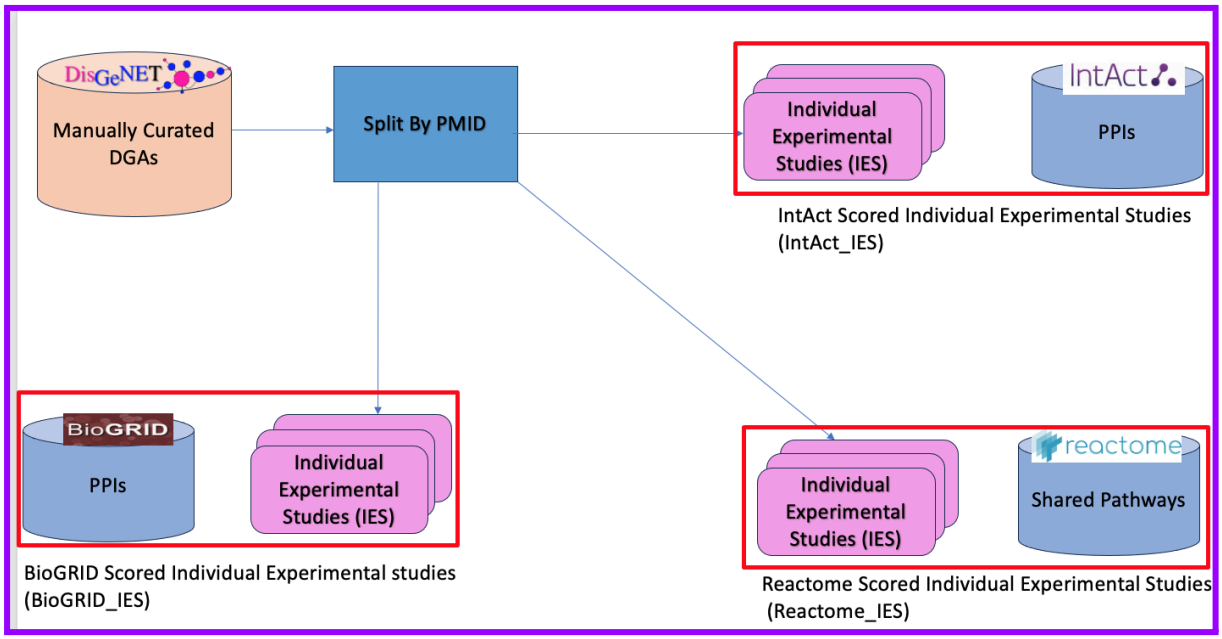


Figure 5.1: Non-DGA gold standard approach for scoring individual DGA experimental studies to build DGA PFINs. Three gold standards—PPI data from BioGRID and IntAct, and shared pathways from Reactome—were used to score studies from DisGeNET using LLS.

Table 5.1 Statistics on the BioGRID database including the total and the type of unique genes and interactions

Type of interaction / Total	High Throughput Interactions	Low Throughput Interactions	Physical Interactions	Genetic Interactions	All Interactions
Unique interactions	700810	107076	775325	15419	789507
Unique genes	24107	17356	26982	6008	27400

On average, the non-DGA gold standards showed a higher overlap with the datasets in terms of individual genes, with approximately 75.55% for BioGRID_IES, around 34.84% for Reactome_IES, and about 9.19% for IntAct_IES. In contrast, the DGA gold standards showed lower gene overlaps, with approximately 54.13% for SSA_SMA, 30.94% for HEL_LEL, 35.00% for MCS_SCS, 28.21% for MGS_IES, and 27.07% for OMIM_IES. However, the non-DGA gold standards had a lower overlap with the datasets in terms of DGA, accounting for approximately 3.20% for BioGRID_IES, 1.48% for Reactome_IES, and 0.12% for IntAct_IES. In contrast, DGA gold standards had higher DGA overlaps, with approximately 8.51% for SSA_SMA, 3.78% for HEL_LEL, 7.72% for MCS_SCS, 3.45% for MG_IES, and 3.59% for OMIM_IES (Table 5.2).

Non-DGA gold standards including PPIs and shared pathways resulted in a high variance in the distribution of confidence scores compared to DGA gold standards including, SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES as illustrated in Figure 5.2. The reduced variance observed in the confidence score distribution for DGA gold standards can be attributed to their higher overlap with individual datasets in terms of associations and the subsequent higher occurrence of infinity scores (Table 5.2). DGA gold standards had a higher rate of infinity scores with 53.45% for SSA_SMA, 74.28% for HEL_LEL, 54.59% for MCS_SCS, 74.41% MG_IES, and 70.00% for OMIM_IES than for none-DGA gold standards with 45.97% for BioGRID_IES, 54.75 for Reactome_IES, and 32.27 for IntAct_IES. These infinity scores contribute to minimising the variability in the LLS across the datasets, as they all receive the maximum score and hold the same rankings. The results indicated that one of the major issues of the identification of the gold standard in the previous

chapter was the high rate of infinity scores (Section 4.3.2.1). However, in this research, it was found that using non-DGA gold standards can help to reduce the high rate of the infinity scores.

The non-DGA gold standards exhibited a lower average in the LLS distribution, with averages of approximately 7.47% for BioGRID_IES, 7.62% for Reactome_IES, and 4.19% for IntAct_IES. In contrast, the DGA gold standards displayed a higher average, with approximately 13.89% for SSA_SMA, 16.33% for HEL_LEL, 17.50% for MCS_SCS, 18.28% for MGS_EIS, and 18.83% for OMIM_IS (Figure 5.2). The lower average LLS distribution for the non-DGA gold standards can be attributed to the limited overlap between the gold standards and the datasets. This lower overlap is a result of using different data types for the datasets (DGAs) and the gold standards (none-DGAs). While in DGA gold standards, the same data types were employed for both datasets and the gold standards by scoring DGA data against DGA data, in non-DGA gold standards, different data types were used for the gold standards and the datasets. DGA data was scored against PPI data and shared pathway data. Furthermore, in the case of non-DGA gold standards, the scoring involved assessing gene-gene associations based on their disease similarity, where two genes are considered to interact if they are involved in the same disease. This may contrast with scoring physical interactions which are based on the direct evidence (BioGRID_IES) or pathway-based interactions (Reactome_IES). It is important to note that while two genes may be associated with the same disease, they may not necessarily exhibit physical interactions, which BioGRID scores for. Additionally, two genes associated with the same disease may not always be part of the same pathways, as assessed by Reactome_IES.

During the scoring process, datasets may receive scores of zero or negative if they do not have any overlap with the gold standard, resulting in data loss. Using non-DGA gold standards led to a notably higher rate of data loss, with 97% for BioGRID_IES, 97% for Reactome_IES, and 99.84% for IntAct_IES. In comparison, the DGA gold standards experienced a lower rate of data loss, with rates like 81.53% for SSA_SSA, 82.12% for HEL_LEL, 61.18% for MCS_SCS, 70.00% for MG_IES and 66.03% for OMIM_IES (Table 5.3). The high rate of data loss is a significant challenge observed in non-DGA gold standards, stemming from various factors. One common issue shared with DGA gold standards is the definition of the individual datasets. Treating individual studies as distinct

datasets introduces complications, as the focus of a dataset (or study) may differ from that of the gold standards. This approach to dataset identification has a limitation, as it may fail to capture the overlap required for accurate scoring. Furthermore, non-DGA gold standards face a specific challenge related to the difference in data types utilised in both the gold standards (non-DGA data) and the datasets (DGA data). This mismatch may create an issue, rendering it impossible to effectively score certain datasets. For instance, datasets containing only a single gene or isolated associations, missing common diseases shared between two genes, pose a significant scoring challenge. In fact, a high number of approximately 34,942 datasets could not be scored against non-DGA gold standards due to this limitation. Even among the datasets that can be scored (i.e., those containing multiple genes and common diseases), there remains a risk of data loss during scoring. This risk arises from discrepancies between the genes present in the datasets and those represented in the gold standards. It's important to note that while genes may share common diseases, they are not necessarily required to physically interact (as measured by PPI gold standards) or belong to the same pathways (as measured by shared pathways gold standard). This disparity further complicates the scoring process and contributes to the high rate of data loss observed in non-DGA gold standards.

Table 5.2. Gold Standard overlap. The overlap in terms of, genes, and associations of the scored datasets with non-DGA gold standard data including BioGRID scored Individual Experimental Studies (BioGRID_IES), Reactome pathway scored Individual Experimental Studies (Reactome_IES), and IntAct scored Individual Experimental StudiesInt (Act_IES) and with DGA gold standard data including SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES.

	DisGeNET	BioGRID_ IES	Reactome_ IES	IntAct_ IES	SSA_ SMA	HEL_ LEL	MCS_ SCS	MG_ IES	OMIM_ IES
Genes (Datasets)	9469	9469	9469	9469	7669	8256	8654	8890	8994
Diseases (Datasets)	10563	10563	10563	10563	7950	7866	8722	9663	9789
Associations (Datasets)	79104	79104	79104	79104	66022	66234	68839	76655	76865
Genes (Gold standard)	-	16427	4856	1414	5540	3394	3436	2714	2490
Diseases (Gold standard)	-	-	-	-	6530	5642	6636	3358	3075
Associations (Gold standard)	-	106411	19599	7997	16486	13632	13355	3449	3219
Overlap (Genes)%	-	75.55	34.84	9.19	54.13	30.94	35.00	28.21	27.07
Overlap (Diseases)	-	-	-	-	51.52	39.59	57.00	25.27	26.21
Overlap (Associations)%	-	3.20	1.48	0.12	8.51	3.78	7.72	3.45	3.59

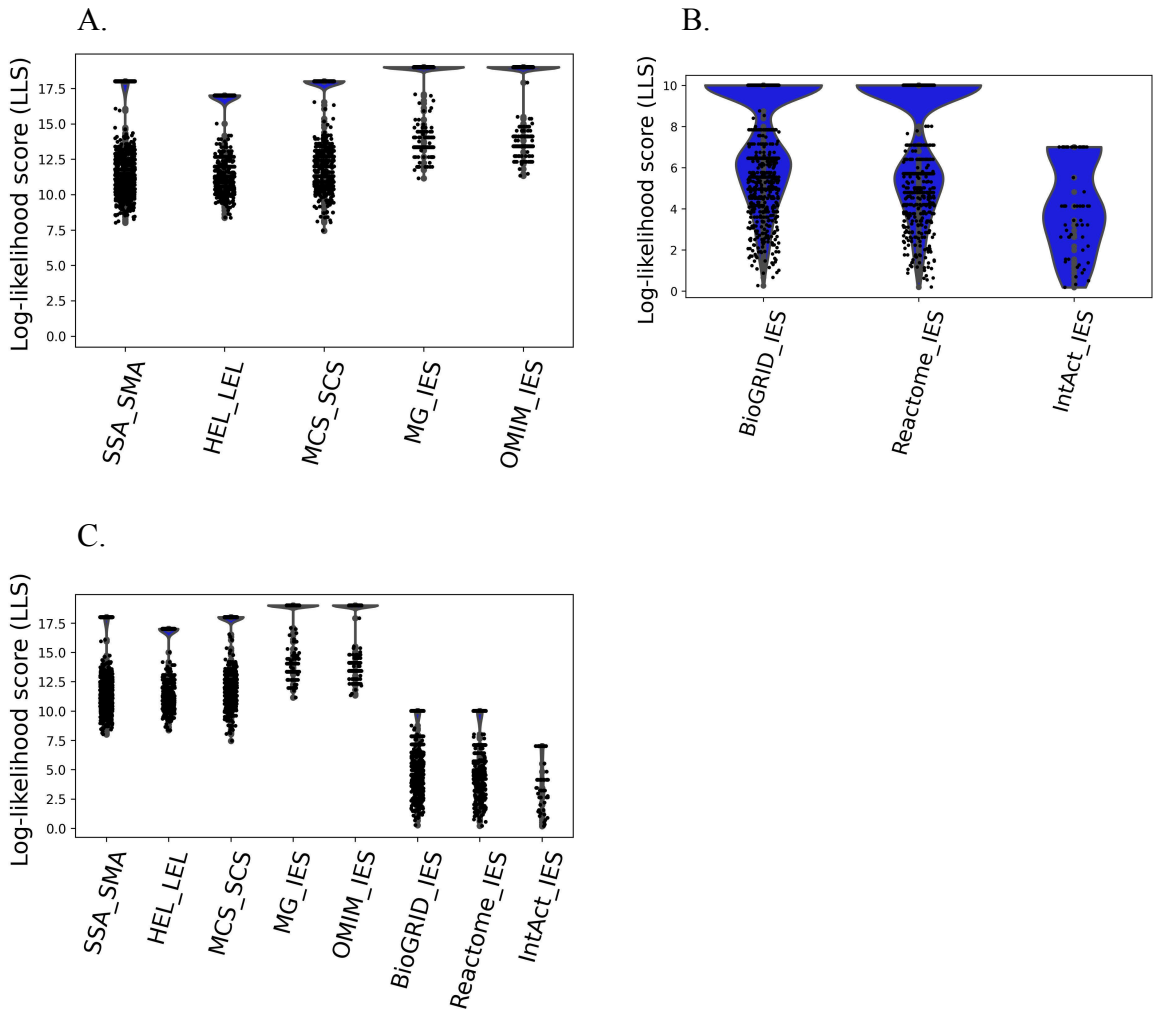


Figure 5.2. A. LLS distributions for DGA and non-DGA gold standards. (A) DGA gold standards: SSA_SSA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES. (B) Non-DGA gold standards: BioGRID_IES, Reactome_IES, IntAct_IES. (C) Combined view. DGA gold standards showed higher LLS, indicating stronger alignment with DGA studies, as they reflect the same data type. In contrast, lower LLS with non-DGA gold standards suggests that disease-associated genes may not interact physically or share pathways. Non-DGA standards also showed greater LLS variability.

Table 5.3. Level of individual study loss during scoring. Datasets are lost if they have negative scores (poor overlap with the Gold Standard) or no score (no overlap with the Gold Standard). Datasets containing a single DGA could not be scored using non-DGA gold standards including the BioGRID, Reactome, and IntAct since a minimum of two genes are required to assess a dataset using these gold standards. A subset of the data is excluded for use as a gold standard in some DGA gold standards including SSA_SMA, HEL_LEL, and MCS_SCS networks.

Approaches	Network	Negative score	No score	Gold Standard Subset	Unscored	Total
DGA gold standards	SSA_SMA	5003	5542	21719	-	32264(81.53%)
	HEL_LEL	5162	15273	12064	-	32499(82.12%)

	MCS_SCS	6098	12371	5744	-	24213(61.18%)
	MG_IES	338	27019	-	-	27357 (70.00%)
	OMIM_IES	418	25713	-	-	26131(66.03%)
Non-DGA gold standard	BioGRID_IS	2816	600	-	34942	38358(97.00%)
	Reactome_IES	1757	1885	-	34942	38584(97.00%)
	IntAct_IES	275	4295	-	34942	39512(99.84%)

The integration of datasets followed Lee and colleagues' method, which involves a weighted sum approach (Equation 3.2). The integration proceeded from the highest confidence scores to the lowest, utilising D values of 1.2, 1.3, and 1.3 for BioGRID_IES, Reactome_IES, and IntAct_IES respectively (Figure 5.3.D). For comparison and evaluation purposes, three DGA PFINs were constructed: BioGRID scored individual experimental studies (BioGRID_IES network), Reactompathway scored individual experimental studies (Reactome_IES network), and IntAct scored individual experimental studies (IntAct_IES network). These DGA PFINs were then compared with those generated in the previous chapter, which included the SSA_SMA network, HEL_LEL network, MCS_SCS network, MG_IES network, and OMIM_IES network (Section 4.3.2.2). The edge weights in these networks varied, reflecting differences in the confidence score distributions and the number of datasets incorporated in each network due to the use of different data types of gold standards during scoring (as shown in Figure 5.3).

The non-DGA-scored networks exhibited lower average edge weight distributions, with averages of approximately 4.52% for BioGRID_IES, 4.21% for Reactome_IES, and 1.85% for IntAct_IES. In contrast, DGA-scored networks had higher average edge weight distributions, with averages of approximately 16.20% for SSA_SSA, 21.68% for HEL_LEL, 24.84% for MCS_SCS, 31.79% for MG_IES, and 30.01% for OMIM_IES (as shown in Figure 5.3).

DGA-scored networks showed a high correlation with each other in terms of the weighted sum, which represents the edge weight of the networks, while none-DGA-scored networks are highly correlated with each other, indicating that networks scored using the same type of gold standard (either DGA or non-DGA) produce similar edge weights (Figure 5.4). For instance, OMIM_IES and MG_IES had a correlation of 0.76, while BioGRID_IES and

Reactome_IES showed a similar correlation of 0.73. This correlation suggests that each type of gold standard measures different aspects of associations between diseases and genes. Specifically, DGA gold standards primarily capture biomarker associations, whereas non-DGA gold standards focus on molecular mechanisms underlying disease associations. DGA gold standards emphasise the identification of genetic biomarkers, while non-DGA gold standards prioritise the investigation of molecular mechanisms involved in disease associations, uncovering the biological processes and pathways through which genes contribute to the development or progression of diseases. The low correlation between DGA-scored networks and non-DGA-scored networks, for instance, BioGRID_IES and MG_IES, suggests that despite genes being associated with the same disease, they may not necessarily physically interact sharing direct evidence of interactions or involvement in the same pathways or biological processes.

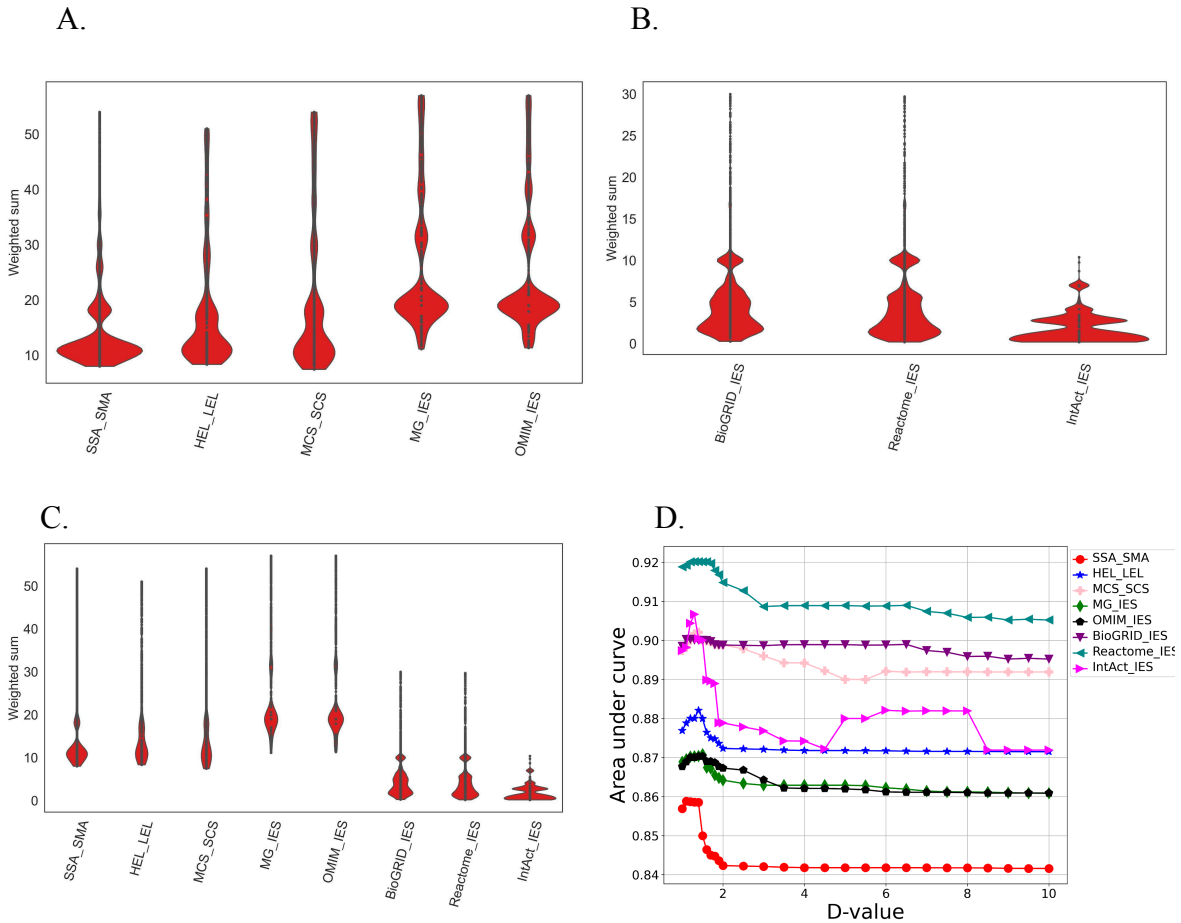


Figure 5.3. Weighted sum distribution and D values. (A–B) Weighted sums for DGA (SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES) and non-DGA (BioGRID_IES, Reactome_IES, IntAct_IES) gold standards. (C) Combined view shows higher edge weights for DGA gold standards, likely due to scoring with the same data type. Lower values in non-DGA networks suggest limited overlap in physical interaction or pathways. (D) D values used to integrate datasets, selected based on AUC performance in link prediction: 1.1–1.5 across networks.

While using non-DGA gold standards did lead to a higher rate of data loss in terms of datasets, the non-DGA scored networks exhibited more edges, such as 23,148 edges for BioGRID_IES and 18,954 edges for Reactome_IES, in comparison to the DGA-scored networks with 7220 for HEL_LEL, 12495 for MCS_SCS, and 7792 for MG_IES, and 11080 for OMIM_IES. These results can be attributed to the fact that DGA gold standards tended to score smaller datasets compared to non-DGA gold standards. Furthermore, the non-DGA scored networks demonstrated greater connectivity, having fewer isolated connected components, such as 253 connected components for BioGRID_IES, 260 for Reactome_IES, and four for IntAct_IES in contrast to the DGA scored networks with 476 connected components for SSA_SMA, 674 for HEL_LEL, 1,218 for MCS_SCS, and 1254 for MG_IES and 1038 for OMIM_IES. Table 5.4 shows statistics on the DGA and non-DGA-scored networks.

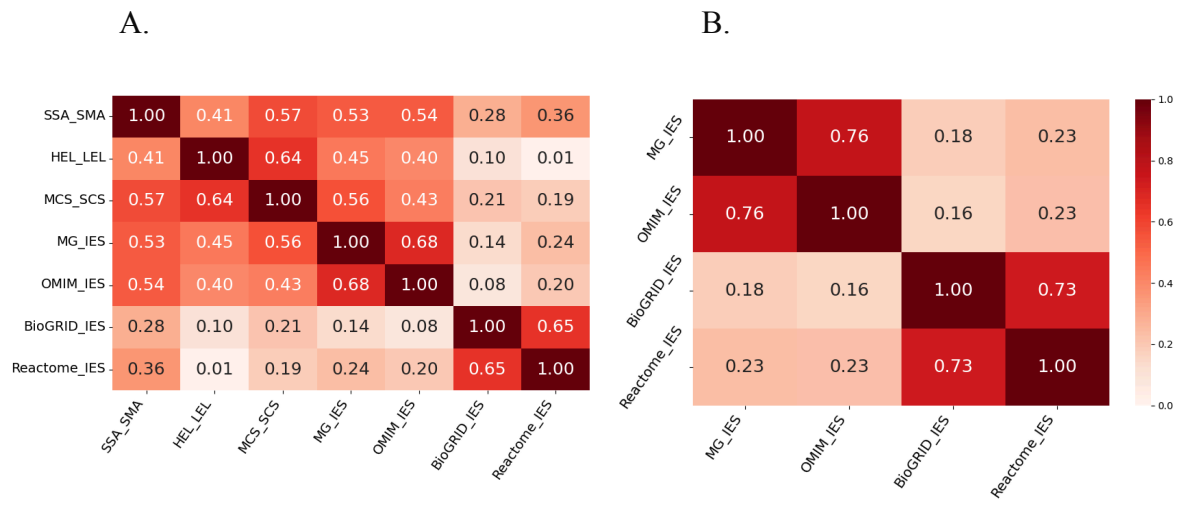


Figure 5.4. Correlation of DGA-scored networks by weighted sum. (A) Correlation across all networks. (B) Correlation between DGA-scored (e.g., MG_IES, OMIM_IES) and non-DGA-scored networks (e.g., BioGRID_IES, Reactome_IES).

The variations in how datasets were scored based on DGAs and non-DGA gold standards had noticeable effects on the content and structure of the networks. Network sizes differed between the two types of gold standards (as indicated in Table 5.4). Networks scored using DGA gold standards featured more disease nodes but fewer genes while non-DGA-scored networks had more genes but fewer diseases. On average, despite having fewer scored datasets, non-DGA-scored networks, such as the BioGRID_IES and Reactome_IES, emerged

as larger networks in terms of the total associations than DGA-scored networks. This result can signify that a DGA gold standard tends to score disease-related nodes in the network, while non-DGA gold standards tend to score gene-related nodes.

Table 5.4: Network statistics. Topological characteristics for the non_DGA scored networks and DGA scored networks: BioGRID_IES, Reactome_IES, IntAct_IES. Statistics were calculated using the Cytoscape NetworkAnalyser plugin and clustering was carried out using the Markov Clustering Algorithm (MCL).

		DGA-based networks					non-DGA-based networks		
Topological Structure	**DisGeNET (Curated DGAs)**	**SSA_ SMA**	**HEL_ LEL**	**MCS_ SCS**	**MG_ IES**	**OMIM_ IES**	**BioGRID_ IES**	**Reactome_ IES**	**IntAct_ IES**
Number of disease nodes	10563	6382	3050	5341	4911	5816	2149	2005	157
Number of gene nodes	9469	5020	2069	3430	2511	2909	4852	4360	1880
Number of DGAs	79104	24775	7220	12495	7792	11080	23148	18954	4787
Average number of neighbours	8.386	4.674	3.664	3.719	2.780	3.120	8.530	8.231	4.700
Average gene degree	8.354	4.431	3.490	3.643	3.103	3.809	4.771	4.347	3.123
Average disease degree	7.489	3.152	2.367	2.339	1.587	1.905	10.772	9.453	6.267
Network diameter	14	18	22	20	26	22	16	14	11
Network radius	7	1	1	1	1	1	1	1	1
Connected components	511	476	674	1218	1254	1038	253	260	4
Characteristic path length	4.766	5.497	7.328	7.215	8.410	7.695	4.621	4.833	4.039
Clusters	1764	1345	593	957	915	1156	448	433	45

5.3.1.2 Network Evaluation

The non-DGA-based networks were evaluated using the same techniques used in the previous chapter through two different network analysis methods: link prediction and network clustering analysis (Section 4.3.3).

Link Prediction

The JI algorithm was used for link prediction (Section 3.3.4, Equation 3.13). The networks were filtered to eliminate edges with low confidence by applying edge weight thresholds. These thresholds were determined based on both the average edge weight distribution (Figure 5.3) and AUC optimization. Specifically, thresholds of 3, 2, and 0.50 were selected for the BioGRID_IES network, Reactome_IES network, and IntAct_IES network, respectively. To assess the performance, the AUC values of the non-DGA-based networks were compared to those of the DGA-based networks, which included the SSA_SMA network, the HEL_LEL network, the MCS_SCS network, the MG_IES network, and the OMIM_IES network. Notably, the non-DGA-based networks outperformed the DGA-based networks, with AUC values of 0.9091 for the BioGRID_IES, 0.9242 for the Reactome_IES, and 0.9030 for the IntAct_IES, compared to AUC values of 0.8607 for the SSA_SMA, 0.8821 for the HEL_LEL, 0.9021 for the MCS_SCS, 0.8713 for the MG_IES, and 0.8702 for the OMIM_IES for the DGA-based networks (Figure 5.5). These enhancements were statistically significant, as indicated by the standard error of the Wilcoxon statistic.

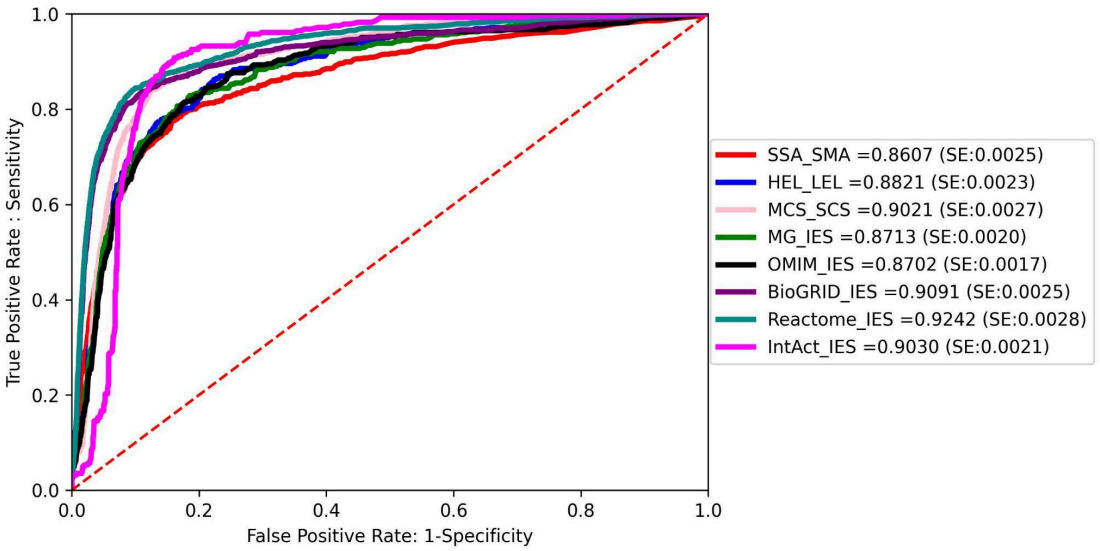


Figure 5.5: ROC curves for link prediction. Non-DGA networks (BioGRID_IES, Reactome_IES, IntAct_IES) achieved higher AUCs (e.g., 0.9242 for Reactome_IES) than DGA networks (e.g., 0.8607 for SSA_SMA), indicating better performance. Differences were statistically significant based on the Wilcoxon test.

Clustering

The MCL algorithm (Section 3.3.2) was applied to cluster the non-DGA-based networks. The purpose of this analysis was to evaluate the connectivity of clusters, based on the hypothesis that diseases strongly associated with their related genes (indicated by a high confidence

score)should be grouped within the same cluster. To test this hypothesis, the cohesion of the network clusters was analysed by examining the relationship between diseases and their associated genes.

The assessment of average cohesion involved calculating the average number of related genes per disease within a cluster, the average number of related genes per cluster across the entire network, and the overall level of network interconnectedness (Section 3.3.3.1). This approach aimed to assess the extent to which diseases and their associated genes tended to cluster together within the network. The analysis revealed that the clustering patterns observed in non-DGA-based networks were less cohesive than those in DGA-based networks. Specifically, the DGAs within the clusters in non-DGA-based networks were less related. In contrast, the clusters in DGA-based networks tended to contain more closely related DGAs, indicating higher cohesion and stronger associations between diseases and genes within these clusters. Specifically, the non-DGA-based networks resulted in a smaller number of clusters, but these clusters were larger. For instance, BioGRID_IES had 448 clusters, Reactome_IES had 433 clusters, and IntAct_IES had 45 clusters, with sizes ranging from 150 to three nodes. In contrast, the DGA-based networks produced a higher number of clusters, with examples including SSA_SMA contained 1,345 clusters ranging from 468 to three nodes, the HEL_LEL network had 593 clusters ranging from 113 nodes to three nodes, MCS_SCS had 957 clusters ranging from 150 nodes to three nodes. MG_IES had 915 clusters ranging from 75 nodes to three nodes, and OMIM_IES had 1,156 clusters ranging from 136 nodes to three nodes. The reason for the smaller number of clusters in the non-DGA networks can be attributed to their higher level of connectivity compared to the DGA-based networks. The non-DGA-based networks exhibited lower average cluster cohesiveness, with values of 0.56, 0.45, and 0.25 for BioGRID_IES, Reactome_IES, and IntAct_IES, respectively, in contrast to the DGA-based networks which showed higher average cluster cohesiveness with values of 0.81, 0.85, 0.84, 0.86, 0.87 for SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, respectively (Table 5.5). This low average of cluster cohesion in the non-DGA-based networks can be attributed to the distinct data types used between the individual datasets (DGAs) and the gold standards (non-DGAs) data. In the DGA-based networks, we assessed genes based on their association with the same diseases, either through physical interactions (BioGRID_IES) or shared pathway interactions (Reactome_IES). It is important to note that genes within a cluster may indeed be associated with the same disease,

but they might not necessarily physically interact with each other or belong to the same pathways. This discrepancy led to the lower average cluster connectedness in terms of the cohesiveness of genes and their association with related diseases.

Table 5.5. A summary of the average cluster Cohesiveness for the DGA-based networks and the non-DGA-based networks. The DGA-based networks have a higher average than the non-DGA-based networks.

Network Type	Network Name	Average Cluster Cohesiveness
DGA-based networks	SSA_SMA	0.81
	HEL_LEL	0.85
	MCS_SCS	0.84
	MGS_IES	0.86
	OMIM_IES	0.87
Non-DGA-based networks	BioGRID_IES	0.56
	Reactome_IES	0.45
	IntAct_IES	0.25

It was hypothesised that when a disease is linked to a gene with a high-confidence score (indicated by a high edge weight), this gene and its linked disease should cluster together within a single cluster. Consequently, the average cohesiveness of these clusters should increase when applying edge weight thresholds, which essentially reduces the number of edges with lower confidence scores. To test this hypothesis, we introduced edge weight thresholds to the scored networks and calculated the average cluster cohesiveness at various levels of these thresholds to assess any improvements. The determination of the edge weight threshold for each network was conducted individually, taking into account several factors such as the network's average edge weight distribution, its tendency, the distribution of cluster sizes, and the number of clusters (Figure 5.6). The goal was to select a threshold that would not significantly affect the network's size, as changes in network size could potentially lead to an artificial improvement in cluster connectedness. This precaution was taken to prevent any potential scenario where an improvement in the average connectedness of clusters could be misconstrued as solely resulting from the reduction of lowest confidence scores, when in reality, it might be due to the resizing of the network. As a result, specific

thresholds were chosen for each network: thresholds of 3, 2, and 0.50 were selected for the BioGRID_IES network, Reactome_IES network, and IntAct_IES network, respectively.

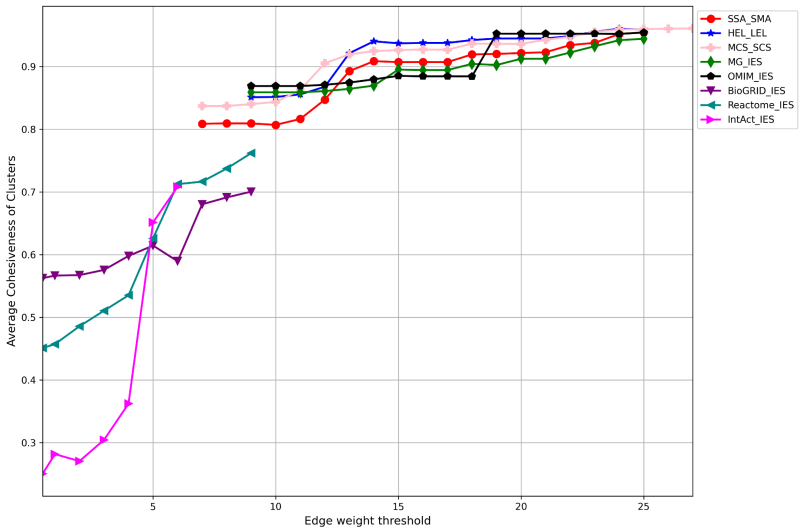


Figure 5.6. Average cluster connectedness at different thresholds. DGA-based networks (SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, OMIM_IES) show higher connectedness than non-DGA networks (BioGRID_IES, Reactome_IES, IntAct_IES).

The results suggest that non-DGA-based networks performed well in the task of link prediction but exhibited poorer performance in cluster analysis. In contrast, DGA-based networks excelled in cluster analysis but showed a lower average performance in link prediction. Consequently, each approach has its own strengths and can be valuable for specific applications. For instance, non-DGA based networks can be advantageous in link prediction applications, while DGA-based networks can be particularly useful in cluster analysis applications.

5.3.2 Addressing Limitations in Network Analysis Techniques

Network performance analysis techniques, such as link prediction for bipartite networks, present certain limitations (4.3.3.1). To address these limitations, the bipartite networks were collapsed into unipartite networks, simplifying the evaluation process. In this section, the DGA-based networks and non-DGA-based networks were collapsed into Gene-Gene association (GGA) networks and into Disease-Disease Association (DDA) networks (Figure 5.7). Subsequently, MCL clustering algorithm was independently applied on these networks to identify the densely interconnected clusters. The disease clusters were evaluated by testing

if they participate in the same biological processes, share the same drugs, share the same genes, have disease semantic similarity than randomly selected diseases (Section 3.3.3.2). The gene clusters were then assessed using gene enrichment analysis (Sections 3.3.3.3). The investigation included the examination of both functional homogeneity, which assesses the functional consistency within clusters, and functional specificity, which gauges the uniqueness of a cluster's function in relation to other clusters. The functional homogeneity and specificity of the clusters were evaluated in terms of MF, BP, and CC.

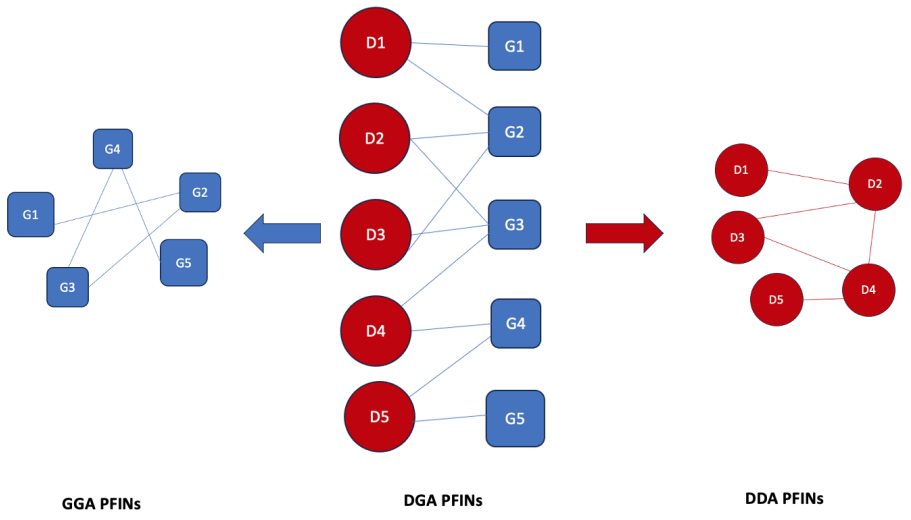


Figure 5.7. Collapsing DGA PFINs into GGA and DDA PFINs. In GGA PFINs, nodes are genes linked by shared disorders; in DDA PFINs, nodes are disorders linked by shared genes. Confidence scores are based on integrated DGA evidence, ranked from highest to lowest.

5.3.2.1 Collapsing Disease-Gene Association PFINs to Gene-Gene Association PFINs and Disease-Disease Association PFINs

Starting from the DGA scored datasets including SSA_SMA, HEL_LEL, MCS_SCS, MG_IES, and OMIM_IES, and the non-DGA scored datasets including, BioGRID_IES, Reactome_IES, and IntAct_IES, GGA PFINs, and DDA PFINs were generated. In the GGA PFINs, nodes represent genes, and two genes are connected to each other if they share at least one disorder in which disorders are associated with both genes. In the DDA PFINs, nodes represent disorders, and two disorders are connected to each other if they share at least one gene in which genes are associated with both disorders. The weight of an edge was generated by integrating the collapsed evidence (DGAs) that are implicated in both genes in the case of GGA PFINs and in both diseases in the case of DDA PFINs, from the highest score to the lowest scores using a weighted sum developed by Lee and colleagues (Equation 3.2). Figure

5.8 shows the edge weight distribution of collapsed GGA PFINs: collapsed DGA-based networks including the GGA_SSA_SMA, the GGA_HEL_LEL, the GGA_MCS_SCS, the GGA_MG_IES, and the GGA_OMIM_IES, and the collapsed non-DGA-based networks including the GGA_BioGRID_IES, the GGA_Reactome_IES, and the GGA_ IntAct_IES. The GGA PFINs and the DDA_PFINs were clustered using the weighted MCL clustering algorithm (Section 3.3.2). Subsequently, The disease clusters and gene clusters were evaluated (Section 3.3.3).

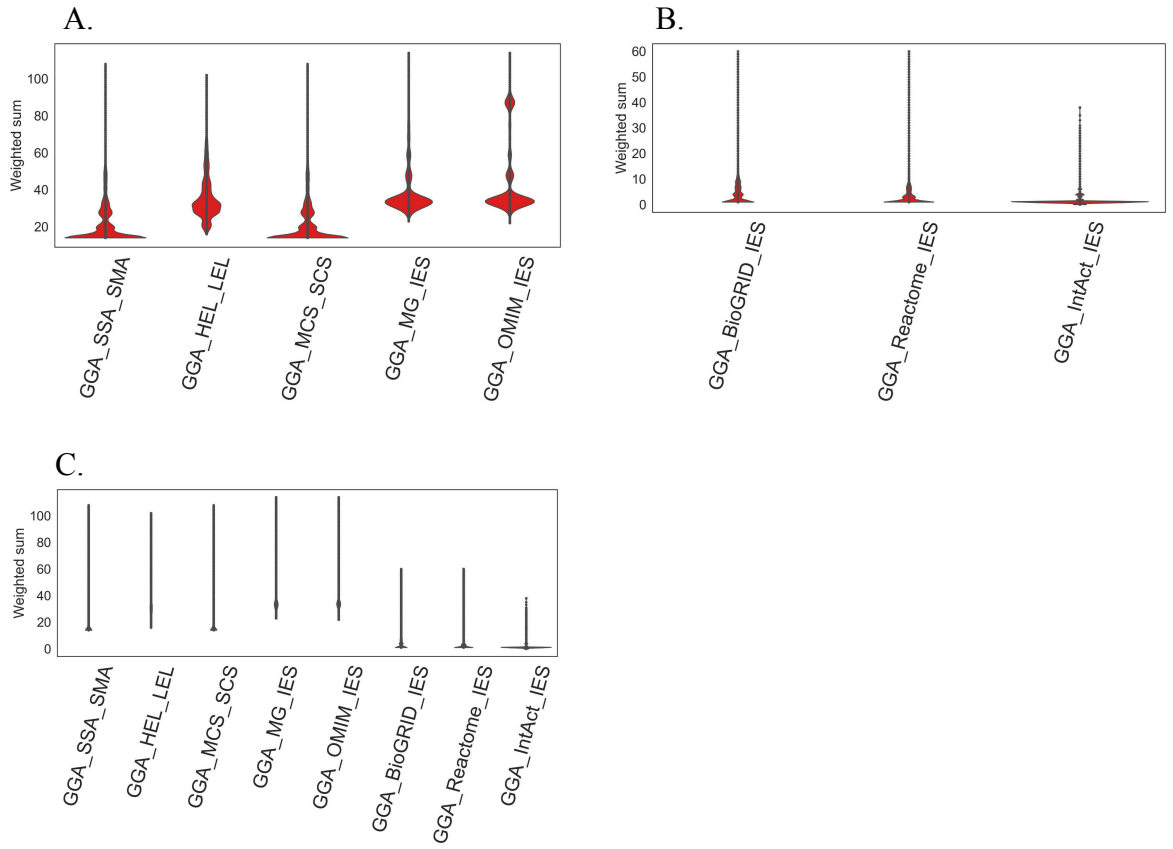


Figure 5.8. Edge weight distribution in collapsed networks. (A) DGA-based networks: GGA_SSA_SSA, GGA_HEL_LEL, GGA_MCS_SCS, GGA_MG_IES, GGA_OMIM_IES. (B) Non-DGA-based networks: GGA_BioGRID_IES, GGA_Reactome_IES, GGA_IntAct_IES.

5.3.2.2 Gene Cluster Evaluation

To assess the cohesiveness of gene clusters, the functional enrichment analysis was performed in order to find the GO terms in biological processes, which are significantly represented (enriched) by the genes in the resulting clusters. Functional enrichment analyses and measures for functional specificity were used to evaluate functional relevance and

specificity of clusters of the GGA PFINs (Section 3.3.3.3). Functional homogeneity, functional heterogeneity and functional specificity were evaluated. Table 5.6. The number and the average size of clusters for the collapsed GGA PFINs.

The clusters of the collapsed GGA PFINs were detected using MCL algorithm and analysed to investigate how biological processes affect the functional properties of the clusters. The clusters were evaluated in terms of functional coherence and specificity by using measures developed in [135]. The number of clusters for the collapsed GGA PFINs varied among the DGA-based networks and non-DGA-based networks considerably. It was found that the number of clusters for the collapsed DGA-based networks (101 for GGA_SSA_SMA, 108 for GGA_ HEL_LEL, 101 for GGA MCS_SCS, 156 GGA_ MG_IES, 188 GGA_ OMIM_IES), is much more than that of non-DGA-based networks (42 for GGA_ BioGRID_IES, 69 for GGA_ Reactome_IES, 211 for GGA_ IntAct_IES). The size of clusters for the collapsed GGA PFINs range from two to 3823. Clusters of size two were not considered in the cluster evaluation. Figure 5.9 shows cluster size distribution for the collapsed GGAPFINs. The average size of the clusters for DGA-based networks (46 for collapsed SSA_SMA, 13 for GGA_HEL_LEL, 46 for GGA_ MCS_SCS, 8 GGA_MGS_IES, and 10 GGA_OMIM_IES) is smaller than that of non-DGA-based networks (109 for GGA_BioGRID_IES, 60 for GGA_Reactome_IES, 9 for GGA_ IntAct_IES). The lower number and bigger size of clusters for collapsed non-DGA-based networks compared to collapsed DGA-based networks is due to the fact that the collapsed non-DGA-based networks were more connected than the collapsed DGA-based networks.

Table 5.6. A summary of the average cluster sizes for the DGA-based networks and the non-DGA-based networks. The DGA-based networks have a higher average than the non-DGA-based networks.

Network Name	Number of Clusters	Average Cluster Size
GAA_SSA_SMA	101	46
GGA_HEL_LEL	108	13
GGA_MCS_SCS	101	46
GGA_MGS_IES	156	8
GGA_OMIM_IES	188	10
GGA_ BioGRID_IES	42	109
GGA_ Reactome_IES	69	60
GGA_ IntAct_IES	211	9

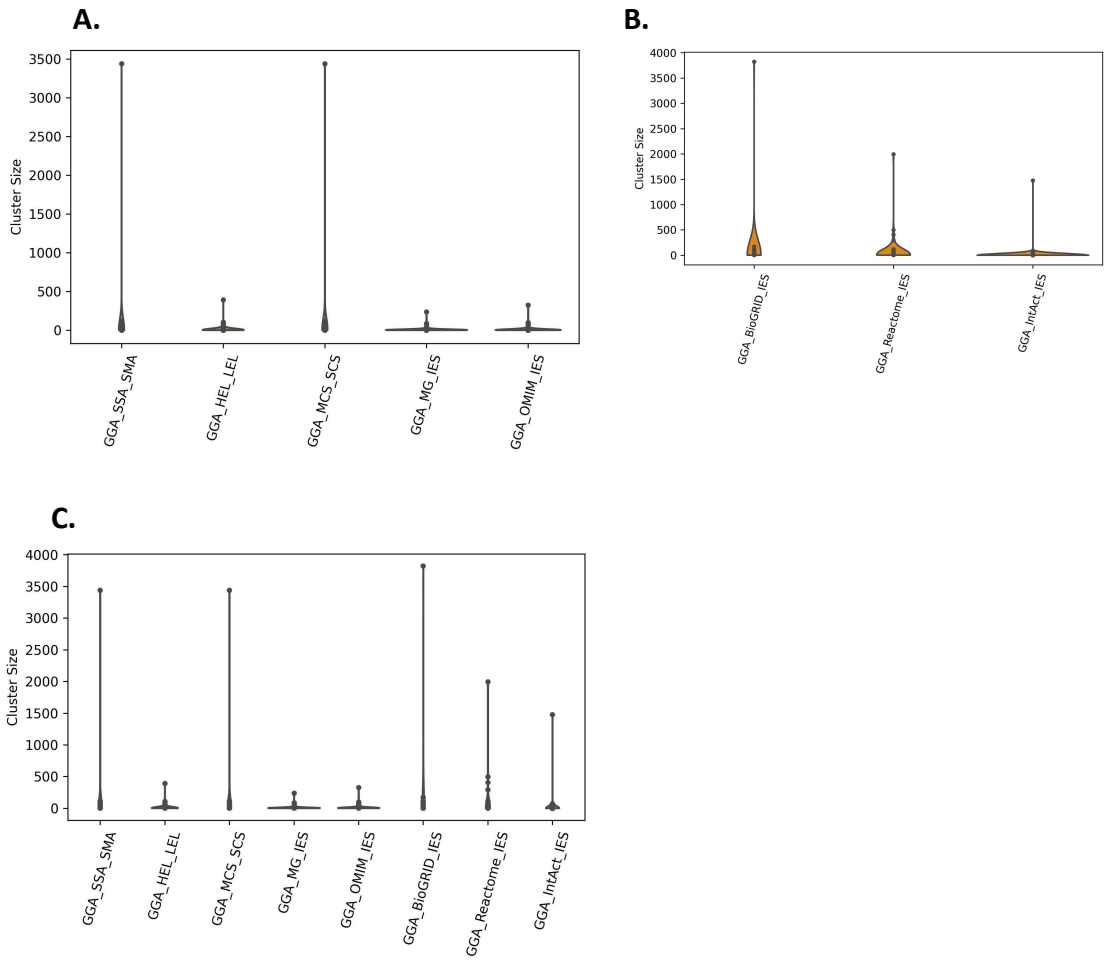


Figure 5.9. A. Cluster size distribution for the collapsed DGA-based networks. B. The average size of the cluster for collapsed non-DGA networks. C. Cluster size distribution for all GGA networks.

Genes clusters tend to share the same functional profile**.** To study functional relevance of these clusters, the clusters from all the collapsed GGA PFINs were tested for their biological relevance by using functional enrichment analysis. The enriched function set F is given by the union of all significantly enriched functions across clusters and functional specificities of the set of enriched functions were computed for each collapsed GGA PFINs. Three measures developed in [135] to evaluate the gene clusters: functional homogeneity, functional heterogeneity, and function specificity (Table 5.7). Functional homogeneity of a cluster quantifies functional consistency of a cluster as defined by the maximal fraction of genes associated with a biological function. The homogeneity ranges from 0 to 1 where a value of 1 indicates that all genes in the cluster exhibit that function. A cluster's heterogeneity value estimates how specific a function is for a particular cluster. Functional specificity value

measures how exclusively the cluster is enriched by the specific biological function (Section 3.3.3.3).

Table 5.7 A summary of the average cluster Homogeneity, Heterogeneity, and Specificity for the gene clusters of the collapsed DGA-based networks and the collapsed non-DGA-based networks. The collapsed DGA-based networks have a higher average than the collapsed non-DGA-based networks.

	Collapsed GGA Networks	Average of Cluster Homogeneity	Average of Cluster Heterogeneity	Average of Cluster Specificity
DGA-based Networks	**GGA_ SSA_SMA**	0.50	0.022	57.501
	GGA_ HEL_LEL	0.51	0.018	55.512
	GGA MCS_SCS	0.50	0.018	57.501
	GGA_ MG_IES	0.65	0.012	87.653
	GGA_ OMIM_IES	0.57	0.011	93.572
Non-DGA-based Networks	**GGA_BioGRID_IES**	0.51	0.042	20.511
	GGA_IntAct_IES	0.22	0.32	4.552
	GGA_Reactome_IES	0.58	0.022	45.583

The collapsed DGA-based networks produced averages of biological process homogeneity values with 0.50, 0.51, 0.50, 0.65. and 0.57 for GGA_SMA_SSA, GGA_HEL_LEL, GGA_MCS_SCS, GGA_MG_IES, and GGA_OMIM_IES, respectively). The collapsed non-DGA-based networks produced averages of biological process homogeneity with 0.58 , 0.51, and 0.26 for GGA_BioGRID_IES, GGA_Reactome_IES, and GGA_ IntAct_IES). On average, the GGA_MG_IES and the GGA_OMIM_IES, GGA_Reactome_IES_were observed to be the highest homogeneity of the biological processes (0.65, 0.58, 0.57). Figure 5.10 shows the distribution of gene cluster homogeneity.

The collapsed DGA-based networks produced lower average of biological process heterogeneity values (with 0.022, 0.018, 0.018, 0.012, 0.011, and 0.022 for GGA_SMA_SSA, GGA_HEL_LEL, GGA_MCS_SCS, GGA_MG_IES, and GGA_OMIM_IES) respectively, than collapsed non-DGA-based networks (with 0.042 , 0.022, and 0.32 for GGA_ BioGRID_IES, GGA_Reactome_IES, and GGA_ IntAct_IES). On average, GGA_MG_IES,

GGA_OMIM_IES were observed to be the lowest heterogeneity (0.0.12 and 0.011) . Figure 5.11 shows the distribution of genes clusters hetergenity. The collapsed DGA-based networks produced higher average of biological process specificity values (with 57.50%, 55.51%, 57.50%, 87.65%, and 93.57%, for GGA_SMA_SSA, GGA_HEL_LEL, GGA_MCS_SCS, GGA_MG_IES, and GGA_OMIM_IES, respectively) than collapsed non-DGA-based networks (20.51% , 45.58%, and 4.55% for GGA_BioGRID_IES, GGA_Reactome_IES, and GGA_IntAct_IES) (Figure 5.12).

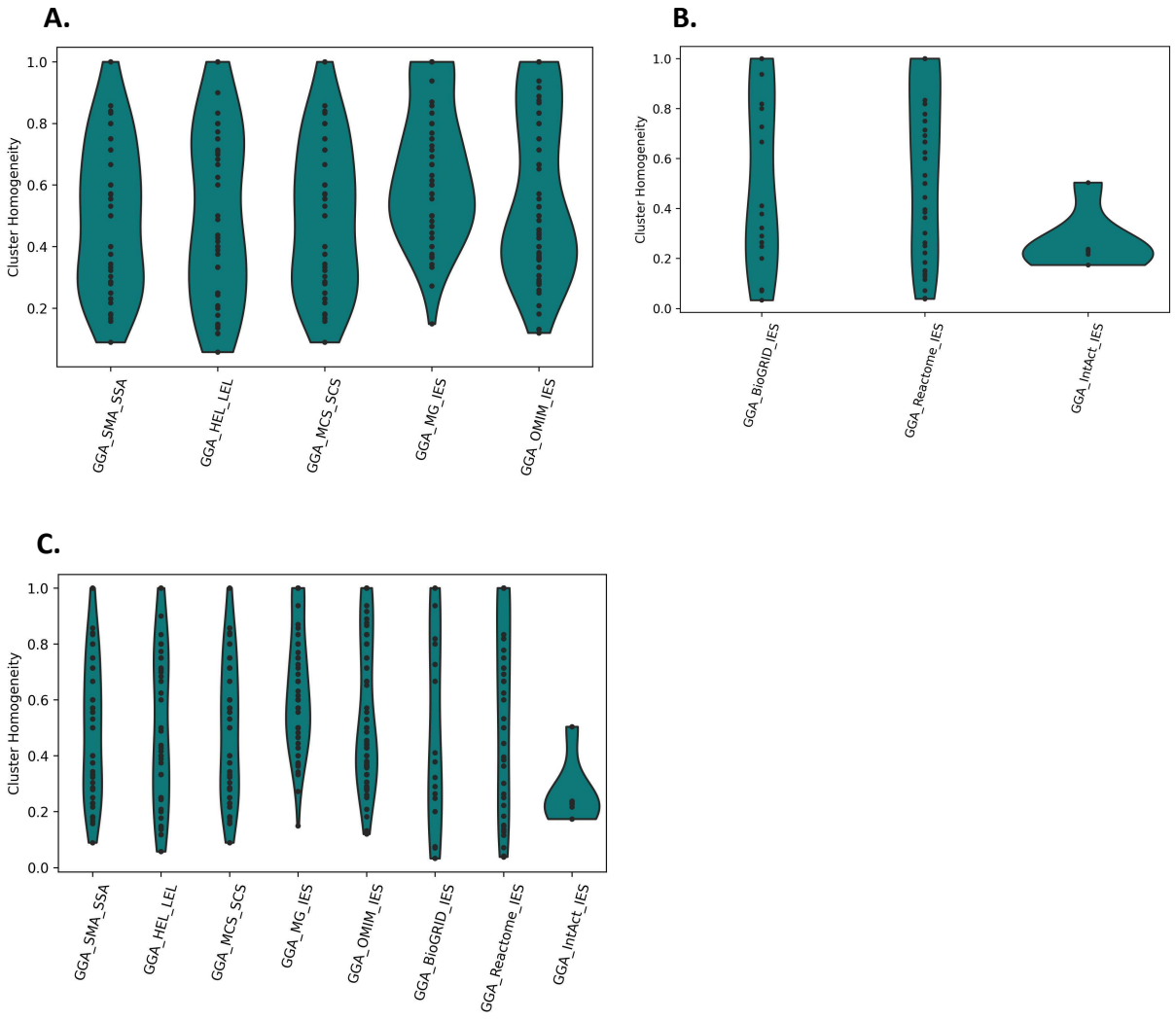


Figure 5.10. Gene cluster homogeneity (biological processes) in collapsed GGA networks. (A) DGA-based, (B) non-DGA-based, (C) all networks.

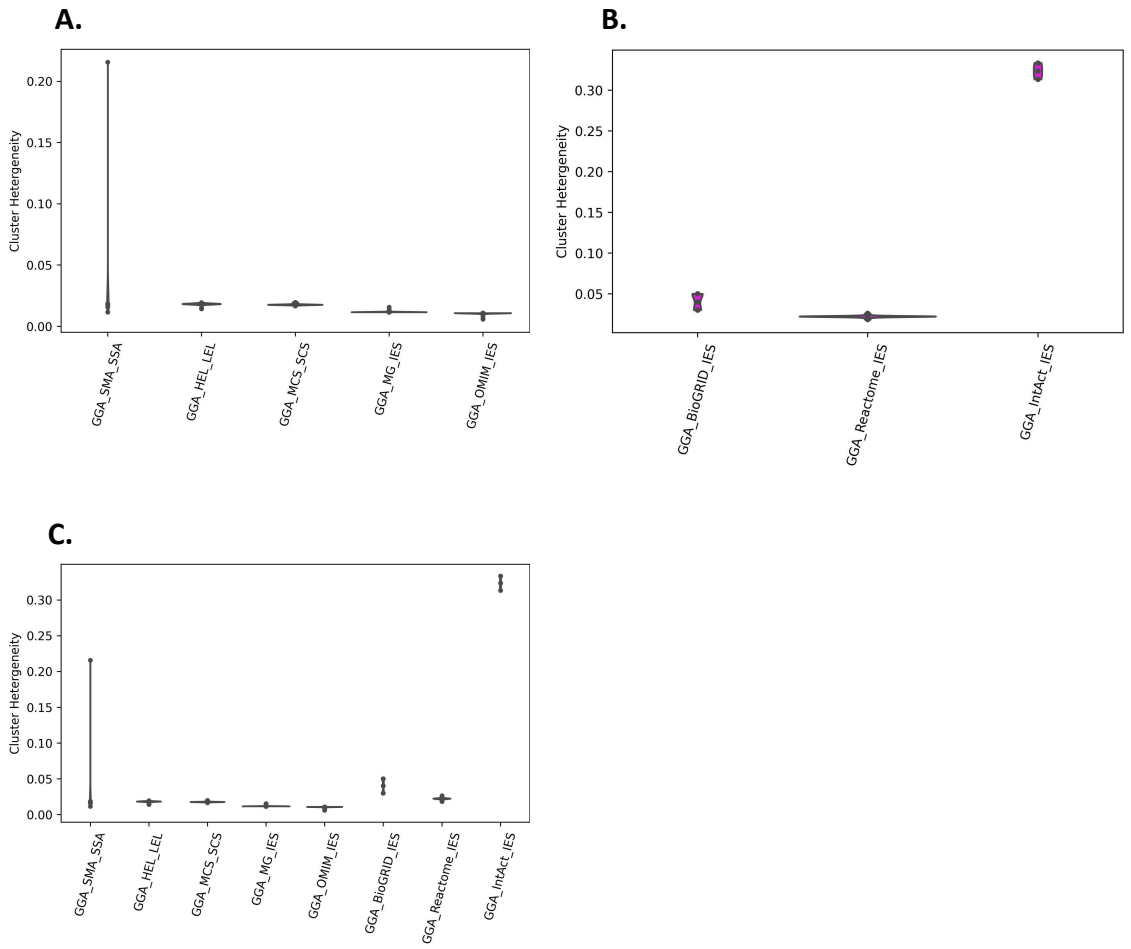
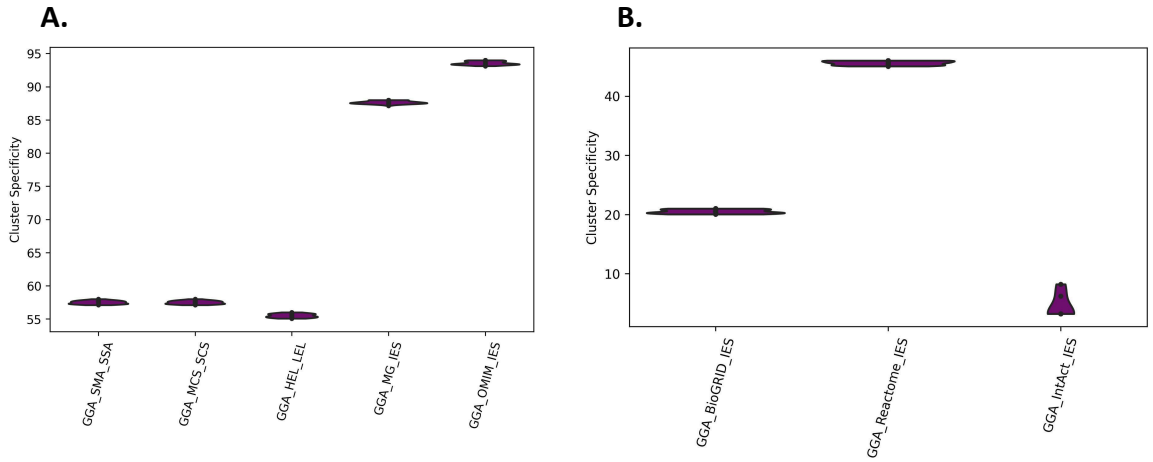


Figure 5.11. Distribution of gene cluster heterogeneity (biological processes) in collapsed GGA networks: (A) DGA-based, (B) non-DGA-based, and (C) all networks.



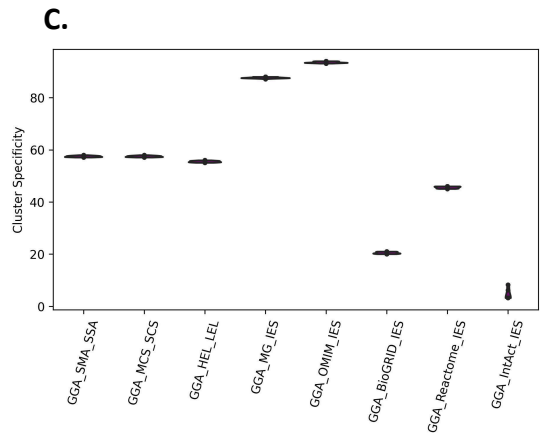


Figure 5.12. Gene cluster specificity (biological processes) in collapsed GGA networks. (A) DGA-based, (B) non-DGA-based, (C) all networks.

5.3.2.3 Disease Cluster Evaluation

The disease clusters, identified from the collapsed DDA networks, were analysed using the four similarity measures developed in [260]: the disease semantic similarity, disease genetic similarity, disease biological process similarity and disease treatment similarity (Section 3.3.3.2). To assess the cohesiveness of the disease clusters, we looked for similarities between disease pairs within clusters. The hypothesis was that if two diseases are associated with a high confidence score (edge weight), then they should be clustering together in one community, and therefore they should have common biological pathways and genes. To test this hypothesis, the average of biological process similarity, shared genes, disease semantic similarity, and shared drugs were calculated for each disease pair in each cluster, and then the overall average of similarity of the network was calculated. For each collapsed DDA network, a random network was built with the same network structure, and preserved degree and edge weight distributions using the configuration mode method [266], [267]. For each collapsed DDA network, we generated 100 random networks and computed the average shared genes, shared drugs, biological process similarity and semantic similarity in each cluster for each network for comparison.

The disease clusters of the collapsed DGA-based networks, including DDA_SMA_SSA, DDA_HEL_LEL, DDA_MCS_SCS, DDA_MG_IES, DDA_OMIM_IES, had a higher average of similarity measures compared to non-DGA-based networks, including

DDA_BioGRID_IES, DDA_Reactome_IES, and DDA_IntAct_IES. The DDA_OMIM_IES had the highest average of shared genes with 0.78 , shared drugs with 0.69, disease semantic similarity with 0.27, and genetic similarity with 0.74. Table 5.8 shows statistics for shared genes, drugs, disease semantic similarity, and biological process similarity of disease clusters of the collapsed DDA networks and the random disease clusters. The disease semantic similarity of the clusters in all networks is low, with specific values being 0.21 for DDA_SMA_SSA, 0.23 for DDA_HEL_LEL, 0.17 for DDA_MCS_SCS, 0.25 for DDA_MG_IES, 0.27 for DDA_OMIM_IES, 0.14 and 0.15 for DDA_BioGRID_IES, and 0.12 for DDA_IntAcr_IES. These low values are due to the limitations in disease identifier mapping between DisGeNET, which uses UMLS, and Disease Ontology, which uses Disease Ontology identifiers. Only 0.24% of DisGeNET diseases were mapped to the Disease Ontology. Consequently, some clusters had no mapped diseases and were not considered in the evaluation of disease semantic similarity.

Table 5.8. Statistics for shared genes, shared drugs, disease semantic similarity, and biological process similarity for the disease clusters of the collapsed DGA-based networks and the collapsed non-DGA networks.

Networks	Collapsed DDA networks	Similarity Measures	Average similarity for disease clusters	Average similarity for random clusters
DGA-based networks	DDA_SMA_SSA	Shared genes	0.57	1.2E-3
		Shared drugs	0.49	0.06
		Disease semantic similarity	0.21	0.15
		Gene semantic similarity	0.54	3.2E-6
	DDA_HEL_LEL	Shared genes	0.67	2.8E-4
		Shared drugs	0.50	0.10
		Disease semantic similarity	0.23	0.18
		Gene semantic similarity	0.52	4.2E-4
	DDA_MCS_SCS	Shared genes	0.71	4.3E-5
		Shared drugs	0.44	0.08
		Disease semantic similarity	0.17	0.12

		Gene semantic similarity	0.63	6.3E-5
	DDA_MG_IES	**Shared genes**	0.74	1.9E-5
		Shared drugs	0.68	0.03
		Disease semantic similarity	0.25	0.11
		Gene semantic similarity	0.64	3.2E-6
	DDA_OMIM_IES	**Shared genes**	0.78	0.004
		Shared drugs	0.69	0.16
		Disease semantic similarity	0.27	0.14
		Gene semantic similarity	0.74	2.4E-5
non-DGA-based networks	**DDA_BioGRID_IES**	**Shared genes**	0.42	1.3E-7
		Shared drugs	0.36	0.07
		Disease semantic similarity	0.14	0.09
		Gene semantic similarity	0.39	1.3E-8
	DDA_Reactome_IES	**Shared genes**	0.51	7.9E-8
		Shared drugs	0.44	0.06
		Disease semantic similarity	0.15	0.04
		Gene semantic similarity	0.49	3.5E-3
	DDA_IntAct_IES	**Shared genes**	0.42	7.9E-4
		Shared drugs	0.32	0.09
		Disease semantic similarity	0.12	0.03
		Gene semantic similarity	0.35	1.4E-3

5.3.3 Addressing Limitations in Individual Datasets Definition

The results indicated that an individual experimental study-based approach has major drawbacks, including a high rate of data loss and infinity scores in both types of gold standard data, including the DGA gold standards and the non-DGA gold standards. Even though in non-DGA gold standards, the rate of infinity scores was reduced, the high rate of

data loss is due to the fact that an individual DGA experimental study focuses on a single disease or a group of related diseases. Treating an individual study as a single dataset leads to data loss when scoring against the gold standard data if the gold standard does not contain the study's diseases of interest. Using non-DGA gold standards by treating diseases from an individual study as a dataset, and thereby scoring all DGAs for a disease on how well they reflect known biology, led to a high rate of data loss. The majority of the DGA experimental studies contain only a single gene, making it impossible to score them against non-DGA gold standards. Therefore, this section provides solutions to the limitations of the dataset definition.

5.3.3.1 Identification of individual datasets based on Disease-Gene Association Type Ontology

In this section, disease-gene association type ontology was used to define the individual datasets for DGAs. The DisGeNET 7.0 SQLite[57] database was downloaded to extract the disease-gene association type for gene-disease associations since these are not available in the TSV files DisGeNET releases. The "geneDiseaseNetwork" table was extracted. Figure B.1 represents the SQL schema of the DisGeNET sqlite database.

Then, curated DGAs from DisGeNET were split by the "associationType" to define separate DGA datasets. The DisGeNET association type ontology was developed by DisGeNET. All DGA types from the original data sources are formally constructed from a parent GeneDiseaseAssociation class if there is an association between the gene and the disease and represented as ontological classes. The DisGeNET association type ontology is shown in Figure B.2.

Splitting curated DGAs from DisGeNET by association types produced nine datasets, including *ChromosomalRearrangement, Biomarker, SusceptibilityMutation, FusionGene, Therapeutic, GermlineModifyingMutation, GermlineCausalMutation, GeneticVariation,* and *SomaticCausalMutation.* Unfortunately, only two datasets were unique: *Biomarker* and *Therapeutic*, whereas all the remaining association types are subsets of *Biomarker* and *Therapeutic* (Figure 5.13). Therefore, a total of two unique evidence types are only available. The majority of DGA had *Biomarker* association type with 74865 DGAs, and the smallest

[57] https://www.disgenet.org/downloads

dataset was *GermlineModifyingMutation* with 62 DGAs. Since the concept of PFINs is based on integrating multiple and unique evidence types, this approach was not suitable for defining individual datasets. Table 5.9 shows the nine datasets and their size after splitting the DGAs from DisGeNET by association type ontology.

Table 5.9 Nine disease-gene datasets based on DisGeNET disease-gene association types. These datasets were generated by splitting the disease-gene associations according to association type ontology, and their sizes are presented.

Datasets	Size
Chromosomal Rearrangement	188
Biomarker	74865
Susceptibility Mutation	459
Fusion Gene	236
Therapeutic	4662
Germline Modifying Mutation	62
Germline Causal Mutation	4920
GeneticVariation	6034
Somatic Causal Mutation	198

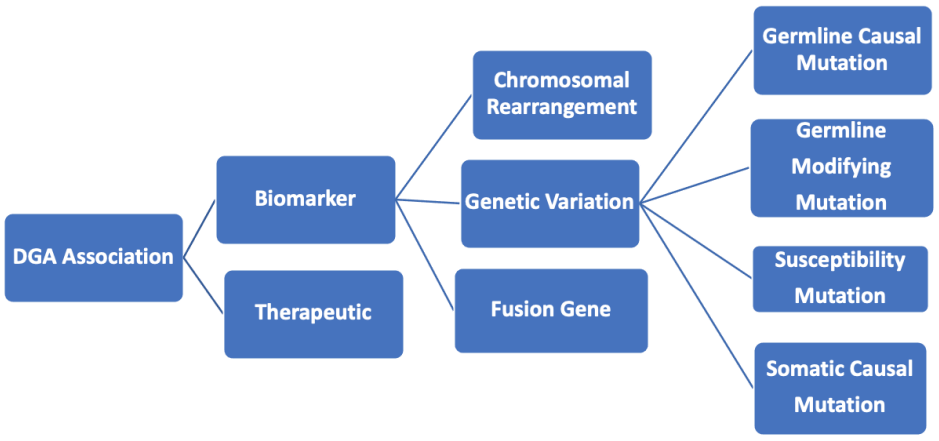


Figure 5.13. Nine datasets from DisGeNET curated DGAs split by association type, identifying two unique types: Therapeutic and Biomarker.

5.3.3.2 Text Mining Approach to Individual Dataset Definition

The findings indicated that employing individual DGA experimental studies as representations of datasets resulted in a significant loss of data for two main key reasons. Firstly, individual DGA studies typically focus on a specific disease or related disease groups that may not align with the Gold Standard data. Secondly, when using the non-DGA gold standards, only a limited number of DGA experimental studies can be scored. This limitation arose because the majority of studies contained either a single DGA or multiple DGAs containing only one single gene. Consequently, utilising this approach was inappropriate for defining individual datasets for the application of the PFIN approach on DGA data. Although the treatment of individual studies as datasets has been employed in PPI PFINs and has demonstrated a minimal data loss rate. Employing multiple experimental studies rather than a single study as treating them as individual datasets may solve the issue of data loss.

In this section, a text mining approach was applied on biomedical literature to define individual datasets. The aim was to extract the experimental techniques utilised in DGA experimental studies. Subsequently, Individual DGA experimental studies were clustered based on the similarity of their experimental techniques. Each cluster contains a group of experimental studies sharing common experimental techniques. Each single cluster was then treated as a single individual dataset representing a separate and unique evidence supporting an DGA.

Two distinct approaches were employed: the abstract-based approach which is based on abstract content and the method-based approach which is based on the method's section's content. Each approach has its benefits and limitations. For instance, in the abstract-based approach, the majority of abstracts in PMC articles are accessible while only a few numbers of articles where the methods section can be accessed since most of the articles are not open access. In contrast, the method-based approach provides more details since the method sections of articles typically contain more details about the experimental techniques used in experimental studies whereas abstracts typically lack detailed information about the methods used in papers, providing only a high-level description.

Extracting Abstracts and Methods from DGA Experimental Studies in the Biomedical Literature

The abstracts and methods sections of the curated individual experimental studies articles from DisGeNET were extracted from the literature. First, the list of PubMed IDs of the curated DGA experimental studies extracted from DisGeNET. Second, the PubMed IDs were used to extract the abstract texts in an XML file from the Pubmed database. Third, the PubMed IDs were mapped to PMC IDs to extract the full-text PMC articles in an XML file from the PMC database. Fourth, The XML files were parsed to extract abstract sections and method sections. The total number of curated experimental studies from DisGeNET was 39574 PubMed IDs. The total number of Pubmed IDs that mapped to PMC IDs was 13616. The number of articles that had Abstracts was 37,826. The number of articles that were open-accessed and available for downloading of the full text as XML files was only 1274. The total number of articles that had abstracts and were open-accessed was 719 (Section 3.5.1 for the details of parsing PMC database).

After the text for the abstracts and methods were obtained from the literature, the next step was to construct a dictionary including the common terms associated with experimental techniques employed in DGA studies. This dictionary was used for mining DGA experimental techniques employed in biomedical literature. The next section describes the creation of a robust dictionary containing relevant terms of DGA experimental techniques.

Construction of DGA Experimental Techniques Dictionary

The Experimental Factor Ontology (EFO) and the Ontology of bioscientific data analysis and data management (EDAM) were used to build the dictionary of the DGA experimental technique terms. Through the exploration of these two ontologies using an ontology editing and visualisation tool, it was discovered that the branches containing relevant terms related to DGA experimental techniques correspond to the "*Planned Process*" class in EFO and the "*Topic*" class in the EDAM. The *Topic* class from EDAM had 263 terms whereas the *Planned Process* class from EFO had 848 terms and the total unique terms extracted from EDAM and EFO was 1093. The two ontologies contain equivalent terms with different names. For instance, in the EFO, there is a term labelled '*polymerase chain reaction*,' while in the

EDAM, a similar term is denoted as *'PCR experiment*.' These types of terms were mapped to a unique term using equivalent synonyms (Section 3.5.2 for the details of ontology parsing).

Mining DGA Experimental Techniques from Abstracts and Methods Sections

The 37,825 articles containing abstracts were mined for abstract DGA experimental technique terms using the DGA experimental techniques dictionary, resulting in 19,215 articles with abstract-mined DGA experimental technique terms. Additionally, the 1,274 full-text articles were mined for method section DGA experimental technique terms using the same dictionary, yielding 1,233 articles with method-mined DGA experimental technique terms. Among these, 719 articles had both abstracts and methods sections, containing abstract-mined DGA experimental technique terms and method-mined DGA experimental technique terms. Table 5.10 shows statistics on the articles with abstracts and methods sections and their mined DGA experimental technique terms. Figure 5.14 and 5.15 shows the frequency of the abstract-mined DGA experimental technique terms and the method-mined DGA experimental technique terms, respectively.

Table 5.10. Statistics on the abstract and method sections of PMC articles in DisGeNET.

Number of articles from DisGeNET	39574
Number of articles with abstracts	37825
Number of articles with mined terms in their abstracts	19215
Number of articles with methods	1274
Number of articles with mined terms in their methods	1233
Number of articles with abstract and methods	719

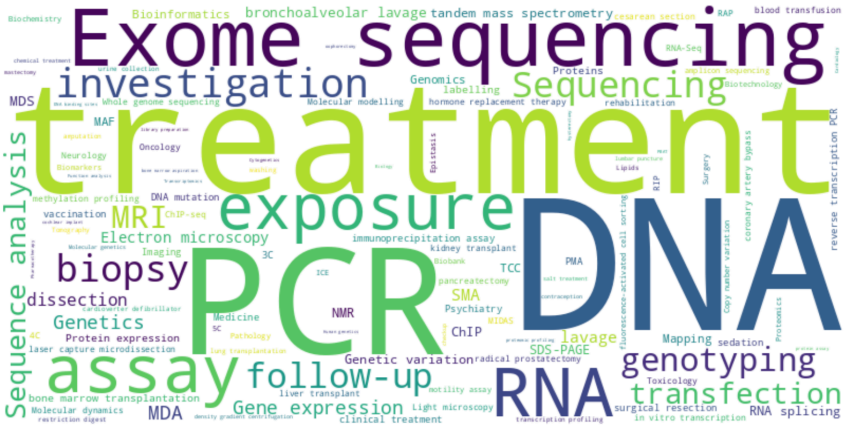


Figure 5.14 Frequency of abstract-mined DGA experimental technique terms. 'DNA', 'treatment', and 'PCR' are most frequent; terms like 'vaccination' and 'clinical treatments' appear less often.

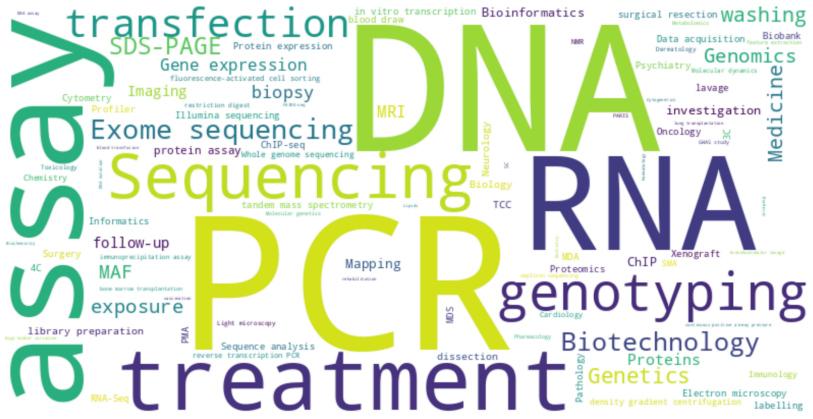


Figure 5.15. Frequency of method-mined DGA technique terms. 'DNA', 'RNA', and 'PCR' are most common; 'TCC' and 'Immunology' appear less frequently.

For articles with both abstracts and methods sections, the consistency between the sets of terms mined from the abstracts and those mined from the methods was examined. A high degree of consistency was observed , as the majority of terms in the abstract-mined set were also found in the method-mined set for the same article. In contrast, few terms present in the method-mined set were found in the abstract-mined set and this due to the high level of detail provided in the methods sections, resulting in a more extensive list of experimental techniques. Nevertheless, overall consistency was maintained, as all terms identified in the abstract mined set were also present in the method-mined set. To assess the consistency of

abstract-mined DGA experimental technique terms with method-mined DGA experimental technique terms, the ratio of abstract-mined DGA experimental technique terms identified in the method-mined set to the total number of terms in the abstract-mined set was computed for each article. Conversely, to measure the consistency of method-mined DGA experimental technique terms with abstract-mined DGA experimental technique terms, we calculated the ratio of method-mined DGA experimental technique terms found in the abstract-mined set to the total number of terms in the method-mined set. For an overarching measure of overall consistency, it was determined the ratio of the total number of common terms between the abstract-mined set and the method-mined set to the total number of terms in the combined abstract-mined and method-mined sets (Figure 5.16).

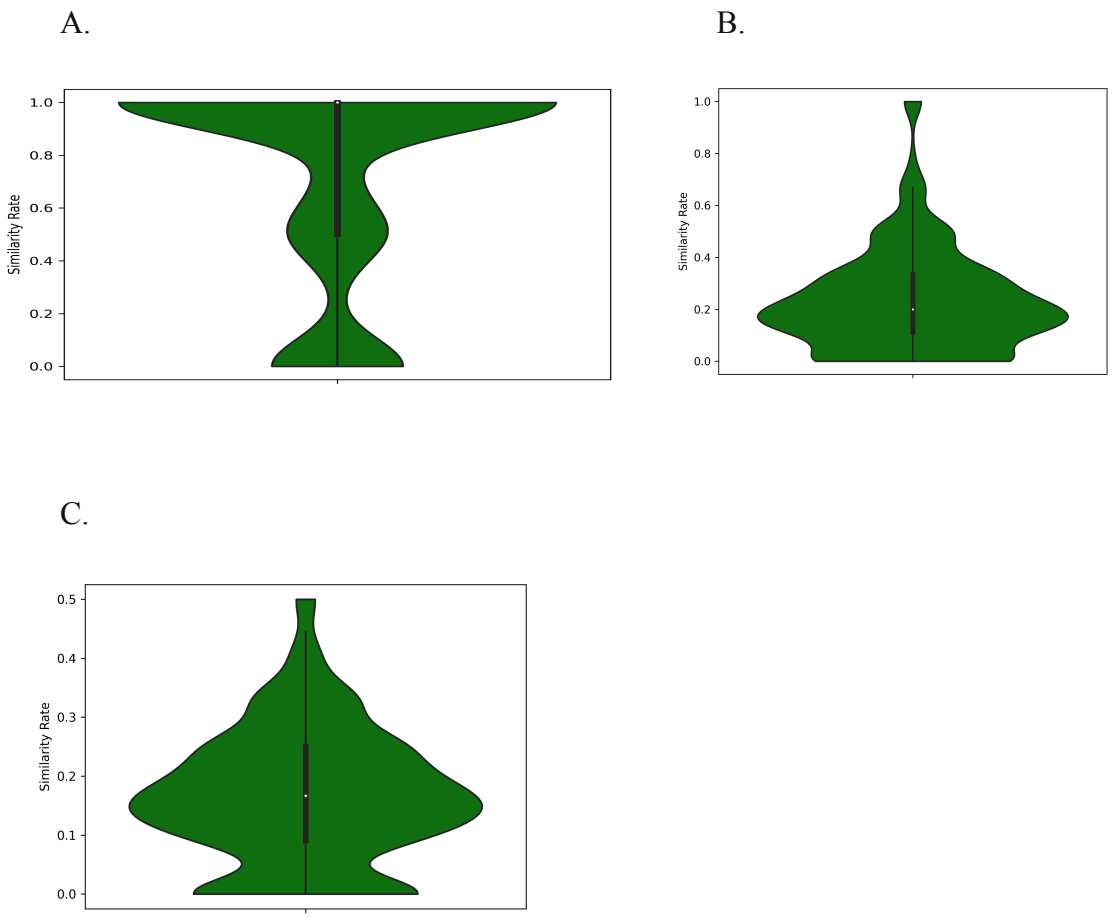


Figure 5.16. Ratio distribution of abstract-mined and method-mined DGA technique terms. (A) Abstract terms found in the method-mined set. (B) Method terms found in the abstract-mined set. (C) Common terms relative to the total combined terms.method-mined set.

Clustering DGA Experimental Studies Based on Experimental Techniques Extracted from Abstract

The 19215 experimental studies with abstract-minded terms were clustered based on their abstract-mined DGA experimental technique terms using the hierarchical clustering method. Hamming distance was used as the metric distance because it resulted in the highest values of the CPCC (Section 3.5.3) as well as Hamming distance is suitable for binary classification, aligning with the binary representation of our data. Table 5.11 shows the values of CPCC when using different distance metrics with Hamming resulting in the highest values of CPCC (Section 3.5.3 for the details of the clustering method). Regarding the selection of the linkage method, Ward's linkage method was used due to its ability to yield higher CPCC values, as demonstrated in Table 5.12. In hierarchical clustering, various linkage methods are used to determine how clusters are formed. The choice of linkage method can significantly impact the resulting hierarchical clustering.

Table 5.11. Systematic analysis of distance metric selection in the hierarchical clustering of the experimental studies based on abstract-mined DGA experimental technique terms. The table shows a systematic analysis of the chosen distance metric. Hamming distance was selected as it yielded higher values of PCC which was used as an evaluation metric for the quality of clusters. Furthermore, the choice of Hamming distance aligns well with our binary data representation.

Metric distance	CPCC
Hamming	**0.60**
Euclidean	0.48
Jaccard	0.46
Cosine	0.47
Correlation	0.48

Table 5.12. Systematic Analysis of linkage method selection in hierarchical clustering of experimental studies based on abstract-mined DGA experimental technique terms. The table shows a systematic analysis of the chosen linkage method. Ward's linkage was selected as it yielded higher values of PCC which was used as an evaluation metric for the quality of clusters.

Linkage Methods	CPCC
Ward's Linkage	**0.60**
Single	0.55
Complete	0.57
Average	0.82 (only one cluster)

Centroid	0.81 (only one cluster)
Median	0.43
Weighted	0.56

To choose the distance threshold for clustering, systematic analysis of the clusters based on Silhouette scores was applied (Section 3.5.3.1). Silhouette score compares each study within the clusters with other studies in the same cluster, minimising the distance between studies in one cluster and comparing the studies in the cluster with other studies in the other clusters, maximising the distance between studies in different clusters. A distance threshold cutoff of 0.02 was selected, as it produced higher Silhouette score values which suggested well-matched studies within their own clusters while maintaining a distinct separation from neighbouring clusters as shown in table 5.13. A good number of clusters with an appropriate size of studies were identified. Choosing a threshold of 0.00 gave the highest results; however, some clusters contained only a single study, which is not suitable for representing individual datasets. Figure 5.17 shows the hierarchical clustering of the experimental studies based on the abstract-mined DGA experimental technique terms using hamming distance and the threshold distance between data points (studies) in a single cluster was 0.02 .

Table 5.13. Systematic analysis of distance threshold selection of the hierarchy clustering of the experimental studies based on abstract-mined DGA experimental technique terms. The table presents a systematic analysis of the chosen distance threshold based on a Silhouette Score which minimises distances within studies clustered together and maximises distances between studies in different clusters.

Thresholds	**#Clusters**	**Silhouette Score**
0.00	1311	0.96
0.01	701	0.93
0.02	321	**0.88**
0.03	186	0.83
0.04	132	0.80
0.05	100	0.77
0.06	79	0.73
0.07	67	0.70
0.08	52	0.67
0.09	47	0.66

0.1	36	0.63
0.2	18	0.50
0.3	11	0.44
0.4	7	0.35
0.5	6	0.33

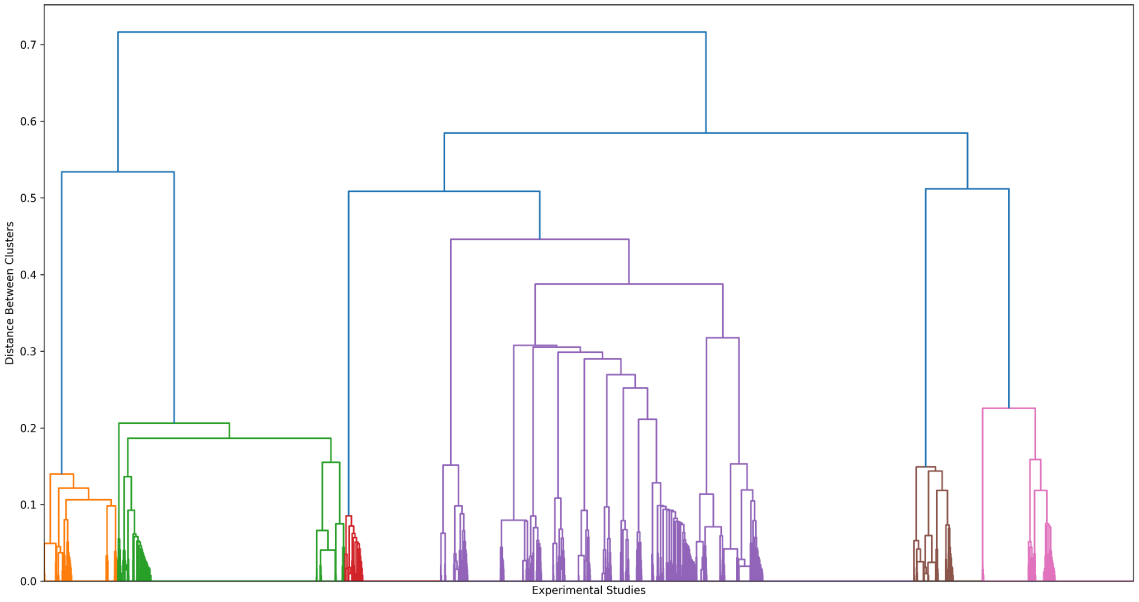


Figure 5.17. Hierarchical clustering of DisGeNET DGA experimental studies using abstract-mined technique terms. A cutoff distance of 0.02 was used to optimise cluster separation, resulting in 321 clusters. Clustering was based on Hamming distance, with vertical lines representing datasets and colours indicating clusters. The vertical axis shows linkage distance (dissimilarity).

Clustering DGA Experimental Studies based on Experimental Techniques extracted from Methods Sections

Similarly, the 1233 experimental studies with method-mined DGA experimental technique terms were also clustered based on their method-mined DGA experimental technique terms. The hierarchical clustering method was chosen to group studies with similar terms together into cohesive clusters. Hamming distance was used as the metric distance, due to the highest values of CPCC. Table 5.14 shows the values of CPCC when using different distance metrics with Hamming resulting in the highest values of CPCC. Regarding the selection of the linkage method, Ward's linkage method was used due to its ability to yield higher CPCC values, as demonstrated in Table 5.15.

Table 5.14. Systematic analysis of distance metric selection in the hierarchical clustering of the experimental studies based on method-mined DGA experimental technique terms. The table shows a systematic analysis of the chosen distance metric. Hamming distance was selected as it yielded higher values of PCC which was used as an evaluation metric for the quality of clusters. The choice of Hamming distance aligns well with our binary data representation.

Metric distance	CPCC
Hamming	**0.35**
Euclidean	0.31
Jaccard	0.19
Cosine	0.19
Correlation	0.27

Table 5.15 Systematic Analysis of linkage method selection in the hierarchical clustering of the experimental studies based on method-mined DGA experimental technique terms. The table shows a systematic analysis of the chosen linkage method. Ward's linkage was selected as it yielded higher values of PCC which was used as an evaluation metric for the quality of clusters.

Linkage Methods	CPCC
Ward's Linkage	**0.40**
Single	0.30
Complete	0.27
Average	0.18
Centroid	0.16
Median	0.14
Weighted	0.11

To choose the distance threshold for clustering, systematic analysis of the clusters based on Silhouette scores was applied. A distance threshold cutoff of 0.04 was selected, as it produced higher Silhouette score values which suggested well-matched studies within their own clusters while maintaining a distinct separation from neighbouring clusters as shown in table 5.16. Figure 5.18 shows the hierarchical clustering of the experimental studies based on the abstract-mined DGA experimental technique terms using hamming distance and the threshold distance between data points (studies) in a single cluster was 0.04 .

Table 5.16. Systematic Analysis of Distance Threshold Selection of the hierarchy clustering of the experimental studies based on method-mined DGA experimental technique terms. The table presents a systematic analysis of the chosen distance threshold based on a Silhouette Score which minimises distances within studies clustered together and maximises distances between studies in different clusters.

Thresholds	#Clusters	Silhouette Score
0.00	242	0.24
0.01	212	0.21
0.02	154	0.18
0.03	109	0.15
0.04	81	**0.12**
0.05	58	0.10
0.06	39	0.09
0.07	30	0.07
0.08	23	0.06
0.09	18	0.05
0.1	12	0.04
0.2	3	0.04
0.3	2	0.05
0.4	1	0.10
0.5	1	0.10

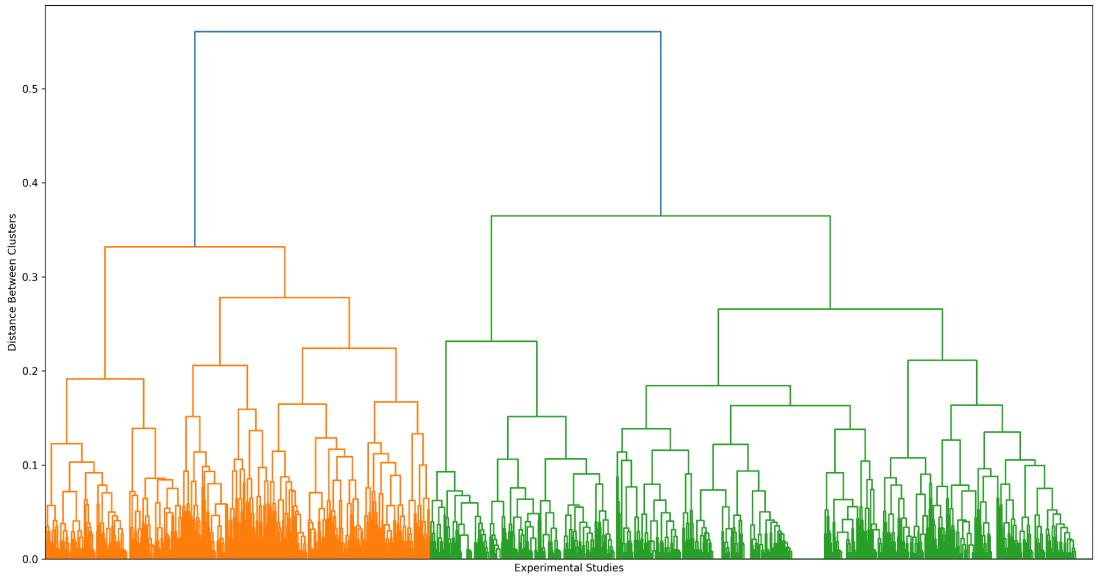


Figure 5.18. Hierarchical clustering of experimental studies using method-mined DGA technique terms. A cutoff distance of 0.04 was applied, resulting in 81 clusters representing individual datasets.

To explore the consistency between clusters of experimental studies based on abstract-mined DGA experimental technique terms and those based on method-mined DGA experimental technique terms, a subset of experimental studies with both abstract-mined and method-mined DGA experimental technique terms (total of 719 studies) were subjected to a two-step clustering process. Initially, the experimental studies were clustered based on abstract-mined DGA experimental technique terms, followed by a separate clustering based on method-mined DGA experimental technique terms. Subsequently, an examination was conducted to assess the consistency between clusters derived from abstract-based clustering and those resulting from method-based clustering. Figure 5.19 shows the clustering of these studies based on abstract-mined DGA experimental technique terms and method-mined DGA experimental technique terms.

The consistency between clusters based on abstract-mined DGA experimental technique terms and method-mined DGA experimental technique terms was not high. This discrepancy arises because abstracts generally lack detailed information about experimental techniques compared to the methods section. Consequently, the list of terms extracted from abstracts is shorter and less comprehensive than that from the methods section. This difference in information can influence clustering results, leading to variations in the clusters formed from the same experimental studies due to differing lists of experimental techniques used in the clustering. Figure 5.20 illustrates the similarity between clusters of the experimental studies based on abstract terms and the clusters of the experimental studies based on method terms.

Furthermore, the number of articles with open access is only 1233, limiting the pool of articles with accessible methods sections from which to extract experimental techniques. Another issue is the construction of the dictionary used for text mining. Dictionary construction is one of the most essential steps in a text mining approach. Unfortunately, the current dictionary contains noise terms such as '*Biochemistry*,' '*Biology*,' '*Chemistry*,' and '*Drug discovery.*' These terms are too general and do not accurately reflect the specific experimental techniques used. Building a dictionary is complex and requires manual curation by experts. The accuracy and comprehensiveness of the dictionary significantly affect the results of the text mining approach. Therefore, developing a precise and extensive dictionary is imperative and would form the basis of future studies. Ensuring that the dictionary

accurately represents the experimental techniques will enhance the reliability of clustering results and improve the overall effectiveness of the text mining process.

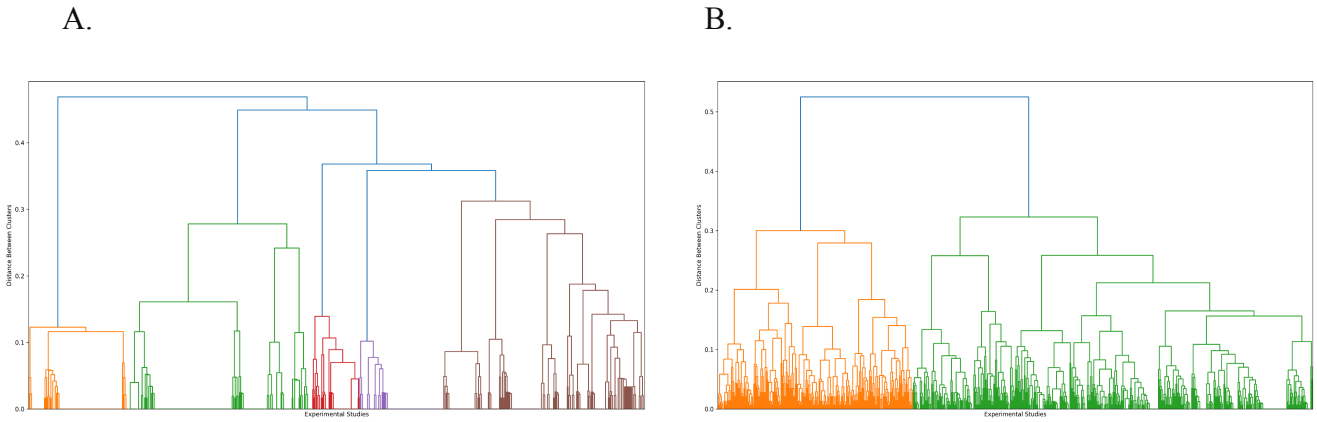


Figure 5.19. Clustering of experimental studies with both abstract- and method-mined DGA technique terms. (A) Clustering based on abstract-mined terms. (B) Clustering based on method-mined terms.

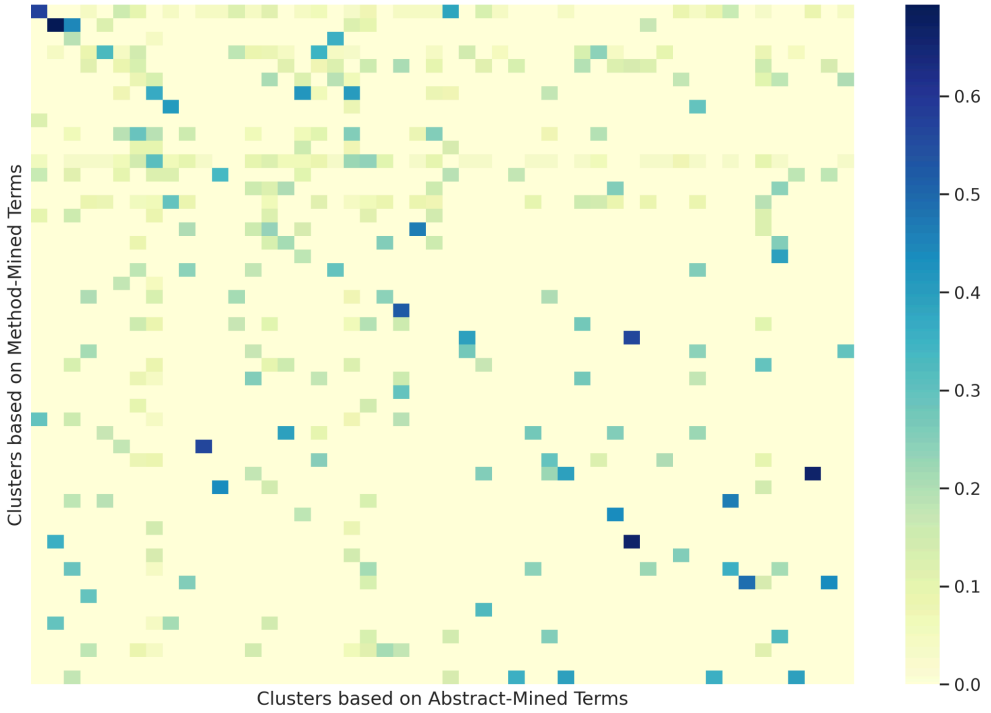


Figure 5.20. Similarity between clusters based on abstract- and method-mined DGA technique terms. Heatmap colour intensity indicates the degree of similarity.

5.4 Conclusion

In this chapter, the limitations identified in the previous chapters were addressed by focusing on three main aspects: the definition of gold standard data, the definition of individual datasets, and the evaluation of network performance.

For the gold standard identification, we introduced a new type of gold standard data. Specifically, non-DGA gold standards, including PPI data and pathways data, were utilised to score DGA datasets. This approach reduced the incidence of infinite scores. However, the rate of data loss remained high. Using non-DGA gold standards introduced issues where the individual datasets cannot be scored against the gold standard. This occurs when a dataset contains only a single gene associated with multiple diseases or when the diseases in the dataset have no common genes, making it impossible to score against PPI or pathways gold standard data. Another issue that leads to a high rate of data loss arises when using non-DGA gold standards due to the discrepancy between the data types of the individual dataset (DGA data) and the gold standard (non-DGA data). The scoring may be reduced because the genes involved in the same disease might not share the same pathway (as scored by pathways data) or might not physically interact (as scored by PPI data).

The use of individual experimental studies as separate datasets contributed to data loss because DGA experimental studies often focus on a single disease or a related group of diseases, which may not be represented in the gold standard. Since it is impractical to have a DGA gold standard that encompasses all diseases, using multiple experimental studies as individual datasets can help mitigate data loss. Consequently, we introduced a novel text mining approach to group experimental studies into individual datasets based on shared experimental techniques. This method aimed to reduce data loss by creating more comprehensive datasets in terms of diseases. This novel approach was introduced that involved extracting DGA experimental techniques employed in DGA experimental studies from biomedical literature. This approach applied a text mining methodology using the terms defined in the EFO and EDAM ontologies. Furthermore, the DGA experimental studies were clustered based on their similarity in experimental techniques

Despite the initial results from text mining demonstrating promising outcomes, this approach has certain limitations. First, mining experimental techniques from abstracts does not yield accurate results because abstracts tend to provide summaries rather than detailed descriptions of the experimental techniques used. Many abstracts only include one or two sentences describing the methods, which is insufficient for detailed analysis. Second, not all articles are open access, limiting our ability to mine the more detailed methods sections. Third, constructing a dictionary that accurately represents the experimental techniques used to generate DGA data is complex and requires manual curation to ensure it includes the specific terms used in biomedical literature. Existing experimental ontologies often contain noise and general terms, which do not accurately reflect the experimental techniques used. The effectiveness of the text mining approach is highly dependent on the quality and comprehensiveness of the dictionary used, which contains terms for experimental techniques applied to generate DGAs. Dictionary construction is regarded as one of the most critical steps in the text mining process. A well-constructed and comprehensive dictionary that covers most experimental terms can significantly enhance performance. However, the inclusion of unrelated terms in the dictionary can negatively impact results. To address these issues, future studies could implement a manual curation process to build a robust dictionary of the most common experimental techniques found in biomedical literature. This curated dictionary will enhance the accuracy and effectiveness of our text mining approach, leading to more reliable clustering results.

Synonym matching was employed to map equivalent terms across different ontologies by using synonyms defined within the ontologies themselves. While other methods were not explored in this work, future research may investigate more advanced approaches, such as machine learning-based term alignment or embedding-based similarity techniques. For example, in the context of experimental ontologies, the term "*Western Blot*" may be mapped to "*Immunoblotting,*" or "*Gene Knockout*" to "*Gene Inactivation*," ensuring consistency in terminology across ontologies.

Additionally, in this work, ontologies including EFO and EDAM were utilised to build a dictionary of experimental techniques used to generate DGAs. This dictionary was then employed to extract relevant techniques from DGA biomedical literature by matching terms in the dictionary with the text in articles, including abstracts and methods sections. The

current approach relied on term matching. In future work, machine learning and natural language processing techniques, such as the Named Entity Recognition (NER) model, may be explored. These models could be trained on DGA biomedical literature to automatically extract experimental techniques from DGA-related articles.

In summary, the solutions presented in this chapter focus on addressing the limitations in gold standard data, dataset definitions, and network performance evaluation techniques. In the future, we aim to enhance the overall effectiveness of text mining approaches to individual datasets identification. Future work will prioritise refining the dictionary through manual curation to better capture the specific experimental techniques, thereby improving the robustness of our methodologies.

Building DGA PFINs requires two main components: individual datasets that represent separate evidence of DGAs, which are scored and subsequently integrated into a network based on the order of their confidence scores, and a gold standard that assesses the quality of these datasets and assigns their confidence scores. In this work, a novel text mining approach was proposed to identify individual datasets by grouping multiple DGA experimental studies based on the similarity of their experimental techniques. Each cluster represents one DGA individual dataset, which is scored against the gold standard. This approach could address the challenge associated with treating a single DGA experimental study as an individual dataset, which resulted in significant data loss. Thus far, individual datasets that represent DGAs were identified. However, due to time constraints, the construction of the PFINs based on the text mining approach and the assessment of network performance have not yet been completed. These aspects of the research will be prioritised in future work.

Chapter 6

A computational Approach to Drug Repurposing Incorporating Graph Neural Networks and Probabilistic Functional Integrated Networks focusing on Disease-gene Association Data

6.1 Introduction

Deep learning has gained considerable attention in biomedical applications with the revolution in publicly available biomedical data sources and the notable improvement in computational techniques [65]. As an inspiring machine learning division, deep learning has significantly boosted and emerged as the leading technique for drug repurposing and development in the most recent published studies [15], [18], [66], [67], [132], [241] (Computational methods to drug repurposing are discussed in 2.7.1 and deep learning approaches in drug repurposing discussed in Section 2.7.1.1). Previous studies showed that DNN-based repurposing methods outperform ML-based methods such as SVM or RF [132]. Moreover, deep learning methods have been reported to be more effective for drug target prediction [133]. Among various deep-learning models, GNNs [25], [64], [67], [68], [69], [70], [71], [311] have attracted increasing attention from the drug repurposing field due to the power of graph modelling in biomedical data [25], [128], [134] (GNNs are discussed in Section 2.6, Graph modelling applications in biomedical data are discussed in Section 2.2.1, and GNNs applications in drug repurposing discussed in Section 2.6.1).

Although the application of GNN approaches for computational drug repurposing have produced very promising results [26], [64], [67], [68], [69], [70], [71], [311], there are still several limitations that need to be addressed. First, GNN-based methods are dependent on the graphs used to train them. GNNs applied to drug repurposing are often trained on incomplete and noisy biomedical knowledge graphs. To-date, many GNN-based studies have been developed around knowledge graphs with limited nodes and edge types and with restricted node and edge features. This limited scope can restrict the ability of GNN models to capture the full complexity of drug-disease relationships. These approaches tend to aggregate information from directly connected nodes restricted in a drug-protein-gene-disease, ignoring the other types of biological entities such as pathways, tissues, signatures, side effects, and

phenotypes, that contain rich information about graphs in drug repurposing applications. The inclusion of additional relevant biomedical entities in a knowledge graph could potentially contribute to enriching the graph structure which might lead the GNN models to predict better over other cases. Recently, new approaches to computationally scalable data integration have led to the availability of a new generation of knowledge graphs with a much richer set of node and edge types and the inclusion of a wider range of node and edge features. These approaches facilitate the aggregation of information not just from directly connected nodes in a drug-protein-gene-disease complex, but also from other types of biological entities. Second, most existing GNN approaches to drug repurposing applications often do not consider data quality issues at the same level as the problem of missing data. Some existing approaches of GNNs to drug repurposing deal well with incomplete datasets, assuming that the existing data is high-quality, whereas biological data, in general, has a high rate of false results (Section 2.3.1) [32], [33]. However, the success of computational techniques is highly dependent on data quality [19], [21]. Incorporating PFIN approaches with biomedical knowledge graphs could greatly reduce noise since PFINs typically reduce noise during integration which might lead to improved GNN performance. Third, most of the current studies neglect the inclusion of node features within GNNs due to the lack of these features [64]. The incorporation of node features could improve the performance of GNNs. In drug repurposing, node features may contain valuable information about graph structures that are not leveraged by existing methods. Finally, many of the previous studies focused on specific diseases, for example, these studies targeting COVID-19; constructed graphs containing information exclusively related to it [25], [26], [68], [68], [311].

In this chapter, the limitations discussed above were addressed by incorporating GNNs within a recently developed extended biomedical knowledge graph and PFINs, in an attempt to bridge these gaps to generate drug repurposing hypotheses using link prediction between drugs and diseases. Overall, the main contributions of this chapter can be summarised as follows:

1. It was hypothesised that incorporating more extensive node and edge types into a GNN could improve its performance. A GNN framework was trained using a newly developed, rich, Heterogeneous Biomedical Knowledge Graph (HBKG), called NeDRex. NeDRex is an extended graph including multiple diverse types of nodes and

edges, consisting of 11 different types of nodes and 19 different types of edges. The NeDRex graph is also a general graph containing different types of diseases, disorders, or syndromes.

2. Many of the currently available HBKGs lack predefined and meaningful node and edge features. It was incorporated a range of node features into the GNNs since a node in NeDRexDB is labelled with a set of attributes, such as the node's name and description. The NeDRex graph contains rich and meaningful features about nodes and edges. In this chapter, specifically, node features were incorporated within GNNs.

3. The PFIN approach was incorporated within the GNN technique. One of the DGA PFINs constructed in the previous chapter (chapter 5) was selected based on its best performance among the others. The subset of DGAs in the extended integrated heterogeneous biomedical knowledge graph (NeDRex graph) was replaced by the DGA PFIN, which had already been scored against high-quality gold standard data, with only the DGAs that passed the scoring included. The inclusion of the DGA PFIN could potentially enhance GNN performance, as PFINs typically reduce data noise.

6.2 Data Sources

NeDRexDB[58] [48] was used as a data source for this research. NeDRexDB was used to build the heterogeneous biomedical knowledge graph that was used for GNN training and testing. NeDRex is a platform in network medicine that allows for the identification of disease modules and drug repurposing. NeDRex was built of three primary elements: a knowledge base (NeDRexDB), a Cytoscape app (NeDRexApp), and an API (NeDRexAPI) and is available as an open version and a licensed version. The open version does not contain OMIM [88] and DrugBank [221] due to licensing restrictions. In this chapter, the licensed version was used. NeDRexDB includes information from 19 existing databases including NCBI [256], DrugCentral [220], CTD [261], DrugBank, SIDER [312], HPO [223], Mondo [282], OMIM, DisGeNET [141], HPA [218], BioGRID [87], Reactome [253], UniProt [217], GO [264], BioOntology [313], Repotrial, UniChem [314], and Clinvar [189] (For more details about the original databases in NeDRexDB, refer to Section 2.5). The NeDRexDB

[58] https://api.nedrex.net/.

used for the present work was queried in December 2022. Statistics for nodes and edges sourced from NeDRexDB can be seen in Tables 6.1 and 6.2, respectively.

RepoDB[59] [257] was used to form a specific evaluation dataset, and drug-disease indications from RepoDB were extracted. RepoDB is a gold standard database for drug repurposing, including a collection of documented instances of both successful and unsuccessful drug repositioning. It can serve as a valuable benchmark for assessing computational repositioning methods. The data within RepoDB has been sourced from DrugCentral and ClinicalTrials[60].

Table 6.1. Statistics on data sources, types, attributes, and number of nodes in the NeDRex database. The features that were excluded from the feature matrix are highlighted using bold formatting.

Node Type	**Data Source**	**#Nodes**	**Attributes of nodes**
Disorder	Mondo Repotrial	22384	**Description**, Name, **Domainids**, Synonyms, Type
Drug	Drugbank Unichem	14315	CasNumber, **Description**, Name, **Domainids**, DrugCategories, **Indication**, DrugGroups, Sequence, Synonyms, Type
Gene	NCBI	81491	ApprovedSymbol, Chromosome, Description, DisplayName, **DomainIds**, GeneType, MapLocation, Symbols, Synonyms, Type
Protein	UniProt	204961	Name, **Domainids**, GeneName, Sequence, Synonyms, **Taxid**, Type
Genomic_variant	Clinvar	1475162	AlternativeSequence, **Chromosome**, **Domainids,** Position, ReferenceSequence, VariantType, Type
GO annotation	GO	43558	Description, Name, **Domainids**, Synonyms, Type
Pathway	Reactom	2566	Name, **Domainids**, Species, Type
Side-effect	Bioontology .org	15095	Name, **Domainids**, Type
Phenotype	HPO	16614	Name, Synonyms, **Domainids**, Description, Type
Signature	UNIPROT	42199	Name, **Domainids**, Type
Tissue	Uberon	64	Name, **Domainids**, Type

[59] http://apps.chiragjpgroup.org/repoDB/

[60] https://clinicaltrials.gov/

Table 6.2. Statistics on data sources, types, attributes, and number of edges in the NeDRexDB database.

EdgeType	**Data Source**	**#Edges**	**Attributes of Edges**
Drug_has_indication	Drugcentral CTD	16837	Type
Drug_has_target	Drugbank drugcentral	26537	Actions, Tags, Type,
Drug_has_side_effect	Sider	217739	Maximum_frequency, Minimum_frequency, Type
Drug_has_contraindication	DrugCentral	13788	Type
Molecule_similarity_molecule	Repotrial	188199	Maccs, Type, Morgan_r1, Morgan_r2, Morgan_r3, Morgan_r4
Disorder_has_phenotype	HPO	216172	Type
Disorder_is_subtype_of_disorder	Mondo	35512	Type
Gene_associated_with_disorder	DisGeNET OMIM	30252	Type, Score
Gene_expressed_in_tissue	HPA	1133252	TPM, nTPM, pTPM, Type
Protein_interacts_with_protein	BioGRID	2346570	BrainTissue, DevelopmentStages, EvidenceType, JointTissues, Methods, SubcellularLocations, Tissues, Type
Protein_encoded_by_gene	UniProt	32930	Type
Protein_in_pathway	Reactom	121759	Type
Protein_has_signature	UniProt	1826518	Type
Protein_expressed_in_tissue	HPA	632173	Type, Level
Protein_has_go_annotation	GO	297003	Qualifiers, Type
Go_is_subtype_of_go	GO	70058	Type
Side_effect_same_as_phenotype	Bioontology	1242	Type
Variant_affects_gene	Clinvar	1596637	Type, DataSources
Variant_assocaited_with_disorder	Clinvar	1104119	Accession, Effects, ReviewStatus, Type

6.3 Results and Discussion

6.3.1. Drug-Disease Link Prediction

The drug repurposing problem was approached as a link prediction task in a biomedical knowledge graph of diseases and their interactions with different biomedical entities such as

drugs, genes, proteins, pathways, phenotypes, signatures, side effects, or tissues. The task involved predicting links between disease and drug. The outcomes of the model can be used to measure the confidence of the link so that a link, with a score of 0 meaning low confidence and a score of 1 meaning high confidence. The problem was addressed as a binary classification task where the classes are negative class; non-existent (0) and positive class; existent (1).

6.3.2 Construction of the Heterogenous Biomedical Knowledge Graphs from NeDRexDB

Five versions of the HBKGs were constructed to test five distinct hypotheses. Firstly, it was hypothesised that enhancing the graph structure by incorporating additional pertinent biomedical entities, such as pathways, tissues, signature, side_effects, GO_annotations, genome variants, and phenotypes, could significantly enrich the HBKG, potentially leading to improved performance of the GNN. To assess this hypothesis, a GNN model was applied to two distinct versions of the graph: the reduced version of the NeDRex graph reducing types of nodes and edges (Reduced_NeDRex), which contains a subset of node and edge types focused on drug-protein-gene-disease relationships, and the complete version of NeDRex (Complete_NeDRex), containing 11 different node types and 19 unique edge types. Secondly, it was hypothesised that enhancing GNN performance could be achieved by incorporating node features. To test this hypothesis, GNN models were applied on both versions, the Reduced_NeDRex and the Complete_NeDRex incorporating node features in NeDRexDB (Reduced_NeDRex_With_NodeFeatures and Complete_NeDRex_With_NodeFeatures). Lastly, it was hypothesised that incorporating the PFIN approach within the GNN technique could potentially enhance GNN performance. To test this hypothesis, a GNN was applied to a new version of the Complete_NeDRex graph (Complet_NeDRex_PFIN) which replaced DGA edges in the Complete_NeDRex version with the DGA in the PFIN constructed in the previous chapter. All versions of the HBKGs were constructed from NeDRexDB using the Pytorch Geometric Python library[61].

6.3.2.1 Reduced_ NeDRex

The Reduced_NeDRex was constructed as a subgraph of NeDRexDB, by including only specific types of nodes and edges. In this version, the GNN only aggregates information from

[61] https://pytorch-geometric.readthedocs.io/en/latest/install/installation.html

directly connected nodes that describe drug-protein-gene-disorder relationships, ignoring other types of biological entities such as pathways, phenotypes, tissues, signatures, side_ effects, genomic_variants, and GO annotations. The Reduced_NeDRex graph contains four node types: *drug, disorder, gene, and protein*, and five edge types: *drug-has-indication, drug-has-target, gene_associated_with_disorder, protein_interacts_with_protein, and protein_encoded_by_gene.* Figure 6.1 shows the schema of the Reduced_NeDRex graph.

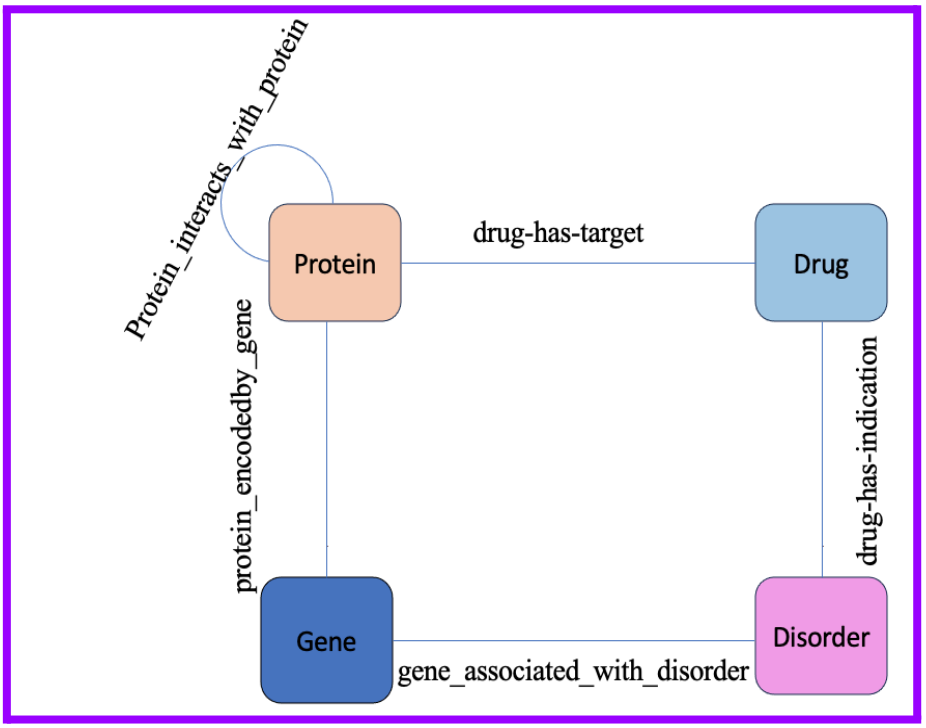


Figure 6.1. Schematic of the Reduced_NeDRex graph from NeDRexDB, containing four node types (drug, disorder, gene, protein) and five edge types. It is a subgraph of NeDRexDB.

6.3.2.2 Complete_NeDRex

The complete_NeDRex consisted of 11 types of nodes including *disorder, drugs, gene, protein, genomic_variant, go annotation, pathways, side effect, phenotype, tissue, and signature*. The set of edges consisted of 19 types including *drug-has-indication, drug-has-target, drug_has_side-effect, drug_has_contraindication, protein_has_signature, molecule_similarity_molecule, disorder_has_phenotype, disorder_is_subtype_of_disorder, gene_associated_with_disorder, protein_interacts_with_protein, gene_expressed_in_tissue, protein_in_pathway, variant_assocaited_with_disorder, side_effect_same_as_phenotype, protein_expressed_in_tissue,* go_is_subtype_of_go, *protein_encoded_by_gene,*

protein_has_go_annotation, variant_affects_gene. Figure 6.2 shows the schema of the Complete_NeDRex graph.

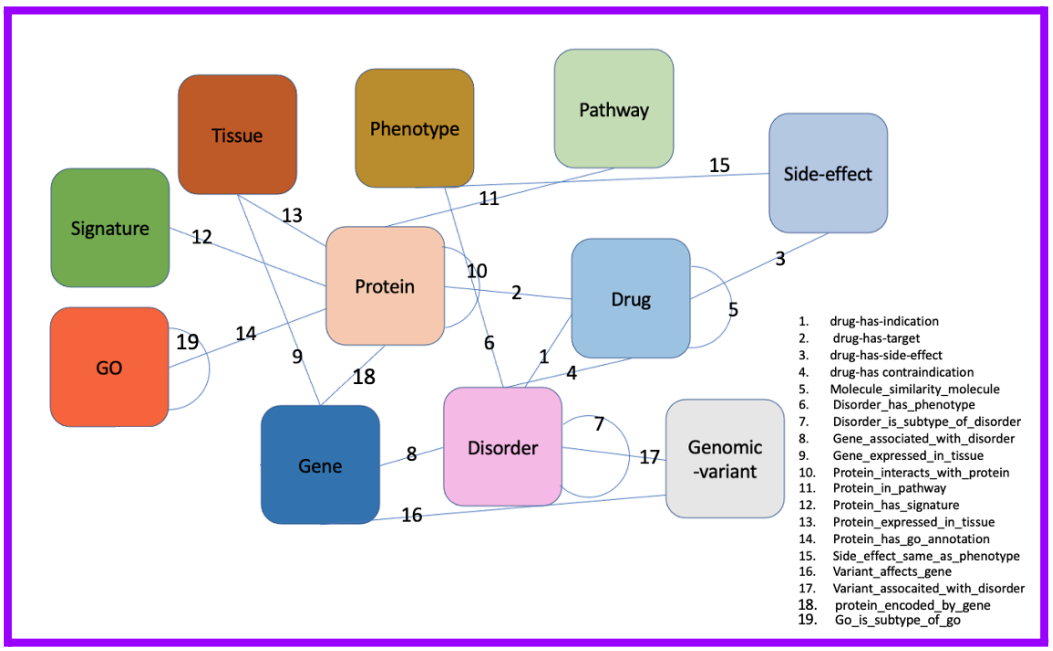


Figure 6.2. Schematic of Complete_NeDRex, featuring 11 node types and 19 edge types. Numbers on edges indicate semantic edge types between node pairs.

6.3.2.3 Reduced_NeDRex_With_NodeFeatures and Complet_NeDRex_With_NodeFeatues

To investigate the hypothesis that incorporating node features with GNN improves performance, two versions of the graphs were constructed: Reduced_NeDRex_With_NodeFeatures and Complete_NeDRex_With_NodeFeatures. NeDRexDB has a rich set of node and edge features (see Tables 6.1 and 6.2, respectively) that were relevant to use in this study. For instance, a node categorised as a *gene* has attributes such as *Chromosome*, *Description*, *Name*, *GeneType*, *MapLocation*, *Symbols*, *Synonyms*, and *Type*. Additionally, each edge in NeDRexDB is characterised by attributes such as edge confidence and evidence type. As an illustration, an edge of type "*protein_interacts_with_protein*" is associated with attributes such as *BrainTissue*, *DevelopmentStages*, *EvidenceType*, *JointTissues*, *Methods*, *SubcellularLocations*, *Tissues*, and *Type*. It is important to mention that not all node and edge features were included in the node features matrix and edge features matrix. Some features were omitted because they do not reflect data useful for this study. For instance, features like *Domainids* merely list node

IDs from various databases, providing little significance compared to features like *Name* or *Description*. Similarly, features that describe the source of edges such as *Datasources* were excluded, as they do not contribute any additional useful information. While some features are informative, they were left out of the node features matrix due to missing data. For example, the *Description* feature for drug nodes carries meaning, but it had to be removed due to that it is missing for some nodes. The features that were excluded from the features matrix are highlighted using a bold formatting style in Table 6.1.

The concept of encoders was utilised to define how the values of specific features should be encoded into a numerical feature representation. For example, a sentence encoder was applied to encode raw feature strings into low-dimensional embeddings. For this, the sentence-transformers library[62], which provides many state-of-the-art pretrained natural language processing embedding models, was used [315]. The SequenceEncoder was used to encode a list of strings features of nodes into node features matrix. Each node and edge in NeDRexDB is annotated with features, which correspond to their included attributes (see Tables 6.1 and 6.2, respectively). Edge features were not incorporated with the GNN in this work. To assess how the inclusion of node features influences the performance of the GNN, we apply the GNN on each version of the graph with/without node features. In the case of the graph without node features, the node features vector in the matrix was set to a constant value of one.

6.3.2.4 Complete_NeDRex_PFIN

To assess the impact of data noise on the performance of GNNs, the final version of the graph (Complete_NeDRex_PFIN) was constructed by substituting the DGAs in the Complet_NeDRex with one of the DGA PFINs constructed in the previous chapter; the BioGRID_IES network (Section 5.3.1.1 for the PFINs construction using non-DGA gold standard). The reason for choosing the BioGRID_IES network is that, from the previous chapter, it was observed that non-DGA-based networks including the BioGRID_IES network, demonstrated superior performance in link prediction analysis when compared to DGA-based networks which performed well in network clustering analysis (Section 5.3.1.2 for the

[62] https://www.sbert.net/

network evaluation). Consequently, it was concluded that none-DGA-based-networks may produce better performance in link prediction applications, while DGA-based networks may be useful in network clustering tasks. Therefore, given that the drug repurposing problem tackled in this chapter was framed as a link prediction problem, it was opted to replace the DGAs in the Complete_NeDRex with the BioGRID_IES network.

The Complete_NeDRex and the Complete_NeDRex_PFIN exhibited variations in both content and topological structures in terms of the DGA subgraph, despite that the DGAs in both versions of the graph originated from the same source which is DisGeNET. This discrepancy arose from the inclusion of distinct subsets of DGAs in the NeDRexDB and the BioGRID_IES. Specifically, the DGAs in the BioGRID_IES are the subset from DisGeNET that passed scoring against the non-DGA gold standard data (BioGRID). In contrast, DGAs within the NeDRexDB constitute a subset of DisGeNET and OMIM that are chosen by NeDRexDB's team including manually curated data only. The Complete_NeDRex contained 30,252 DGAs while the Complete_NeDRex_PFIN contained 23,085 DGAs. The PFIN approach is designed to systematically reduce noise during integration. Therefore, it is reasonable to posit that DGAs in Complete_NeDRex_PFIN may exhibit a lower level of noise compared to DGAs present in the Complete_NeDRex graph which may lead to better performance for the Complete_NeDRex_PFIN graph over the Complete_NeDRex.

The topological structure of a graph profoundly influences how GNNs operate, impacting the aggregation of neighbour information, the flow of information, and the learning of representations. As a result, differences in topological structures can lead to different results in GNN performances. GNNs operate by aggregating information from neighbouring nodes. If the topological structures differ, the set of neighbours for each node changes, impacting the information available for aggregation, and affecting the learning of node representations and, consequently, the overall performance of the GNN. The variations in the number and type of nodes, edges, and overall graph density can affect the scale and complexity of the learning task. GNNs adapt their parameters based on the graph structure, and differences in size and density may lead to different learning dynamics.

6.3.3 Heterogeneous Link-Level Graph Neural Model Architecture

The objective of the drug repurposing exercise was to estimate the probability of a link existing between a given disease and drug pair, using a GNN model. The architecture of the model includes two main parts: an encoder and a decoder (Figure 6.3). The encoder part generates representations of the nodes (nodes embeddings), producing a vector embedding for each node in the graph. The encoder captures the essential features of a node from its neighbours in the NeDRex graph. The decoder part utilises these embeddings to predict the probability of an edge's presence between the drugs and the disorders; the '*drug-has-indication*' edge. This problem is treated as an end-to-end task, where both the encoder and decoder are optimised in tandem. The encoder contains two layers of a graph convolution network. After each of them, a batch normalisation takes place to reduce the variance between layers during the training process [316]. Between the layers, there is a leaky rectified linear unit (LReLU), providing non-linearity, and a dropout layer, to reduce variance [317]. The GNN's Layers use GraphSAGE as a framework [280]. The GraphSAGE generates node representations by sampling and aggregating features from a node's neighbourhood (Section 2.6) . For each node in the graph, a fixed-size neighbourhood is sampled. This involves selecting a set of neighbouring nodes for each target node. After sampling neighbours, an aggregation function was applied to gather information from the sampled nodes. The aggregation function combines the feature information from the target node and its neighbours to create a new representation for the target node. This aggregated information was used to update the node's representation (Equation 3.15) [280]. The GraphSAGE model was chosen because it is effective in handling large graphs and achieving good generalisation performance for inductive learning on graph-structured data. The final embeddings were produced by utilising 2-hop neighbourhood features since we have only implemented two layers of GraphSAGE. After generating the drug and disorder embeddings, the decoder utilised these embeddings to estimate the likelihood of an edge between a disorder and a drug. This estimation is computed by applying the sigmoid function to the dot product of the embedding of the disorders and drugs, represented in Equation 3.16.

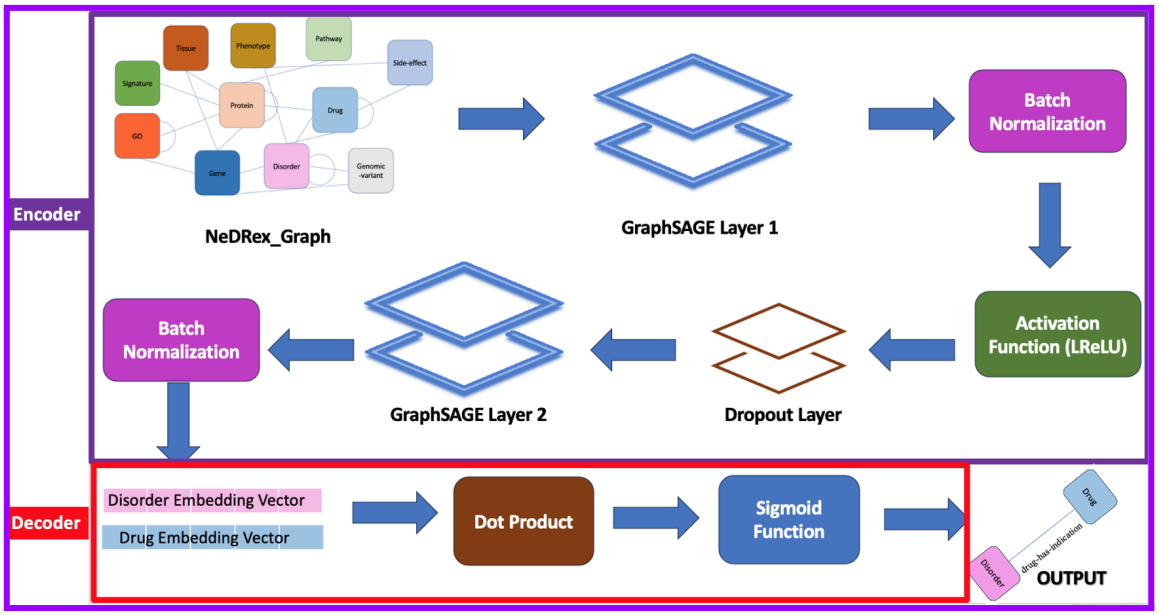


Figure 6.3. Structure of the Graph Neural Network encoder and decoder. A batch normalisation follows two graph convolution layers. A non-linear activation function (LReLU) and a dropout layer were applied between the two GraphSAGE layers.

The construction of all versions of HBKG, the model implementation and the training phase were developed in Python 3.8.10, using the PyTorch Geometric(PyG) framework. The PyG framework was employed for these tasks. The experiments were executed on the Google Colab server. The code can be accessed on GitHub:
https://colab.research.google.com/drive/1_REVe42P7H6USan0Z1GsLPve_AmiamEx

6.3.4 Model Training

The five models, including the Reduced_NeDRex, the Complete_NeDRex, the Reduced_NeDRex_With_NodeFeatures, the Complete_NeDRex_With_NodeFeatures, and the Complete_NeDRex_PFIN were trained separately. During each training phase, the graph was split into training, validation, and test sets ensuring that no data about edges used throughout the evaluation phase was leaked into the training phase. The edges in the `("drug_has_indication")` were randomly divided into training, validation and test edges. The edge-level random split was performed such that each edge was exclusively either in the training, validation, or test split. To train each model, a distinct training set comprising 80% of the original graph was utilised. The remaining 20% of the data was divided, with 10% allocated for internal validation purposes such as monitoring overfitting and underfitting, tuning hyperparameters, and ensuring no data leakage, while the remaining 10% was allocated for testing.

NeDRexDB contains only positive edges. For example, it is stated that a "drug_has_indication", but there is no edge type for "*drug_does_not_have_indication*". A lack of such an edge between a "*drug*" and a "*disorder*" could mean that a drug does not have an indication for a disorder, however, It is not known whether there is evidence/no evidence that the drug does not have an indication. One option is to make a closed-world assumption that states that no relationship between two nodes means a true negative relationship. We have 16837 positive samples ("*Drug_has_indication*") in the NeDRexDB. Negative sampling was implemented such that for every positive edge in the graph, a random edge (referred to as a negative edge) was sampled. These negative edges were not originally present in the graph and the ratio of these edges to positive edges is one-to-one.

Each version of the graph was trained using mini-batch stochastic gradient descent by dividing the training set into batches. A mini-batch loader was developed to generate subgraphs that were used as input into the GNN models. While this process is not required for small graphs such as the reduced version of NeDRex, it is necessary in the case of large graphs such as the complete version of NeDRex that do not fit onto GPU memory otherwise. Therefore, multiple hops were sampled from both ends of an edge and a subgraph was created from these samples. Negative sampling was generated using a closed-world assumption during the mini-batch creation process, thereby preserving a class ratio of one (the ratio of negative samples to positive samples) in each batch.

The model parameters were optimised by computing the loss value from the ground-truth labels and the obtained predictions. The ground truth labels refer to the labels of the real edges in the graph between drug and disorder; "*drug_has_indication*". The true positive edges were labelled with one and the true negative edges were labelled with zero. These labels were compared with the predicted edges which were labelled with the predicted probability scores. The model parameters were optimised by minimising the loss value, which is obtained by measuring the difference between the predicted scores and the ground-truth labels. The Binary Cross Entropy with Logit Loss was used as the loss function to adjust model parameters through back-propagation and stochastic gradient descent (Equation 3.17). The BCELogitLoss was chosen because it is suitable for binary classification

problems and the drug repurposing problem was treated in this work as a binary classification (Figure 6.4).

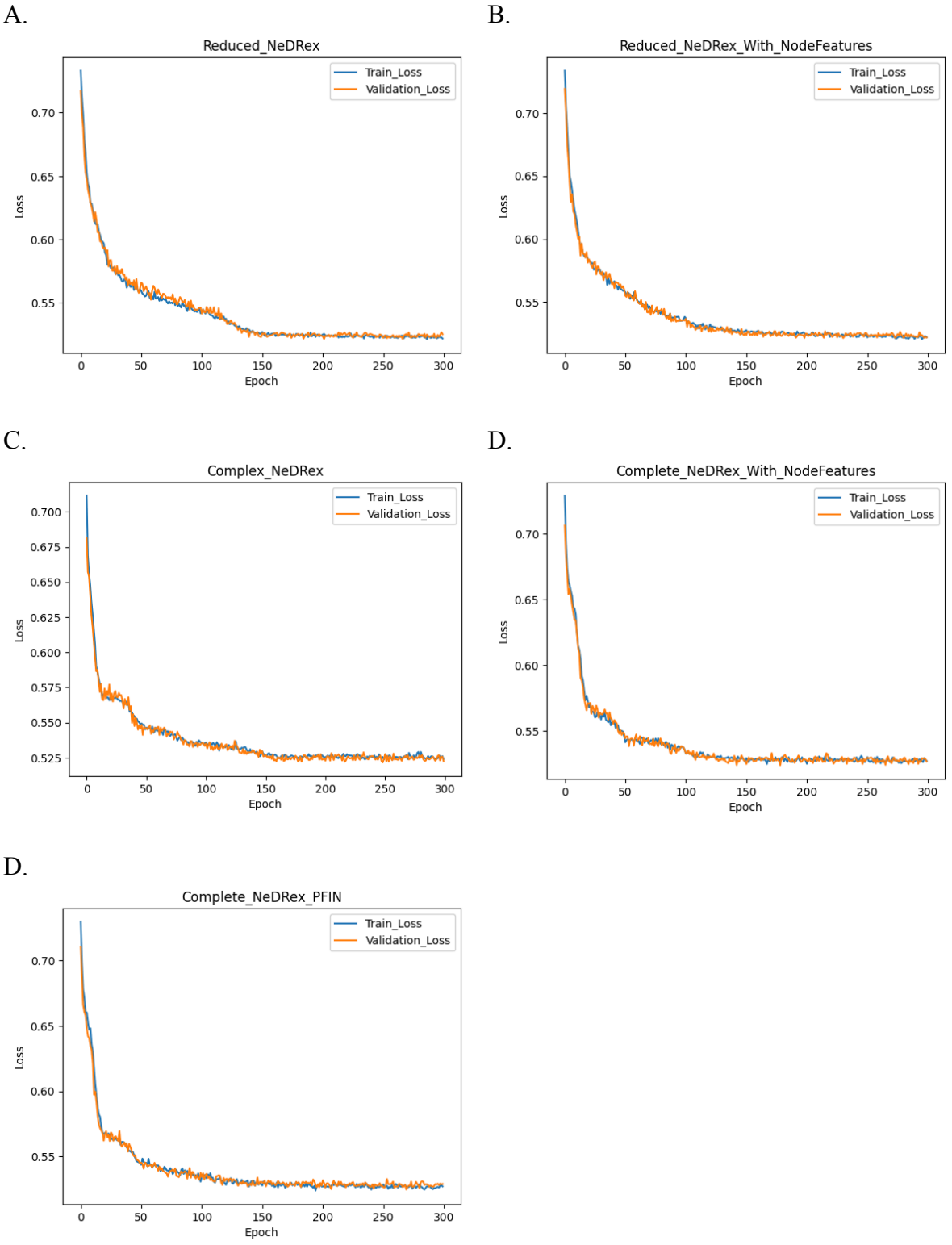


Figure 6.4 Training and validation loss curves for five model configurations: A. Reduced_NeDRex, B. Complete_NeDRex, C. Reduced_NeDRex_With_NodeFeatures, D. Complete_NeDRex_With_NodeFeatures, and E. Complete_NeDRex_PFIN. Each curve shows the loss values over 300 training epochs.

The model was optimised with the Adam optimiser [318], and the hyperparameter tuning encompassed key factors such as epochs (training iterations), weight decay, learning rate, and hidden embeddings. To optimise the model's performance, hyperparameter tuning was carried out using the Weight & Biases platform[63]. The selected set of hyperparameters was determined based on achieving the highest Accuracy, AUROC, and AUCPR while minimising the average loss values on the training, test and validation sets and producing optimal curves(no overfitting, no underfitting). Table 6.3 shows the final chosen hyperparameter configuration for each version of the graph.

Table 6.3. The final hyperparameter configuration selected for each graph version is determined by the optimal performance achieved by the five models.

Hyperparameters	**Reduced_ NeDRex**	**Reduced_ NeDRex_ With_NodeFeatures**	**Complete_ NeDRex**	**Complete_ NeDRex_ With_NodeFeatures**	**NedRex_ PFIN**
Activation function	Leaky_ relu	Leaky_relu	Leaky_ relu	Leaky_relu	Leaky_relu
Batch_size	1024	1024	1024	1024	1024
Dropout	0.60	0.60	0.80	0.80	0.80
Learning rate	0.01	0.01	0.01	0.01	0.01
Weight decay	0.0001	0.0001	0.001	0.001	0.001
Hidden channels	80	80	100	100	100
Aggregation function	mean	mean	mean	mean	mean
Number layers	2	2	2	2	2

6.3.5 Model Evaluating and Validating

It is important to mention that the data utilised for training, testing, and validation were kept strictly separate. After training, the models were evaluated on unseen data coming from the test and the validation sets. Two separate steps were applied to evaluate the model. Firstly, the testing set, consisting of 10% of the original graph, was utilised to calculate traditional evaluation metrics including the Area Under the Receiver Operator Characteristics Curve (AUROC) [319], the Area Under the Precision-Recall Curve (AUPRC) [320], and overall

[63] https://wandb.ai/site

accuracy which are widely used for drug indication prediction tasks. Secondly, further validation was performed with the validation set, consisting of 10% of the original graph. For this, we define a new `mini batch Loader` to generate subgraphs for validation. The validation loop iterates through mini-batches within the validation set, generating predictions for validation edges using the models. It evaluates the performance of the model by calculating the AUC score across the predictions and their associated ground-truth edges, including both positive and negative edges.

6.3.6 Further Validation

For added validation, a secondary evaluation was conducted using a Gold Standard dataset for drug repurposing, RepoDB [257]. The drug-disease indications provided by RepoDB were used to validate the predictions made by the trained model. The disease IDs were mapped from UMLS IDs to Mondo IDs since NeDRexDB used mondo IDs for diseases while RepoDB used UMLS IDs. From these indications, 3, 812 drug-disease pairs were selected for repositioning after mapping UMLS to Mondo. Out of these, 2562 pairs are already present in NedRexDB and are therefore excluded. As for the nodes, we selected only the cases from RepoDB where both the drug and the disease node were already in the graph. Therefore, 1250 pairs were chosen and then supplied to the trained models, with the expectation that each pair would result in a prediction score of one. Additionally, an additional negative set of 1250 randomly generated drug-disease pairs was introduced. In this case, the expectation is that all pairs would receive a prediction score of zero, given that the likelihood of randomly selecting a valid pair for repositioning is notably low.

To test the hypothesis that adding extra types of nodes and edges to the graph, including node features, as well as incorporating PFIN approach, could improve the performance of GNNs, five versions of the graph were created using NeDRexDB. These versions were Reduce_NeDRex and Complete_NeDRex which looked at how adding extra nodes and edges affected performance. It was also created Reduce_NeDRex_with_NodeFeatures and Complete_NeDRex_with_NodeFeatures to see if adding features to the nodes improved GNN performance. NeDRex_PFIN were also created to test the hypothesis that incorporating PFIN with the HBKG could reduce noise within the HBKG, potentially enhancing the performance

of the GNN. The five models were trained and tested on Reduced_NedRex, Complete_NedRex,Reduced_NedRex_With_NodeFeatures,Complete_NedRex_With_NodeFeatures. Overall, the results indicated that a model trained on the complete version of the graph performed better than the reduced version. Significant results were observed for models trained with both versions of the graph; the reduced and the complete graph, with node features. Moreover, some metrics showed better performance in the reduced graph when incorporating node features compared to the complete graph with no node features. For example, the accuracy of the Reduced_NeDRex_With_NodeFeatures model improved compared to the Complete_NeDRex model. The results indicate that node features have a considerable impact and have significantly improved the performance of the GNN in both cases. The Complete_NeDRex model improved the Reduced_NeDRex model by, approximately, 0.09 for the Accuracy, 0.10 for the AUPRC, and 0.07 for the AUCROC (Table 6.4). More importantly, training with the Complete_NeDRex_With_NodeFeatures improved on training with the complete version without nodes features by approximately 0.09 for the Accuracy, 0.04 for the AUPRC, and 0.07 for the AUCROC. Complete_NeDRex_With_NodeFeatures model outperformed all the proposed models across all presented evaluation metrics (See bold results in Table 6.4). This result indicated that adding additional types of nodes and edges as well as incorporating node features could significantly improve the performance of GNNs. Table 6.5 contains the comparison of the results of the four versions of the graph. Figure 6.5 shows the AUCROC for the models trained on the four versions of the graph.

In the context of Complete_NeDRex, it is important to highlight that the enhancement primarily focuses on incorporating additional data of diverse types, rather than simply increasing quantity. While there is a common concern that adding more data could introduce noise into the knowledge graph, in this specific case, the augmentation is specifically tailored to include various types of valuable information. Importantly, the size of each association remains relatively modest when compared to documented associations in existing literature. For example, the *Drug_has_indication* edge comprises only 14,315 edges, significantly smaller than the numerous edges found in the literature. Similarly, the *Gene_associated_with_disorder* consists of 30,252 edges, a fraction of the often million-edged literature counterparts. This is attributed to NeDRex's database integrating data

from reliable primary resources, emphasising manual curation, as opposed to indiscriminately including all available data sources.

Table 6.4 Comparative analysis of the results for the five graph versions tested on RepoDB. Metrics including Loss, Accuracy, AUCROC, and AUCPR are presented for models trained on Reduced_NeDRex, Complete_NeDRex, Reuced_NeDRex_With_NodeFeatures, Complet_NeDRex_With_NodeFeatures, and Complete_NeDRex_PFIN. Results highlighted in bold indicate the best-performing outcome.

Graph version	Loss	Accuracy	AUCPR	AUCROC
Reduced_NeDRex	0.66	0.57	0.789	0.79
Reduced_NeDRex_with_NodesFeatures	0.60	0.71	0.87	0.85
Complete_NeDRex	0.63	0.66	0.890	0.863
Complete_NeDRex_with_NodesFeatures	**0.58**	**0.75**	**0.93**	**0.930**
Complete_NeDRex_PFIN	0.66	0.57	0.77	0.78

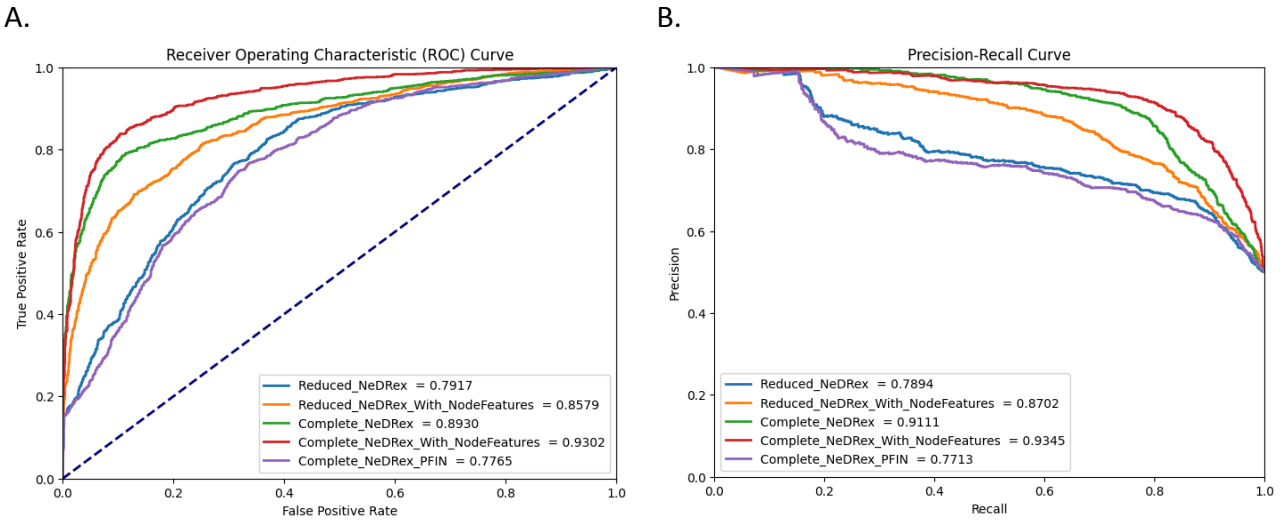


Figure 6.5. (A) ROC curves and (B) Precision-Recall curves for models trained on various NeDRex graph versions, validated using the drug repositioning gold standard, RepoDB.

Moreover, the intricate interconnectedness within the biological system necessitates a thorough examination of all connected components to obtain a holistic understanding. Consider the following example: *Drug1* targets *Protein1*, which is part of *Pathway1*, and *Disorder1* is associated with *gene1*, encoding *Protein2*, also within *Pathway1*. In the context

of Complete_NeDRex, one of the augmented potential paths to gather information from neighbours to generate node embedding involves *Drug1* ➔ *Protein1* ➔ *Pathway1* for *Drug1 and Disorder1* ➔ *Gene1* ➔ *Protein2* ➔ *Pathway1* for *Disorder2*. In contrast, within the Reduced_NeDRex where the Pathway node type is absent, there may be minimal to no similarity in node embeddings between *Drug1* and *Disorder1* due to the lack of pathway data. However, in the Reduced_NeDRex case, a conspicuous similarity emerges in node embeddings due to the shared pathway, highlighting the significance of considering the broader network for a more comprehensive understanding. Figure 6.6 and 6.7 illustrates how the absence of critical data types, such as the pathway node, can impact node embeddings.

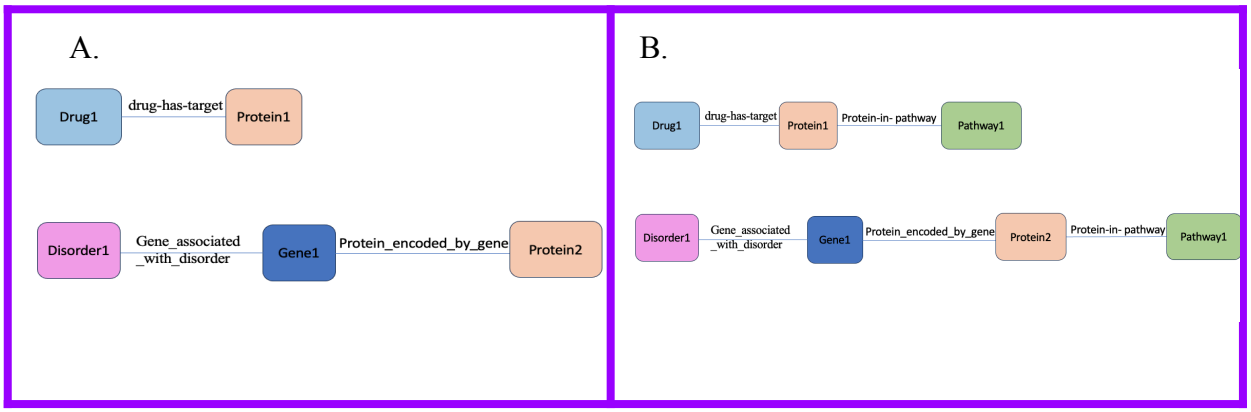


Figure 6.6 Impact of missing pathways on node embeddings. (A) In Reduced_NeDRex, Drug1 and Disorder1 show no embedding similarity due to the absence of pathway nodes. (B) In Complete_NeDRex, including pathway nodes enables potential embedding similarity.

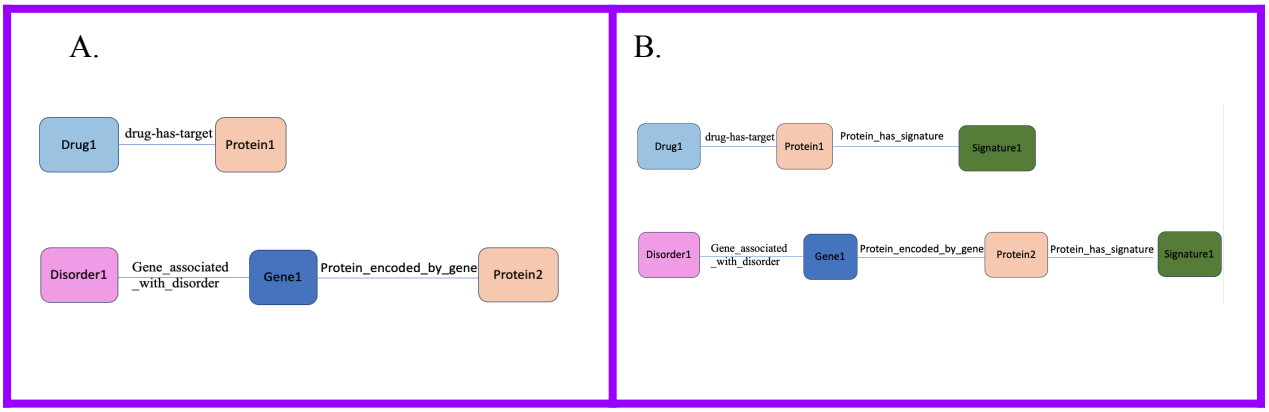


Figure 6.7 Impact of missing data types on node embeddings. (A) In Reduced_NeDRex, Drug1 and Disorder1 show no embedding similarity due to the absence of the signature node. (B) In Complete_NeDRex, the inclusion of the signature node enables potential embedding similarity.

The Complete_NeDRex_PFIN produced the lowest performance among the five versions of the graph. However, this outcome can be attributed to a specific limitation. The necessity for disorder ID mapping arose because the NeDRex graph employed Mondo identifiers for disorders, whereas the PFIN network used UMLs identifiers. The total number of DGAs in PFINS was 23,085, and the ID mapping process resulted in only 7,524 DGAs being successfully mapped, indicating a relatively high rate of DGA loss. Despite this challenge, it is worth noting that PFIN typically contributed to noise reduction during integration. Therefore, incorporating PFINs into the knowledge graph may lead to an overall improvement in performance. Therefore, once the ID mapping issue is addressed, incorporating PFINs into the knowledge graph could potentially lead to overall performance improvement.

6.3.7 Novel Drug-Disease Predictions

The model trained on the Complete_NeDRex_With_NodeFeatures was used to predict novel drug-disease indications. To filter the prediction for further investigation, a filter was employed to choose cases with a prediction score of 1. Another filter was also applied to select diseases that had no existing treatments in the original graph. After applying these filters, a set of novel predictions generated by the model was chosen for further analysis, as shown in Table 6.5.

Table 6.5. The novel predictions of the Complete_NeDRex_With_NodeFeatures model and the number of previous clinical trials reported these indications. The details of some predictions are provided in the text below. Figure C.1 shows the subgraphs of the rest of these predictions.

Drug ID	**Drug name**	**Disorder ID**	**Disorder name**	**Number of Clinical Trials**
Drugbank.DB00073	Rituximab	mondo.0002280	Anemia	42
drugbank.DB00147	Pyridoxal	mondo.0018076	tuberculosis	38

drugbank.DB 0073	Rituximab	mondo.0001106	Kidney failure	25
drugbank.DB 00005	Etanercept	mondo.0002280	anemia	2
drugbank.DB 00041	Aldesleukin	mondo.0005148	type 2 diabetes mellitus	2 note: Aldesleukin is currently undergoing clinical trials in Phase II for the treatment of Type 1 Diabetes, Updated March 19, 2024[64]
drugbank.DB 00009	Alteplase	mondo.0004995	Cardiovascular disorder	317
drugbank.DB 00762	Irinotecan	mondo.0002691	liver cancer	118
Drugbank.DB 00030	insulin human	mondo.0005252	Heart Failure	58
drugbank.DB 00030	Insulin human	mondo.0005406	Gestational diabetes	138
drugbank.DB 00013	Urokinase	mondo.0004995	Cardiovascular disorder	42
drugbank.DB 01914	D-glucose	mondo.0005267	heart disorder	586

Table 6.6 Novel predictions of the Complete_NeDRex_With_NodeFeatures model, with literature supporting these indications, specifically focused on Alzheimer's disease.

64 https://www.pharmaceutical-technology.com/data-insights/aldesleukin-iltoo-pharma-type-1-diabetes-juvenile-diabetes-likelihood-of-approval/#:~:text=Aldesleukin%20is%20under%20clinical%20development,for%20progressing%20into%20Phase%20III.

Disease ID	Disease Name	Drug ID	Drug Name	Previous Studies Supporting Prediction
Mondo.0004975	Alzheimer disease	drugbank.DB03366	Imidazole	Dhingra *et al.* 2022 [321]
		drugbank.DB01225	Enoxaparin	Bergamaschini *et a*l. 2004 [322]
		drugbank.DB00490	Buspirone	Desai *et al.* 2003 [323]
		drugbank.DB00671	Cefixime	Nassar *et al.* 2023 [324]
		drugbank.DB00543	Amoxapine	Li *et al.* 2017 [325]
		drugbank.DB00649	Stavudine	La Rosa *et al.* 2022 [326]
		drugbank.DB00638	Inulin	Hoffman *et al.* 2019 [327]
		drugbank.DB00683	Midazolam	Wang *et al*. 2022 [328]
		drugbank.DB00700	Eplerenone	Hira *et al.* 2020 [329]

Rituximab for Anemia

The Complete_NeDRex_With_NodeFeatures model suggested that Rituximab may have potential applications in treating anemia. The prediction is bolstered by findings from 42 clinical trials. Rituximab is a monoclonal antibody medication that targets a specific protein called CD20, which is found on the surface of B cells. B cells are a subset of white blood cells that play a role in the immune system. Rituximab works by binding to CD20, leading to the destruction of B cells. Anemia is a health condition marked by a decrease in the level of haemoglobin in the bloodstream. Haemoglobin is a protein found in red blood cells which binds with oxygen and transports it to the tissues of the body. Anemia can result in a reduced ability of the blood to carry oxygen to various parts of the body. Anemia is treated by Cyclophosphamide. Cyclophosphamide is also used to treat Granulomatosis with Polyangiitis (GPA) and Microscopic Polyangiitis (MPA) [330]. The advent of Rituximab has prompted inquiries into the necessity of Cyclophosphamide in the management of GPA and MPA, suggesting that newer treatment options are under consideration and assessment. Puéchal and co-workers conducted a comparative study that showed that rituximab is more effective than

cyclophosphamide for the treatment of GPA [331]. The study found a significantly higher remission rate with Rituximab induction therapy compared to cyclophosphamide, supporting the notion that Rituximab may be a more favourable choice for remission induction in patients with GPA. Brodsky et al. found that high-dose cyclophosphamide is a highly efficient treatment for severe aplastic anemia [332]. Rodrigo et al. systematically evaluated the therapeutic efficacy of rituximab, particularly in the treatment of different types of autoimmune hemolytic anemia (AIHA) [333]. AIHA is a specific type of anemia that occurs when the body's immune system mistakenly targets and destroys its own red blood cells. They found that there was enough evidence to endorse rituximab as a second-line treatment for AIHA either as monotherapy or combined therapy. Fattizzo et al. investigated the efficacy of low-dose rituximab in the treatment of AIHA [334]. The study utilised a fixed dose of 100 mg of rituximab administered once weekly for 4 weeks, combined with a short course of prednisone. The results showed that this Rituximab regimen led to response rates of over 80% within the first three years of treatment. The findings suggested that low-dose Rituximab could be an effective treatment option for AIHA, with the potential to reduce the need for steroids. Figure 6.8 shows the subgraph containing anemia and Rituximab.

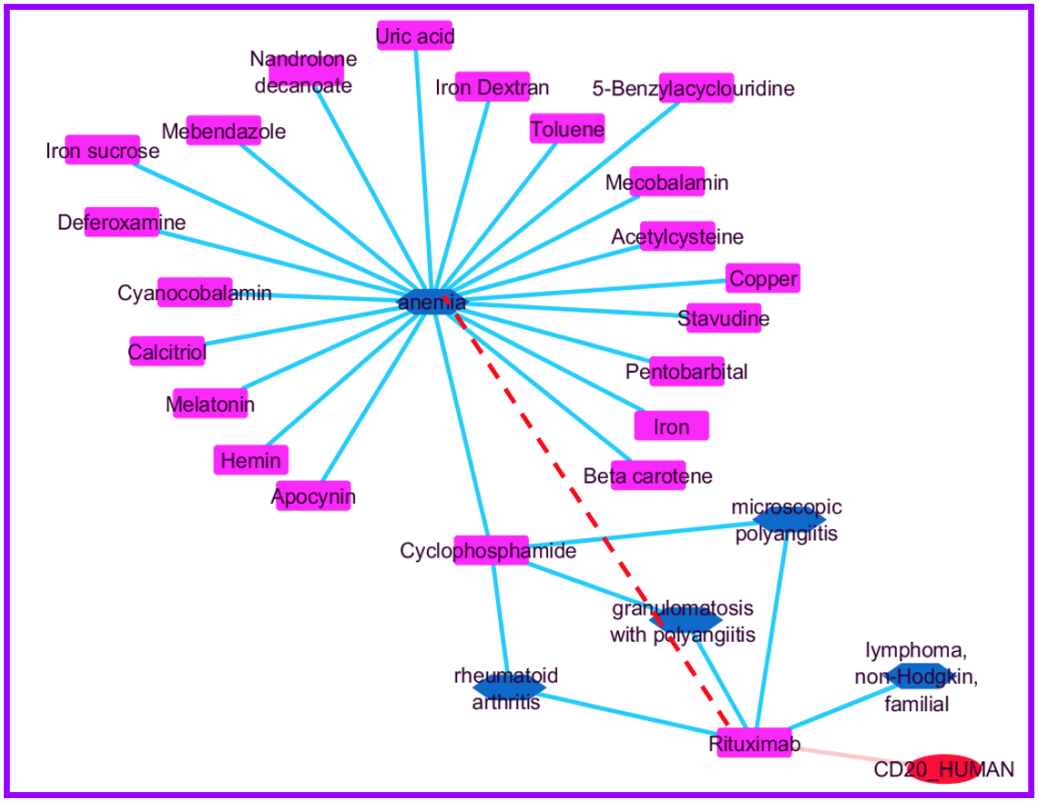


Figure 6.8 Predicted association between Rituximab and anemia. Pink node = Rituximab; red/blue nodes = proteins/diseases. Dashed edge shows predicted link. Based on existing links with Cyclophosphamide (used for anemia,

GPA, and MPA), and Rituximab's greater effectiveness in GPA, Rituximab may be a stronger candidate for anemia treatment. Blue edges = drug-indication; pink edges = drug-target.

Pyridoxal for tuberculosis

The Complete_NeDRex_With_NodeFeatures model also suggested that Pyridoxal may be a potential treatment for tuberculosis. This prediction is supported by 38 clinical studies. Pyridoxal is one of the forms of vitamin B6, which is a water-soluble vitamin needed for diverse biological processes. Pyridoxal is a coenzyme in many enzymatic reactions involved in amino acid metabolism. It is involved in the conversion of amino acids, especially in the synthesis and breakdown of certain neurotransmitters. Tuberculosis (TB) is an infectious disease caused by the bacterium Mycobacterium tuberculosis. TB spreads through the air when infected individuals cough or sneezes. In the 1960s, the Tuberculosis Chemotherapy Center in Madras conducted two randomised controlled trials to investigate the impact of pyridoxal on isoniazid-induced peripheral neuropathy [335]. The first trial involved patients on isoniazid who displayed neuropathy symptoms and did not initially receive pyridoxal. Those patients were subsequently administered either pyridoxal or other B vitamins. Results revealed the effectiveness of pyridoxal in treating neuropathy. The second trial randomised isoniazid patients into groups receiving 6 mg of pyridoxal alone, 6 mg of pyridoxal within a vitamin B complex supplement, 48 mg of pyridoxine alone, or a B complex supplement without pyridoxal. They observed that neuropathy occurred only in the group lacking pyridoxal which indicates the efficacy of pyridoxal in isoniazid-induced peripheral neuropathy. The study investigated the use of pyridoxal as a supplement in addressing the side effects associated with tuberculosis medications. Additionally, Dick and co-workers investigated the impact of vitamin B6 biosynthesis in *Mycobacterium tuberculosis* [336]. The study aimed to identify potential targets for novel antimycobacterial agents. The researchers discovered a heteromeric PLP synthase, which consists of Pdx1 and Pdx2, responsible for synthesising PLP in the pathogen. Disruption of the *pdx1* gene led to a strictly B6 auxotrophic *M. tuberculosis* mutant, whcih emphasises the importance of de novo PLP synthesis for bacterial growth and survival. The study combined *in vitro* and *in vivo* models to demonstrate the dependence of bacterium on PLP biosynthesis for regrowth during dormancy and revealing a severe growth defect of the *Δpdx1* mutant in immunocompetent mice. The findings emphasised the critical role of vitamin B6 biosynthesis in *M. tuberculosis*

survival and they proposed this pathway as a potential target for the development of new antitubercular agents, particularly in the context of multidrug-resistant TB. Figure 6.9 shows the subgraph containing Pyridoxal and tuberculosis.

Furthermore, Chen et al. discovered that verapamil, an FDA-approved calcium channel blocker, enhances the effectiveness of various anti-tuberculosis drugs [337]. Additionally, verapamil is recognized as the most potent inhibitor of aldehyde oxidase AOXA [338]. Pyridoxal is also oxidised by liver aldehyde oxidase, indicating that aldehyde oxidase is involved in the processing of pyridoxal [339]. Given that both verapamil and pyridoxal target the same protein, AOXA, and considering the proven efficacy of verapamil in enhancing the effectiveness of anti-tuberculosis drugs, pyridoxal may hold potential as an enhancer of anti-tuberculosis drugs. Moreover, both verapamil and pyridoxal are indicated for the same disorder [340], [341], visual epilepsy, and since verapamil is already used to treat tuberculosis, It may suggest a potential edge between pyridoxal and tuberculosis. Figure 6.9 shows the subgraph containing these semantic relationships.

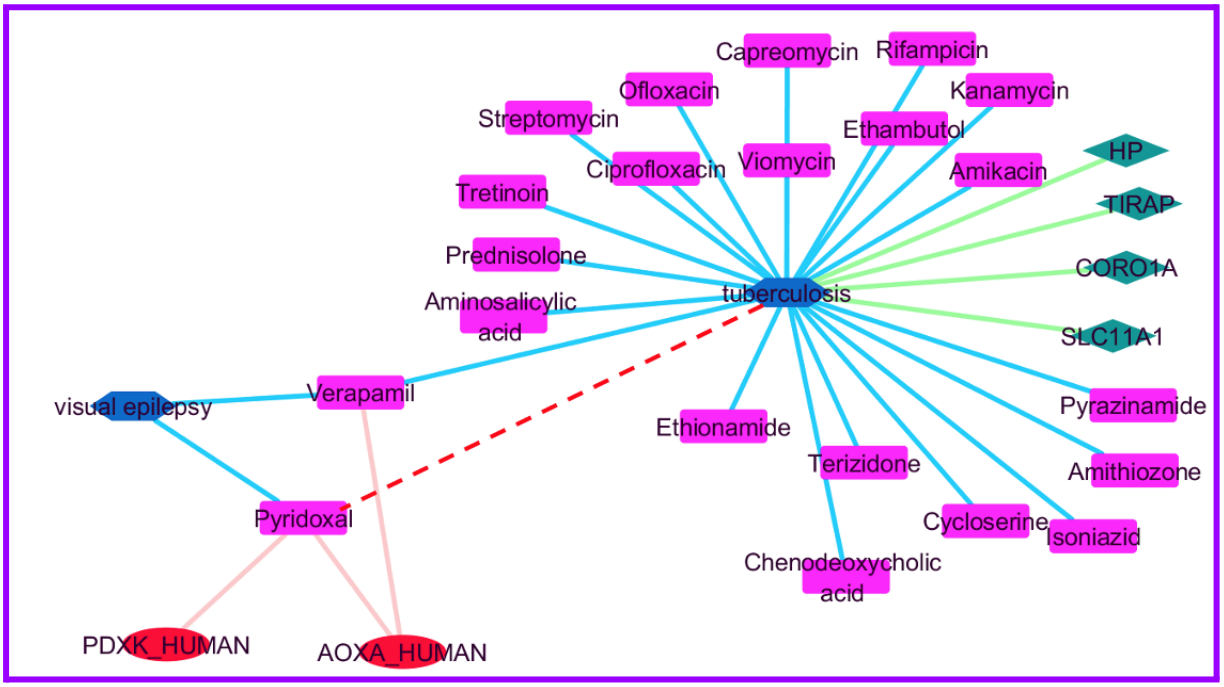


Figure 6.9. Predicted indication link between Pyridoxal and tuberculosis. Pink = drugs, red = proteins, blue = disorders, green = genes. Dashed line shows the predicted edge. Edge types: blue = indication, pink = target, green = gene-disorder. The prediction path: Verapamil treats tuberculosis and targets AOXA; Pyridoxal also targets AOXA.

Rituximab for Kidney Failure

The model predicted that Rituximab may be a potential treatment for kidney failure. This prediction is supported by 25 clinical studies. Kidney failure refers to a medical condition where the kidneys are not able to filter the blood from the waste. Guo and co-workers conducted a study to evaluate the efficiency and the safety of rituximab in treating membranous nephropathy patients with kidney insufficiency [342]. The study included 35 patients treated with rituximab. Patients were monitored at intervals of one to three months for six months and the clinical data was recorded. The results showed that 20% of the patients demonstrated full or partial response six months after receiving rituximab treatment. Significant improvements in anti-PLA2R antibody titer, proteinuria, serum albumin, and glomerular filtration rate, were observed. Responders maintained a stable kidney function throughout the study period. The study suggested that rituximab might be considered as a substitute medication for decreasing proteinuria and maintaining renal function in membranous nephropathy patients, even those with kidney deficiency. Figure 6.10 shows the subgraph illustrating the predicted link between Rituximab and kidney failure.

A study found that patients with end-stage renal disease undergoing hemodialysis exhibit a decreased inducibility of interferon-gamma mRNA (IFNG) [343]. Additionally, several studies have shown the roles of the IFNG gene in rheumatoid arthritis [344]. Also, studies found that Rituximab is an effective agent in the treatment of rheumatoid arthritis [345]. Therefore, it can be concluded that Rituximab may be a potential treatment for kidney failure since kidney failure shares the same gene with rheumatoid arthritis.

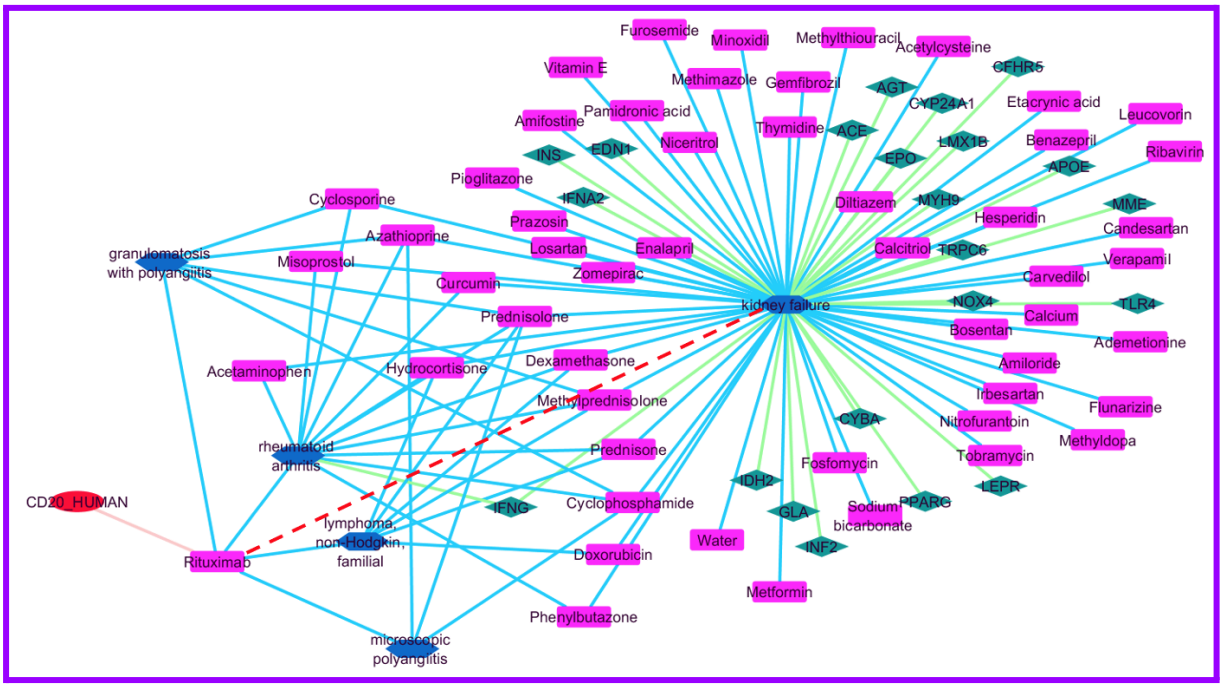


Figure 6.10 Predicted association between Rituximab and kidney failure. Pink = drug, blue = disorders, red = proteins, green = genes. Edge types: blue = indication, pink = target, green = gene-disorder. The predicted link is based on: IFNG associated with both kidney failure and rheumatoid arthritis, and Rituximab indicated for rheumatoid arthritis, suggesting Rituximab may also be indicated for kidney failure.

Irinotecan for liver cancer

The model proposed that the drug Irinotecan may be an effective treatment for liver cancer, and this prediction is reinforced by evidence from 114 clinical trials. Liver cancer, also known as hepatic cancer, refers to the development of malignant tumours within the liver [346]. The liver, a vital organ located in the upper right side of the abdomen, plays an important role in various bodily functions, including metabolism, detoxification, and the production of proteins [347]. Major risk factors for liver cancer include chronic infection with hepatitis B or C viruses, non-alcoholic fatty liver disease, and certain genetic conditions [346]. Irinotecan is a chemotherapy medication used in the treatment of various types of cancer [348]. It belongs to a class of drugs known as topoisomerase inhibitors. Irinotecan is commonly employed in the treatment of colorectal cancer, and it may also be used for other types of cancer, such as lung and pancreatic cancer Irinotecan is commonly employed in the treatment of colorectal cancer, and it may also be used for other types of cancer, such as lung and pancreatic cancer [348]. The mechanism of action of irinotecan involves interference with the DNA replication process in rapidly dividing cells, ultimately leading to cell death. It inhibits an enzyme called topoisomerase I, which is involved in the relaxation of DNA during replication. Additionally, Irinotecan targets Top1_Human, a protein encoded by the Top1

gene, which is known to be involved in liver cancer [349]. Figure 6.11 provides a visual representation of the subgraph highlighting the connection between liver cancer and its associated genes including Top1 as well as Irinotecan and its targets. Top1_Human Protein, also known as DNA topoisomerase 1, is an enzyme that controls and alters the topological states of DNA during transcription. Top1_Human Protein catalyses the transient breaking and rejoining of a single strand of DNA, allowing for the necessary unwinding during processes like replication. The Top1_Human protein is encoded by the TOP1 gene. Liu et al. provided evidence supporting the association of TOP1 with liver cancer [349]. The research utilised immunohistochemistry staining and data mining from microarray and demonstrated significantly higher expression of TOP1 at the protein and mRNA levels in liver cancer tissues compared to control non-tumor tissues. The specific target of the Top1_Human protein is the TOP1-DNA cleavage complex which is a specific target of TOP1 inhibitors, such as irinotecan [350]. Irinotecan is an inhibitor which disrupts the catalytic activity of Top1, forming persistent DNA cleavage complexes and inducing DNA damage, ultimately impeding cancer cell proliferation. TOP1 has been identified as a target for cancer treatment [350].

Pozzo et al. conducted a study aimed at observing the effects of neoadjuvant therapy with irinotecan and 5-fluorouracil (5-FU)/folinic acid (FA) on the resection rate and survival of colorectal cancer patients with initially unresectable hepatic metastases [351]. Forty patients received neoadjuvant chemotherapy, and the objective response rate was 47.5%, with two complete responses and disease stabilisation in 27.5% of patients. Thirteen patients (32.5%) underwent potentially curative liver resection following chemotherapy. The treatment was well-tolerated, with typical adverse events associated with the chemotherapy agents used. The study suggested that neoadjuvant therapy with irinotecan combined with 5-FU/FA enabled a significant proportion of patients with initially unresectable liver metastases to undergo surgical resection.

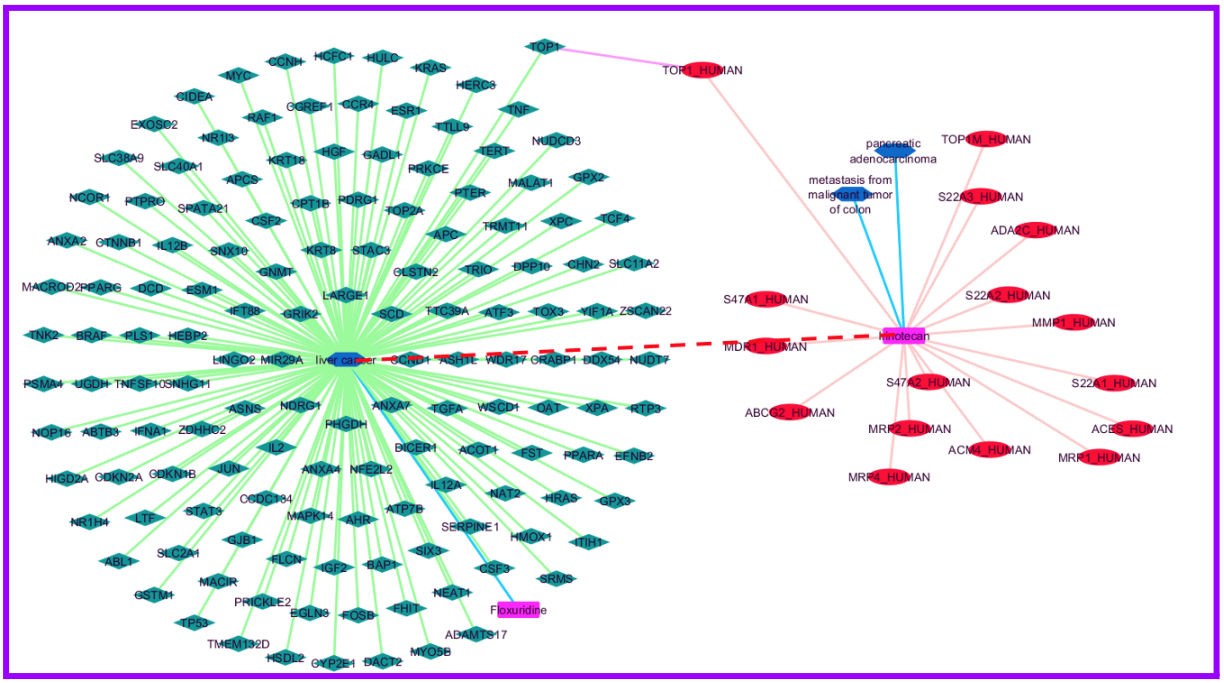


Figure 6.11 Predicted link between Irinotecan and liver cancer. Blue = disorder, green = gene, pink = drug, red = protein. The dashed red edge shows the predicted association. Irinotecan targets TOP_HUMAN, encoded by TOP1, which is involved in liver cancer, suggesting a potential indication. Edge types: blue = indication, pink = target, green = gene-disorder, purple = protein-gene.

6.4 Conclusion

The results presented here provide evidence that incorporating deep learning with a biomedical knowledge graph could be an effective method for helping to tackle the drug repurposing problem. Training GNNs on an extended knowledge graph incorporating node features and applying the model to link predictions could have the potential to speed up the drug repurposing process by reducing costs and time by highlighting potential repurposing opportunities that may not have been highlighted by other approaches. The main contribution of the work presented here was to develop GNN models for drug repurposing by training with a newly developed graph, NeDRex. Different views of this knowledge base were used; an extended graph containing an extended set of node types, and edge types as well as a rich set of features for nodes and edges. The analysis of some novel predictions generated by the Complete_NeDRex_With_NodeFeatures model indicated its ability to produce potentially valuable drug repurposing hypotheses, further supported by the results of testing with both the test graph and the RepoDB test. In addition, the set of novel predictions validated the potential of this method as these novel predictions have been reported by previous studies (Table 6.4). However, it is important to emphasise that the predictions should not be taken as

fact and clinical and experimental validation are required. The Complete_ NeDRex_With_NodeFeatures model can be used to propose hypotheses about potential drug repurposing indications that need further investigation. Despite the significant results of this research work, some limitations, and opportunities for future research, were presented. Firstly, the validation process utilised RepoDB but the need for Disease ID mapping between RepoDB and NedrexDB led to a restricted validation set comprising only 1250 drug-disease indications mapped from UMLS to MONDO. The smaller size of this dataset may have the potential to influence the effectiveness of the validation process. A more extensive mapping from UMLS to MONDO would be desirable. More sophisticated validation approaches such as integrating multiple gold standard datasets for drug repurposing and using them as an external source of evaluation, could be an interesting direction for future work. Secondly, as it is difficult to find negative drug-disease indications, random negative samples were generated [352]. This was done based on the closed-world concept, where edges not present in the training set are considered negative. However, it is important to note that there is no assurance that these edges may not be positive in the validation and test sets. New literature-scale representations of pharmacological knowledge may help to resolve this issue highlighting data conflicts and potentially providing negative relationships between concepts [353]. Thirdly, randomly dividing the graph into three subgraphs for training, testing, and validation might result in isolated subgraphs or edges. Nodes with numerous connections in the graph are likely to receive higher scores compared to isolated nodes. Consequently, the model exhibits bias towards nodes with greater connectivity. In the future, we aim to tackle this issue using a graph pruning technique by selectively removing certain elements, typically isolated nodes or edges with low connectivity. By eliminating isolated nodes or edges that contribute minimally to the overall connectivity and structure of the graph. This process helps mitigate biases that may arise during model training, especially when randomly dividing the graph into subsets for tasks such as training, testing, and validation.

An overall finding of this study is that adding more node types and edge types, assuming that all added data is of good quality, improves the GNN performance, however, some node and edge types may have more significant effects on the performance of the model than other and some of them may reduce the performance. In the future, we intend to tackle this issue by systematically analysing the impact on predicted capabilities of the model of different classes

of nodes and edges. This will involve experimenting with the removal and addition of specific types of nodes and edges to observe how these modifications impact the overall performance. In the context of Complet_NedRex_With_NodeFeatures, it is important to highlight that the enhancement primarily focuses on incorporating additional data of diverse types, rather than simply increasing quantity. While there is a common concern that adding more data could introduce noise into the knowledge graph, in this specific case, the augmentation is specifically tailored to include various types of valuable information. Importantly, the size of each association in the NedRex database is relatively modest when compared to existing databases for drug repurposing. For example, the *Drug_has_indication* edge comprises only 14,315 edges, significantly smaller than the numerous edges found in the literature. Similarly, the *Gene_associated_with_disorder* consists of 30,252 edges, a fraction of the often million-edged literature counterparts. This is attributed to NedRex's database integrating data from reliable primary resources, emphasising manual curation, as opposed to indiscriminately including all available data sources. Moreover, the intricate interconnectedness within biological systems necessitates a thorough examination of all connected components to obtain a holistic understanding. The research findings indicated that integrating node features could enhance the GNN outcomes. Moreover, node features have significantly improved the GNN performance in both versions of the graph. Graphs with node features performed better than graphs without node features. One limitation of this work is the missing of some important node features that hold significant meaning for some nodes. Consequently, these features have been excluded from the graph. For instance, the '*description*' feature for the '*drug*' node was missing for some nodes, therefore, this feature was excluded from the node matrix features. In the future, we plan to address this limitation by either augmenting the dataset by completing these important features or removing nodes lacking these essential features. Nevertheless, the question of determining which node features to include remained unresolved. A potential future direction involves conducting a systematic analysis of feature selection to identify the most important features influencing GNN performance. An intriguing avenue for future research involves enhancing GNN performance through the incorporation of edge-level features. The NeDRex database offers an extensive list of edge features, as illustrated in Table 6.2. Leveraging these edge features, which encompass valuable information about graphs, could have the potential to significantly improve performance.

The work is limited too in the relationships it can predict, and its use cases, since the model is just trained to predict drug-disease links. The rich heterogeneous graph allows holding many relationship types in the same data structure. We could make the model take advantage of all the available information and predict any edge type existing in NeDRexDB, see edge types in Table 6.2. For example, the model could be trained to predict disease-gene associations, drug-target interaction or any type of edges available in NeDRexDB. However, for this work, we gave more importance to the specialisation of drug repurposing applications.

Including PFIN in a GNN has the potential to enhance performance, as PFIN is known for reducing noise during integration. This improvement has been demonstrated in applications such as protein function prediction. In the current study, a DGA PFIN was constructed with the specific goal of developing computational approaches for drug repurposing, with a focus on disease-gene associations.

In the future, several other improvements are also possible, such as a much wider exploration of the hyperparameter space. Among the other future research directions, the model's behaviour under different circumstances. Testing different types of encoders, such as the use of Graph attention network (GAT), other types of decoders, and other types of loss functions could provide potential improvements.

Chapter 7

Conclusions and Future Work

7.1 Introduction

Drug repurposing provides a cost-effective and time-saving strategy for developing new treatments [229], [230]. Computational approaches facilitate this process by rapidly generating hypotheses [8]. Among these approaches, network integration is particularly notable in the most recent *in silico* drug repurposing [12], [27], [45]. Network integration combines diverse data sources into a unified network, yielding new insights into potential drug repurposing opportunities. This method is essential for examining the relationships among various components of biological systems, such as proteins, genes, diseases, and drugs [12], [48], [64]. A thorough understanding of these interconnected components aids in comprehending complex biological systems and producing precise results [48], [64].

DGA networks serve as one of the foundational elements of these integrated networks and are extensively used in drug repurposing applications. However, current DGA networks often contain significant noise due to false results in biological data [163], [164]. PFINs have demonstrated efficacy in reducing noise within PPI networks [57], though their application to DGA networks remains limited. This project examined the use of the PFIN approach to enhance the accuracy and reliability of DGA networks.

Accurate networks are important for computational methods such as deep learning, which has emerged as a leading technique in computational drug repurposing. Graph Neural Networks (GNNs), in particular, have gained considerable attention. GNNs operate on integrated heterogeneous networks, with their performance highly dependent on the quality of these networks. Therefore, the accuracy and reliability of the included data are critical.

This research investigated and developed computational approaches to drug repurposing by integrating PFINs with GNNs, focusing on DGA data. First, PFIN approaches to DGA data were examined and refined. Second, GNN models were developed to predict links between drugs and diseases using heterogeneous integrated networks that incorporate DGA networks. The efficacy of deep learning models, such as GNNs, relies heavily on the quality of the

training data. Integrating PFIN approaches can reduce noise in these networks, thus enhancing the performance of GNNs.

In the concluding chapter, the key findings and contributions of each research chapter were reviewed, contextualising them within the study's original objectives. The implications of these conclusions for current research practices were discussed, address the limitations of this work and their impact on the findings, and propose potential directions for future research.

7.2 What Has Been Achieved?

The contribution of the work presented can be categorised into two main parts:

7.2.1 Part One: Investigation of Novel Approaches to Apply the PFIN Approach to DGA Data

Identifying DGAs is a key approach in the field of drug repurposing to help understand disease mechanisms. Computational approaches to identifying DGAs that rely on a single data type have various limitations (Section 2.3). By adopting a comprehensive approach and considering all available evidence, more reliable inferences can be drawn. Although methods like DisGeNET score and integrate DGAs, their scoring can be arbitrary. For instance, DisGeNET scores DGAs based on the evidence supporting them, using a cumulative metric that includes manually curated sources, model organism databases, and text-mined associations (Section 2.3). This approach does not account for duplicated evidence across the three evidence types which introduces bias to the confidence scores. Existing methods for integrating and scoring DGAs are based on heuristic methods [108], [141], [160]. However, these heuristic methods often fail to account for duplicate data [141], [159] . Duplicate data can significantly inflate confidence scores, compromising the accuracy of the predictions. For instance, the confidence scores developed by DisGeNET were upweighted due to the duplicate evidence across curated, animal model, inferred, and literature sources (Section 4.3.5). However, after removing these duplicates, the DisGeNET's confidence scores were diminished, resulting in more accurate confidence scores (Section 4.3.5). The correlation between DisGeNET's DGA scores with and without duplicate data illustrated the impact, as scores with duplicates consistently appeared higher than those without (Figure 4.28). This

consistent upward shift demonstrated the influence of duplicated data on the confidence scores.

The PFIN approach detailed in Chapters 4 and 5 offers a more unbiased and comprehensive view of DGAs. Unlike DisGeNET, the probabilistic approach in this work derives associations from separate, independent sources of evidence, reducing the bias caused by duplicated associations. By utilising diverse and independent evidence sources, PFIN provides a more robust and thorough evaluation of DGAs, ensuring that duplicated associations do not unduly influence the scores. This method could enhance prediction reliability, making it an essential tool in drug repurposing. In the approach presented in this work, for example, the data-source-based approach began by removing duplicate evidence among the curated data sources as the initial step in generating DGA confidence scores. This step is important, as PFIN integrates DGAs from distinct and independent sources (Section 4.3.1.1). Unlike existing heuristic methods, which generate DGA confidence scores susceptible to bias from duplicate evidence, the approach in this work is more robust due to duplicate removal. It was found that duplicate evidence significantly impacts both LLS scores and weighted sums, which, in turn, could influence PFIN performance (see Figure 4.15).

The PFIN approach has been widely applied in PPI networks to generate PPI confidence scores [56], [57], [354]. In PPI networks, there are many available sources of individual datasets and gold standards [87], [183], [185], [186]. However, in the case of DGA networks, heuristic methods have been widely used to generate confidence scores [141], [160], while the use of probabilistic methods has remained limited. For DGA networks, defining individual datasets and determining a reliable gold standard is less straightforward and can depend on the network's intended use. A network constructed to study a specific group of diseases may have different considerations and validation criteria than one aimed at drawing broader, more global inferences across diseases. In Chapter 4, novel strategies were described to investigate the identification of gold standard data and individual datasets to build DGA PFINs. Two main approaches were developed to define the gold standard for DGA data and the individual datasets representing separate evidence of DGAs: the data source-based approach and the individual study-based approach. The results indicated that the individual study-based approach outperformed the data source-based approach in network analysis techniques. However, treating individual experimental studies for DGA as individual datasets

resulted in a high rate of data loss. This is because DGA experimental studies often focus on specific diseases or related groups of diseases, which may not be covered in the gold standard data. Consequently, there is a lack of overlap between the datasets and the gold standard, leading to data loss. A key question was how to define individual datasets for scoring?. In PPI networks, treating individual experimental studies as distinct datasets yielded sensible results and led to the exclusion of only a few datasets [50], [57], [90], [192]. However, in DGA data, this approach may be suboptimal because most DGA studies focus on a single disease. This bias results in more datasets being excluded if their diseases are not in the gold standard. Creating a gold standard for DGAs that covers every possible disease is currently not feasible, highlighting the need for an approach that reduces data loss. Evaluations of other gold standard data were conducted in chapter 5. A potential approach was to use gold standard data of a different type, such as shared pathway or high confidence PPIs. In this case datasets were scored on how well they reflect known DGAs.

In Chapter 5, novel approaches were described to address the limitations presented in Chapter 4: limitations in gold standard identification, individual dataset identification, and network evaluation techniques. To solve the limitation in identifying gold standard data, a new type of gold standard was introduced using non-DGA gold standard data, including PPI gold standard data and pathway interaction gold standards. It was found that using non-DGA gold standards resulted in high data loss due to two main reasons: studies with single genes could not be scored against non-DGA gold standards, and scoring DGAs against physical PPIs or shared pathways led to loss because two genes may be involved in the same disease but not physically interact or be in the same pathway. PPI gold standards, such as BioGRID, have been used to score PPIs for building PFINs, resulting in minimal data loss [50], [57]. However, in the work presented in Chapter 5, a high rate of data loss was observed due to the nature of DGAs compared to PPIs. Existing approaches typically score PPI datasets against PPI gold standards, whereas this work scored genes associated with the same diseases against both PPIs pathways. This method showed that two genes may be implicated in the same disease without sharing common biological connections; they might not share the same pathways or engage in physical interactions. Despite the high data loss, this approach could be particularly valuable for understanding the underlying biology of diseases.

To address the limitations in identifying individual datasets, where using a single experimental study as an individual dataset led to high data loss, a novel text mining approach was introduced. This approach grouped multiple experimental studies based on their experimental techniques, creating more diverse individual datasets and reducing both low (which led to data loss) and high overlap (which led to infinity scores). While several text mining approaches have been developed to extract DGAs , they have not focused on the experimental techniques that generated these associations. Understanding the experimental techniques employed to generate DGAs is valuable information that can help define the evidence supporting these associations. The novel text mining approach to extract experimental DGA techniques was developed in Chapter 5 and has not been applied to extract DGA experimental techniques from the biomedical literature. However, this approach has limitations, primarily in dictionary construction. The performance of the text mining approach is highly dependent on the accuracy of the dictionary used to mine the DGA literature. Building an accurate and noise-free dictionary is important. Future work could focus on manually cleaning the dictionary or creating a curated dictionary for DGA experimental techniques. In the work reported in chapter 5, the experimental techniques in EFO and EDAM ontologies were used to build a dictionary, which was then applied to extract the techniques by matching terms in the dictionary with biomedical literature. However, a more robust technique that could be explored in the future is a machine learning approach using a Named Entity Recognition model to extract experimental techniques. Training the model on annotated text, which includes experimental techniques, is currently a work in progress and has not been completed yet. NER models have previously been trained to extract DGAs from biomedical literature [108] but have not been applied to extract experimental techniques associated with DGAs.

The text mining approach clustered DGA experimental studies based on their common experimental techniques. To evaluate the quality of these clusters, existing techniques, such as the Silhouette score, were used, indicating that studies grouped in the same cluster are more related in terms of experimental techniques. Future work could involve systematic analysis to score these clusters against gold standards, build PFINs by integrating these datasets based on their scores, and evaluate the resulting networks using network evaluation techniques.

To address limitations in network analysis techniques (Section 4.3.3.1), a projection approach was introduced, which involved collapsing DGA networks to GGA and DDA networks, as analysis techniques for unipartite networks are more available and accurate. This approach involved collapsing scored datasets and integrating them based on their scores, then using unipartite network evaluation techniques. However, some limitations remain, such as datasets that cannot be collapsed to GGA networks if they contain only single genes or to DDA networks if they contain only single diseases. Existing methods for bipartite network evaluation, including link prediction, often adapt unipartite link prediction techniques for weighted bipartite networks [270], [271]. These methods, however, have their own limitations (Section 4.3.3.1). Some existing methods for DGA network analysis often involve projecting the DGA network into unweighted unipartite networks, such as GGA (gene-gene association) or DDA (disease-disease association) networks [63], [260]. Once projected, these unipartite networks are clustered, and the resulting clusters are then evaluated for their biological properties [135], [260]. However, the use of unweighted networks means that edge confidence levels are not incorporated, which may limit the precision and relevance of the identified clusters. In this work, the weighted DGA PFINs were projected into GGA and DDA networks, and clusters were formed based on edge weights to create gene and disease clusters. Additionally, edge weights were used to threshold the network, filtering out low-confidence edges before clustering to enhance cluster quality. This approach enhances the biological relevance of the clusters and provides a more nuanced view of gene and disease associations.

In summary, this research presented a comprehensive approach to improving the reliability of DGA identification in drug repurposing. By addressing the limitations of current methods used in DGA networks and introducing novel strategies to generate DGA confidence scores, this work enhances the robustness and accuracy of computational predictions, paving the way for more effective drug repurposing efforts.

7.2.2 Part two: Investigation of Novel Deep Learning Approaches to Drug Repurposing based on DGA Data Using GNN Models

In Chapter 6, a set of hypotheses was stated and tested to improve GNNs performance. First, the investigation focused on whether training GNNs on extended heterogeneous biomedical knowledge graphs, which include multiple and diverse types of nodes and edges, could enhance GNN performance. Two versions of the graphs were implemented, one with reduced types of nodes and edges and the second with extended types of nodes and edges to see the impact of the additional types of edges and nodes on the performance of the GNNs. Results demonstrated that incorporating a greater variety of node and edge types, given that the data is of high quality, enhances GNN performance. However, the influence of different node and edge types on the model's performance can vary, with some potentially reducing effectiveness. In the future, the plan to address this issue is by systematically analysing the impact of different classes of nodes and edges on the model's predictive capabilities. This will involve experimenting with the addition and removal of specific node and edge types to observe how these changes affect overall performance. Existing GNN-based methods operate within constrained knowledge graphs, characterised by limited nodes and edge types [67], [68], [69]. This limited scope can restrict the model's ability to capture the full complexity of drug-disease relationships. These approaches aggregate information from directly connected nodes restricted in a drug-protein-gene-disease, ignoring the other types of biological entities such as pathways, tissues, side effects, phenotypes, that contain rich information about graphs in drug repurposing applications. The results showed that including other relevant biomedical entities to the Knowledge graph, like the NeDRex graph, could greatly contribute to enrich the graph structure and semantics which could improve the GNN prediction.

Secondly, the study examined whether incorporating node features within GNNs could lead to performance improvements. The research demonstrated that adding node features could improve GNN performance. Graphs that included node features performed significantly better than those without them. However, one limitation was the absence of some critical node features, which led to their exclusion from the graph. In future work, the plan is to overcome this limitation by either supplementing the dataset to include these essential features or by removing nodes that lack them. Deciding which node features to include is still an open question. A possible future direction is to systematically analyse feature selection to

pinpoint the most important features for GNN performance. Additionally, enhancing GNN performance by incorporating edge-level features presents an interesting research avenue. The NedRex database provides a rich set of edge features. Utilising these edge features could significantly boost performance. Previous studies have lacked the incorporation of node features in GNN models due to the absence of predefined features for nodes. For instance, in work focused on training a GNN model using a heterogenous biomedical knowledge graph [64], the researchers noted one limitation: they had not included node features because these features were unavailable. They recommended that future work incorporate node features to improve model performance.

Third, the research investigated whether incorporating PFIN approaches within GNNs could improve performance, given that PFINs typically reduce noise in integrated networks. However, it was found that this approach resulted in lower performance compared to graphs without filtering DGAs using the PFIN approach. This outcome was attributed to a specific limitation: the necessity for disorder identifier mapping between the NeDRex graph, which used Mondo identifiers for disorders, and the PFIN network, which used UMLS identifiers. Due to this mapping process, a significant number of DGAs in the PFIN were lost. To address the limitation of identifier mapping in future work, several strategies can be employed. Firstly, developing or adopting more advanced methods for mapping disorder identifiers between databases, such as Mondo and UMLS, could improve accuracy. This could involve creating a comprehensive mapping database or utilising machine learning techniques. Secondly, augmenting the datasets by filling in missing identifiers through cross-referencing additional databases, using text mining techniques to extract identifiers from scientific literature, or manually curating the data to ensure completeness could help. Thirdly, implementing hybrid approaches that combine multiple mapping strategies might reduce data loss.

Although existing GNN approaches for computational drug repurposing have achieved remarkable performance, they often prioritise handling missing data while assuming existing data is of high quality. This assumption can limit the effectiveness of these models, as their success relies heavily on the quality and reliability of the data used. Addressing data quality issues to the same extent as data completeness may therefore enhance the robustness and applicability of these methods. Despite this challenge, PFINs generally contributed to noise

reduction during integration and could potentially lead to overall performance improvement. Including PFIN in a GNN has the potential to enhance performance, as PFIN is known for reducing noise during integration. This improvement has been demonstrated in applications such as protein function prediction. In the current study, a DGA PFIN was constructed with the specific goal of developing computational approaches for drug repurposing, focusing on DGAs. As a future avenue of research, other types of PFINs, such as PPI PFINs and DTI PFINs, could be explored to further expand the scope and capabilities of the approach.

Many previous studies have focused on specific diseases. For example, research targeting COVID-19 constructed graphs specifically populated with COVID-19-related information [25], [68], [311]. In contrast, the NeDRex biomedical knowledge graph includes a variety of diseases, enabling a more heterogeneous approach. Training the GNN model on a diverse biomedical knowledge graph like NeDRex facilitated predictions across a broad range of drugs and diseases, beyond just COVID-19 (See novel predictions in Section 6.3.7). This approach broadens the scope of drug repurposing applications, potentially extending their relevance to a wider array of diseases.

The model with the best performance was used to predict links between drugs and disorders, and these predictions were validated through literature and previous studies. However, these predictions require further validation in vitro and in vivo and cannot be considered completely accurate at this stage. The ultimate aim of any drug repositioning project is to advance promising hits to clinical trials to benefit patients. Drug repositioning is more complex than often imagined, with many projects halting at the in vitro stage. Validation of candidate hits for preclinical drug evaluation necessitates in vitro and in vivo models. Additionally, the selection of appropriate hits for validation is critical, as factors such as high toxicity, cost, and low bioavailability can influence the choice of drugs. Clinical trial costs are also a significant consideration, especially for off-patent drugs. Identifying potential collaborations is essential for advancing promising hits.

Dosing is a particularly challenging aspect of drug repositioning, as the therapeutic threshold and half-life of drugs vary widely. Drugs are approved at specific dosage strengths, and any deviation requires substantial development costs. For instance, sildenafil was reformulated at different dosages for its repositioning from treating erectile dysfunction to pulmonary arterial hypertension. Caution is needed when in vitro concentrations exceed clinically observed

levels. Therefore, dose levels must be carefully considered before advancing repositioning opportunities. Pharmacological modelling can predict dose ranges, which is important for ranking lead optimization and designing early clinical trials.

7.3 Conclusion

The approaches researched and developed in this work addressed several limitations in existing methods for drug repurposing. One of the most common strategies in computational drug repurposing is network integration, which combines diverse heterogeneous biomedical data into a single network. DGA networks are considered one of the most frequent and fundamental subnetworks within these large, integrated heterogeneous networks. Despite their importance in understanding disease mechanisms and facilitating drug repurposing, DGA networks have notable limitations. Existing DGA networks often contain high levels of noise due to false results, particularly from high-throughput data. Additionally, these networks are typically either unweighted or weighted with unreliable confidence scores.

To address these issues, an accurate integrated network was constructed using the PFIN approach. This method reduces noise in DGA networks, enhancing their biological accuracy. Building DGA PFINs requires two main components: a high-confidence gold standard for DGAs and individual datasets representing DGAs. The work presented focused on developing strategies to identify high-confidence gold standards and compile reliable individual datasets for constructing accurate DGA PFINs.

The resultant accurate DGA PFINs can be integrated with other biomedical networks and mined for drug repurposing opportunities. Many computational approaches have been developed to mine networks and infer links between drugs and diseases. The quality of network mining results heavily depends on the quality of these networks. GNNs are commonly used to mine relationships between drugs and related entities such as diseases and targets. These deep learning models work on integrated networks. However, existing GNNs suffer from two main issues: low data quality or missing data. They often operate on noisy networks with high false result rates or on limited heterogeneous integrated biological networks that exclude important biological entities and relationships.

To address these gaps, GNN models were applied to extended heterogeneous integrated networks with various types of nodes and edges, reducing noise by incorporating the PFIN

approach. These enhanced GNN models are powerful tools for predicting links between drugs and diseases.

In conclusion, while this research has contributed towards addressing several limitations of current *in silico* drug repurposing methods by focusing on improving the data quality of DGA networks, it also raises new research questions as discussed through this section. Addressing these questions will lead to more accurate and comprehensive results. The contributions of this work have mitigated some of the limitations in existing approaches, particularly concerning the data quality of DGA networks, which are the foundation of most biological networks used in drug repurposing. By improving the quality of biological data, this work promises to help enhance the effectiveness of computational methods in drug repurposing.

Appendix A
Investigating the Applicability of Probabilistic Functional Integrated Networks to Disease-Gene Networks

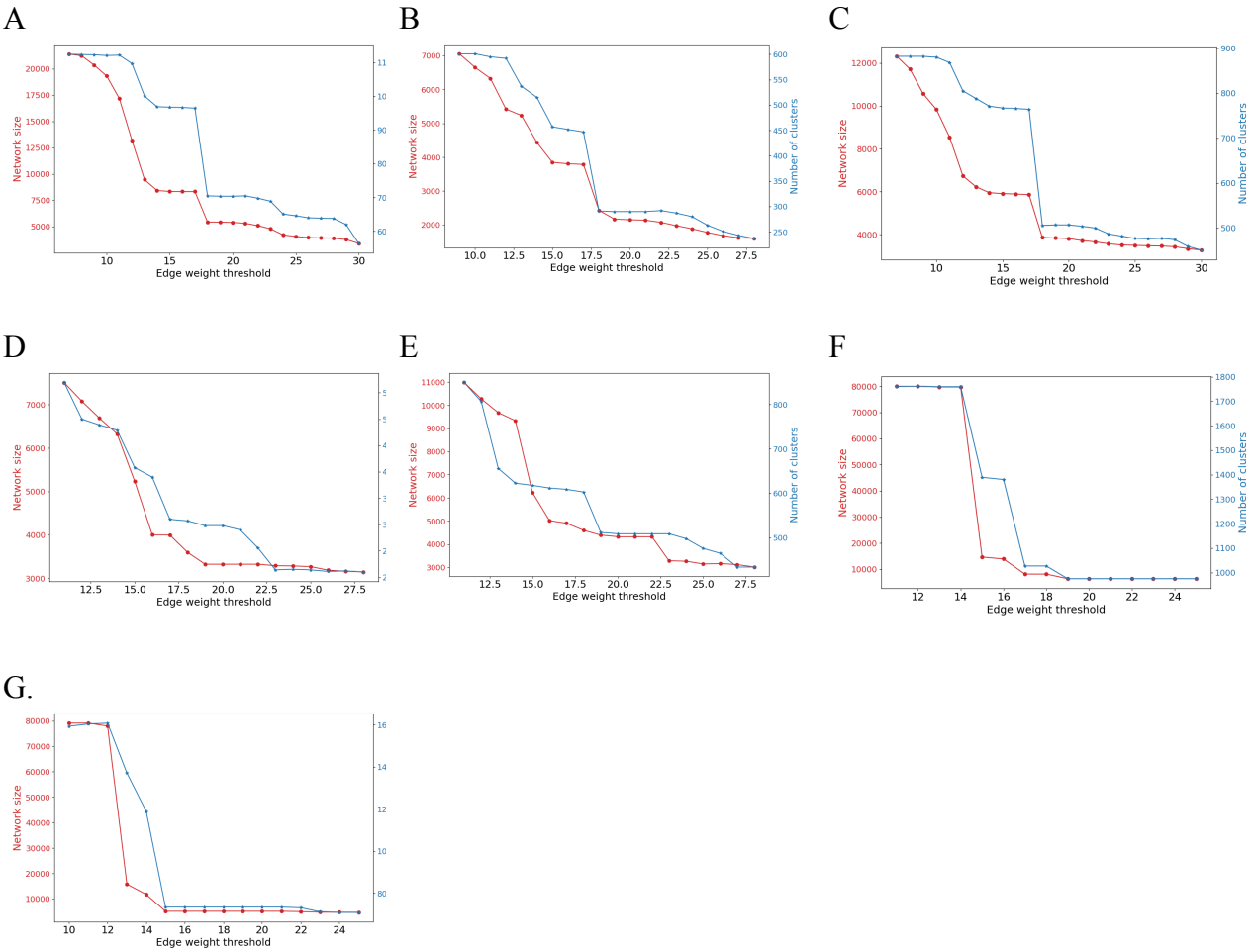


Figure A.1. Network size and cluster count at selected edge weight thresholds, chosen based on weight distribution and highest average cluster connectedness. Thresholds: 12 for SSA_SMA, HEL_LEL, MCS_SCS; 13 for MG_IES, OMIM_IES, UniProt_CDS; 15 for OMIM_CDS.

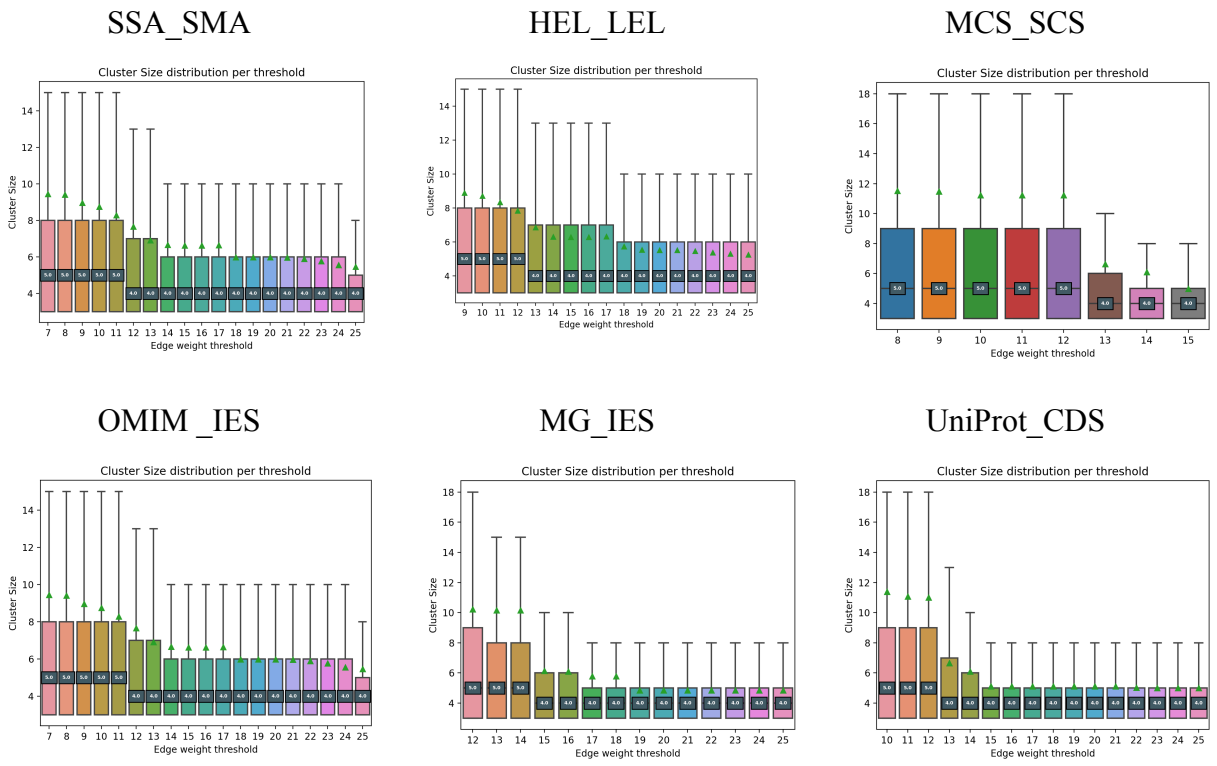


Figure A.2. The cluster size distribution across different edge weight thresholds. A corresponds to SSA_SMA, B to HEL_LEL, C to MCS_SCS, D to MG_IES, E to OMIM_IES, F to UniProt_CDS, and G to OMIM_CDS

Appendix B

Constructing Disease-Gene Association PFINs with Gene-Gene Association Gold Standards

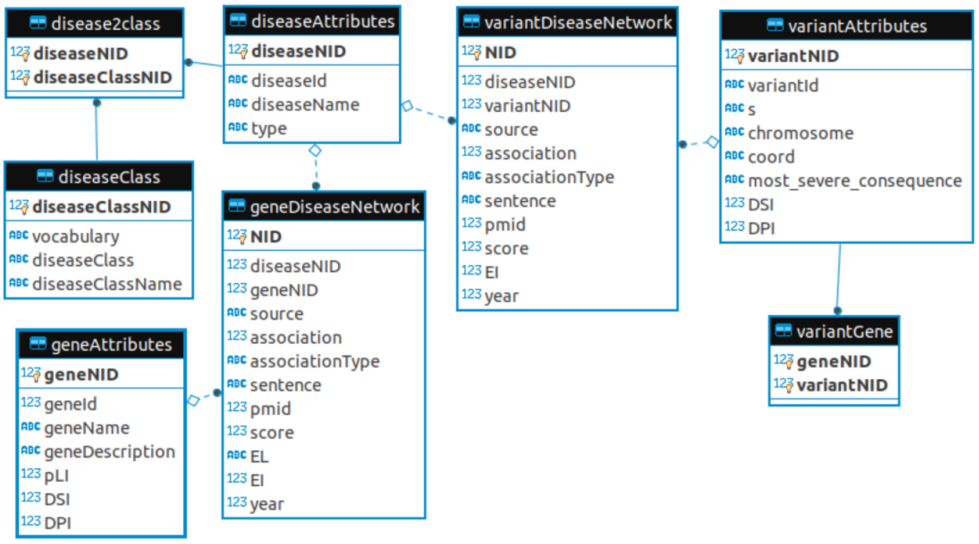


Figure B.1. SQL schema of the DisGeNET sqlite database, from https://www.disgenet.org/app.

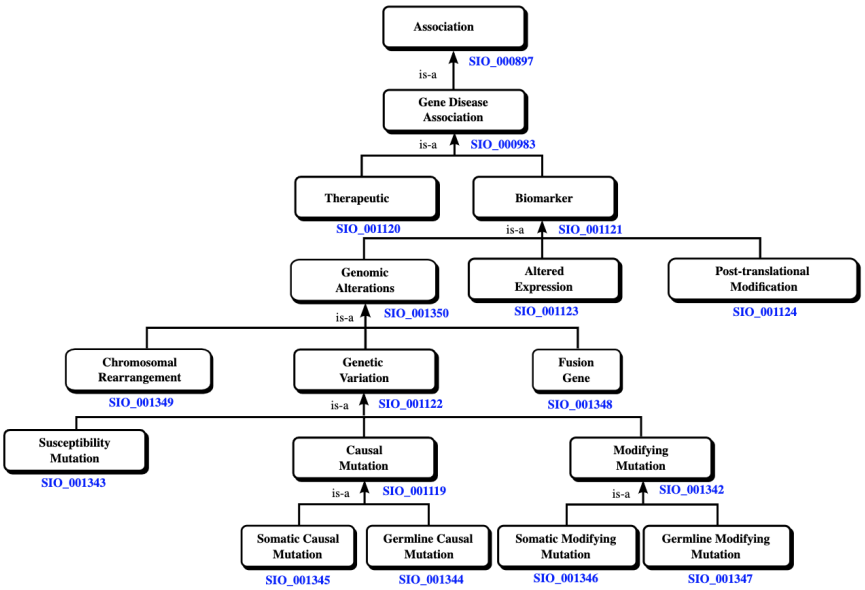
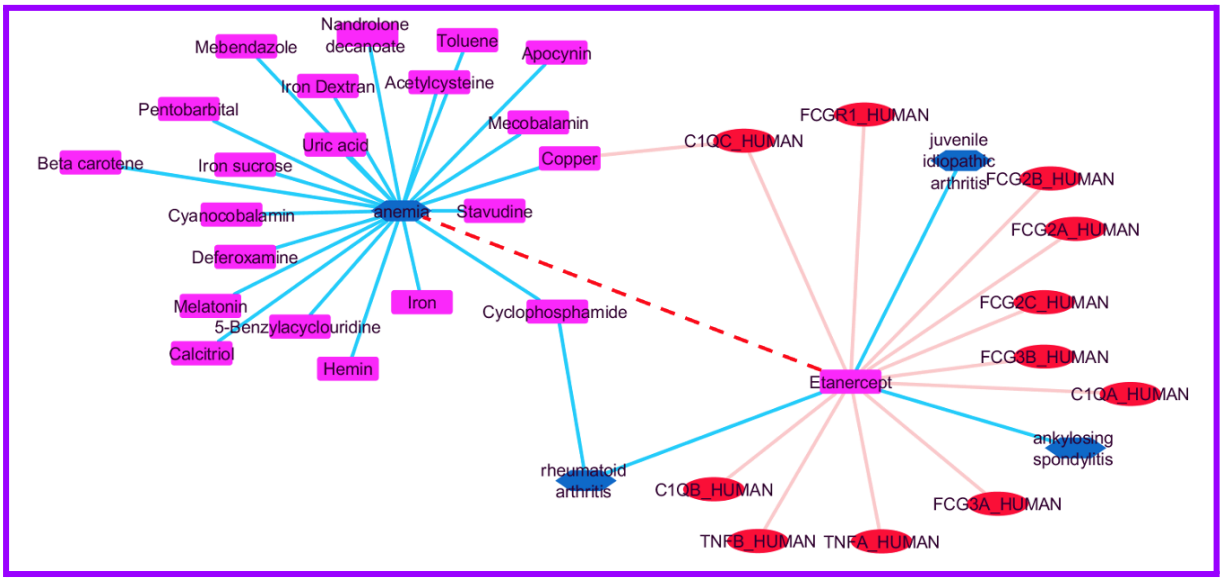
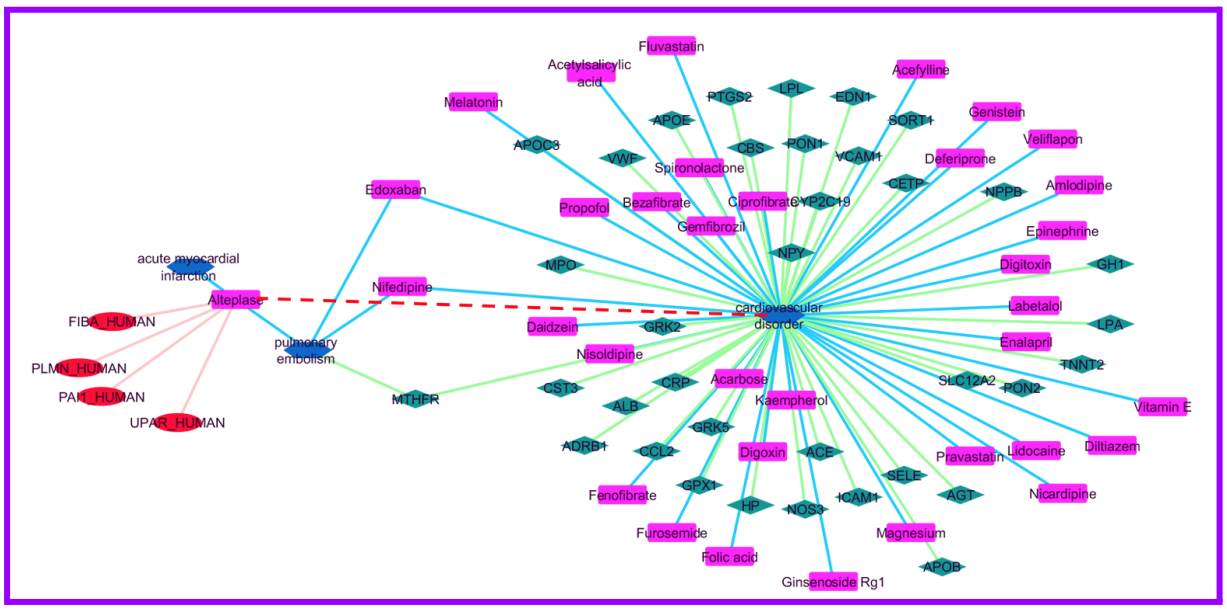


Figure B.2. DisGeNET association type ontology, from, (https://www.disgenet.org/dbinfo)

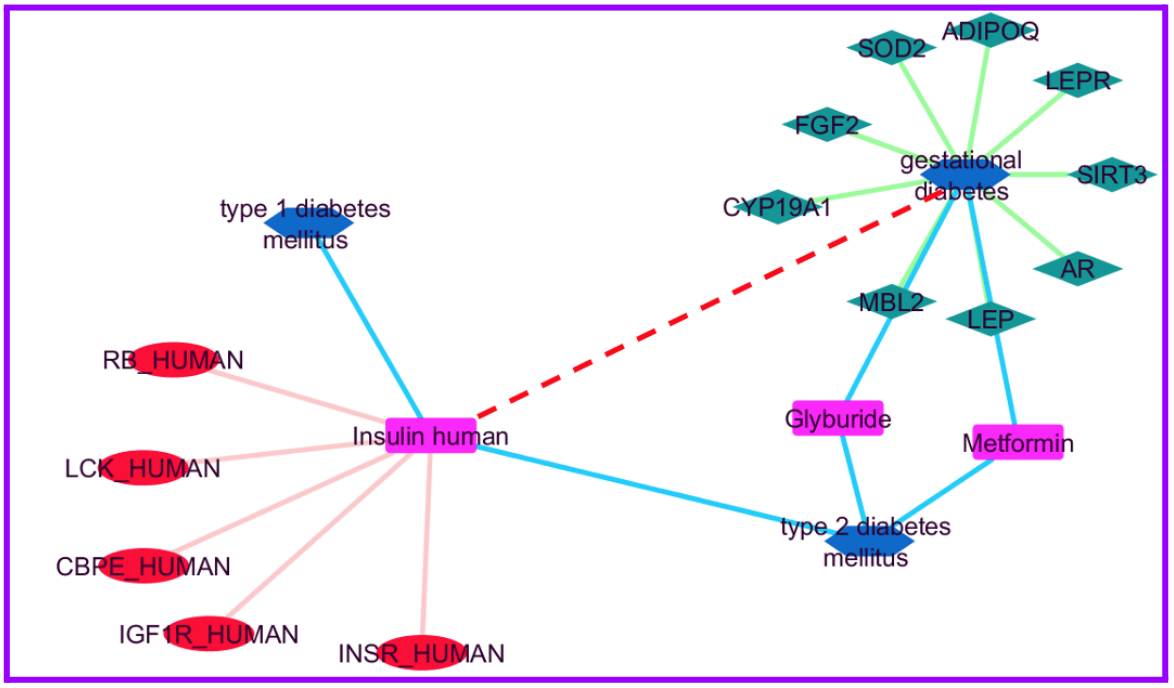
Appendix C

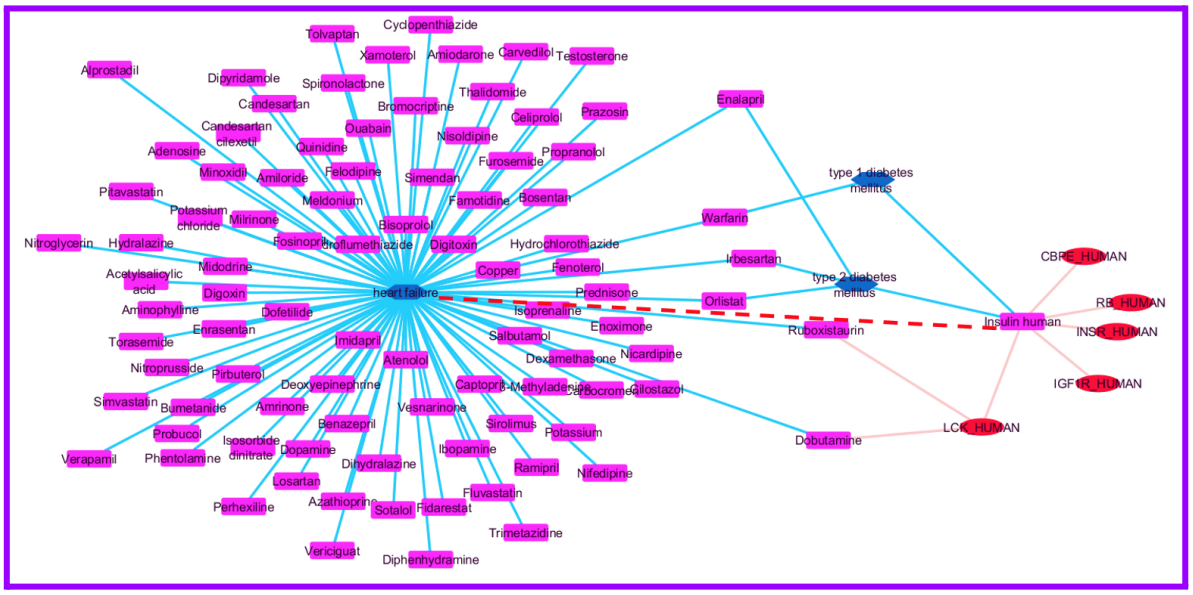
A computational Approach to Drug Repurposing Incorporating Graph Neural Networks and Probabilistic Functional Integrated Networks focusing on Disease-gene Association Data



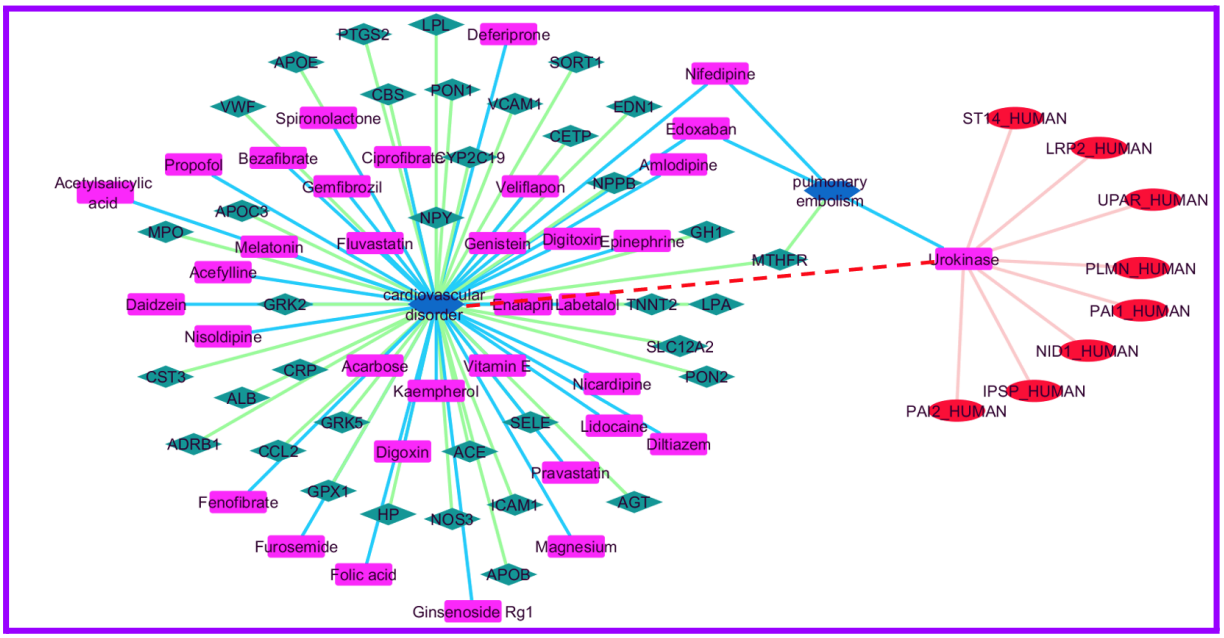


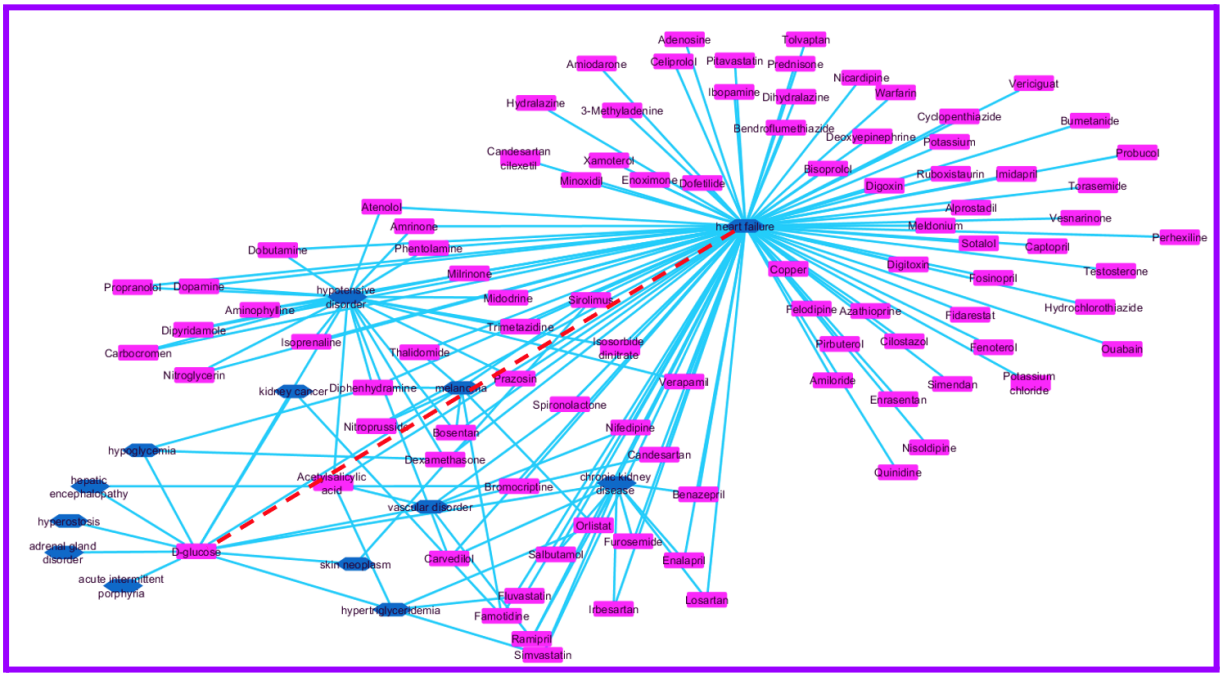
Appendix C


ADIPOQ
SOD2
LEPR
FGF2
gestational diabetes
SIRT3
CYP19A1
AR
MBL2
LEP
type 1 diabetes mellitus
RB_HUMAN
Insulin human
Glyburide
Metformin
LCK_HUMAN
type 2 diabetes mellitus
CBRE_HUMAN
IGF1R_HUMAN
INSR_HUMAN


Cyclopenthiazide
Tolvaptan
Xamoterol
Amiodarone
Carvedilol
Testosterone
Alprostadil
Dipyridamole
Spironolactone
Thalidomide
Candesartan
Bromocriptine
Celiprolol
Prazosin
Enalapril
Candesartan cilexetil
Ouabain
Nisoldipine
Adenosine
Quinidine
Propranolol
Minoxidil
Amiloride
Felodipine
Furosemide
Simendan
type 1 diabetes mellitus
Pitavastatin
Meldonium
Famotidine
Bosentan
Potassium chloride
Milrinone
Warfarin
Nitroglycerin
Hydralazine
Bisoprolol
Fosinopril
Hydroflumethiazide
Digitoxin
Hydrochlorothiazide
Irbesartan
CBRE_HUMAN
Acetylsalicylic acid
Midodrine
Copper
Fenoterol
type 2 diabetes mellitus
Digoxin
heart failure
Prednisone
Orlistat
RB_HUMAN
Aminophylline
Dofetilide
Isoprenaline
Insulin human
Enoximone
Ruboxistaurin
INSR_HUMAN
Torasemide
Enrasentan
Imidapril
Salbutamol
Dexamethasone
Nicardipine
Atenolol
Nitroprusside
Pirbuterol
Deoxyepinephrine
Captopril
Methyladenine
Carbocromen
Allostazol
IGF1R_HUMAN
Simvastatin
Bumetanide
Amrinone
Vesnarinone
Sirolimus
Potassium
LCK_HUMAN
Probucol
Benazepril
Dobutamine
Isosorbide dinitrate
Isopamine
Verapamil
Phentolamine
Dopamine
Dihydralazine
Ramipril
Nifedipine
Losartan
Fluvastatin
Perhexiline
Azathioprine
Sotalol
Fidarestat
Trimetazidine
Vericiguat
Diphenhydramine

Appendix C


PTGS2
LPL
Deferiprone
APOE
SORT1
Nifedipine
CBS
PON1
VWF
Spironolactone
VCAM1
EDN1
ST14_HUMAN
CETP
Edoxaban
LRP2_HUMAN
Propofol
Bezafibrate
Ciprofibrate
CYP2C19
Amlodipine
Acetylsalicylic acid
Gemfibrozil
Veliflapon
NPPB
pulmonary embolism
UPAR_HUMAN
APOC3
NPY
MPO
Fluvastatin
Genistein
Digitoxin
Epinephrine
GH1
Melatonin
MTHFR
Urokinase
PLMN_HUMAN
Acefylline
Daidzein
GRK2
cardiovascular disorder
Enalapril
Labetalol
TNNT2
LPA
PAI1_HUMAN
Nisoldipine
SLC12A2
NID1_HUMAN
CST3
CRP
Acarbose
Vitamin E
PON2
IPSP_HUMAN
ALB
Kaempherol
Nicardipine
PAI2_HUMAN
GRK5
SELE
Lidocaine
ADRB1
CCL2
Digoxin
ACE
Diltiazem
Pravastatin
Fenofibrate
GPX1
ICAM1
AGT
HP
NOS3
Furosemide
Magnesium
Folic acid
APOB
Ginsenoside Rg1


Adenosine
Tolvaptan
Amiodarone
Celiprolol
Pitavastatin
Prednisone
Nicardipine
Isopamine
Warfarin
Vericiguat
Hydralazine
3-Methyladenine
Dihydralazine
Cyclopenthiazide
Bumetanide
Bendroflumethiazide
Deoxyepinephrine
Potassium
Candesartan cilexetil
Xamoterol
Probucol
Bisoprolol
Minoxidil
Enoximone
Dofetilide
Digoxin
Ruboxistaurin
Imidapril
Atenolol
Torasemide
Amrinone
heart failure
Aprostadil
Meldonium
Vesnarinone
Dobutamine
Phentolamine
Perhexiline
Sotalol
Captopril
Propranolol
Dopamine
Milrinone
Copper
Digitoxin
Testosterone
hypotensive disorder
Posinopril
Aminophylline
Midodrine
Sirolimus
Felodipine
Azathioprine
Fidarestat
Hydrochlorothiazide
Dipyridamole
Trimetazidine
Pirbuterol
Cilostazol
Ouabain
Isoprenaline
Isosorbide dinitrate
Carbocromen
Thalidomide
Fenoterol
Nitroglycerin
Prazosin
Amiloride
Potassium chloride
kidney cancer
Diphenhydramine
melanoma
Verapamil
Simendan
Spironolactone
Enrasentan
Nitroprusside
Nifedipine
Bosentan
hypoglycemia
Candesartan
Nisoldipine
Dexamethasone
hepatic encephalopathy
Acetylsalicylic acid
Bromocriptine
chronic kidney disease
Quinidine
Benazepril
vascular disorder
hyperoxaluria
Orlistat
adrenal gland disorder
Furosemide
D-glucose
Carvedilol
Salbutamol
Enalapril
skin neoplasm
acute intermittent porphyria
Fluvastatin
Losartan
hypertriglyceridemia
Irbesartan
Famotidine
Ramipril
Simvastatin

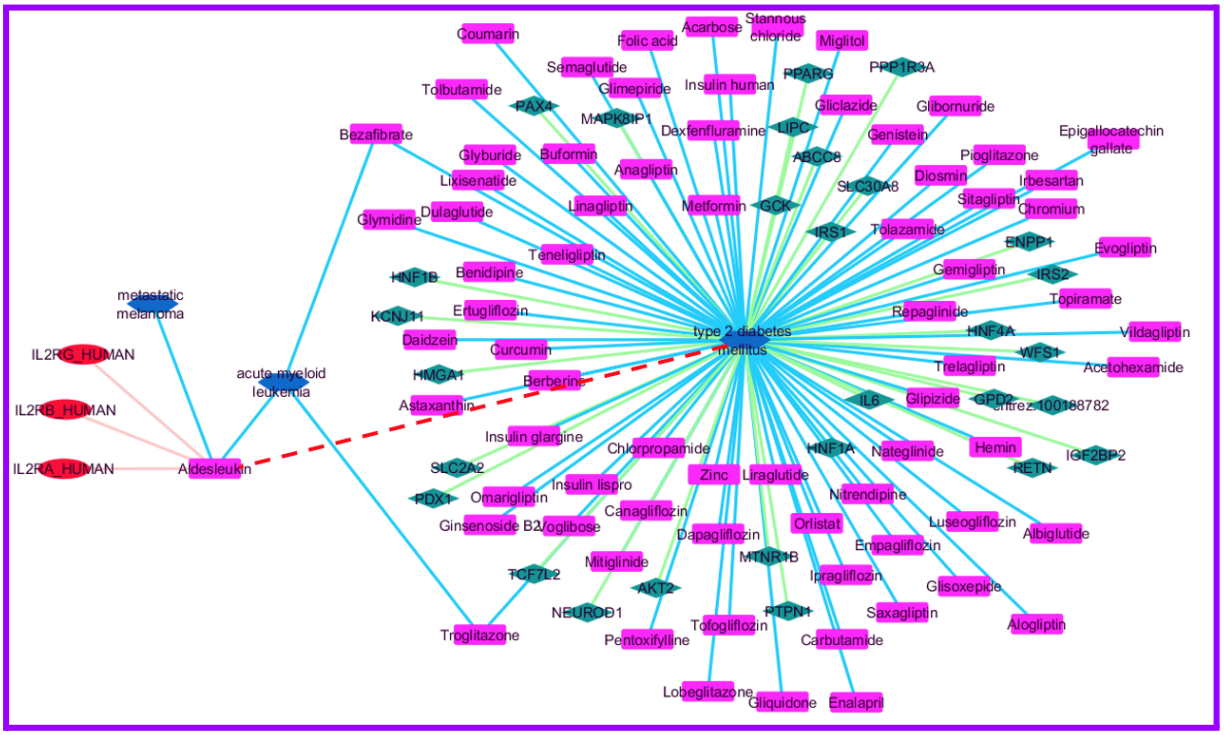


Figure C.1. Semantic subgraph showcasing some predictions. In this visualisation, nodes are colour-coded for easy identification: blue nodes denote entities of type 'disorder,' green nodes represent 'gene' entities, pink nodes signify 'drug' entities and red nodes indicate 'protein' entities. Notably, the dashed red edge symbolises the predicted link between the drug and the disease as inferred by the model.

References

[1] C. Leopold, J. D. Chambers, and A. K. Wagner, 'Thirty Years of Media Coverage on High Drug Prices in the United States—A Never-Ending Story or a Time for Change?', *Value in Health*, vol. 19, no. 1, pp. 14–16, Jan. 2016, doi: 10.1016/j.jval.2015.10.008.

[2] J. D. Chambers, T. Thorat, C. L. Wilkinson, and P. J. Neumann, 'Drugs Cleared Through The FDA's Expedited Review Offer Greater Gains Than Drugs Approved By Conventional Process', *Health Affairs*, vol. 36, no. 8, pp. 1408–1415, Aug. 2017, doi: 10.1377/hlthaff.2016.1541.

[3] *Roundtable on Translating Genomic-Based Research for Health; Board on Health Sciences Policy; Institute of Medicine. Drug Repurposing and Repositioning: Workshop Summary. Washington (DC): National Academies Press (US); 2014 Aug 8. 6, Increasing the Efficiency and Success of Repurposing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK235866/*.

[4] C. L. Bellera *et al.*, 'Computer-guided drug repurposing: Identification of trypanocidal activity of clofazimine, benidipine and saquinavir', *European Journal of Medicinal Chemistry*, vol. 93, pp. 338–348, Mar. 2015, doi: 10.1016/j.ejmech.2015.01.065.

[5] Z. Liu *et al.*, 'In silico drug repositioning – what we need to know', *Drug Discovery Today*, vol. 18, no. 3–4, pp. 110–115, Feb. 2013, doi: 10.1016/j.drudis.2012.08.005.

[6] M. Rudrapal, S. J. Khairnar, and A. G. Jadhav, 'Drug Repurposing (DR): An Emerging Approach in Drug Discovery', in *Drug Repurposing - Hypothesis, Molecular Aspects and Therapeutic Applications*, F. A. Badria, Ed., IntechOpen, 2020. doi: 10.5772/intechopen.93193.

[7] A. V. Sadybekov and V. Katritch, 'Computational approaches streamlining drug discovery', *Nature*, vol. 616, no. 7958, pp. 673–685, Apr. 2023, doi: 10.1038/s41586-023-05905-z.

[8] T. N. Jarada, J. G. Rokne, and R. Alhajj, 'A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions', *J Cheminform*, vol. 12, no. 1, p. 46, Dec. 2020, doi: 10.1186/s13321-020-00450-7.

[9] K. Park, 'A review of computational drug repurposing', *Transl Clin Pharmacol*, vol. 27, no. 2, p. 59, 2019, doi: 10.12793/tcp.2019.27.2.59.

[10] D. J. Rigden and X. M. Fernández, 'The 2022 *Nucleic Acids Research* database issue and the online molecular biology database collection', *Nucleic Acids Research*, vol. 50, no. D1, pp. D1–D10, Jan. 2022, doi: 10.1093/nar/gkab1195.

[11] T. Barrett *et al.*, 'NCBI GEO: mining tens of millions of expression profiles--database and tools update', *Nucleic Acids Research*, vol. 35, no. Database, pp. D760–D765, Jan. 2007, doi: 10.1093/nar/gkl887.

[12] 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, doi: 10.7717/peerj.1558.

[13] J. Goecks, V. Jalili, L. M. Heiser, and J. W. Gray, 'How Machine Learning Will Transform Biomedicine', *Cell*, vol. 181, no. 1, pp. 92–101, Apr. 2020, doi: 10.1016/j.cell.2020.03.022.

[14] Y. Luo *et al.*, 'A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information', *Nat Commun*, vol. 8, no. 1, p. 573, Sep. 2017, doi: 10.1038/s41467-017-00680-8.

[15] A. Aliper, S. Plis, A. Artemov, A. Ulloa, P. Mamoshina, and A. Zhavoronkov, 'Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data', *Mol. Pharmaceutics*, vol. 13, no. 7, pp. 2524–2530, Jul. 2016, doi: 10.1021/acs.molpharmaceut.6b00248.

[16] M. P. Menden *et al.*, 'Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties', *PLoS ONE*, vol. 8, no. 4, p. e61318, Apr. 2013, doi: 10.1371/journal.pone.0061318.

[17] F. Napolitano *et al.*, 'Drug repositioning: a machine-learning approach through data integration', *J Cheminform*, vol. 5, no. 1, p. 30, Dec. 2013, doi: 10.1186/1758-2946-5-30.
[18] H. Chen, O. Engkvist, Y. Wang, M. Olivecrona, and T. Blaschke, 'The rise of deep learning in drug discovery', *Drug Discov Today*, vol. 23, no. 6, pp. 1241–1250, Jun. 2018, doi: 10.1016/j.drudis.2018.01.039.
[19] L. Carius and R. Findeisen, 'The impact of experimental data quality on computational systems biology and engineering', *IFAC-PapersOnLine*, vol. 49, no. 26, pp. 140–146, 2016, doi: 10.1016/j.ifacol.2016.12.116.
[20] L. Cai and Y. Zhu, 'The Challenges of Data Quality and Data Quality Assessment in the Big Data Era', *CODATA*, vol. 14, no. 0, p. 2, May 2015, doi: 10.5334/dsj-2015-002.
[21] V. Sessions and M. Valtorta, 'The Effects of Data Quality on Machine Learning Algorithms.', *ICIQ*, vol. 6, pp. 485–498, 2006.
[22] L. Budach *et al.*, 'The Effects of Data Quality on Machine Learning Performance', 2022, doi: 10.48550/ARXIV.2207.14529.
[23] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, 'A survey on missing data in machine learning', *J Big Data*, vol. 8, no. 1, p. 140, Oct. 2021, doi: 10.1186/s40537-021-00516-9.
[24] G. Canbek, 'Gaining insights in datasets in the shade of "garbage in, garbage out" rationale: feature-space distribution fitting', In Review, preprint, Feb. 2022. doi: 10.21203/rs.3.rs-1369128/v1.
[25] J. Al-Saleem *et al.*, 'Knowledge Graph-Based Approaches to Drug Repurposing for COVID-19', *J. Chem. Inf. Model.*, vol. 61, no. 8, pp. 4058–4067, Aug. 2021, doi: 10.1021/acs.jcim.1c00642.
[26] S. Sadeghi, J. Lu, and A. Ngom, 'An Integrative Heterogeneous Graph Neural Network–Based Method for Multi-Labeled Drug Repurposing', *Front. Pharmacol.*, vol. 13, p. 908549, Jul. 2022, doi: 10.3389/fphar.2022.908549.
[27] 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, no. 5, p. e0155811, 2016, doi: 10.1371/journal.pone.0155811.
[28] C. Daraio, S. Di Leo, and M. Scannapieco, 'Accounting for quality in data integration systems: a completeness-aware integration approach', *Scientometrics*, vol. 127, no. 3, pp. 1465–1490, Mar. 2022, doi: 10.1007/s11192-022-04266-0.
[29] Q. Chen *et al.*, 'Quality Matters: Biocuration Experts on the Impact of Duplication and Other Data Quality Issues in Biological Databases', *Genomics Proteomics Bioinformatics*, vol. 18, no. 2, pp. 91–103, Apr. 2020, doi: 10.1016/j.gpb.2018.11.006.
[30] A. Bernasconi, 'Data quality-aware genomic data integration', *Computer Methods and Programs in Biomedicine Update*, vol. 1, p. 100009, 2021, doi: 10.1016/j.cmpbup.2021.100009.
[31] Z. Zhang, X. Xiao, W. Zhou, D. Zhu, and C. I. Amos, 'False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy', *Human Molecular Genetics*, vol. 31, no. 1, pp. 146–155, Dec. 2021, doi: 10.1093/hmg/ddab203.
[32] Y.-H. Kim *et al.*, 'False-negative errors in next-generation sequencing contribute substantially to inconsistency of mutation databases', *PLoS ONE*, vol. 14, no. 9, p. e0222535, Sep. 2019, doi: 10.1371/journal.pone.0222535.
[33] X. Shen and Ö. Carlborg, 'Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability', *Front. Genet.*, vol. 4, 2013, doi: 10.3389/fgene.2013.00093.
[34] Q. Chen, J. Zobel, and K. Verspoor, 'Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study', *Database*, vol. 2017, p. baw163, 2017, doi: 10.1093/database/baw163.
[35] Qingyu Chen, 'DUPLICATION IN BIOLOGICAL DATABASES: DEFINITIONS, IMPACTS AND METHODS', Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy, 2017. [Online]. Available:

https://minerva-access.unimelb.edu.au/handle/11343/197466
[36] M. Masseroli, G. Ghisalberti, and L. Tettamanti, 'Detection of Errors and Inconsistencies in Biomolecular Databases through Integrative Approaches and Quality Controls', in *2010 IEEE International Conference on BioInformatics and BioEngineering*, Philadelphia, PA, USA: IEEE, 2010, pp. 294–295. doi: 10.1109/BIBE.2010.60.
[37] L. Holman, M. L. Head, R. Lanfear, and M. D. Jennions, 'Evidence of Experimental Bias in the Life Sciences: Why We Need Blind Data Recording', *PLoS Biol*, vol. 13, no. 7, p. e1002190, Jul. 2015, doi: 10.1371/journal.pbio.1002190.
[38] Bleier, Ruth, 'Bias in Biological and Human Sciences: Some Comments', 1978, [Online]. Available: https://eric.ed.gov/?id=EJ189742
[39] J. E. Flores, D. M. Claborne, Z. D. Weller, B.-J. M. Webb-Robertson, K. M. Waters, and L. M. Bramer, 'Missing data in multi-omics integration: Recent advances through artificial intelligence', *Front. Artif. Intell.*, vol. 6, p. 1098308, Feb. 2023, doi: 10.3389/frai.2023.1098308.
[40] C. Srivastava, 'Biological Data Analysis: Error and Uncertainty', *wjcat*, vol. 1, no. 3, pp. 67–74, Nov. 2013, doi: 10.13189/wjcat.2013.010302.
[41] S. Tandy-Connor *et al.*, 'False-positive results released by direct-to-consumer genetic tests highlight the importance of clinical confirmation testing for appropriate patient care', *Genetics in Medicine*, vol. 20, no. 12, pp. 1515–1521, Dec. 2018, doi: 10.1038/gim.2018.38.
[42] A. Stroud, A. Gamblin, P. Birchall, S. Harbison, and S. Opperman, 'A comprehensive study into false positive rates for "other" biological samples using common presumptive testing methods', *Science & Justice*, vol. 63, no. 3, pp. 414–420, May 2023, doi: 10.1016/j.scijus.2023.04.006.
[43] J. Mante *et al.*, 'A heuristic approach to handling missing data in biologics manufacturing databases', *Bioprocess Biosyst Eng*, vol. 42, no. 4, pp. 657–663, Apr. 2019, doi: 10.1007/s00449-018-02059-5.
[44] T. Das, K. Bhattarai, S. Rajaganapathy, L. Wang, J. R. Cerhan, and N. Zong, 'Leveraging multi-source to resolve inconsistency across pharmacogenomic datasets in drug sensitivity prediction', Health Informatics, preprint, Jun. 2023. doi: 10.1101/2023.05.25.23290546.
[45] F. Vitali *et al.*, 'A Network-Based Data Integration Approach to Support Drug Repurposing and Multi-Target Therapies in Triple Negative Breast Cancer', *PLoS ONE*, vol. 11, no. 9, p. e0162407, Sep. 2016, doi: 10.1371/journal.pone.0162407.
[46] Aoesha Alsobhe, P. Gater, K. James, S. Cockell, and A. Wipat, 'A computational approach to drug repurposing through graph neural networks and biomedical knowledge graphs to predict drug-disease indications', 2024, doi: 10.13140/RG.2.2.36572.40324.
[47] 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*, vol. 24, no. 13, pp. i232-240, Jul. 2008, doi: 10.1093/bioinformatics/btn162.
[48] S. Sadegh *et al.*, 'NeDRex - an integrative and interactive network medicine platform for drug repurposing', in *NeDRex - an integrative and interactive network medicine platform for drug repurposing*, ScienceOpen, Aug. 2022. doi: 10.14293/S2199-1006.1.SOR-.PPPY90R8.v1.
[49] A. Pavel *et al.*, 'Integrated network analysis reveals new genes suggesting COVID-19 chronic effects and treatment', *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1430–1441, Mar. 2021, doi: 10.1093/bib/bbaa417.
[50] D. J. Skelton *et al.*, 'Drug repurposing prediction for COVID-19 using probabilistic networks and crowdsourced curation', 2020, doi: 10.48550/ARXIV.2005.11088.
[51] I. A. Kovács *et al.*, 'Network-based prediction of protein interactions', *Nat Commun*, vol. 10, no. 1, p. 1240, Mar. 2019, doi: 10.1038/s41467-019-09177-y.
[52] 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, no. 6, p. e20284, Jun. 2011, doi: 10.1371/journal.pone.0020284.
[53] F. Azuaje, 'Drug interaction networks: an introduction to translational and clinical applications',

Cardiovascular Research, vol. 97, no. 4, pp. 631–641, Mar. 2013, doi: 10.1093/cvr/cvs289.
[54] J. T. Dudley, T. Deshpande, and A. J. Butte, 'Exploiting drug-disease relationships for computational drug repositioning', *Briefings in Bioinformatics*, vol. 12, no. 4, pp. 303–311, Jul. 2011, doi: 10.1093/bib/bbr013.
[55] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, 'A Probabilistic Functional Network of Yeast Genes', *Science*, vol. 306, no. 5701, pp. 1555–1558, Nov. 2004, doi: 10.1126/science.1099511.
[56] A. G. Fraser and E. M. Marcotte, 'A probabilistic view of gene function', *Nat Genet*, vol. 36, no. 6, pp. 559–564, Jun. 2004, doi: 10.1038/ng1370.
[57] K. James, A. Alsobhe, S. J. Cockell, A. Wipat, and M. Pocock, 'Integration of probabilistic functional networks without an external Gold Standard', *BMC Bioinformatics*, vol. 23, no. 1, p. 302, Dec. 2022, doi: 10.1186/s12859-022-04834-4.
[58] I. Lee, Z. Li, and E. M. Marcotte, 'An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae', *PLoS ONE*, vol. 2, no. 10, p. e988, Oct. 2007, doi: 10.1371/journal.pone.0000988.
[59] James, K., Lycett, S.J., Wipat, A., Hallinan, J.S. 'Multiple Gold Standards address bias in functional network integration'. Newcastle University School of Computing Science Technical Report Series 2011; CS-TR-1302.
[60] A. Suratanee and K. Plaimas, 'Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions', *PLoS ONE*, vol. 13, no. 6, p. e0199435, Jun. 2018, doi: 10.1371/journal.pone.0199435.
[61] V. D. Tran, A. Sperduti, R. Backofen, and F. Costa, 'Heterogeneous networks integration for disease–gene prioritization with node kernels', *Bioinformatics*, vol. 36, no. 9, pp. 2649–2656, May 2020, doi: 10.1093/bioinformatics/btaa008.
[62] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, 'Network medicine: a network-based approach to human disease', *Nat Rev Genet*, vol. 12, no. 1, pp. 56–68, Jan. 2011, doi: 10.1038/nrg2918.
[63] 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, no. 6, p. e20284, Jun. 2011, doi: 10.1371/journal.pone.0020284.
[64] A. Ayuso-Muñoz, L. Prieto-Santamaría, E. Ugarte-Carro, E. Serrano, and A. Rodríguez-González, 'Uncovering hidden therapeutic indications through drug repurposing with graph neural networks and heterogeneous data', *Artificial Intelligence in Medicine*, vol. 145, p. 102687, Nov. 2023, doi: 10.1016/j.artmed.2023.102687.
[65] C. Cao *et al.*, 'Deep Learning and Its Applications in Biomedicine', *Genomics, Proteomics & Bioinformatics*, vol. 16, no. 1, pp. 17–32, Feb. 2018, doi: 10.1016/j.gpb.2017.07.003.
[66] X. Pan *et al.*, 'Deep learning for drug repurposing: Methods, databases, and applications', *WIREs Comput Mol Sci*, vol. 12, no. 4, Jul. 2022, doi: 10.1002/wcms.1597.
[67] S. Doshi and S. P. Chepuri, 'A computational approach to drug repurposing using graph neural networks', *Computers in Biology and Medicine*, vol. 150, p. 105992, Nov. 2022, doi: 10.1016/j.compbiomed.2022.105992.
[68] K. Hsieh *et al.*, 'Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence', *Sci Rep*, vol. 11, no. 1, p. 23179, Nov. 2021, doi: 10.1038/s41598-021-02353-5.
[69] M. Jiang *et al.*, 'Drug–target affinity prediction using graph neural network and contact maps', *RSC Adv.*, vol. 10, no. 35, pp. 20701–20712, 2020, doi: 10.1039/D0RA02297G.
[70] J. Lim, S. Ryu, K. Park, Y. J. Choe, J. Ham, and W. Y. Kim, 'Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation', *J. Chem. Inf. Model.*, vol. 59, no. 9, pp. 3981–3988, Sep. 2019, doi: 10.1021/acs.jcim.9b00387.
[71] W. Torng and R. B. Altman, 'Graph Convolutional Neural Networks for Predicting Drug-Target Interactions', *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4131–4149, Oct. 2019, doi: 10.1021/acs.jcim.9b00628.

[72] R. PalSingh and V. Vandana, 'Application of Graph Theory in Computer Science and Engineering', *IJCA*, vol. 104, no. 1, pp. 10–13, Oct. 2014, doi: 10.5120/18165-9025.
[73] G. A. Pavlopoulos *et al.*, 'Using graph theory to analyze biological networks', *BioData Mining*, vol. 4, no. 1, p. 10, Dec. 2011, doi: 10.1186/1756-0381-4-10.
[74] S.-Y. Chao, 'Graph Theory and Analysis of Biological Data in Computational Biology', in *Advanced Technologies*, K. Jayanthakumaran, Ed., InTech, 2009. doi: 10.5772/8205.
[75] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, 'Increasing confidence of protein interactomes using network topological metrics', *Bioinformatics*, vol. 22, no. 16, pp. 1998–2004, Aug. 2006, doi: 10.1093/bioinformatics/btl335.
[76] V. Lapatas, M. Stefanidakis, R. C. Jimenez, A. Via, and M. V. Schneider, 'Data integration in biological research: an overview', *J of Biol Res-Thessaloniki*, vol. 22, no. 1, p. 9, Dec. 2015, doi: 10.1186/s40709-015-0032-5.
[77] D. Gomez-Cabrero *et al.*, 'Data integration in the era of omics: current and future challenges', *BMC Syst Biol*, vol. 8, no. Suppl 2, p. I1, 2014, doi: 10.1186/1752-0509-8-S2-I1.
[78] M. D. M. R. García, J. García-Nieto, and J. F. Aldana-Montes, 'An ontology-based data integration approach for web analytics in e-commerce', *Expert Systems with Applications*, vol. 63, pp. 20–34, Nov. 2016, doi: 10.1016/j.eswa.2016.06.034.
[79] A. Bahga and V. K. Madisetti, 'Healthcare Data Integration and Informatics in the Cloud', *Computer*, vol. 48, no. 2, pp. 50–57, Feb. 2015, doi: 10.1109/MC.2015.46.
[80] T. D. T. Oyedotun and H. Burningham, 'The need for data integration to address the challenges of climate change on the Guyana coast', *Geography and Sustainability*, vol. 2, no. 4, pp. 288–297, Dec. 2021, doi: 10.1016/j.geosus.2021.11.003.
[81] W. Lemahieu, S. vanden Broucke, and B. Baesens, *Principles of database management: the practical guide to storing, managing and analyzing big and small data*. Cambridge: Cambridge university press, 2018.
[82] A. H. Chen, W. Ge, W. Metcalf, E. Jakobsson, L. S. Mainzer, and A. E. Lipka, 'An assessment of true and false positive detection rates of stepwise epistatic model selection as a function of sample size and number of markers', *Heredity*, vol. 122, no. 5, pp. 660–671, May 2019, doi: 10.1038/s41437-018-0162-2.
[83] B. Wang *et al.*, 'Network enhancement as a general method to denoise weighted biological networks', *Nat Commun*, vol. 9, no. 1, p. 3108, Dec. 2018, doi: 10.1038/s41467-018-05469-x.
[84] M. A. Mahdavi and Y.-H. Lin, 'False positive reduction in protein-protein interaction predictions using gene ontology annotations', *BMC Bioinformatics*, vol. 8, p. 262, Jul. 2007, doi: 10.1186/1471-2105-8-262.
[85] K. James, A. Alsobhe, S. J. Cockell, A. Wipat, and M. Pocock, 'Integration of probabilistic functional networks without an external Gold Standard', *bioRxiv*, p. 2021.10.01.462727, Jan. 2021, doi: 10.1101/2021.10.01.462727.
[86] K. James, 'Knowledge derivation and data mining strategies for probabilistic functional integrated networks'. https://theses.ncl.ac.uk/jspui/handle/10443/1436. Published 2012.
[87] R. Oughtred *et al.*, 'The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions', *Protein Science*, vol. 30, no. 1, pp. 187–200, Jan. 2021, doi: 10.1002/pro.3978.
[88] J. S. Amberger, C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh, 'OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders', *Nucleic Acids Research*, vol. 43, no. D1, pp. D789–D798, Jan. 2015, doi: 10.1093/nar/gku1205.
[89] J. Yu and R. L. Finley, 'Combining multiple positive training sets to generate confidence scores for protein–protein interactions', *Bioinformatics*, vol. 25, no. 1, pp. 105–111, Jan. 2009, doi: 10.1093/bioinformatics/btn597.
[90] 'James K, Wipat A, Hallinan J. Integration of Full-Coverage Probabilistic Functional Networks with Relevance to Specific Biological Processes. In: Paton N.W., Missier P., Hedeler C. (eds) Data

Integration in the Life Sciences. DILS 2009. Lecture Notes in Computer Science, vol 5647. Springer, Berlin, Heidelberg.'.

[91] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, 'A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae)', *Proc Natl Acad Sci U S A*, vol. 100, no. 14, pp. 8348–8353, Jul. 2003, doi: 10.1073/pnas.0832373100.

[92] J. Wang, Y. Zuo, L. Liu, Y. Man, M. G. Tadesse, and H. W. Ressom, 'Identification of Functional Modules by Integration of Multiple Data Sources Using a Bayesian Network Classifier', *Circ Cardiovasc Genet*, vol. 7, no. 2, pp. 206–217, Apr. 2014, doi: 10.1161/CIRCGENETICS.113.000087.

[93] C. Xing and D. B. Dunson, 'Bayesian Inference for Genomic Data Integration Reduces Misclassification Rate in Predicting Protein-Protein Interactions', *PLoS Comput Biol*, vol. 7, no. 7, p. e1002110, Jul. 2011, doi: 10.1371/journal.pcbi.1002110.

[94] A. Alexeyenko and E. L. L. Sonnhammer, 'Global networks of functional coupling in eukaryotes from comprehensive data integration', *Genome Res.*, vol. 19, no. 6, pp. 1107–1116, Jun. 2009, doi: 10.1101/gr.087528.108.

[95] K. Xia, D. Dong, and J.-D. J. Han, 'IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model', *BMC Bioinformatics*, vol. 7, no. 1, p. 508, Dec. 2006, doi: 10.1186/1471-2105-7-508.

[96] T. P. O'Connor and R. G. Crystal, 'Genetic medicines: treatment strategies for hereditary disorders', *Nat Rev Genet*, vol. 7, no. 4, pp. 261–276, Apr. 2006, doi: 10.1038/nrg1829.

[97] A. Devaprasad, T. R. Radstake, and A. Pandit, 'Integration of immunome with disease-gene network reveals common cellular mechanisms between IMIDs and drug repurposing strategies', Bioinformatics, preprint, Dec. 2019. doi: 10.1101/2019.12.12.874321.

[98] S. Sadegh *et al.*, 'Network medicine for disease module identification and drug repurposing with the NeDRex platform', *Nat Commun*, vol. 12, no. 1, p. 6848, Nov. 2021, doi: 10.1038/s41467-021-27138-2.

[99] M. Dawn Teare and J. H. Barrett, 'Genetic linkage studies', *Lancet*, vol. 366, no. 9490, pp. 1036–1044, Sep. 2005, doi: 10.1016/S0140-6736(05)67382-5.

[100] E. Uffelmann *et al.*, 'Genome-wide association studies', *Nat Rev Methods Primers*, vol. 1, no. 1, p. 59, Aug. 2021, doi: 10.1038/s43586-021-00056-9.

[101] J. C. Simpson, 'Functional Assays', in *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine*, Springer Berlin Heidelberg, 2006, pp. 617–620. doi: 10.1007/3-540-29623-9_3700.

[102] M. Boutros and J. Ahringer, 'The art and design of genetic screens: RNA interference', *Nat Rev Genet*, vol. 9, no. 7, pp. 554–566, Jul. 2008, doi: 10.1038/nrg2364.

[103] M. Reza Khorramizadeh and F. Saadat, 'Animal models for human disease', in *Animal Biotechnology*, Elsevier, 2020, pp. 153–171. doi: 10.1016/B978-0-12-811710-1.00008-2.

[104] G. Valentini, A. Paccanaro, H. Caniza, A. E. Romero, and M. Re, 'An extensive analysis of disease-gene associations using network integration and fast kernel-based gene prioritization methods', *Artificial Intelligence in Medicine*, vol. 61, no. 2, pp. 63–78, Jun. 2014, doi: 10.1016/j.artmed.2014.03.003.

[105] M. Asif, H. F. M. C. M. Martiniano, A. M. Vicente, and F. M. Couto, 'Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology', *PLoS ONE*, vol. 13, no. 12, p. e0208626, Dec. 2018, doi: 10.1371/journal.pone.0208626.

[106] M. Kang, S. Kim, D.-B. Lee, C. Hong, and K.-B. Hwang, 'Gene-specific machine learning for pathogenicity prediction of rare BRCA1 and BRCA2 missense variants', *Sci Rep*, vol. 13, no. 1, p. 10478, Jun. 2023, doi: 10.1038/s41598-023-37698-6.

[107] S. Azadifar and A. Ahmadi, 'A novel candidate disease gene prioritization method using deep graph convolutional networks and semi-supervised learning', *BMC Bioinformatics*, vol. 23, no. 1, p. 422, Oct. 2022, doi: 10.1186/s12859-022-04954-x.

[108] 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*, vol. 74, pp. 83–89, Mar. 2015, doi: 10.1016/j.ymeth.2014.11.020.
[109] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, and D. S. Wishart, 'PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites', *Nucleic Acids Res*, vol. 36, no. Web Server issue, pp. W399-405, Jul. 2008, doi: 10.1093/nar/gkn296.
[110] R. M. Piro and F. Di Cunto, 'Computational approaches to disease-gene prediction: rationale, classification and successes', *FEBS J*, vol. 279, no. 5, pp. 678–696, Mar. 2012, doi: 10.1111/j.1742-4658.2012.08471.x.
[111] A. M. Bianco, A. Marcuzzi, V. Zanin, M. Girardelli, J. Vuch, and S. Crovella, 'Database tools in genetic diseases research', *Genomics*, vol. 101, no. 2, pp. 75–85, Feb. 2013, doi: 10.1016/j.ygeno.2012.11.001.
[112] D. Tamborero *et al.*, 'Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations', *Genome Med*, vol. 10, no. 1, p. 25, Dec. 2018, doi: 10.1186/s13073-018-0531-8.
[113] S. Aymé, B. Dallapiccola, and D. Donnai, 'Orphanet Journal of Rare Diseases: Launch Editorial', *Orphanet J Rare Dis*, vol. 1, no. 1, pp. 1, 1750-1172-1–1, Dec. 2006, doi: 10.1186/1750-1172-1-1.
[114] A. Gutiérrez-Sacristán *et al.*, 'PsyGeNET: a knowledge platform on psychiatric disorders and their genes: Table 1.', *Bioinformatics*, vol. 31, no. 18, pp. 3075–3077, Sep. 2015, doi: 10.1093/bioinformatics/btv301.
[115] Y. Kim, J.-H. Park, and Y.-R. Cho, 'Network-Based Approaches for Disease-Gene Association Prediction Using Protein-Protein Interaction Networks', *IJMS*, vol. 23, no. 13, p. 7411, Jul. 2022, doi: 10.3390/ijms23137411.
[116] J. Piñero *et al.*, 'DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants', *Nucleic Acids Res*, vol. 45, no. D1, pp. D833–D839, Jan. 2017, doi: 10.1093/nar/gkw943.
[117] M. R. Montinari, S. Minelli, and R. De Caterina, 'The first 3500 years of aspirin history from its roots – A concise summary', *Vascular Pharmacology*, vol. 113, pp. 1–8, Feb. 2019, doi: 10.1016/j.vph.2018.10.008.
[118] J. Miner and A. Hoffhines, 'The discovery of aspirin's antithrombotic effects', *Tex Heart Inst J*, vol. 34, no. 2, pp. 179–186, 2007.
[119] C. H. Hennekens, M. L. Dyken, and V. Fuster, 'Aspirin as a Therapeutic Agent in Cardiovascular Disease: A Statement for Healthcare Professionals From the American Heart Association', *Circulation*, vol. 96, no. 8, pp. 2751–2753, Oct. 1997, doi: 10.1161/01.CIR.96.8.2751.
[120] P. L. R. Andrews, R. S. B. Williams, and G. J. Sanger, 'Anti-emetic effects of thalidomide: Evidence, mechanism of action, and future directions', *Current Research in Pharmacology and Drug Discovery*, vol. 3, p. 100138, 2022, doi: 10.1016/j.crphar.2022.100138.
[121] W. Rehman, L. M. Arfons, and H. M. Lazarus, 'The rise, fall and subsequent triumph of thalidomide: lessons learned in drug development', *Therapeutic Advances in Hematology*, vol. 2, no. 5, pp. 291–308, Oct. 2011, doi: 10.1177/2040620711413165.
[122] P. Mehta and M. Hussein, 'Thalidomide as anti-inflammatory therapy for multiple myeloma', *Leukemia*, vol. 17, no. 11, pp. 2237–2238, Nov. 2003, doi: 10.1038/sj.leu.2403118.
[123] *BarabÃ¡si A, PÃ³sfai M. Network Science. Cambridge: Cambridge University Press; 2016.*
[124] Z.-H. Chen, Z.-H. You, Z.-H. Guo, H.-C. Yi, G.-X. Luo, and Y.-B. Wang, 'Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model', *Front. Bioeng. Biotechnol.*, vol. 8, p. 338, Jun. 2020, doi: 10.3389/fbioe.2020.00338.
[125] Y. Xiong *et al.*, 'Heterogeneous network embedding enabling accurate disease association predictions', *BMC Med Genomics*, vol. 12, no. S10, p. 186, Dec. 2019, doi: 10.1186/s12920-019-0623-3.

[126] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, 'A survey of heterogeneous information network analysis', *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 1, pp. 17–37, Jan. 2017, doi: 10.1109/TKDE.2016.2598561.
[127] X.-M. Zhang, L. Liang, L. Liu, and M.-J. Tang, 'Graph Neural Networks and Their Current Applications in Bioinformatics', *Front. Genet.*, vol. 12, p. 690049, Jul. 2021, doi: 10.3389/fgene.2021.690049.
[128] Y. Jiaxuan, 'EMPOWERING DEEP LEARNING WITH GRAPHS', STANFORD UNIVERSITY, 2021. [Online]. Available: https://purl.stanford.edu/mz469rn9516
[129] J. Zhou *et al.*, 'Graph neural networks: A review of methods and applications', *AI Open*, vol. 1, pp. 57–81, 2020, doi: 10.1016/j.aiopen.2021.01.001.
[130] S. Doshi and S. P. Chepuri, 'A computational approach to drug repurposing using graph neural networks', *Computers in Biology and Medicine*, vol. 150, p. 105992, Nov. 2022, doi: 10.1016/j.compbiomed.2022.105992.
[131] F. Zhang, W. Hu, and Y. Liu, 'GCMM: graph convolution network based on multimodal attention mechanism for drug repurposing', *BMC Bioinformatics*, vol. 23, no. 1, p. 372, Sep. 2022, doi: 10.1186/s12859-022-04911-8.
[132] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, 'Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets', *Mol Pharm*, vol. 14, no. 12, pp. 4462–4475, Dec. 2017, doi: 10.1021/acs.molpharmaceut.7b00578.
[133] F. Li, Z. Zhang, J. Guan, and S. Zhou, 'Effective drug–target interaction prediction with mutual interaction neural network', *Bioinformatics*, vol. 38, no. 14, pp. 3582–3589, Jul. 2022, doi: 10.1093/bioinformatics/btac377.
[134] K. Hänsel, S. N. Dudgeon, K.-H. Cheung, T. J. S. Durant, and W. L. Schulz, 'From Data to Wisdom: Biomedical Knowledge Graphs for Real-World Data Insights', *J Med Syst*, vol. 47, no. 1, p. 65, May 2023, doi: 10.1007/s10916-023-01951-2.
[135] R. Kaalia and J. C. Rajapakse, 'Functional homogeneity and specificity of topological modules in human proteome', *BMC Bioinformatics*, vol. 19, no. S13, p. 553, Feb. 2019, doi: 10.1186/s12859-018-2549-8.
[136] R. Ferrari, R. C. Lovering, J. Hardy, P. A. Lewis, and C. Manzoni, 'Weighted Protein Interaction Network Analysis of Frontotemporal Dementia', *J. Proteome Res.*, vol. 16, no. 2, pp. 999–1013, Feb. 2017, doi: 10.1021/acs.jproteome.6b00934.
[137] E. P. García Del Valle, G. Lagunes García, L. Prieto Santamaría, M. Zanin, E. Menasalvas Ruiz, and A. Rodríguez-González, 'Disease networks and their contribution to disease understanding: A review of their evolution, techniques and data sources', *Journal of Biomedical Informatics*, vol. 94, p. 103206, Jun. 2019, doi: 10.1016/j.jbi.2019.103206.
[138] H. Quan *et al.*, 'Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data':, *Medical Care*, vol. 43, no. 11, pp. 1130–1139, Nov. 2005, doi: 10.1097/01.mlr.0000182534.19832.83.
[139] M. Gustafsson *et al.*, 'Modules, networks and systems medicine for understanding disease and aiding diagnosis', *Genome Med*, vol. 6, no. 10, p. 82, Dec. 2014, doi: 10.1186/s13073-014-0082-6.
[140] Z.-C. Li, Y.-H. Lai, L.-L. Chen, Y. Xie, Z. Dai, and X.-Y. Zou, 'Identifying and prioritizing disease-related genes based on the network topological features', *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1844, no. 12, pp. 2214–2221, Dec. 2014, doi: 10.1016/j.bbapap.2014.08.009.
[141] J. Piñero *et al.*, 'The DisGeNET knowledge platform for disease genomics: 2019 update', *Nucleic Acids Research*, p. gkz1021, Nov. 2019, doi: 10.1093/nar/gkz1021.
[142] F. Emmert-Streib, M. Dehmer, and B. Haibe-Kains, 'Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks', *Front. Cell Dev. Biol.*, vol. 2, Aug. 2014, doi: 10.3389/fcell.2014.00038.
[143] V. Lacroix, L. Cottret, P. Thebault, and M.-F. Sagot, 'An Introduction to Metabolic Networks and

Their Structural Analysis', *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 5, no. 4, pp. 594–617, Oct. 2008, doi: 10.1109/TCBB.2008.79.
[144] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, 'Multi-omics Data Integration, Interpretation, and Its Application', *Bioinform Biol Insights*, vol. 14, p. 117793221989905, Jan. 2020, doi: 10.1177/1177932219899051.
[145] V. Gligorijević and N. Pržulj, 'Methods for biological data integration: perspectives and challenges', *J. R. Soc. Interface.*, vol. 12, no. 112, p. 20150571, Nov. 2015, doi: 10.1098/rsif.2015.0571.
[146] C. Szabo, A. Masiello, J. F. Ryan, The BIC Consortium, and L. C. Brody, 'The Breast Cancer Information Core: Database design, structure, and scope', *Hum. Mutat.*, vol. 16, no. 2, pp. 123–131, Aug. 2000, doi: 10.1002/1098-1004(200008)16:2<123::AID-HUMU4>3.0.CO;2-Y.
[147] A. Kumar, A. Bansal, and T. R. Singh, 'ABCD: Alzheimer's disease Biomarkers Comprehensive Database', *3 Biotech*, vol. 9, no. 10, p. 351, Oct. 2019, doi: 10.1007/s13205-019-1888-0.
[148] C. Rubio-Perez *et al.*, 'Genetic and functional characterization of disease associations explains comorbidity', *Sci Rep*, vol. 7, no. 1, p. 6207, Jul. 2017, doi: 10.1038/s41598-017-04939-4.
[149] Y. Ko, M. Cho, J.-S. Lee, and J. Kim, 'Identification of disease comorbidity through hidden molecular mechanisms', *Sci Rep*, vol. 6, no. 1, p. 39433, Dec. 2016, doi: 10.1038/srep39433.
[150] V. Ormazabal, S. Nair, O. Elfeky, C. Aguayo, C. Salomon, and F. A. Zuñiga, 'Association between insulin resistance and the development of cardiovascular disease', *Cardiovasc Diabetol*, vol. 17, no. 1, p. 122, Dec. 2018, doi: 10.1186/s12933-018-0762-4.
[151] E. Barrett-Connor, D. Wingard, N. Wong, and R. Goldberg, 'Heart Disease and Diabetes', in *Diabetes in America*, 3rd ed., C. C. Cowie, S. S. Casagrande, A. Menke, M. A. Cissell, M. S. Eberhardt, J. B. Meigs, E. W. Gregg, W. C. Knowler, E. Barrett-Connor, D. J. Becker, F. L. Brancati, E. J. Boyko, W. H. Herman, B. V. Howard, K. M. V. Narayan, M. Rewers, and J. E. Fradkin, Eds., Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases (US), 2018. Accessed: Jan. 25, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK568001/
[152] J. M. Valderas, B. Starfield, B. Sibbald, C. Salisbury, and M. Roland, 'Defining Comorbidity: Implications for Understanding Health and Health Services', *The Annals of Family Medicine*, vol. 7, no. 4, pp. 357–363, Jul. 2009, doi: 10.1370/afm.983.
[153] K. Wang, M. Li, and H. Hakonarson, 'Analysing biological pathways in genome-wide association studies', *Nat Rev Genet*, vol. 11, no. 12, pp. 843–854, Dec. 2010, doi: 10.1038/nrg2884.
[154] 'National Cancer Institute. "BRCA Mutations: Cancer Risk and Genetic Testing Fact Sheet." National Cancer Institute, https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet. Accessed 21 Feb 2024'.
[155] I. Gorodetska, I. Kozeretska, and A. Dubrovska, '*BRCA* Genes: The Role in Genome Stability, Cancer Stemness and Therapy Resistance', *J. Cancer*, vol. 10, no. 9, pp. 2109–2127, 2019, doi: 10.7150/jca.30410.
[156] Z. Liu *et al.*, 'Similarity-based prediction for Anatomical Therapeutic Chemical classification of drugs by integrating multiple data sources', *Bioinformatics*, vol. 31, no. 11, pp. 1788–1795, Jun. 2015, doi: 10.1093/bioinformatics/btv055.
[157] J. L. Espinoza, N. Shah, S. Singh, K. E. Nelson, and C. L. Dupont, 'Applications of weighted association networks applied to compositional data in biology', *Environmental Microbiology*, vol. 22, no. 8, pp. 3020–3038, Aug. 2020, doi: 10.1111/1462-2920.15091.
[158] J. Pardo-Diaz, P. S. Poole, M. Beguerisse-Díaz, C. M. Deane, and G. Reinert, 'Generating weighted and thresholded gene coexpression networks using signed distance correlation', *Net Sci*, vol. 10, no. 2, pp. 131–145, Jun. 2022, doi: 10.1017/nws.2022.13.
[159] D. Grissa, A. Junge, T. I. Oprea, and L. J. Jensen, 'Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration', *Database*, vol. 2022, p. baac019, Mar. 2022, doi: 10.1093/database/baac019.
[160] N. T. Strande *et al.*, 'Evaluating the Clinical Validity of Gene-Disease Associations: An

Evidence-Based Framework Developed by the Clinical Genome Resource', *The American Journal of Human Genetics*, vol. 100, no. 6, pp. 895–906, Jun. 2017, doi: 10.1016/j.ajhg.2017.04.015.
[161] A. J. Kavran and A. Clauset, 'Denoising large-scale biological data using network filters', *BMC Bioinformatics*, vol. 22, no. 1, p. 157, Dec. 2021, doi: 10.1186/s12859-021-04075-x.
[162] H. M. Colhoun, P. M. McKeigue, and G. D. Smith, 'Problems of reporting genetic associations with complex outcomes', *The Lancet*, vol. 361, no. 9360, pp. 865–872, Mar. 2003, doi: 10.1016/S0140-6736(03)12715-8.
[163] J. P. A. Ioannidis, 'Why Most Published Research Findings Are False', *PLoS Med*, vol. 2, no. 8, p. e124, Aug. 2005, doi: 10.1371/journal.pmed.0020124.
[164] L. R. Jager and J. T. Leek, 'An estimate of the science-wise false discovery rate and application to the top medical literature', *Biostatistics*, vol. 15, no. 1, pp. 1–12, Jan. 2014, doi: 10.1093/biostatistics/kxt007.
[165] M. Schuurbiers *et al.*, 'Biological and technical factors in the assessment of blood-based tumor mutational burden (bTMB) in patients with NSCLC', *J Immunother Cancer*, vol. 10, no. 2, p. e004064, Feb. 2022, doi: 10.1136/jitc-2021-004064.
[166] S. Bustin and J. Huggett, 'qPCR primer design revisited', *Biomol Detect Quantif*, vol. 14, pp. 19–28, Dec. 2017, doi: 10.1016/j.bdq.2017.11.001.
[167] M. Smith, 'Validating Real-Time Polymerase Chain Reaction (PCR) Assays', in *Encyclopedia of Virology*, Elsevier, 2021, pp. 35–44. doi: 10.1016/B978-0-12-814515-9.00053-9.
[168] L. Garibyan and N. Avashia, 'Polymerase Chain Reaction', *Journal of Investigative Dermatology*, vol. 133, no. 3, pp. 1–4, Mar. 2013, doi: 10.1038/jid.2013.1.
[169] J. Tate and G. Ward, 'Interferences in immunoassay', *Clin Biochem Rev*, vol. 25, no. 2, pp. 105–120, May 2004.
[170] P. N. Robinson *et al.*, 'Improved exome prioritization of disease genes through cross-species phenotype comparison', *Genome Res.*, vol. 24, no. 2, pp. 340–348, Feb. 2014, doi: 10.1101/gr.160325.113.
[171] X. Lin, M. Liu, and X. Chen, 'Assessing reliability of protein-protein interactions by integrative analysis of data in model organisms', *BMC Bioinformatics*, vol. 10, no. S4, p. S5, Apr. 2009, doi: 10.1186/1471-2105-10-S4-S5.
[172] B. D. Fulcher, A. Arnatkeviciute, and A. Fornito, 'Overcoming false-positive gene-category enrichment in the analysis of spatially resolved transcriptomic brain atlas data', *Nat Commun*, vol. 12, no. 1, p. 2669, May 2021, doi: 10.1038/s41467-021-22862-1.
[173] E. Cano-Gamez and G. Trynka, 'From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases', *Front. Genet.*, vol. 11, p. 424, May 2020, doi: 10.3389/fgene.2020.00424.
[174] C. Quick, X. Wen, G. Abecasis, M. Boehnke, and H. M. Kang, 'Integrating comprehensive functional annotations to boost power and accuracy in gene-based association analysis', *PLoS Genet*, vol. 16, no. 12, p. e1009060, Dec. 2020, doi: 10.1371/journal.pgen.1009060.
[175] J. Chen, W. Hsu, M. L. Lee, and S.-K. Ng, 'Discovering reliable protein interactions from high-throughput experimental data using network topology', *Artif Intell Med*, vol. 35, no. 1–2, pp. 37–47, Oct. 2005, doi: 10.1016/j.artmed.2005.02.004.
[176] X. Lei and Y. Zhang, 'Predicting disease-genes based on network information loss and protein complexes in heterogeneous network', *Information Sciences*, vol. 479, pp. 386–400, Apr. 2019, doi: 10.1016/j.ins.2018.12.008.
[177] Y. Zhang *et al.*, 'Prioritizing disease genes with an improved dual label propagation framework', *BMC Bioinformatics*, vol. 19, no. 1, p. 47, Dec. 2018, doi: 10.1186/s12859-018-2040-6.
[178] J. Choi and H.-J. Kwon, 'The Information Filtering of Gene Network for Chronic Diseases: Social Network Perspective', *International Journal of Distributed Sensor Networks*, vol. 11, no. 9, p. 736569, Sep. 2015, doi: 10.1155/2015/736569.
[179] O. Kuchaiev, M. Rasajski, D. J. Higham, and N. Przulj, 'Geometric de-noising of protein-protein interaction networks', *PLoS Comput Biol*, vol. 5, no. 8, p. e1000454, Aug. 2009, doi:

10.1371/journal.pcbi.1000454.
[180] J. J. R. Burns *et al.*, 'Addressing noise in co-expression network construction', *Briefings in Bioinformatics*, vol. 23, no. 1, p. bbab495, Jan. 2022, doi: 10.1093/bib/bbab495.
[181] B. Klein, A. Swain, T. Byrum, S. V. Scarpino, and W. F. Fagan, 'Exploring noise, degeneracy and determinism in biological networks with the einet package', *Methods Ecol Evol*, vol. 13, no. 4, pp. 799–804, Apr. 2022, doi: 10.1111/2041-210X.13805.
[182] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, 'Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes', *The American Journal of Human Genetics*, vol. 78, no. 6, pp. 1011–1025, Jun. 2006, doi: 10.1086/504300.
[183] M. Ashburner *et al.*, 'Gene Ontology: tool for the unification of biology', *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.
[184] R. Goel, H. C. Harsha, A. Pandey, and T. S. K. Prasad, 'Human Protein Reference Database and Human Proteinpedia as resources for phosphoproteome analysis', *Mol. BioSyst.*, vol. 8, no. 2, pp. 453–463, 2012, doi: 10.1039/C1MB05340J.
[185] M. Kanehisa, 'KEGG: Kyoto Encyclopedia of Genes and Genomes', *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[186] A. Fabregat *et al.*, 'Reactome pathway analysis: a high-performance in-memory approach', *BMC Bioinformatics*, vol. 18, no. 1, p. 142, Dec. 2017, doi: 10.1186/s12859-017-1559-2.
[187] S. Orchard *et al.*, 'The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases', *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D358-363, Jan. 2014, doi: 10.1093/nar/gkt1115.
[188] N. Campbell, 'Genetic association database', *Nat Rev Genet*, vol. 5, no. 2, pp. 87–87, Feb. 2004, doi: 10.1038/nrg1288.
[189] M. J. Landrum *et al.*, 'ClinVar: improving access to variant interpretations and supporting evidence', *Nucleic Acids Research*, vol. 46, no. D1, pp. D1062–D1067, Jan. 2018, doi: 10.1093/nar/gkx1153.
[190] M. J. Li *et al.*, 'GWASdb v2: an update database for human genetic variants identified by genome-wide association studies', *Nucleic Acids Res*, vol. 44, no. D1, pp. D869–D876, Jan. 2016, doi: 10.1093/nar/gkv1317.
[191] J. Weile *et al.*, 'Bayesian integration of networks without gold standards', *Bioinformatics*, vol. 28, no. 11, pp. 1495–1500, Jun. 2012, doi: 10.1093/bioinformatics/bts154.
[192] K. James, A. Wipat, and J. Hallinan, 'Is newer better?—evaluating the effects of data curation on integrated analyses in Saccharomyces cerevisiae', *Integr. Biol.*, vol. 4, no. 7, pp. 715–727, 2012, doi: 10.1039/C2IB00123C.
[193] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, 'Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns', *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, no. 50, pp. 19033–19038, Dec. 2006, doi: 10.1073/pnas.0609152103.
[194] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, 'GeneRank: Using search engine technology for the analysis of microarray experiments', *BMC Bioinformatics*, vol. 6, no. 1, p. 233, Sep. 2005, doi: 10.1186/1471-2105-6-233.
[195] M. Zitnik, M. Agrawal, and J. Leskovec, 'Modeling polypharmacy side effects with graph convolutional networks', *Bioinformatics*, vol. 34, no. 13, pp. i457–i466, Jul. 2018, doi: 10.1093/bioinformatics/bty294.
[196] Z. Ahmed, S. Zeeshan, D. Mendhe, and X. Dong, 'Human gene and disease associations for clinical-genomics and precision medicine research', *Clinical and Translational Medicine*, vol. 10, no. 1, pp. 297–318, Mar. 2020, doi: 10.1002/ctm2.28.
[197] M. F. Berger and E. R. Mardis, 'The emerging clinical relevance of genomics in cancer medicine', *Nat Rev Clin Oncol*, vol. 15, no. 6, pp. 353–365, Jun. 2018, doi: 10.1038/s41571-018-0002-6.
[198] L. H. Goetz and N. J. Schork, 'Personalized medicine: motivation, challenges, and progress',

Fertility and Sterility, vol. 109, no. 6, pp. 952–963, Jun. 2018, doi: 10.1016/j.fertnstert.2018.05.006.
[199] A. Andermann and I. Blancquaert, 'Genetic screening: A primer for primary care', *Can Fam Physician*, vol. 56, no. 4, pp. 333–339, Apr. 2010.
[200] K. Sonehara and Y. Okada, 'Genomics-driven drug discovery based on disease-susceptibility genes', *Inflamm Regener*, vol. 41, no. 1, p. 8, Dec. 2021, doi: 10.1186/s41232-021-00158-7.
[201] J.-L. E. Pritchard, T. A. O'Mara, and D. M. Glubb, 'Enhancing the Promise of Drug Repositioning through Genetics', *Front. Pharmacol.*, vol. 8, p. 896, Dec. 2017, doi: 10.3389/fphar.2017.00896.
[202] A. Bodaghi, N. Fattahi, and A. Ramazani, 'Biomarkers: Promising and valuable tools towards diagnosis, prognosis and treatment of Covid-19 and other diseases', *Heliyon*, vol. 9, no. 2, p. e13323, Feb. 2023, doi: 10.1016/j.heliyon.2023.e13323.
[203] T. M. Frayling, 'Genome-wide association studies provide new insights into type 2 diabetes aetiology', *Nat Rev Genet*, vol. 8, no. 9, pp. 657–662, Sep. 2007, doi: 10.1038/nrg2178.
[204] K. Opap and N. Mulder, 'Recent advances in predicting gene-disease associations', *F1000Res*, vol. 6, p. 578, 2017, doi: 10.12688/f1000research.10788.1.
[205] M. Oti, S. Ballouz, and M. A. Wouters, 'Web tools for the prioritization of candidate disease genes', *Methods Mol Biol*, vol. 760, pp. 189–206, 2011, doi: 10.1007/978-1-61779-176-5_12.
[206] L.-C. Tranchevent, F. B. Capdevila, D. Nitsch, B. De Moor, P. De Causmaecker, and Y. Moreau, 'A guide to web tools to prioritize candidate genes', *Brief Bioinform*, vol. 12, no. 1, pp. 22–32, Jan. 2011, doi: 10.1093/bib/bbq007.
[207] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, 'Similarity computation strategies in the microRNA-disease network: a survey', *Brief Funct Genomics*, vol. 15, no. 1, pp. 55–64, Jan. 2016, doi: 10.1093/bfgp/elv024.
[208] A. Krishnan *et al.*, 'Genome-wide prediction and functional characterization of the genetic basis of autism spectrum disorder', *Nat Neurosci*, vol. 19, no. 11, pp. 1454–1462, Nov. 2016, doi: 10.1038/nn.4353.
[209] 'A deep learning approach to identify association of disease–gene using information of disease symptoms and protein sequences. Analytical Methods'.
[210] K. Lage *et al.*, 'A human phenome-interactome network of protein complexes implicated in genetic disorders', *Nat Biotechnol*, vol. 25, no. 3, pp. 309–316, Mar. 2007, doi: 10.1038/nbt1295.
[211] Q. Zou, J. Li, C. Wang, and X. Zeng, 'Approaches for Recognizing Disease Genes Based on Network', *BioMed Research International*, vol. 2014, pp. 1–10, 2014, doi: 10.1155/2014/416323.
[212] X. Chen *et al.*, 'A deep learning approach to identify association of disease–gene using information of disease symptoms and protein sequences', *Anal. Methods*, vol. 12, no. 15, pp. 2016–2026, 2020, doi: 10.1039/C9AY02333J.
[213] J. Li, X. Zhu, and J. Y. Chen, 'Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts', *PLoS Comput Biol*, vol. 5, no. 7, p. e1000450, Jul. 2009, doi: 10.1371/journal.pcbi.1000450.
[214] X. Zeng, Y. Liao, Y. Liu, and Q. Zou, 'Prediction and Validation of Disease Genes Using HeteSim Scores', *IEEE/ACM Trans. Comput. Biol. and Bioinf.*, vol. 14, no. 3, pp. 687–695, May 2017, doi: 10.1109/TCBB.2016.2520947.
[215] X. Zhang *et al.*, 'The expanded human disease network combining protein–protein interaction information', *Eur J Hum Genet*, vol. 19, no. 7, pp. 783–788, Jul. 2011, doi: 10.1038/ejhg.2011.30.
[216] V. Renganathan, 'Text Mining in Biomedical Domain with Emphasis on Document Clustering', *Healthc Inform Res*, vol. 23, no. 3, p. 141, 2017, doi: 10.4258/hir.2017.23.3.141.
[217] M. Magrane and U. Consortium, 'UniProt Knowledgebase: a hub of integrated protein data', *Database*, vol. 2011, no. 0, pp. bar009–bar009, Mar. 2011, doi: 10.1093/database/bar009.
[218] P. J. Thul and C. Lindskog, 'The human protein atlas: A spatial map of the human proteome',

Protein Science, vol. 27, no. 1, pp. 233–244, Jan. 2018, doi: 10.1002/pro.3307.
[219] H. L. Rehm *et al.*, 'ClinGen — The Clinical Genome Resource', *N Engl J Med*, vol. 372, no. 23, pp. 2235–2242, Jun. 2015, doi: 10.1056/NEJMsr1406261.
[220] O. Ursu *et al.*, 'DrugCentral: online drug compendium', *Nucleic Acids Res*, vol. 45, no. D1, pp. D932–D939, Jan. 2017, doi: 10.1093/nar/gkw993.
[221] D. S. Wishart *et al.*, 'DrugBank: a knowledgebase for drugs, drug actions and drug targets', *Nucleic Acids Research*, vol. 36, no. suppl_1, pp. D901–D906, Jan. 2008, doi: 10.1093/nar/gkm958.
[222] D. Welter *et al.*, 'The NHGRI GWAS Catalog, a curated resource of SNP-trait associations', *Nucl. Acids Res.*, vol. 42, no. D1, pp. D1001–D1006, Jan. 2014, doi: 10.1093/nar/gkt1229.
[223] S. Köhler *et al.*, 'The Human Phenotype Ontology in 2021', *Nucleic Acids Research*, vol. 49, no. D1, pp. D1207–D1217, Jan. 2021, doi: 10.1093/nar/gkaa1043.
[224] M. Kotlyar, C. Pastrello, N. Sheahan, and I. Jurisica, 'Integrated interactions database: tissue-specific view of the human and model organism interactomes', *Nucleic Acids Res*, vol. 44, no. D1, pp. D536–D541, Jan. 2016, doi: 10.1093/nar/gkv1115.
[225] T. N. Kipf and M. Welling, 'Semi-Supervised Classification with Graph Convolutional Networks', 2016, doi: 10.48550/ARXIV.1609.02907.
[226] W. L. Hamilton, R. Ying, and J. Leskovec, 'Inductive Representation Learning on Large Graphs', 2017, doi: 10.48550/ARXIV.1706.02216.
[227] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, 'Graph Attention Networks', 2017, doi: 10.48550/ARXIV.1710.10903.
[228] 'Deotarse P. P.1, Jain A. S.1, Baile. M. B, et al. Drug repositioning: a review. Int J Pharma Res Rev. 2015;4:51–58'.
[229] 'Sertkaya A, Birkenbach A, Berlind A, Eyraud J. Examination of clinical trial costs and barriers for drug development. US Department of health and human services, office of the assistant secretary for planning and evaluation report. 2014;1:1–92'.
[230] A. Mullard, 'New drugs cost US$2.6 billion to develop', *Nat Rev Drug Discov*, vol. 13, no. 12, pp. 877–877, Dec. 2014, doi: 10.1038/nrd4507.
[231] D. Sun, W. Gao, H. Hu, and S. Zhou, 'Why 90% of clinical drug development fails and how to improve it?', *Acta Pharmaceutica Sinica B*, vol. 12, no. 7, pp. 3049–3062, Jul. 2022, doi: 10.1016/j.apsb.2022.02.002.
[232] E. Kim, J. Yang, S. Park, and K. Shin, 'Factors Affecting Success of New Drug Clinical Trials', *Ther Innov Regul Sci*, vol. 57, no. 4, pp. 737–750, Jul. 2023, doi: 10.1007/s43441-023-00509-1.
[233] C. A. Umscheid, D. J. Margolis, and C. E. Grossman, 'Key concepts of clinical trials: a narrative review', *Postgrad Med*, vol. 123, no. 5, pp. 194–204, Sep. 2011, doi: 10.3810/pgm.2011.09.2475.
[234] D. Sardana, C. Zhu, M. Zhang, R. C. Gudivada, L. Yang, and A. G. Jegga, 'Drug repositioning for orphan diseases', *Brief Bioinform*, vol. 12, no. 4, pp. 346–356, Jul. 2011, doi: 10.1093/bib/bbr021.
[235] T. T. Ashburn and K. B. Thor, 'Drug repositioning: identifying and developing new uses for existing drugs', *Nat Rev Drug Discov*, vol. 3, no. 8, pp. 673–683, Aug. 2004, doi: 10.1038/nrd1468.
[236] H. Xue, J. Li, H. Xie, and Y. Wang, 'Review of Drug Repositioning Approaches and Resources', *Int. J. Biol. Sci.*, vol. 14, no. 10, pp. 1232–1244, 2018, doi: 10.7150/ijbs.24612.
[237] T. Pillaiyar, S. Meenakshisundaram, M. Manickam, and M. Sankaranarayanan, 'A medicinal chemistry perspective of drug repositioning: Recent advances and challenges in drug discovery', *European Journal of Medicinal Chemistry*, vol. 195, p. 112275, Jun. 2020, doi: 10.1016/j.ejmech.2020.112275.
[238] M. K. Tripathi, S. Sharma, T. P. Singh, A. S. Ethayathulla, and P. Kaur, 'Computational Intelligence in Drug Repurposing for COVID-19', in *Computational Intelligence Methods in COVID-19: Surveillance, Prevention, Prediction and Diagnosis*, vol. 923, K. Raza, Ed., in Studies in

Computational Intelligence, vol. 923. , Singapore: Springer Singapore, 2021, pp. 273–294. doi: 10.1007/978-981-15-8534-0_14.
[239] A. Cakmak and G. Ozsoyoglu, 'Mining biological networks for unknown pathways', *Bioinformatics*, vol. 23, no. 20, pp. 2775–2783, Oct. 2007, doi: 10.1093/bioinformatics/btm409.
[240] 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, no. 5, p. e0155811, May 2016, doi: 10.1371/journal.pone.0155811.
[241] R. Liu, L. Wei, and P. Zhang, 'A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data', *Nat Mach Intell*, vol. 3, no. 1, pp. 68–75, Jan. 2021, doi: 10.1038/s42256-020-00276-w.
[242] F. Yang *et al.*, 'Machine Learning Applications in Drug Repurposing', *Interdiscip Sci Comput Life Sci*, vol. 14, no. 1, pp. 15–21, Mar. 2022, doi: 10.1007/s12539-021-00487-8.
[243] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, 'PREDICT: a method for inferring novel drug indications with application to personalized medicine', *Mol Syst Biol*, vol. 7, no. 1, p. 496, Jan. 2011, doi: 10.1038/msb.2011.26.
[244] 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, no. 11, p. e78518, Nov. 2013, doi: 10.1371/journal.pone.0078518.
[245] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, 'DeepTox: Toxicity Prediction using Deep Learning', *Front. Environ. Sci.*, vol. 3, Feb. 2016, doi: 10.3389/fenvs.2015.00080.
[246] J. Xiong, Z. Xiong, K. Chen, H. Jiang, and M. Zheng, 'Graph neural networks for automated de novo drug design', *Drug Discovery Today*, vol. 26, no. 6, pp. 1382–1393, Jun. 2021, doi: 10.1016/j.drudis.2021.02.011.
[247] F. Cheng *et al.*, 'Prediction of drug-target interactions and drug repositioning via network-based inference', *PLoS Comput Biol*, vol. 8, no. 5, p. e1002503, 2012, doi: 10.1371/journal.pcbi.1002503.
[248] C. Wu, R. C. Gudivada, B. J. Aronow, and A. G. Jegga, 'Computational drug repositioning through heterogeneous network clustering', *BMC Syst Biol*, vol. 7, no. S5, p. S6, Dec. 2013, doi: 10.1186/1752-0509-7-S5-S6.
[249] H. Haeberle, J. T. Dudley, J. T. C. Liu, A. J. Butte, and C. H. Contag, 'Identification of cell surface targets through meta-analysis of microarray data', *Neoplasia*, vol. 14, no. 7, pp. 666–669, Jul. 2012, doi: 10.1593/neo.12634.
[250] B. Chen, Y. Ding, and D. J. Wild, 'Assessing Drug Target Association Using Semantic Linked Data', *PLoS Comput Biol*, vol. 8, no. 7, p. e1002574, Jul. 2012, doi: 10.1371/journal.pcbi.1002574.
[251] J. Li and Z. Lu, 'Systematic identification of pharmacogenomics information from clinical trials', *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 870–878, Oct. 2012, doi: 10.1016/j.jbi.2012.04.005.
[252] F. Ehrhart, E. L. Willighagen, M. Kutmon, M. van Hoften, L. M. G. Curfs, and C. T. Evelo, 'A resource to explore the discovery of rare diseases and their causative genes', *Sci Data*, vol. 8, no. 1, p. 124, May 2021, doi: 10.1038/s41597-021-00905-y.
[253] M. Gillespie *et al.*, 'The reactome pathway knowledgebase 2022', *Nucleic Acids Research*, vol. 50, no. D1, pp. D687–D692, Jan. 2022, doi: 10.1093/nar/gkab1028.
[254] J. Malone *et al.*, 'Modeling sample variables with an Experimental Factor Ontology', *Bioinformatics*, vol. 26, no. 8, pp. 1112–1118, Apr. 2010, doi: 10.1093/bioinformatics/btq099.
[255] M. Black *et al.*, 'EDAM: the bioscientific data analysis ontology (update 2021)', 2022, doi: 10.7490/F1000RESEARCH.1118900.1.
[256] NCBI Resource Coordinators, 'Database Resources of the National Center for Biotechnology Information', *Nucleic Acids Res*, vol. 45, no. D1, pp. D12–D17, Jan. 2017, doi: 10.1093/nar/gkw1071.
[257] A. S. Brown and C. J. Patel, 'A standard database for drug repositioning', *Sci Data*, vol. 4, no. 1, p. 170029, Mar. 2017, doi: 10.1038/sdata.2017.29.

[258] P. Shannon *et al.*, 'Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks', *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.
[259] '59. Dongen S M V. Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000(http://www.library.uu.nl/digiarchief/dip/diss/1895620/full.pdf or http://micans.org/mcl/lit/svdthesis.pdf.gz'.
[260] C. Zheng and R. Xu, 'Large-scale mining disease comorbidity relationships from post-market drug adverse events surveillance data', *BMC Bioinformatics*, vol. 19, no. S17, p. 500, Dec. 2018, doi: 10.1186/s12859-018-2468-8.
[261] A. P. Davis *et al.*, 'The Comparative Toxicogenomics Database: update 2019', *Nucleic Acids Research*, vol. 47, no. D1, pp. D948–D954, Jan. 2019, doi: 10.1093/nar/gky868.
[262] L. M. Schriml *et al.*, 'The Human Disease Ontology 2022 update', *Nucleic Acids Research*, vol. 50, no. D1, pp. D1255–D1261, Jan. 2022, doi: 10.1093/nar/gkab1063.
[263] C. Guangchuang Yu [Aut, *DOSE*. (2017). [object Object]. doi: 10.18129/B9.BIOC.DOSE.
[264] The Gene Ontology Consortium, 'The Gene Ontology Resource: 20 years and still GOing strong', *Nucleic Acids Research*, vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.
[265] C. Guangchuang Yu [Aut, *GOSemSim*. (2017). [object Object]. doi: 10.18129/B9.BIOC.GOSEMSIM.
[266] G. Tosadori, I. Bestvina, F. Spoto, C. Laudanna, and G. Scardoni, 'Creating, generating and comparing random network models with Network Randomizer', *F1000Res*, vol. 5, p. 2524, Oct. 2016, doi: 10.12688/f1000research.9203.1.
[267] F. Iorio, M. Bernardo-Faura, A. Gobbi, T. Cokelaer, G. Jurman, and J. Saez-Rodriguez, 'Efficient randomization of biological networks while preserving functional characterization of individual nodes', *BMC Bioinformatics*, vol. 17, no. 1, p. 542, Dec. 2016, doi: 10.1186/s12859-016-1402-1.
[268] A. Garcia-Moreno *et al.*, 'Functional Enrichment Analysis of Regulatory Elements', *Biomedicines*, vol. 10, no. 3, p. 590, Mar. 2022, doi: 10.3390/biomedicines10030590.
[269] S. F. [Ctb] Robert Gentleman [Aut], *GOstats*. (2017). [object Object]. doi: 10.18129/B9.BIOC.GOSTATS.
[270] O. Kart, E. Hayirci, A. Kut, and Z. Isik, 'Supervised Link Prediction Developed For Bipartite Social Networks', in *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, Istanbul Turkey: ACM, Oct. 2019, pp. 14–17. doi: 10.1145/3369114.3369134.
[271] Y. Lu, Y. Guo, and A. Korhonen, 'Link prediction in drug-target interactions network using similarity indices', *BMC Bioinformatics*, vol. 18, no. 1, p. 39, Jan. 2017, doi: 10.1186/s12859-017-1460-z.
[272] 'Kohavi, Ron. "A study of cross-validation and bootstrap for accuracy estimation and model selection." Ijcai. Vol. 14. No. 2. 1995.',
[273] M. Stone, 'Cross-Validatory Choice and Assessment of Statistical Predictions', *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 36, no. 2, pp. 111–133, Jan. 1974, doi: 10.1111/j.2517-6161.1974.tb00994.x.
[274] K. H. Zou, A. J. O'Malley, and L. Mauri, 'Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models', *Circulation*, vol. 115, no. 5, pp. 654–657, Feb. 2007, doi: 10.1161/CIRCULATIONAHA.105.594929.
[275] F. Nielsen, 'Hierarchical Clustering', in *Introduction to HPC with MPI for Data Science*, in Undergraduate Topics in Computer Science. , Cham: Springer International Publishing, 2016, pp. 195–211. doi: 10.1007/978-3-319-21903-5_8.
[276] R. Vijay, P. Mahajan, and R. Kandwal, 'Hamming Distance based Clustering Algorithm':, *International Journal of Information Retrieval Research*, vol. 2, no. 1, pp. 11–20, Jan. 2012, doi: 10.4018/ijirr.2012010102.
[277] P. R. Carvalho, C. S. Munita, and A. L. Lapolli, 'Validity studies among hierarchical methods of cluster analysis using cophenetic correlation coefficient', *Braz. J. Rad. Sci.*, vol. 7, no. 2A, Feb. 2019, doi: 10.15392/bjrs.v7i2A.668.

[278] K. R. Shahapure and C. Nicholas, 'Cluster Quality Analysis Using Silhouette Score', in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, sydney, Australia: IEEE, Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.

[279] Angur Mahmud Jarman, 'Hierarchical Cluster Analysis: Comparison of Single linkage,Complete linkage, Average linkage and Centroid Linkage Method', 2020, doi: 10.13140/RG.2.2.11388.90240.

[280] W. L. Hamilton, R. Ying, and J. Leskovec, 'Inductive Representation Learning on Large Graphs', 2017, doi: 10.48550/ARXIV.1706.02216.

[281] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[282] N. A. Vasilevsky *et al.*, 'Mondo: Unifying diseases for the world, by the world', Health Informatics, preprint, Apr. 2022. doi: 10.1101/2022.04.13.22273750.

[283] O. Bodenreider, 'The Unified Medical Language System (UMLS): integrating biomedical terminology', *Nucleic Acids Research*, vol. 32, no. 90001, pp. 267D – 270, Jan. 2004, doi: 10.1093/nar/gkh061.

[284] B. Chen, J. Wang, M. Li, and F.-X. Wu, 'Identifying disease genes by integrating multiple data sources', *BMC Med Genomics*, vol. 7, no. S2, p. S2, Dec. 2014, doi: 10.1186/1755-8794-7-S2-S2.

[285] U. M. Singh-Blom, N. Natarajan, A. Tewari, J. O. Woods, I. S. Dhillon, and E. M. Marcotte, 'Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses', *PLoS ONE*, vol. 8, no. 5, p. e58977, May 2013, doi: 10.1371/journal.pone.0058977.

[286] I. Barrio-Hernandez, J. Schwartzentruber, A. Shrivastava et al. "Network expansion of genetic associations defines a pleiotropy map of human cell biology". Nat Genet 55, 389–398 (2023). https://doi.org/10.1038/s41588-023-01327-9

[287] J. He et al., 'Enhancing disease risk gene discovery by integrating transcription factor-linked trans-variants into transcriptome-wide association analyses', Nucleic Acids Res, vol. 53, no. 1, p. gkae1035, Jan. 2025, doi: 10.1093/nar/gkae1035.

[288] G. Galindez, S. Sadegh, J. Baumbach, T. Kacprowski, and M. List, 'Network-based approaches for modeling disease regulation and progression', Comput Struct Biotechnol J, vol. 21, pp. 780–795, 2023, doi: 10.1016/j.csbj.2022.12.022.

[289] Jochen Weile, Katherine James, Jennifer Hallinan, Simon J. Cockell, Phillip Lord, Anil Wipat, Darren J. Wilkinson, 'Bayesian integration of networks without gold standards', Bioinformatics, Volume 28, Issue 11, 1 June 2012, Pages 1495–1500, https://doi.org/10.1093/bioinformatics/bts154'.

[290] Katherine James, Anil Wipat, Jennifer Hallinan, 'Is newer better?—evaluating the effects of data curation on integrated analyses in Saccharomyces cerevisiae', Integrative Biology, Volume 4, Issue 7, July 2012, Pages 715–727, https://doi.org/10.1039/c2ib00123c.

[291] I. Condò, 'Rare Monogenic Diseases: Molecular Pathophysiology and Novel Therapies', *IJMS*, vol. 23, no. 12, p. 6525, Jun. 2022, doi: 10.3390/ijms23126525.

[292] J. R. Cardoso, L. M. Pereira, M. D. Iversen, and A. L. Ramos, 'What is gold standard and what is ground truth?', *Dental Press J. Orthod.*, vol. 19, no. 5, pp. 27–30, Oct. 2014, doi: 10.1590/2176-9451.19.5.027-030.ebo.

[293] A. K. Bajpai *et al.*, 'Systematic comparison of the protein-protein interaction databases from a user's perspective', *Journal of Biomedical Informatics*, vol. 103, p. 103380, Mar. 2020, doi: 10.1016/j.jbi.2020.103380.

[294] J. Zhao, Y. Lei, J. Hong, C. Zheng, and L. Zhang, 'AraPPINet: An Updated Interactome for the Analysis of Hormone Signaling Crosstalk in Arabidopsis thaliana', *Front. Plant Sci.*, vol. 10, p. 870, Jul. 2019, doi: 10.3389/fpls.2019.00870.

[295] S. Raj, A. Vishnoi, and A. Srivastava, 'GOLD standard dataset for Alzheimer genes', *Data in Brief*, vol. 30, p. 105439, Jun. 2020, doi: 10.1016/j.dib.2020.105439.

[296] N. Soltani, K. A. Stevens, V. Klaassen, M.-S. Hwang, D. A. Golino, and M. Al Rwahnih, 'Quality Assessment and Validation of High-Throughput Sequencing for Grapevine Virus Diagnostics', *Viruses*, vol. 13, no. 6, p. 1130, Jun. 2021, doi: 10.3390/v13061130.
[297] F. F. D. R. Vicente, F. M. Lopes, and R. F. Hashimoto, 'Improvement of GNs inference through biological data integration', in *2011 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, San Antonio, TX, USA: IEEE, Dec. 2011, pp. 70–73. doi: 10.1109/GENSiPS.2011.6169446.
[298] Y. Lu, Y. Guo, and A. Korhonen, 'Link prediction in drug-target interactions network using similarity indices', *BMC Bioinformatics*, vol. 18, no. 1, p. 39, Dec. 2017, doi: 10.1186/s12859-017-1460-z.
[299] 'Stijn van Dongen, Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, May 2000. (http://www.library.uu.nl/digiarchief/dip/diss/1895620/full.pdf or http://micans.org/mcl/lit/svdthesis.pdf.gz)'.
[300] M. E. Hatten and M. F. Roussel, 'Development and cancer of the cerebellum', *Trends in Neurosciences*, vol. 34, no. 3, pp. 134–142, Mar. 2011, doi: 10.1016/j.tins.2011.01.002.
[301] Y. H. Youn and Y.-G. Han, 'Primary Cilia in Brain Development and Diseases', *The American Journal of Pathology*, vol. 188, no. 1, pp. 11–22, Jan. 2018, doi: 10.1016/j.ajpath.2017.08.031.
[302] D. N. Louis *et al.*, 'The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary', *Acta Neuropathol*, vol. 131, no. 6, pp. 803–820, Jun. 2016, doi: 10.1007/s00401-016-1545-1.
[303] M. D. Taylor *et al.*, 'Molecular subgroups of medulloblastoma: the current consensus', *Acta Neuropathol*, vol. 123, no. 4, pp. 465–472, Apr. 2012, doi: 10.1007/s00401-011-0922-z.
[304] S. Goutagny *et al.*, 'Long-term follow-up of 287 meningiomas in neurofibromatosis type 2 patients: clinical, radiological, and molecular features', *Neuro-Oncology*, vol. 14, no. 8, pp. 1090–1096, Aug. 2012, doi: 10.1093/neuonc/nos129.
[305] M. Sinyuk, E. E. Mulkearns-Hubert, O. Reizes, and J. Lathia, 'Cancer Connectors: Connexins, Gap Junctions, and Communication', *Front. Oncol.*, vol. 8, p. 646, Dec. 2018, doi: 10.3389/fonc.2018.00646.
[306] E. Peixoto, S. Richard, K. Pant, A. Biswas, and S. A. Gradilone, 'The primary cilium: Its role as a tumor suppressor organelle', *Biochemical Pharmacology*, vol. 175, p. 113906, May 2020, doi: 10.1016/j.bcp.2020.113906.
[307] H.-J. Lee, S.-H. Shim, M.-R. Song, H. Lee, and J. C. Park, 'CoMAGC: a corpus with multi-faceted annotations of gene-cancer relations', *BMC Bioinformatics*, vol. 14, no. 1, p. 323, Dec. 2013, doi: 10.1186/1471-2105-14-323.
[308] K James, A Alsobhe, M Pocock, A Wipat, S. J. Cockell, 'A multi-faceted gold standard reduces data loss during probabilistic integration of disease-gene associations'.
[309] L. Lu and T. Zhou, 'Link Prediction in Complex Networks: A Survey', *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, Mar. 2011, doi: 10.1016/j.physa.2010.11.027.
[310] S. Navlakha and C. Kingsford, 'The power of protein interaction networks for associating genes with diseases', *Bioinformatics*, vol. 26, no. 8, pp. 1057–1063, Apr. 2010, doi: 10.1093/bioinformatics/btq076.
[311] V. N. Ioannidis, D. Zheng, and G. Karypis, 'Few-shot link prediction via graph neural networks for Covid-19 drug-repurposing', 2020, doi: 10.48550/ARXIV.2007.10261.
[312] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, 'The SIDER database of drugs and side effects', *Nucleic Acids Res*, vol. 44, no. D1, pp. D1075–D1079, Jan. 2016, doi: 10.1093/nar/gkv1075.
[313] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy, 'BioPortal as a Dataset of Linked Biomedical Ontologies and Terminologies in RDF', *Semant Web*, vol. 4, no. 3, pp. 277–284, 2013.
[314] J. Chambers *et al.*, 'UniChem: a unified chemical structure cross-referencing and identifier tracking system', *J Cheminform*, vol. 5, no. 1, p. 3, Dec. 2013, doi: 10.1186/1758-2946-5-3.

[315] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', 2018, doi: 10.48550/ARXIV.1810.04805.
[316] S. Ioffe and C. Szegedy, 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift', 2015, doi: 10.48550/ARXIV.1502.03167.
[317] 'Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." The journal of machine learning research 15.1 (2014): 1929-1958.'.
[318] D. P. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization', 2014, doi: 10.48550/ARXIV.1412.6980.
[319] K. A. Spackman, 'SIGNAL DETECTION THEORY: VALUABLE TOOLS FOR EVALUATING INDUCTIVE LEARNING', in *Proceedings of the Sixth International Workshop on Machine Learning*, Elsevier, 1989, pp. 160–163. doi: 10.1016/B978-1-55860-036-2.50047-3.
[320] J. Davis and M. Goadrich, 'The relationship between Precision-Recall and ROC curves', in *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.
[321] A. K. Dhingra, B. Chopra, A. Jain, and J. Chaudhary, 'Imidazole: Multi-targeted Therapeutic Leads for the Management ofAlzheimer's Disease', *MRMC*, vol. 22, no. 10, pp. 1352–1373, Jun. 2022, doi: 10.2174/1389557522666220104152141.
[322] L. Bergamaschini *et al.*, 'Peripheral Treatment with Enoxaparin, a Low Molecular Weight Heparin, Reduces Plaques and β-Amyloid Accumulation in a Mouse Model of Alzheimer's Disease', *J. Neurosci.*, vol. 24, no. 17, pp. 4181–4186, Apr. 2004, doi: 10.1523/JNEUROSCI.0550-04.2004.
[323] A. K. Desai and G. T. Grossberg, 'Buspirone in Alzheimer's disease', *Expert Review of Neurotherapeutics*, vol. 3, no. 1, pp. 19–28, Jan. 2003, doi: 10.1586/14737175.3.1.19.
[324] H. Nassar, W. Sippl, R. A. Dahab, and M. Taha, 'Molecular docking, molecular dynamics simulations and *in vitro* screening reveal cefixime and ceftriaxone as GSK3β covalent inhibitors', *RSC Adv.*, vol. 13, no. 17, pp. 11278–11290, 2023, doi: 10.1039/D3RA01145C.
[325] X. Li *et al.*, 'A tricyclic antidepressant, amoxapine, reduces amyloid-β generation through multiple serotonin receptor 6-mediated targets', *Sci Rep*, vol. 7, no. 1, p. 4983, Jul. 2017, doi: 10.1038/s41598-017-04144-3.
[326] F. La Rosa *et al.*, 'Modulation of MAPK- and PI3/AKT-Dependent Autophagy Signaling by Stavudine (D4T) in PBMC of Alzheimer's Disease Patients', *Cells*, vol. 11, no. 14, p. 2180, Jul. 2022, doi: 10.3390/cells11142180.
[327] J. D. Hoffman *et al.*, 'Dietary inulin alters the gut microbiome, enhances systemic metabolism and reduces neuroinflammation in an APOE4 mouse model', *PLoS ONE*, vol. 14, no. 8, p. e0221828, Aug. 2019, doi: 10.1371/journal.pone.0221828.
[328] P. Wang, P. Wang, H. Luan, Y. Wu, and Y. Chen, 'Midazolam alleviates cellular senescence in SH-SY5Y neuronal cells in Alzheimer's disease', *Brain and Behavior*, vol. 13, no. 1, p. e2822, Jan. 2023, doi: 10.1002/brb3.2822.
[329] S. Hira, U. Saleem, F. Anwar, Z. Raza, A. U. Rehman, and B. Ahmad, 'In Silico Study and Pharmacological Evaluation of Eplerinone as an Anti-Alzheimer's Drug in STZ-Induced Alzheimer's Disease Model', *ACS Omega*, vol. 5, no. 23, pp. 13973–13983, Jun. 2020, doi: 10.1021/acsomega.0c01381.
[330] C. A. Langford, 'Cyclophosphamide as induction therapy for Wegener's granulomatosis and microscopic polyangiitis', *Clinical and Experimental Immunology*, vol. 164, no. Supplement_1, pp. 31–34, Mar. 2011, doi: 10.1111/j.1365-2249.2011.04364.x.
[331] X. Puéchal *et al.*, 'Rituximab vs Cyclophosphamide Induction Therapy for Patients With Granulomatosis With Polyangiitis', *JAMA Netw Open*, vol. 5, no. 11, p. e2243799, Nov. 2022, doi: 10.1001/jamanetworkopen.2022.43799.
[332] R. A. Brodsky *et al.*, 'High-dose cyclophosphamide for severe aplastic anemia: long-term follow-up', *Blood*, vol. 115, no. 11, pp. 2136–2141, Mar. 2010, doi: 10.1182/blood-2009-06-225375.

[333] C. Rodrigo, S. Rajapakse, and L. Gooneratne, ‘Rituximab in the treatment of autoimmune haemolytic anaemia’, *Brit J Clinical Pharma*, vol. 79, no. 5, pp. 709–719, May 2015, doi: 10.1111/bcp.12498.
[334] B. Fattizzo, A. Zaninoni, L. Pettine, F. Cavallaro, E. Di Bona, and W. Barcellini, ‘Low-dose rituximab in autoimmune hemolytic anemia: 10 years after’, *Blood*, vol. 133, no. 9, pp. 996–998, Feb. 2019, doi: 10.1182/blood-2018-12-885228.
[335] L. A. Zilber, Z. L. Bajdakova, A. N. Gardasjan, N. V. Konovalov, T. L. Bunina, and E. M. Barabadze, ‘THE PREVENTION AND TREATMENT OF ISONIAZID TOXICITY IN THE THERAPY OF PULMONARY TUBERCULOSIS. 2. AN ASSESSMENT OF THE PROPHYLACTIC EFFECT OF PYRIDOXINE IN LOW DOSAGE’, *Bull World Health Organ*, vol. 29, no. 4, pp. 457–481, 1963.
[336] T. Dick, U. Manjunatha, B. Kappes, and M. Gengenbacher, ‘Vitamin B6 biosynthesis is essential for survival and virulence of Mycobacterium tuberculosis: Vitamin B6 biosynthesis of the tubercle bacillus’, *Molecular Microbiology*, vol. 78, no. 4, pp. 980–988, Nov. 2010, doi: 10.1111/j.1365-2958.2010.07381.x.
[337] C. Chen *et al.*, ‘Verapamil Targets Membrane Energetics in Mycobacterium tuberculosis’, *Antimicrob Agents Chemother*, vol. 62, no. 5, pp. e02107-17, May 2018, doi: 10.1128/AAC.02107-17.
[338] R. S. Obach, P. Huynh, M. C. Allen, and C. Beedham, ‘Human Liver Aldehyde Oxidase: Inhibition by 239 Drugs’, *The Journal of Clinical Pharma*, vol. 44, no. 1, pp. 7–19, Jan. 2004, doi: 10.1177/0091270003260336.
[339] R. Schwartz and N. O. Kjeldgaard, ‘The enzymic oxidation of pyridoxal by liver aldehyde oxidase’, *Biochemical Journal*, vol. 48, no. 3, pp. 333–337, Mar. 1951, doi: 10.1042/bj0480333.
[340] S. Agadi, M. M. Quach, and Z. Haneef, ‘Vitamin-Responsive Epileptic Encephalopathies in Children’, *Epilepsy Research and Treatment*, vol. 2013, pp. 1–8, Jul. 2013, doi: 10.1155/2013/510529.
[341] S. Lakshmikanthcharan, S. K. Chaitanya Juluri, and S. M. Nandakumar, ‘Verapamil as an Adjuvant Treatment for Drug-Resistant Epilepsy’, *Indian Journal of Critical Care Medicine*, vol. 22, no. 9, pp. 680–682, Sep. 2018, doi: 10.4103/ijccm.IJCCM_250_18.
[342] Y. Guo *et al.*, ‘Rituximab in patients with membranous nephropathy and kidney insufficiency’, *Front. Pharmacol.*, vol. 13, p. 1002117, Oct. 2022, doi: 10.3389/fphar.2022.1002117.
[343] L. Gerez *et al.*, ‘Regulation of interleukin-2 and interferon-γ gene expression in renal failure’, *Kidney International*, vol. 40, no. 2, pp. 266–272, Aug. 1991, doi: 10.1038/ki.1991.209.
[344] M. Kato, ‘New insights into IFN-γ in rheumatoid arthritis: role in the era of JAK inhibitors’, *Immunological Medicine*, vol. 43, no. 2, pp. 72–78, Apr. 2020, doi: 10.1080/25785826.2020.1751908.
[345] C. C. Mok, ‘Rituximab for the treatment of rheumatoid arthritis: an update’, *DDDT*, p. 87, Dec. 2013, doi: 10.2147/DDDT.S41645.
[346] J. M. Llovet *et al.*, ‘Hepatocellular carcinoma’, *Nat Rev Dis Primers*, vol. 7, no. 1, p. 6, Jan. 2021, doi: 10.1038/s41572-020-00240-3.
[347] R. Saxena, N. D. Theise, and J. M. Crawford, ‘Microanatomy of the human liver?exploring the hidden interfaces’, *Hepatology*, vol. 30, no. 6, pp. 1339–1346, Dec. 1999, doi: 10.1002/hep.510300607.
[348] K. Fujita, ‘Irinotecan, a key chemotherapeutic drug for metastatic colorectal cancer’, *WJG*, vol. 21, no. 43, p. 12234, 2015, doi: 10.3748/wjg.v21.i43.12234.
[349] L.-M. Liu, D.-D. Xiong, P. Lin, H. Yang, Y.-W. Dang, and G. Chen, ‘DNA topoisomerase 1 and 2A function as oncogenes in liver cancer and may be direct targets of nitidine chloride’, *Int J Oncol*, Aug. 2018, doi: 10.3892/ijo.2018.4531.
[350] A. L. Bodley and L. F. Liu, ‘Topoisomerases as Novel Targets for Cancer Chemotherapy’, *Nat Biotechnol*, vol. 6, no. 11, pp. 1315–1319, Nov. 1988, doi: 10.1038/nbt1188-1315.
[351] C. Pozzo *et al.*, ‘Neoadjuvant treatment of unresectable liver disease with irinotecan and 5-fluorouracil plus folinic acid in colorectal cancer patients’, *Annals of Oncology*, vol. 15, no. 6,

pp. 933–939, Jun. 2004, doi: 10.1093/annonc/mdh217.
[352] D. Fanelli, ‘Negative results are disappearing from most disciplines and countries’, *Scientometrics*, vol. 90, no. 3, pp. 891–904, Mar. 2012, doi: 10.1007/s11192-011-0494-7.
[353] D. N. Sosa and R. B. Altman, ‘Contexts and contradictions: a roadmap for computational drug repurposing with knowledge inference’, *Briefings in Bioinformatics*, vol. 23, no. 4, p. bbac268, Jul. 2022, doi: 10.1093/bib/bbac268.
[354] A. Alexeyenko, T. Schmitt, A. Tjarnberg, D. Guala, O. Frings, and E. L. L. Sonnhammer, ‘Comparative interactomics with Funcoup 2.0’, *Nucleic Acids Research*, vol. 40, no. D1, pp. D821–D828, Jan. 2012, doi: 10.1093/nar/gkr1062.