POSITIVE LABELED AND UNLABELED LEARNING METHODS OF META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS

Capt. Thitipong Kawichai

A Dissertation Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in Applied Mathematics and Computational Science

Department of Mathematics and Computer Science

FACULTY OF SCIENCE

Chulalongkorn University

Academic Year 2020

วิธีการเรียนรู้แบบมีฉลากประเภทบวกและไม่มีฉลากของโพรไฟล์เชิงหน้าที่บนวิถีเมตาสำหรับการทำนายความสัมพันธ์ระหว่างยาและโรค

ร.อ.ฐิติพงษ์ กาวิชัย

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์และวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2563

Thesis Title	POSITIVE LABELED AND UNLABELED LEARNING METHODS OF META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS
By	Capt. Thitipong Kawichai
Field of Study	Applied Mathematics and Computational Science
Thesis Advisor	Assistant Professor KITIPORN PLAIMAS, Dr.rer.nat.
Thesis Co Advisor	Assistant Professor Apichat Suratanee, Dr.rer.nat.

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

..... Dean of the FACULTY OF SCIENCE
(Professor POLKIT SANGVANICH, Ph.D.)

DISSERTATION COMMITTEE

..... Chairman
(Associate Professor Treenut Saithong, Ph.D.)

..... Thesis Advisor
(Assistant Professor KITIPORN PLAIMAS, Dr.rer.nat.)

..... Thesis Co-Advisor
(Assistant Professor Apichat Suratanee, Dr.rer.nat.)

..... Examiner
(Assistant Professor KRUNG SINAPIROMSARAN, Ph.D.)

..... Examiner
(Assistant Professor JIRAPHAN SUNTORNCHOST, Ph.D.)

..... Examiner
(THAP PANITANARAK, Ph.D.)

ฐิติพงษ์ กาวิชัย : วิธีการเรียนรู้แบบมีฉลากประเภทบวกและไม่มีฉลากของโพรไฟล์เชิงหน้าที่บนวิถีเมตาสำหรับการทำนายความสัมพันธ์ระหว่างยาและโรค. (POSITIVE LABELED AND UNLABELED LEARNING METHODS OF META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS) อ.ที่ปรึกษาหลัก : ผศ. ดร.กิติพร พลายมาศ, อ.ที่ปรึกษาร่วม : ผศ. ดร.อภิชาต ศุรธณี

ดรักรีโพสิชันนิ่งหรือการค้นพบข้อบ่งชี้ใหม่สำหรับยาที่มีอยู่แล้วเป็นกลยุทธ์ที่สามารถช่วยลดระยะเวลา ค่าใช้จ่าย และความเสี่ยงในการค้นพบและพัฒนายาได้ วิธีเชิงคำนวณจำนวนมากจึงถูกพัฒนาขึ้นเพื่อใช้ระบุความสัมพันธ์ระหว่างยาและโรคสำหรับการตรวจสอบและพัฒนายาต่อไป แนวทางใหม่ที่มีประสิทธิภาพดีกว่าด้วยการใช้ข้อมูลที่น้อยกว่าคือแนวทางบนวิถีเมตา ซึ่งสร้างข้อมูลเชิงเครือข่ายโดยใช้รูปแบบวิถีจากโหนดยาไปยังโหนดโรค อย่างไรก็ตามวิธีบนวิถีเมตาที่มีอยู่แล้วละทิ้งข้อมูลของโหนดกลางตามวิถี ซึ่งเป็นตัวบ่งชี้ที่สำคัญสำหรับการอธิบายความสัมพันธ์ระหว่างยาและโรค งานวิจัยนี้จึงได้นำเสนอวิธีบนวิถีเมตารูปแบบใหม่ภายใต้การเรียนรู้แบบมีฉลากประเภทบวกและไม่มีฉลาก ยีนออนโทโลยีถูกใช้ในการเชื่อมต่อระหว่างยาและโรคในเครือข่ายไตรภาคีของยา ยีนออนโทโลยี และโรค คุณลักษณะของยาและโรคบนวิถีเมตารูปแบบใหม่หรือโพรไฟล์เชิงหน้าที่บนวิถีเมตาถูกสร้างขึ้นโดยการรวมข้อมูลเชิงยีนออนโทโลยีเข้าไปในโพรไฟล์เชิงหน้าที่ แบบจำลองแบบรวมกลุ่มถูกพัฒนาขึ้นบนโพรไฟล์เชิงหน้าที่ของทั้งตัวอย่างที่มีฉลากประเภทบวกและไม่มีฉลาก วิธีที่นำเสนอมีประสิทธิภาพที่ดีกว่าวิธีอื่นที่มีอยู่แล้วด้วยค่าเฉลี่ยของพื้นที่ใต้โค้งความแม่นยำและเรียกคืนเป็น 0.944 และค่าเฉลี่ยของพื้นที่ใต้โค้งอาร์โอซีเป็น 0.930 นอกจากนี้ความสัมพันธ์ระหว่างยาและโรคที่ถูกค้นพบใหม่ด้วยวิธีที่นำเสนอมากถึง 38% ถูกค้นเจอในฐานข้อมูลของการทดลองทางคลินิก

สาขาวิชา	คณิตศาสตร์ประยุกต์และวิทยาการคณนา	ลายมือชื่อนิสิต .....
ปีการศึกษา	2563	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....
		ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

6172811223 : MAJOR APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCE

KEYWORD: drug-disease association, drug repositioning, gene ontology, meta-path, positive-unlabeled learning, tripartite network

Thitipong Kawichai : POSITIVE LABELED AND UNLABELED LEARNING METHODS OF META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS. Advisor: Asst. Prof. KITIPORN PLAIMAS, Dr.rer.nat. Co-advisor: Asst. Prof. Apichat Suratanee, Dr.rer.nat.

Drug repositioning, discovering new indications for existing drugs, is a competent strategy to reduce time, costs, and risk in drug discovery and development. Many computational methods have been developed to identify new drug-disease associations for further validation and drug development. A recent approach showing superior performance with less required data is a meta-path based approach, which derives network-based information using path patterns from drug to disease nodes. However, existing meta-path based methods discard information of intermediate nodes along paths, which are important indicators for describing relationships between drugs and diseases. With known (positive) and unknown (unlabeled) drug-disease associations, this research proposes a new meta-path based method under positive-unlabeled (PU) learning settings for predicting drug-disease associations. Gene ontology (GO) is utilized to connect between drugs and diseases in a drug-GO-disease tripartite network. From this network, new meta-path based features of drug-disease pairs, or meta-path based functional profiles, are created to incorporate GO information into the functional profiles. An ensemble model is trained on these functional profiles of both positive and unlabeled samples. Consequently, the proposed method significantly outperforms other existing methods with the mean values of Area Under Precision-Recall Curves (AUPRC) of 0.944 and Area Under Receiver Operating Characteristic curves (AUROC) of 0.930. Moreover, up to 38% of new drug-disease associations discovered by the proposed method were found in the database of clinical trials.

Field of Study:	Applied Mathematics and Computational Science	Student's Signature .....
Academic Year:	2020	Advisor's Signature .....
		Co-advisor's Signature .....

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my dissertation advisor, Assistant Professor Dr. Kitiporn Plaimas, for her valuable guidance, continuous support, and encouragement. Also, I would like to sincerely express my appreciation to my dissertation co-advisor, Assistant Professor Dr. Apichat Suratanee, for his kind support and helpful suggestions. With their kind assistance and insightful comments, I can improve the quality of my research and accomplish this dissertation. Furthermore, I would like to acknowledge Associate Professor Dr. Treenut Saithong, Chair External Examiner, and all dissertation committees consisting of Assistant Professor Dr. Krung Sinapiromsaran, Assistant Professor Dr. Jiraphan Suntornchost, and Dr. Thap Panitanarak.

In addition, I would like to thank Ms. Satanat Kitsiranuwat for sharing her valuable ideas, useful codes, and data. Also, I would like to thank all members in my research group for good discussions and relationships. Furthermore, I am grateful to my colleagues in Department of Mathematics and Computer Science, Academic Division, Chulachomklao Royal Military Academy, for their support during the course of my study. Importantly, I would like to thank my parents for their continuous support, taking care of me, and blessing me. Additionally, I would like to thank all other my family members and my friends for their encouragement and support.

Lastly, I would like to express my special thanks to the Development and Promotion of Science and Technology Talents project (DPST) for financial support and National e-Science Infrastructure Consortium (http://www.e-science.in.th) for kindly supporting the high-performance computing resources.

Thitipong Kawichai

TABLE OF CONTENTS

Page

ABSTRACT (THAI)....III

ABSTRACT (ENGLISH)....IV

ACKNOWLEDGEMENTS....V

TABLE OF CONTENTS....VI

LIST OF TABLES....X

LIST OF FIGURES....XI

CHAPTER I INTRODUCTION....1

1.1 Background and rationale....1

1.2 Research objectives....3

1.3 Scopes of the research....3

1.4 Expected outcomes....4

1.5 An overview of the dissertation....5

CHAPTER II BACKGROUND KNOWLEDGE AND RELATED WORKS....6

2.1 *De novo* drug discovery and drug repositioning....6

2.1.1 *De novo* drug discovery and its challenges....6

2.1.2 Drug repositioning....8

2.1.3 Successful cases of drug repositioning....9

2.2 Computational drug repositioning....10

2.2.1 Strategies for computational drug repositioning....11

2.2.2 Genetic variation based repositioning....12

2.2.3 Signature-based repositioning....13

2.2.4 Molecular docking based repositioning 14

2.2.5 Phenotype-based repositioning 14

2.2.6 Similarity-based repositioning 15

2.3 Genes, Proteins, and Gene Ontology (GO) 25

2.3.1 Use of genes and proteins to discover the drug-disease relationships 25

2.3.2 An overview of gene ontology 26

2.3.3 Gene ontology applications in drug repositioning 28

2.4 Meta-paths 29

2.4.1 Basic definitions and concepts 29

2.4.2 Applications of meta-paths in drug repositioning 34

2.5 Positive-Unlabeled (PU) learning 36

2.5.1 Introduction to PU learning 36

2.5.2 Categories of PU learning methods 38

2.5.3 PU learning methods for drug repositioning 41

2.6 Extreme Gradient Boosting (XGBoost) 42

CHAPTER III PROTEINS VERSUS FUNCTIONAL INFORMATION 47

3.1 An overview of this study 47

3.2 Data sets 49

3.3 Methods 50

3.3.1 Preparation of drug-disease, drug-drug, and disease-disease pairs 50

3.3.2 Construction of drug-GO and disease-GO associations 50

3.3.3 Measurement of protein-based and functionality-based similarities 53

3.3.4 Performance measurement 54

3.3.5 Classification of drug-disease, drug-drug, and disease-disease associations 56

3.4 Results 57
3.4.1 Data summarization 57
3.4.2 Investigation of sharing proteins and GO functions among known associations 59
3.4.3 Selection of the most suitable similarity index 64
3.4.4 Comparison of protein-based and functionality-based similarities.....66
3.4.5 Case studies.....71
3.5 Discussions.....73
3.6 Summary 75
CHAPTER IV META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS 77
4.1 An overview of the study 77
4.2 Data sets 78
4.3 Methods.....78
4.3.1 Construction of a drug-GO-disease tripartite network.....78
4.3.2 Generation of meta-path based functional profiles for drug-disease pairs 79
4.3.3 Dimensionality reduction of meta-path based functional profiles 85
4.3.4 A classification model framework 86
4.3.5 Experimental settings and performance evaluation 89
4.4 Results 94
4.4.1 The constructed drug-GO-disease tripartite network.....94
4.4.2 Usage of different meta-path based functional profiles 96
4.4.3 Selected values of model parameters 101
4.4.4 Comparison with other methods 103
4.4.5 Validation of predicted drug-disease associations 108

4.4.6 Case studies 110

4.5 Discussions 113

4.6 Summary 116

CHAPTER V CONCLUSIONS AND FUTURE WORKS 118

5.1 Conclusions 118

5.2 Future works 121

REFERENCES 122

APPENDIX 134

VITA 139

LIST OF TABLES

Page

Table 2.1 Examples of successful repositioned drugs 9

Table 2.2 Summary of drug and disease information used in each similarity-based method.....23

Table 2.3 Path instances corresponding to the given meta-paths 33

Table 3.1 The list of data with their sources and versions.....50

Table 3.2 The total numbers of drugs, diseases, proteins, and GO functions.....57

Table 3.3 Statistical information of drug-protein, drug-GO, disease-protein, and disease-GO associations 58

Table 3.4 The numbers of drug-disease, drug-drug, and disease-disease pairs in each class.....59

Table 3.5 Comparison of the numbers of pairs that share proteins and GO functions in the positive class.....63

Table 3.6 Comparison of the number of pairs that share GO functions in each class 64

Table 4.1 Properties of the drug-GO-disease tripartite network.....95

Table 4.2 The selected latent feature percentage for each functional profile matrix.....97

Table 4.3 The number of positive samples that have all-zero functional profiles in each functional profile matrix.....100

Table 4.4 Performance comparison of the proposed method and other methods.....104

Table 4.5 Summary of candidate and non-candidate drug-disease associations and their supporting evidence 109

Table 4.6 The list of new drugs proposed for esophageal cancer.....111

Table 4.7 Performance comparison of the machine learning methods.....115

LIST OF FIGURES

Page

Figure 1.1 An overview of the dissertation.....5

Figure 2.1 The process of *de novo* drug discovery 6

Figure 2.2 The shortened process of drug repositioning.....8

Figure 2.3 An overview of strategies used in computational drug repositioning 11

Figure 2.4 Various cases to illustrate the guilt-by-association principle 15

Figure 2.5 Examples of GO terms to illustrate their hierarchical structures.....27

Figure 2.6 An example of a heterogeneous network and its network schema.....30

Figure 2.7 Examples of meta-paths in a bibliographic network 32

Figure 2.8 One-class learning versus PU learning.....37

Figure 2.9 Illustration of the spy strategy 38

Figure 2.10 A bagging SVM.....40

Figure 2.11 An example illustrating the score calculation of a tree structure 45

Figure 3.1 Three relationships of drugs and diseases under investigation 47

Figure 3.2 A conceptual diagram depicting an overview of the study 48

Figure 3.3 Schematic diagrams summarizing how to construct drug-GO and disease-GO associations 51

Figure 3.4 An overview of the measurement of functionality-based similarities.....53

Figure 3.5 A confusion matrix 55

Figure 3.6 Coverages of positive pairs according to their distances on the PPI network.....60

Figure 3.7 Coverages of positive pairs that share and do not share their GO functions.....62

Figure 3.8 Areas under the ROC curves (AUROC) of all similarity indices for protein-based and functionality-based similarities 65

Figure 3.9 ROC curves of protein-based and functionality-based similarities 67

Figure 3.10 Precision-recall curves of protein-based and functionality-based similarities 67

Figure 3.11 Confusion matrices of protein-based and functionality-based similarities 68

Figure 3.12 Precision, recall, accuracy, and F_1-score of protein-based and functionality-based similarities 69

Figure 3.13 An inferred association of glimepiride and nicorandil 72

Figure 3.14 An inferred association of Myotonia congenita and Gitelman syndrome 73

Figure 4.1 A schematic diagram providing an overview of this study 77

Figure 4.2 Illustration of a drug-GO-disease tripartite network and its network schema 79

Figure 4.3 A demonstration of generating meta-path based functional profiles and further processes 83

Figure 4.4 A classification model used in the proposed method 87

Figure 4.5 Data manipulation for the experiments of this study 90

Figure 4.6 The constructed drug-GO-disease tripartite network 94

Figure 4.7 Performance comparison of using different meta-path based functional profiles 98

Figure 4.8 Performance comparison of using different aggregate schemes 102

Figure 4.9 Improvements of mean AUPRC values when T was increased and the averaging scheme was used 103

Figure 4.10 Comparison of performance based on top-K ranked predictions 106

Figure 4.11 P-values of Wilcoxon signed rank tests for comparing performance based on top-K ranked predictions 108

CHAPTER I

INTRODUCTION

1.1 Background and rationale

Developing a new drug is a time-consuming, expensive, and risky process. For developing only one novel drug, it normally takes more than 12 years and a billion US dollars on average [1]. Moreover, almost 90% of drug candidates fail and are not introduced into the stage of clinical trials [2] due to their insufficient efficacy and safety. A strategy known to successfully resolve the bottleneck of the drug development is drug repositioning, discovery of new indications for existing drugs. Due to availability of their efficacy and safety information, a repositioned drug can save more than half of time and costs invested for a *de novo* drug [1].

To support a task of drug repositioning, many in silico methods have been proposed to identify candidates of drug-disease associations for further validation by wet lab experiments. Most of them are based on the similarity-based approach, which predicts the same treatments for similar diseases and vice versa using drug-drug and disease-disease similarities. For example, Gottlieb et al. [3] utilized five drug-related properties and two data sets of diseases to compute multiple drug-drug and disease-disease similarity scores for using in the large-scale prediction of drug indications (PREDICT). Wang et al. [4] integrated drug target information, drug chemical structures, disease phenotypes, and drug-disease associations to develop a three-layer heterogeneous network model (TL_HGBI) for predicting links between drugs and diseases. Luo et al. [5] proposed a bi-random walk method (MBiRW) which uses drug-drug similarities based on chemical substructures and disease-disease similarities based on disease phenotypes to infer new drug-disease associations. Liang et al. [6] integrated drug chemical information, protein domains, and gene ontology information to compute similarities among drugs in Laplacian Regularized Sparse Subspace Learning (LRSSL). Zhang et al. [7] utilized several drug features and disease semantic information for computing drug-drug and disease-disease similarities in a Similarity Constrained Matrix Factorization method for Drug-Disease associations (SCMFDD). With the similarity-based approach, several methods have been recently proposed for drug repositioning using network embedding [8, 9] and deep learning [10].

From existing methods, various techniques were deployed to compute drug-drug and disease-disease similarities based on drug and disease information. At present, there is no a standard method to generate drug-drug and disease-disease similarities. The similarity measures created by different methods could be totally incongruent [11]. Furthermore, the unavailability of some drug or disease data may preclude the executions of the methods that require various drug and disease data [12]. More importantly, only confirmed drug-disease associations (positive samples) are affordable, and there are no non-associated pairs of drugs and diseases (negative samples) due to lack of application values [13]. Most supervised learning methods treated all drug-disease pairs out of positive samples (unlabeled samples) as negative ones although they contain both positives and negatives. These contaminated negative samples could lead to an unstable decision boundary of the model and result in the inaccurate predictions of drug-disease associations [14, 15].

To solve the problems, Wu et al. [11] proposed Ensemble Meta-Paths and Singular Value Decomposition (EMP-SVD) for predicting drug-disease associations under a Positive-Unlabeled (PU) learning setting, a learning approach with positive and unlabeled data. Without relying on drug and disease similarities, they utilized meta-paths, path structures for extracting network-based information, to generate valuable features for each drug-disease pair from the drug-protein-disease heterogeneous network. To avoid contaminated negative samples, they utilized the heuristic strategy for selecting reliable negative samples from unlabeled drug-disease pairs. This work shows the efficiency of the meta-path based approach with PU learning by producing the superior performance despite less data used (i.e. drug-protein, disease-protein, and drug-disease associations), when compared to existing methods.

Although the meta-path based approach under the PU learning settings is really competent, to the best of my knowledge, there are few studies incorporating both techniques. Furthermore, most meta-path based methods utilize counts of paths extracted by a meta-path without considering information of intermediate nodes along paths although they are very important indicators, such as drug-associated and disease-associated proteins. In addition, most existing methods employ drug-associated and disease-associated proteins as primary intermediaries to bridge between drugs and diseases. With this granular protein information, a large number of proteins are required for creating accurate predictions.

In this research, gene ontology (GO) terms, biological functions annotated for genes and gene products, are utilized as principal indicators for identifying new drug-disease associations. Initially, the feasibility of using GO-based similarity information, or functionality-based similarity measures, for discovering relationships between drugs and diseases, between drugs, and between diseases was assessed. Next, the novel meta-path based method under the PU learning settings was proposed. In this method, the drug-GO-disease tripartite network was constructed by using drug-GO, disease-GO, and drug-disease associations. From the network, new features of drug-disease pairs were generated by differentiating extracted paths according to their incorporated GO nodes and creating as meta-path based profiles of GO functions for each drug-disease pair, called meta-path based functional profiles. These functional profiles of both positive and unlabeled samples were fed into the PU ensemble model to recognize the positive drug-disease associations from the unlabeled drug-disease pairs. The performance of the proposed method was compared with those of existing methods to evaluate its efficiency. After its satisfied performance was demonstrated, the proposed method was employed to discover potential drug-disease associations from the unlabeled drug-disease pairs.

1.2 Research objectives

1. To investigate the feasibility of utilizing GO functions for discovering relationships between drugs and diseases, between drug, and between diseases
2. To develop a meta-path based method for generating meta-path based functional profiles of drug-disease pairs
3. To propose a PU learning method with meta-path based functional profiles for predicting drug-disease associations
4. To apply the proposed method for discovering potential drug-disease associations

1.3 Scopes of the research

In this research, only approved drugs that are used in humans and interact with human target proteins in DrugBank (version 5.1.3) are included. Diseases and disease-associated proteins are limited to human diseases and human proteins found in DisGeNET (version 6.0). Functional information of drugs and diseases used in this research is only Gene Ontology (GO). Only

functional annotation data of human proteins downloaded from the Gene Ontology Annotation (GOA) database (version 191) are used. GO functions are linked to drugs and diseases through drug-associated and disease-associated proteins, respectively.

In the initial investigation of using GO functions, only similarity measures based on drug-associated and disease-associated GO functions are utilized for predicting drug-disease, drug-drug, and disease-disease associations. To enable further discovering of new drug-disease associations, the drug-drug associations are categorized based on sharing some common diseases between drugs. Similarly, the disease-disease associations are defined based on overlapping some common drugs between diseases. Seven well-known similarity indices are compared (i.e. the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, and derived Jaccard similarity index) to select the most suitable one for computing the similarity scores of the drug-disease, drug-drug, and disease-disease pairs. To assess the performance of using the GO-based similarity measures for drug repositioning, the performance of using the protein-based similarity measures is used as the baseline performance and compared with that of using GO functions.

In the proposed method for predicting drug-disease associations, the meta-path based functional profiles are constructed using the drug-GO, disease-GO, drug-disease associations. To discover novel drug-disease associations, the proposed method is employed to predict the unlabeled drug-disease pairs in only the data set used in this dissertation. Information supporting potential drug-disease associations discovered by the proposed method is searched from ClinicalTrials.gov, Comparative Toxicogenomics Database (CTD), and literature.

1.4 Expected outcomes

This research reveals the feasibility of utilizing GO functions in large-scale predicting drug-disease associations. This will lead GO functions and other functional information to gain more attention in being used to develop more advanced in silico methods for drug repositioning. Among potential drug-disease associations discovered in this research, some of them can be selected for further studies to verify their associations or to gain more understanding about their relationships. By utilizing the proposed method, the large-scale predictions can be conducted on larger real data sets to initially screen potential drug-disease associations for further validation in wet lab experiments and drug development.

1.5 An overview of the dissertation

This dissertation report consists of five chapters as illustrated in Figure 1.1. In Chapter I, background and rationale, objectives, scopes, and expected outcomes of the research are initially introduced. To briefly describe the organization of this dissertation report, an overview of the dissertation is also depicted. In Chapter II, background knowledge and some related works are given to assist in understanding other parts of this dissertation. First, *de novo* drug discovery and drug repositioning are introduced. Next, a review of the computational methods for drug repositioning is provided. Related terminologies are also described with a short review of their uses in drug repositioning, including genes, proteins, GO, meta-paths, and PU learning. Additionally, a mathematical formulation of XGBoost (eXtreme Gradient Boosting), an important model used in this research, is demonstrated. Chapter III includes a feasibility study of utilizing GO functions for uncovering relationships between drugs and diseases. The details of this study which include data sets, methods, results, discussions, and summary are provided in this section. Chapter IV is about another study that proposes a new PU learning method with meta-path based functional profiles. The details of the proposed method in terms of materials, methods, results, and discussions are explained in this chapter. Finally, Chapter V provides conclusions, limitations, and future works of this research.

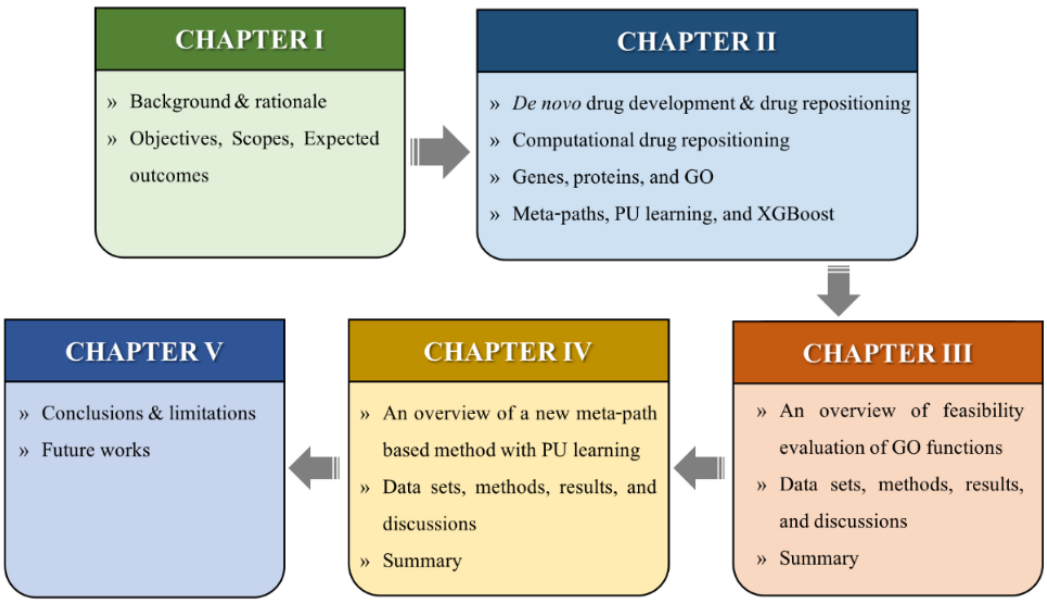


Figure 1.1 An overview of the dissertation

CHAPTER II

BACKGROUND KNOWLEDGE AND RELATED WORKS

In this chapter, a motivation and definition of drug repositioning are provided. Next, reviews of the computational methods developed for drug repositioning are demonstrated. To ease of understanding in the next chapters, some fundamentals about gene ontology (GO), meta-paths, and positive-unlabeled (PU) learning are also given with examples of related studies conducted for drug repositioning. Finally, a mathematical formulation of the extreme gradient boosting (XGBoost) method is described.

2.1 *De novo* drug discovery and drug repositioning

2.1.1 *De novo* drug discovery and its challenges

De novo drug discovery is the process of bringing a new drug to markets (Figure 2.1). Before a drug is widely used to treat patients or sold in markets, it needs to pass several rigorous stages to ensure its safety, efficacy, and appropriate used dosage.

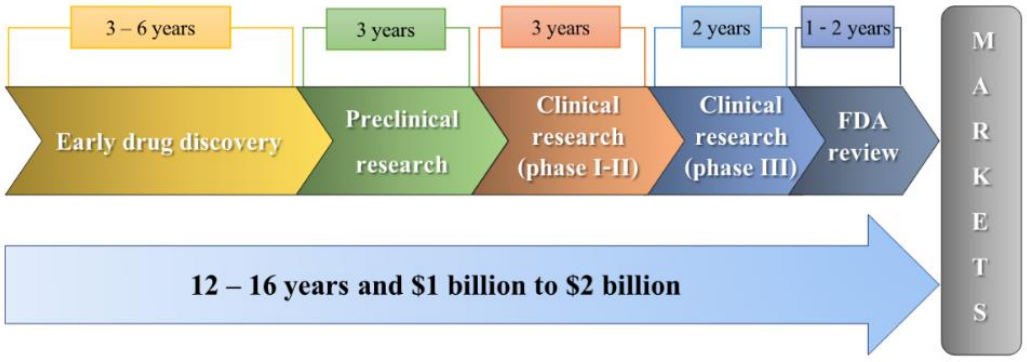


Figure 2.1 The process of *de novo* drug discovery (adapted from [1])

The early stage is to identify a small molecule as a potential candidate compound for further evaluation. At this stage, there are different routes that can lead to drug candidates such as adapting molecular structures of existing drugs, finding new molecules that can interact with a specific target or function in a particular biochemical pathway, and identifying novel compounds in nature [16]. This initial stage takes three to six years to achieve a candidate compound. After

the early drug discovery, preclinical studies are conducted in laboratories and animals to test toxicity and efficacy of the candidate compound [17]. This step takes about three years and is required by the US Food and Drug Administration (FDA) before testing a drug on people [1].

After the preclinical research completed, clinical studies (phase I, II, and III) begin to assess drug safety and efficacy in humans. In the phase I clinical research, 10 to 100 healthy volunteers participate to gather safety information when different doses of the drug are applied [17]. Then, few hundreds of people, including both healthy people and patients, involve in the phase II clinical stage. The aim of this stage is to investigate effectiveness and side effects of the drug [18]. In most clinical studies, the stages of phase I and II takes about three years [19]. Next, the phase III clinical research is conducted to demonstrate efficacy and safety of the drug in a larger group of people. In general, greater than 300 participants from multiple sites are recruited at this stage [17]. Drug side effects that are undiscovered in the phase I and II clinical studies may be detected in the phase III clinical studies. It was reported that greater than 70% of candidate drugs failed at this stage [18]. Normally, the phase III clinical studies are done for three years [19]. After that, all drug safety and efficacy data are submitted to FDA for making a decision to approve or not approve the drug. This process may require one to two years [1].

Due to those several stages to ensure drug safety and efficacy, the overall process to achieve one novel drug to markets takes 12 to 16 years. This very long time could prevent the production of new drugs serving the therapeutic needs of various diseases, such as emerging infectious diseases [20]. Moreover, developing a new drug costs more than a billion US dollar on average, and the costs increase every year [1]. Despite the larger amount of budgets invested in drug development, the trend of the number of FDA approved drugs per dollar are continually decreasing since 1950 [21]. This attrition rate is mainly due to more stringent FDA regulations and unimpressive results of preclinical and clinical studies of investigated drugs [22]. It was estimated that there is a success rate of only 2.01% in *de novo* drug discovery [19]. Most compounds failed to accomplish FDA approval because of their insufficient efficacy and safety [22]. All challenges of *de novo* drug discovery direct researchers and pharmaceutical companies to other approaches that can reduce time, costs, and risk in the drug development.

2.1.2 Drug repositioning

A drug typically interacts with more than one protein and sometimes it binds to unwanted targets [23], called off-target proteins. This could lead to negative or positive pharmacological effects and also provide opportunities of discovering new drug uses. In general, the process of discovering new indications for approved drugs is known as drug repositioning. Sometimes, several terms are used synonymously with drug repositioning, such as drug repurposing, drug redirecting, drug rediscovery, drug reprofiling, drug retasking, drug redirecting, and therapeutic switching [14, 24].

Typically, drug repositioning is conducted in three steps [25] as shown in Figure 2.2. The first important step is to identify new promising relations between approved drugs and diseases. This task is often accomplished by using computational approaches, but sometimes it is conducted by wet lab experiments. This step may include the preclinical testing if there is insufficient information about the approved drugs of interest. Then, the clinical trials are conducted to investigate safety and efficacy of the drugs when they are applied to the new diseases. In some cases of drug repositioning, sufficient information of drug safety and efficacy in preclinical models and humans already exists, leading to starting at the stage of the phase III clinical studies [26]. Despite existing FDA approval of the investigated drugs, before using for the new diseases, the new applications of these drugs need to be approved by FDA again.

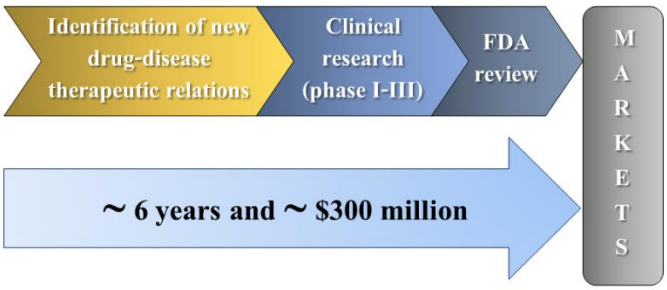


Figure 2.2 The shortened process of drug repositioning

Because safety and efficacy information of approved drugs already exists, drug repositioning can shorten the conventional process of the drug development. To achieve a repositioned drug, it was estimated that drug repositioning takes only six years and costs $300 million on average [1]. These are drastically reduced from those of *de novo* drug discovery.

Moreover, repositioned drug candidates have lower risks to fail, since they have been already approved for their safety in humans [26]. Due to its shorter time, cheaper costs, and lower risks, drug repositioning is known to solve the bottleneck in drug development and has been paid more attention from researchers and pharmaceutical companies during these recent years.

2.1.3 Successful cases of drug repositioning

To explicitly demonstrate the auspiciousness of drug repositioning, some successful cases of repositioned drugs are given as shown in Table 2.1. The most well-known repositioned drug is sildenafil. This drug acts as a phosphodiesterase type 5 (PDE5) inhibitor and was originated to treat angina, a symptom related to coronary heart disease, by Pfizer. Unfortunately, this drug failed in the phase II clinical research due to its insufficient efficacy for angina. Nevertheless, during the clinical trials, sildenafil was fortuitously found to induce penile erections [27]. Then, sildenafil was approved by FDA for the treatment of erectile dysfunction in 1998 and globally sold as Viagra with the total sales in 2012 of $2.05 billion [26].

Table 2.1 Examples of successful repositioned drugs (partially adopted from [26])

Drug name	Original indication	New indication	Year of approval
Aspirin	Inflammation, pain	Recurrent stroke	1980
Zidovudine	Cancer	HIV/ AIDS	1987
Minoxidil	Hypertension	Hair loss	1988
Sildenafil	Angina	Erectile dysfunction	1998
Thalidomide	Morning sickness	Multiple myeloma	2006

Minoxidil was known to alleviate hypertension and approved by FDA since the 1979 [28]. During the clinical studies, hair growth in patients was coincidentally specified as an adverse effect of minoxidil. This drug was further developed and received the new FDA approval for the treatment of hair loss in 1988. It was reported that minoxidil reached the worldwide sales in 2016 of $860 million [26].

Zidovudine was first synthesized to serve as an anticancer agent in 1964, but its low efficacy and high toxicity led zidovudine to fail the treatment of cancer at that time. Until 1984, it

was discovered that zidovudine could inhibit human immunodeficiency viruses (HIV) and raise CD4 cells in patients with Acquired Immunodeficiency Syndrome (AIDS) [29]. These led to the FDA approval of zidovudine as the first drug used against HIV in 1987 [26].

Thalidomide was approved for the treatment of morning sickness in pregnant women and sold in some countries since 1957 [26]. Four years later, thalidomide was withdrawn because it was reported that women who had taken this drug during their pregnancies could procreate children with serious skeletal defects. Serendipitously, it was found that thalidomide could act as an anticancer agent for multiple myeloma, leading to its new FDA approval in 2006 [27].

Aspirin is known to relieve inflammation and pain. Due to the wide range of effects of aspirin, it has been revealed the potentialities for treatments of several diseases, including strokes and colorectal cancer. In 1980, FDA approved the use of aspirin for preventing recurrent strokes after ischemic strokes, a type of strokes [30]. In 2015, US Preventive Services Task Force (UPSTF), an independent association of national experts for issuing prevention recommendations, officially suggested the use of aspirin for preventing colorectal cancer [26]. However, the use of aspirin for the treatment of colorectal cancer has been still under investigation in the stage of clinical studies.

According to the given cases of drug repositioning, it is noticeable that they were coincidentally discovered from the observations during conducting clinical trials or *in vitro* experiments. Sometimes, these experimental approaches take too long to initiate a new drug-disease therapeutic relation for further validation in the drug repositioning process. However, such successful cases motivate researchers to create more systematic and effective approaches which can identify a lot of potential drug-disease treatment relations in a short time.

2.2 Computational drug repositioning

Due to the advent of advanced biotechnologies, a ton of biological data can be generated and publicly provided in many databases, including drug-related and disease-related data. These data give us an opportunity to deploy computational approaches for more efficiently uncovering potential drug-disease treatment relations. Several computational methods have been proposed for identifying drug-disease associations. In this section, an overview of the strategies used in

computational drug repositioning is demonstrated. Then, the reviews of existing methods using each strategy are given.

2.2.1 Strategies for computational drug repositioning

A large number of in silico methods with various strategies have been developed to support drug repositioning. To give an overview of those methods, different perspectives are used for clustering computational methods used in drug repositioning. For example, Xue et al. [19] categorized existing methods into three groups based on their methodologies, which are network-based, text-mining based, and semantic based methods. In this review, strategies used in computational drug repositioning are categorized into five groups which include genetic variation based repositioning, signature-based repositioning, molecular docking based repositioning, phenotype-based repositioning, and similarity-based repositioning, as shown in Figure 2.3.

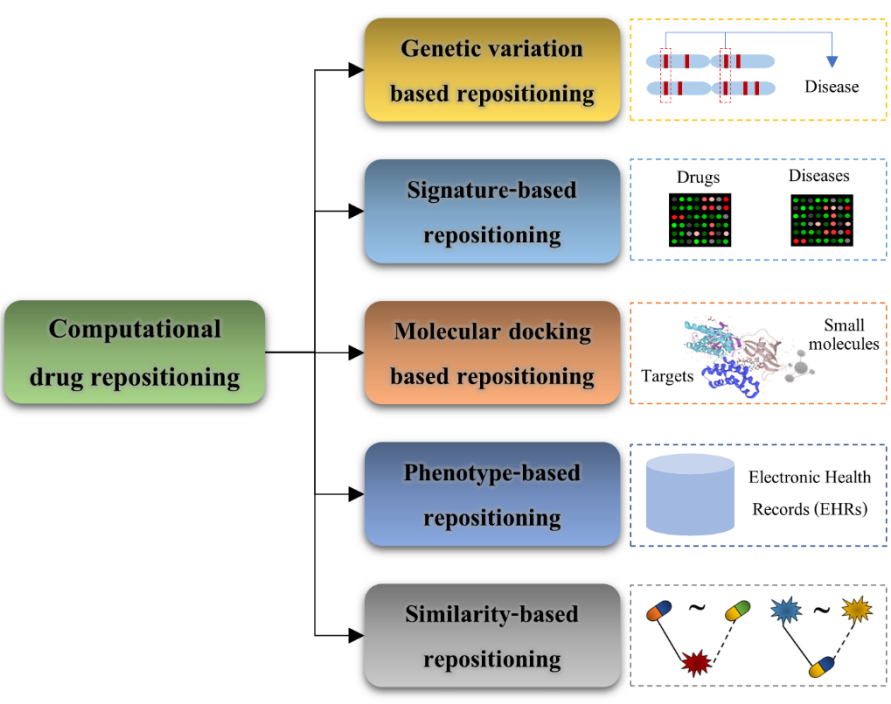


Figure 2.3 An overview of strategies used in computational drug repositioning

In genetic variation based repositioning, genome data from many people are compared to associate particular variations found in deoxyribonucleic acid (DNA) sequences with a disease trait. This approach is known as genome-wide association studies (GWAS). GWAS genes

containing genetic variations which are significantly associated with a disease could be proposed as potential targets for the disease treatment. Signature-based repositioning utilizes transcriptomic data to identify genes that the expression levels are altered due to drug uses or the courses of diseases, called gene signatures. Based on the gene signatures of drugs and diseases, novel drug-disease associations can be inferred. Molecular docking based repositioning typically uses the structures of target proteins and small molecules to simulate how the molecules bind to their targets. Phenotype-based repositioning employs phenotypic data for discovering new drug-disease associations such as electronic health records (EHRs) and clinical data. Similarity-based repositioning presumes that similar drugs tend to have the same indications and similar diseases should have the same treatments.

2.2.2 Genetic variation based repositioning

The progress of DNA sequencing technologies leads the cost for an individual's genome cheaper than it was in the past. This helps to produce a bunch of human genome data and gives an opportunity to study how genotypes link to phenotypic traits on populations. A genome wide association study (GWAS) is an approach that links genetic variants to a particular disease. GWAS can be used to infer new targets for a disease which would be utilized for further drug development [26]. For example, Sanseau et al. [31] utilized GWAS data obtained from the US National Human Genome Research Institute (NHGRI) to construct a collection of genes associated with disease traits. They proposed 991 GWAS genes as potential drug targets for drug development and found that the 155 out of these genes were already in the process of drug development at that time. Interestingly, 92 GWAS genes showed mismatches between their disease traits and old drug indications which indicate the opportunities to use those mismatched genes as promising targets for the new diseases [31]. Okada et al. [32] conducted the GWAS meta-analysis in more than 100,000 European and Asian people to detect the risk loci of rheumatoid arthritis (RA). They discovered 101 risk loci in 98 genes which were further proposed as potential drug target genes for the treatment of RA. By the enrichment analysis, they also demonstrated that those candidate genes significantly overlapped with known drug target genes of RA.

However, GWAS data provide only genomic information associated with diseases, which cannot completely suggest what kind of drugs are potential for the treatment of the diseases (e.g. activator or suppressor) [26]. Thus, integrating GWAS data with other omics data, such as transcriptomic data, would be more promising to perform functional studies, leading to more apparent solutions in drug repositioning.

2.2.3 Signature-based repositioning

Signature-based repositioning utilizes gene expression data related to drugs and diseases to infer new drug-disease associations. A transcriptomic signature of a drug or a disease is specific alteration of gene expression levels due to drug uses or the course of the disease. An example of this strategy is signature reversion. This approach assumes that a drug can treat a disease by reversing the expression levels of genes perturbed under a disease condition to normal expression levels [22]. Therefore, gene expression profiles of drugs and diseases are compared to discover the opposite expression patterns between drugs and diseases. Dudley et al. [33] compared between gene expression profiles of Inflammatory Bowel Disease (IBD) and those of 164 drugs obtained from the Connectivity Map (CMap) database. Based on the comparison of the gene expression profiles, they derived the therapeutic scores of all drugs, which more negative values indicate more anti-correlated patterns of the gene expression between drugs and diseases. Consequently, topiramate was proposed as a new promising drug for the treatment of IBD. Sirota et al. [34] performed a large-scale signature-based drug repositioning by comparing CMap gene expression data of 164 drugs against transcriptomic data of 100 diseases. They discovered more than a thousand potential drug-disease associations and selected the drug cimetidine to be experimentally validated for the treatment of lung adenocarcinoma in mouse models. The results also showed the therapeutic effects of cimetidine for this lung cancer.

Although some studies demonstrate the competence of the signature-based strategy for drug repositioning, this strategy has some drawbacks that should be aware of [22]. First, gene expression data typically contain a large amount of noise resulting in many false positive or negative signatures of drugs and diseases. Furthermore, differential expression of genes sometimes is not directly caused by a disease. Therefore, reversing expression levels of those genes induced by drugs may not be able to treat the disease.

2.2.4 Molecular docking based repositioning

Molecular docking is a computational approach to predict the potentiality of small molecules in binding to their targets. For this strategy, structural data of both drugs and their targets are required. If prior knowledge about an interested target of a disease is known, then virtual screening, computational search in multiple drugs to discover those that can potentially bind to the target protein, can be performed by the molecular docking technique. In reverse, multiple targets can be computationally searched to find the target structure that are most compatible with a particular small molecule [26]. To identify potential drug-disease associations, Dakshanamurthy et al. [35] performed a large-scale molecular docking with 2,335 crystal structures of human proteins and 3,671 FDA approved drugs. Consequently, they discovered anticancer effects of mebendazole, an anti-worm drug. They also validated this association by wet lab experiments and found that mebendazole can inhibit vascular endothelial growth factor receptor 2 (VEGFR2) activity resulting in reducing angiogenesis.

Nevertheless, there are some limitations in the use of the molecular docking based repositioning [26]. Currently, 3D structural data of many target proteins are still unavailable. Lack of the structural data of an interested target impedes the molecular docking based repositioning. In addition, the refined target structures are required to obtain accurate results of molecular docking. Sometimes, only low-resolution structural data are provided, which could lead to high false positive and negative rates in prediction for drug repositioning.

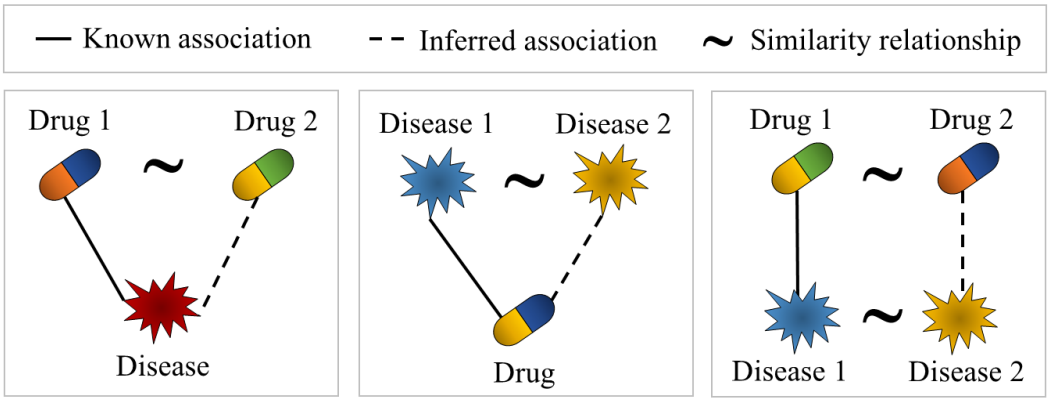
2.2.5 Phenotype-based repositioning

Phenotype-based repositioning are mainly based on phenotypic information such as electronic health records (EHRs) and clinical trial data. EHRs contain medical histories of patients which include diagnoses, symptoms, laboratory results, drug prescriptions, responses, and medical images. Natural language processing can help to mine such large EHR data for drug repositioning. Xu et al. [36] mined EHR data of cancer patients from two large sources and constructed a cohort of patients with both cancers and type 2 diabetes to investigate their drug exposure. Consequently, they discovered the association between decreased mortality rates in cancer patients and the uses of metformin, a drug conventionally used to treat type 2 diabetes. Because EHR data are collected from a large number of patients and for long time periods, it

would be possible to conduct a large-scale drug repositioning on these massive EHR data. In this recent years, researchers and pharmaceutical companies pay more attention on utilizing EHR data for drug repositioning. However, some studies with EHRs are stumbled due to the problems of data privacy, incompleteness, inaccuracy, and incompatibility [22].

2.2.6 Similarity-based repositioning

During many recent years, the vast majority of computational drug repositioning methods have been developed using this strategy. Similarity-based methods discover new drug-disease associations mainly based on similarities between drugs and diseases, between drugs, or between diseases. An important principle underlying similarity-based methods is Guilt-By-Association (GBA) [37]. In drug repositioning, GBA is the principle assumes that similar drugs tend to have the same drug indications or share some common diseases as shown in Figure 2.4(a). Conversely, similar diseases tend to have common treatments or share some common drugs as shown in Figure 2.4(b). Some similarity-based methods utilize both drug-drug and disease-disease similarity to infer new drug-disease associations. According to Figure 2.4(c), if drug 1 is similar to drug 2, disease 1 is also similar to disease 2, and the association between drug 1 and disease 1 is known, the association between drug 2 and disease 2 can be inferred by GBA.



(a) Drug-drug similarity (b) Disease-disease similarity (c) Both similarities

Figure 2.4 Various cases to illustrate the guilt-by-association principle (adapted from [22])

A method that simply applies GBA is the work of Chiang and Butte [38]. They defined disease-disease similarity based on medications that are shared between two diseases. They

assumed that if there is at least one common drug between two diseases, then all drugs of one disease could be used for the treatment of another disease. By investigating FDA approved drugs and drug indications, they could suggest more than 57,000 potential drug-disease associations, and these associations were 12 times more likely to be found in the database of clinical trials than those associations that were not suggested by their proposed method.

Similarity-based methods can be divided into two groups which are those with similarity scores and those without similarity scores. The former is a group of methods that require drug-drug and disease-disease similarity scores for further identifying new drug-disease associations. The latter is a group of methods that do not pre-compute drug-drug and disease-disease similarity scores. Without preparing similarity scores, these methods directly integrate heterogeneous data and use more sophisticated approaches to properly manipulate these integrated data models for predicting drug-disease associations. For instance, a method creates drug features from collected drug properties and uses a machine learning method to predict their associated disease classes. In addition, similarity-based methods can be categorized based on their core models which are machine learning, deep learning, and network models. To ease of understanding, similarity-based methods are mainly separately mentioned according to their core models (i.e. network and machine learning models) and chronologically reviewed in this section.

A network is a powerful tool to represent relationships between one or more than one object type. With a support of integrating multiple objects in a network, network-based approaches have been widely utilized in many applications, including drug repositioning. With linking at least between a drug-drug similarity network and a disease-disease similarity network, GBA can be used to infer new missing links between drugs and diseases in the integrated network. Wang et al. [39] proposed a Heterogeneous Graph Based Inference (HGBI) method for identifying GBA-based missing links between drugs and target proteins. Also, this method can be applied for predicting drug-disease associations. Initially, they constructed a network that integrates drug-drug similarity scores, disease-disease similarity scores, and known drug-disease associations. They computed drug similarity scores based on drug chemical structures and disease similarity scores based on Medical Subject Headings (MeSH) terms describing diseases. They performed an iterative process to propagate existing edge weights throughout the network to generate novel links with their estimated weights between drugs and diseases.

Next, Wang et al. [4] improved HGBI by additionally integrating drug-target interactions and target-target similarity scores based on the target protein sequences into a heterogeneous network and called this new method as the Triple Layer Heterogeneous Graph Based Inference (TL_HGBI). They also adjusted the iterative algorithm to be compatible with this triple layer network. Consequently, they found that TL_HGBI could perform better than HGBI by integrating drug target information.

With the success of using the heterogeneous networks for drug repositioning, Martínez et al. proposed a new method to prioritize drug-disease associations in the heterogeneous network called DrugNet [40]. Their network integrates three subnetworks which include the drug-drug similarity network based on the Anatomical Therapeutic Chemical (ATC) codes, the disease-disease similarity network based on Disease Ontology (DO), and the protein-protein interaction (PPI) network. These subnetworks were connected to one another by known drug-disease, drug-protein, and disease-protein associations. To prioritize new drug-disease associations, they applied a propagation flow algorithm, called ProphNet, which uses a drug and a disease of known drug-disease associations as two query nodes for propagating the nodes' initial scores to their neighbors along intra-connections and inter-connections. As a result, DrugNet can efficiently identify drug-disease associations studied in the clinical trials with the mean Area Under the Receiver Operating Characteristic curve (AUROC) of 0.8364. Moreover, they found that there was a bias of ATC codes where some of them were linked to many drugs. This resulted in the higher ranking of those diseases due to their propagated scores accumulated from multiple drug nodes.

Luo et al. [5] developed the bi-random walk based method (MBiRW) with new similarity measures to predict drug-disease associations. Initially, they computed drug-drug and disease-disease similarity scores based on the drug chemical structures and disease MeSH terms, respectively. By using known drug-disease associations, they improved drug similarity scores based on common diseases between drugs and adjusted disease similarity scores based on common drugs between diseases. Then, they applied a bi-random walk algorithm in the heterogeneous network that integrates the adjusted drug and disease similarity scores with known drug-disease associations. Their results showed that MBiRW significantly outperformed TL_HGBI and DrugNet. However, they also suggested to improve the performance of MBiRW

by using additional information (e.g. target information) for further adjusting similarity scores and building a more complex heterogeneous network.

According to the network-based methods, it is noteworthy that integrating various drug-related and disease-related data could improve the performance of a method. Nevertheless, more complex networks resulting from multiple data integration require more advanced approaches to most advantageously exploit diverse information for making predictions. For many recent years, machine learning is widely used to construct a predictive model for diverse applications, including drug repositioning. In 2011, Gottlieb et al. [3] proposed the most well-known machine learning model for PREdicting Drug IndiCaTions (PREDICT). They utilized five drug-related data (i.e. chemical structures, side effects, protein sequences, GO annotation, and PPI network based information of drug targets) to construct a drug-drug similarity matrix. For measuring disease-disease similarity, they used both phenotypic and genotypic data set. The phenotypic data are disease terms including MeSH and Human Phenotype Ontology (HPO) terms. The genetic data are the data about disease gene signatures obtained from the gene expression analysis, including sequences, PPI-based information, and GO annotation of the disease genes. In PREDICT, features of drug-disease pairs were obtained by combining all drug and disease similarity scores, and these features were fed into a logistic regression model. By 10-fold cross validation, PREDICT performed better than the method of Chiang and Butte [38] with the AUROC value of 0.91. In addition, drug-disease associations proposed by PREDICT were significantly found in the list of drug-disease associations under investigation in clinical studies.

To predict drug multi-therapeutic classes according to the ATC codes, Napolitano et al. [41] developed a multi-class support vector machine (SVM). They derived drug-drug similarity scores from chemical structures, PPI-based information of drug targets, and gene expression profiles after drug uses. Then, they integrated all drug similarity scores by averaging to obtain one drug similarity matrix. They compared the performance of both individual and integrated drug similarity measures. As a result, they found that the integrated similarity measure produces the highest AUROC value. Moreover, their classifier achieves the accuracy of 78%.

Due to availability of diverse drug and disease information, many methods integrated multiple similarity measures of drugs and diseases to advantageously exploit different views of heterogeneous data, possibly leading to an improved accuracy in predictions [42]. Without using

a direct operation (e.g. averaging), another approach which efficiently integrates multiple similarity matrices on a latent feature space is a matrix factorization based approach with optimization techniques. Zhang et al. [43] proposed the DDR method which were created as a nonlinear optimization model to identify new drug-disease associations. They used chemical structures, side effects, and target protein sequences to create drug-drug similarity scores and employed MeSH, DO, and sequences of disease-associated genes to derive disease-disease similarity scores. In DDR, multiple drug and disease similarity measures were integrated with different weights through an optimization model resulting in the latent drug clustering and disease clustering matrix, respectively. Consequently, DDR outperformed PREDICT and could identify the drugs (i.e. nelfinavir and leflunomide) under investigation in clinical trials for the treatment of Systematic Lupus Erythematosus (SLE).

Liang et al. [6] also proposed another optimization model for drug repositioning called Laplacian Regularized Sparse Subspace Learning (LRSSL). In LRSSL, three drug feature profiles, including the profiles of drug chemical structures, target domains, and target GO annotation, were integrated on a common latent subspace. Then, they formulated an optimization model with Laplacian regularization for predicting new drug indications. As a result, LRSSL could efficiently identify many new drug indications which were supported by evidence in public databases and literature. Moreover, they found that the target protein domain and the functional annotation could suggest to underlying mechanisms of the predicted drugs.

Zhang et al. [7] developed the Similarity Constrained Matrix Factorization method for predicting Drug-Disease associations (SCMFDD). In measuring drug-drug similarities, they utilized several drug features, including chemical structures, drug target interactions, enzymes, pathways, and drug-drug interactions. They calculated disease semantic similarity scores based on MeSH terms. Different combinations between one drug similarity and the disease semantic similarity scores were employed as constraints for the matrix factorization of known drug-disease associations. By comparing with other methods, SCMFDD noticeably outperformes PREDICT and LRSSL. Although all different combinations of drug and disease similarities showed better performance than those of other methods, the question which drug and disease features should be combined to create the best SCMFDD model for new data sets is still difficult to answer.

With the progress of computing technologies and big data generation, deep learning has become popular and also has been applied for drug repositioning. For example, Zeng et al. [9] proposed a deep learning method for drug repositioning (deepDR). They used multiple drug features which include drug-drug interactions, drug targets, side effects, chemical structures, ATC-based drug indications, target sequences, and GO annotation. They fused multiple drug features and created common low-dimensional representations of drugs using multi-modal autoencoder. After that, they employed the collective variational autoencoder (cVAE) for classifying drug-disease associations. By 5-fold cross validation, deepDR achieves the high AUROC value of 0.908. With an external data set, deepDR reaches the AUROC value of 0.826. Furthermore, they suggested to integrate disease-related data to improve the performance of deepDR.

In the similarity-based methods, various drug and disease data were included, and different techniques were applied to integrate multiple similarity measures. It was found in the study of Campillos et al. [22] that some similar drugs based on the drug side effects did not share their drug targets or indications. This suggests that drug (disease) similarity based on some drug (disease) properties cannot totally point to accurate indications (treatments). With these properties, multiple data integration may conceal signals of potential drug-disease associations. To avoid this problem, recent methods were proposed with less data required, but they can produce better performance than those of existing methods that uses more data. Zhou et al. [8] developed a Network Embedding based method for predicting Drug-Disease associations (NEDD). This method utilizes the same heterogeneous network as that of HGBI and MBiRW, which integrates a drug similarity measure based on chemical structures and a disease similarity measure based on MeSH terms with known drug-disease associations. They used HIN2vec (Heterogeneous Information Network to Vector) [44], a network representation learning method based on neural networks, to create network representation features of drug and disease nodes. They utilized a random forest model to classify drug-disease pairs based on those network embedding features. Despite the same drug and disease features used, NEDD outperformes the existing network-based methods, including HGBI and MBiRW, due to the advantages of a network embedding approach.

Another example is the work of Tian et al. [45] who introduced a new method called HeteSim_DrugDisease or HSDD for predicting drug-disease associations. They also employed the same heterogeneous network as that used in HGBI, MBiRW, and NEDD. They measure relatedness between drug and disease nodes through diverse meta-paths, path structures defined for extracting semantic information between a pair of node types. HeteSim scores, meta-path based scores, were calculated for all drug-disease pairs and used to classify drug-disease associations. As a result, HSDD performs better than HGBI and MBiRW. Moreover, HSDD outperforms DrugNet, which requires more data than those used in HSDD.

In addition to less information required, recent machine learning based methods were developed to deal with lack of negative samples in drug repositioning. In nature, only known (positive) drug-disease associations are available, but no non-associated (negative) pair between drugs and diseases is identified due to its lack of applications [13]. To temporarily unlock this limitation, most supervised learning based methods, including deep learning methods, treated all unknown (unlabeled) drug-disease associations as negatives. This could result in an unstable and unreliable classifier obtained due to making use of contaminated negative samples [46].

To solve this problem, Wu et al. [11] proposed the Ensemble Meta-Paths and Singular Value Decomposition (EMP-SVD) method with a heuristic strategy for selecting reliable negatives. They used only drug-protein, disease-protein, and drug-disease associations to construct a heterogeneous network. Without drug and disease similarity scores, they utilized multiple meta-paths to generate network-based features of drug-disease pairs. They selected negative samples from unlabeled drug-disease pairs which have no common interacting proteins. Based on each meta-path, latent features of both positive and negative drug-disease pairs obtained from the singular value decomposition (SVD) method were used to generate a random forest classifier. Finally, they aggregated multiple meta-path based classifiers as an ensemble model. As a result, EMP-SVD achieves the greatest values of AUPRC (0.956) and AUROC (0.951) when compared to state-of-the-art methods, including PREDICT, TL_HGBI, MBiRW, LRSSL, and SCMFDD.

Recently, Liu et al. [47] also developed an improved version of EMP-SVD, called Topological Similarity and Singular Value Decomposition (TS-SVD). Based on the same network as that of EMP-SVD, they constructed drug-drug and disease-disease topological similarity

matrices. To create latent features of drug-disease pairs, they performed SVD on the topological similarity matrices of drugs and diseases. They also improved the negative selection strategy of EMP-SVD by selecting unlabeled drug-disease pairs with no k-step paths in the network linking between drugs and diseases, where k = 1, 2, and 3. The results showed that they can greatly enhance the performance of EMP-SVD with the AUROC value of 0.966 and the AUPRC value of 0.974.

The summary of drug and disease data used in each similarity-based method mentioned in this section is shown in Table 2.2. It is noteworthy that most existing similarity-based methods utilize diverse drug and disease information, such as drug chemical structures, drug-associated proteins, drug indications, disease phenotypic terms, and disease-associated proteins, with the hope to enhance performance in predicting drug-disease associations. However, multiple data integration in similarity-based methods usually raises two difficult questions. First, which should drug and disease information be included in the methods? It should be noted that useless drug or disease properties included may decrease signals of potential drug-disease associations. In addition, the more the drug and disease information required for a method are diverse, the harder the execution of that method is. Second, what is the most appropriate way to integrate all drug and disease data? Currently, there is no a standard method to integrate diverse similarity measures. Different techniques could lead to discrepant results, and no one knows which one is more accurate. To avoid these issues, some recent methods were proposed with less but important drug and disease information. Despite less data used, these methods can accomplish better performance when compared to other methods with the multiple data integration. Moreover, with less drug and disease information required, a larger number of drugs and diseases can be probably incorporated into a method, enabling a large-scale prediction of drug-disease associations.

Table 2.2 Summary of drug and disease information used in each similarity-based method

Type	Method (Year of publication)	Drug information					Drug target information							Disease information			
		Structures	Side effects	Signatures	Indications	DDIs	DTIs	Sequences	PPIs	Domains	GO	Enzymes	Pathways	Terms	Genes	Sequences	Signatures
Network-based methods	HGBI (2013)	✓			✓									✓			
	TL_HGBI (2014)	✓			✓		✓	✓						✓			
	DrugNet (2015)				✓		✓		✓					✓	✓		
	MBiRW (2016)	✓			✓									✓			
	HSDD (2018)	✓			✓									✓			
Machine & deep learning	PREDICT (2011)	✓	✓		✓			✓	✓		✓			✓			✓
	Napolitano et al. (2013)	✓		✓	✓				✓								
	DDR (2014)	✓	✓		✓			✓						✓		✓	
	LRSSL (2017)	✓			✓					✓	✓						
	SCMFDD (2018)	✓			✓	✓	✓					✓	✓	✓			

Note that DDIs are drug-drug interactions, DTIs are drug-target interactions, and Disease terms include MeSH, HPO, and DO terms.

Table 2.2 Summary of drug and disease information used in each similarity-based method (continued)

Type	Method (Year of publication)	Drug information					Drug target information							Disease information			
		Structures	Side effects	Signatures	Indications	DDIs	DTIs	Sequences	PPIs	Domains	GO	Enzymes	Pathways	Terms	Genes	Sequences	Signatures
Machine & deep learning	deepDR (2019)	✓	✓		✓	✓	✓	✓			✓						
	EMP-SVD (2019)				✓		✓								✓		
	NEDD (2020)	✓			✓									✓			
	TS-SVD (2020)				✓		✓								✓		

Note that DDIs are drug-drug interactions, DTIs are drug-target interactions, and Disease terms include MeSH, HPO, and DO terms.

2.3 Genes, Proteins, and Gene Ontology (GO)

2.3.1 Use of genes and proteins to discover the drug-disease relationships

Genes are portions of a genome which can be encoded to produce gene products, such as ribonucleic acids (RNAs) and proteins, which mostly are relevant to some biological functions in the cells of organisms. Human diseases can be caused by abnormalities in gene and protein functionalities. For example, mutations in the factor VIII gene (F8) or the factor IX gene (F9) result in deficiencies of coagulation factor VIII or factor IX which can cause hemophilia type A and B [48]. In many studies, genes and proteins associated with diseases are often utilized to produce disease-disease similarity measures for predicting disease-disease associations. A high score of a disease-disease association could point to a shared cause and treatment of diseases [49]. Goh et al. [50] constructed a human disease network (HDN) where connects any two diseases by a weighted link based on the number of genes overlapping between both diseases. In HDN, they found that diseases in the same disease class are usually grouped in the same network cluster due to the greater number of genes shared among the diseases. Since similar diseases are often treated by the same drugs, drug repositioning hypotheses could be simply generated within the same disease cluster of HDN. However, some disease-disease associations which have no common genes cannot be detected using HDN. Zheng et al. [51] constructed expanded HDN (eHDN) by integrating PPIs with disease-associated genes. When compared to HDN, they discovered many new links between diseases which could benefit more insights regarding disease relations and treatments.

Drugs interact with their targets, mostly proteins, at the molecular level to affect the downstream biological processes for disease treatments. Based on drug-interacting proteins, high similarity between two drugs may suggest that both drugs have the same mechanisms of action and some common drug indications [52]. Yildirim et al. [53] proposed a drug target network which links two drugs if they have at least one common target proteins between drugs. They found that drugs which are annotated with similar ATC codes or have similar indications tend to be clustered together in the drug target network. Huang et al. [54] developed a database of drug-protein effects called Drug-protein connectivity MAP (DMAP). Based on these drug-affected proteins, they computed drug-drug similarity scores using the Tanimoto cofficients for predicting drug-disease associations. When compared to the work that used Connectivity MAP (CMAP)

data, they can improve the predictions of drug-disease associations by using DMAP data with drug similarity scores. Moreover, they found that nearly a half of their discovered drug-disease associations had at least one supporting article in PubMed.

Without drug-drug or disease-disease similarity measures, drug-associated and disease-associated genes or proteins can also be utilized to directly infer drug-disease associations. By using the curated data of disease-gene and chemical-gene associations, Davis et al. [55] inferred chemical-disease associations based on some common genes between chemicals and diseases. From a ton of their data, they obtained about 77,000 inferred chemical-disease relationships in total. Yu et al. [56] mapped all proteins associated with a drug and a disease into a PPI network to compute a module distance based score of each drug-disease pair for predicting drug-disease associations. Based on the module distance based scores, they found significant overlapping between their predicted drug-disease associations and those in Comparative Toxicogenomics Database (CTD) and literature.

In summary, it is noteworthy that integrating other useful information (e.g. PPI network information) with drug-associated and disease-associated proteins (or genes) could improve identifying drug-disease associations. Sun et al. [57] investigated GWAS genes of five disease groups and target proteins of drugs used to treat those diseases in the PPI network. They revealed that only a small proportion of known drug-disease associations intersect disease genes and drug target proteins. Rutherford et al. [58] predicted the drug-disease relations based on the direct interactions between drug target proteins and disease-associated genes in the PPI network. They found that this method can identify only a few known drug-disease associations. To improve their predictions, they suggested that indirect interactions, paths of more than one PPI step linking between drug target proteins and disease genes, should be included in the study. According to both studies, it can be implied that many relationships between drugs and diseases are more complex than interacting with the same proteins, and they are hardly detected by solely using gene and protein information.

2.3.2 An overview of gene ontology

Gene Ontology (GO) terms are controlled statements for describing biological functions of genes and gene products [59], such as RNAs and proteins. All GO terms are categorized into

three non-overlapping classes, also known as GO aspects. They are Cellular Component (CC), Molecular Function (MF), and Biological Process (BP). A CC GO term describes a cellular location where a gene product is active or functions, such as the plasma membrane region (GO: 0098590) and cytoplasm (GO: 0005737). An MF GO term indicates a molecular activity that a gene product operates, such as the ribonuclease activity (GO: 0004540) and the peptidase activity (GO: 0008233). A BP GO term depicts a molecular process or biological pathway which typically involves with a series of molecular activities, such as the protein folding (GO: 0006457) and the lipoprotein biosynthetic process (GO: 0042158).

GO terms are organized in a hierarchical structure as demonstrated in Figure 2.5. Links indicate relations between GO terms. The roots of this structure are always three GO classes which are cellular component (GO: 0005575), biological process (GO: 0008150), and molecular function (GO: 0003674). In general, a GO term in a higher level provides broader descriptions for a gene product than that in a lower level. For instance, a CC GO term intracellular organelle (GO: 0043229) is a child of CC GO term organelle (GO: 0043226). For each gene product, a set of particular GO terms can be annotated for describing its functionality. Based on a hierarchy, all parental GO terms of the annotated terms are automatically associated with that gene product.

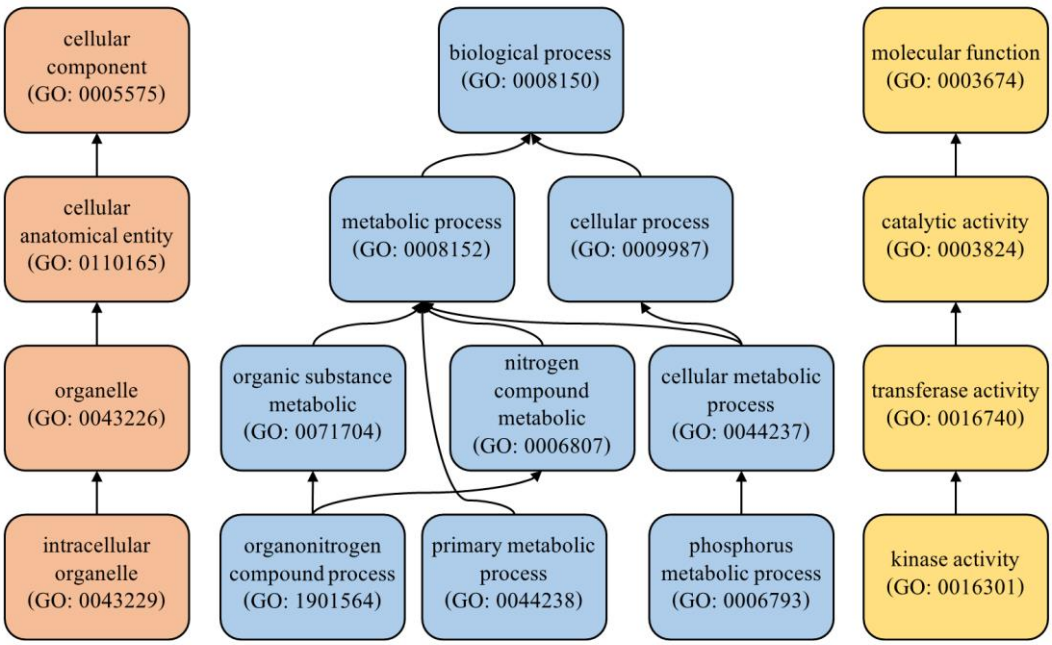


Figure 2.5 Examples of GO terms to illustrate their hierarchical structures

2.3.3 Gene ontology applications in drug repositioning

Gene ontology or GO is often used to enable understanding about molecular mechanisms related to a gene product or a set of gene products. For example, with a set of genes differentially expressed in a specific condition (e.g. a disease and a drug use), a GO analysis or an enrichment analysis is performed to find overrepresented GO terms among those genes, which could imply to significant molecular processes relevant to the interested condition. Due to usefulness of GO terms, there are some studies that utilized GO terms, especially those of BP and MF, for drug repositioning. Mathur and Dinakarpandian [60] used disease-gene associations and BP GO annotation data to create associations between diseases and BP GO terms. They also generate primary associations between drugs and BP GO terms through overlapping genes between drug target genes and disease genes. They analyzed roles of drug targets in gene networks of each annotated BP terms to refine drug-BP associations. Based on their disease-BP and drug-BP associations, 2,078 drug-disease associations were discovered, and 18% of them were found in several clinical studies.

Li et al. [61] utilized BP GO functions related to the autoimmune disease, Myasthenia Gravis (MG), to find new candidate drugs for MG. They collected 464 drugs which overlap their target genes with MG-associated genes. Then, they constructed the drug-GO function network and the GO function network by conducting hypergeometric tests. Based on both networks and the MG-GO associations, they discovered five promising drugs for the treatment of MG, and two of them were under investigation at that time. Passi et al. [62] developed a drug repurposing framework based on MF GO functions for tuberculosis (TB). They created the enhanced drug-target interaction (DTI) network by combining known DTIs with MF GO mapping based DTIs. To identify new DTIs, they utilized the network based inference algorithm and a combined evidence based method. As a result, they discovered four significant TB targets and inferred some novel drugs which are promising for the treatment of TB.

In addition, GO functions are utilized in several similarity-based methods for computing drug-drug similarity scores. For example, Gottlieb et al. [3] used GO functions of any aspects annotated for drug target genes to calculate drug similarity scores in PREDICT. Liang et al. [6] utilized MF and BP GO functions annotated for drug targets to form a type of drug feature profiles and integrated them with other drug feature profiles in LRSSL.

According to the existing GO-based methods, most of them take advantages of MF and BP GO terms but not CC GO terms. In similarity-based methods, GO functions are mostly integrated with other drug and disease information for computing drug-drug similarity scores rather than solely using drug target proteins. Due to usefulness of GO functions shown in the GO-based studies, it would be of great interest to solely make use of GO functions of any aspects for predicting drug-disease associations. Although there are some GO-based approaches already proposed for drug repositioning, they differently utilize GO functions in their own ways, suggesting that there is still much more room for novel GO-based methods.

2.4 Meta-paths

In this section, the definitions of meta-paths and other relevant terms are introduced to mathematically describe what a meta-path is. Also, some basic concepts regarding how to utilize meta-paths are demonstrated. Then, a review of meta-path based methods for drug repositioning is narrated.

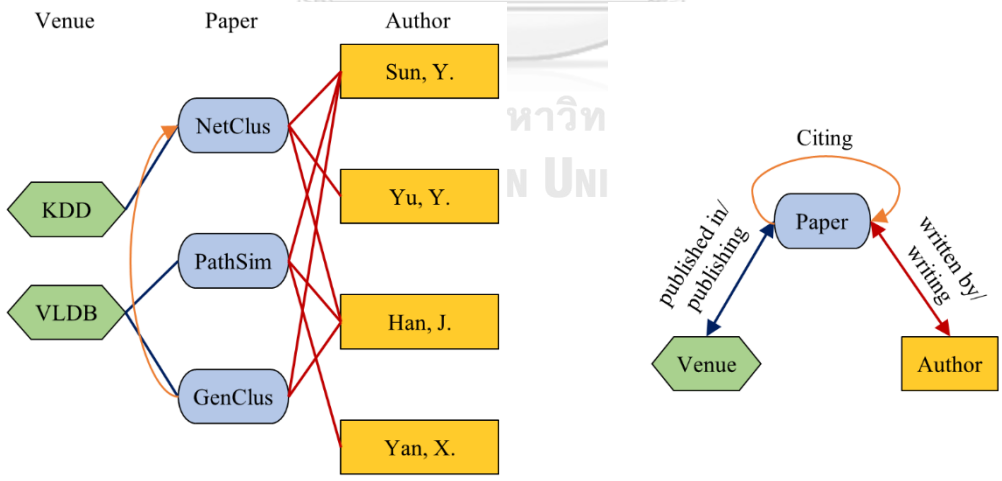
2.4.1 Basic definitions and concepts

The following definitions and concepts are mainly referred to [63].

Definition 2.1 (Heterogeneous network). Let $G = (V, E)$ represent a network, where V and E are the set of nodes and links of the network G, respectively, with a node type mapping function $\rho : V \rightarrow A$ and a link type mapping function $\varphi : E \rightarrow R$. A and R are the set of node types and link types, respectively. The network G is called a heterogeneous network if the total number of node types $|A| > 1$ or the total number of link types $|R| > 1$. Otherwise, it is called a homogeneous network.

Definition 2.2 (Network schema). The network schema of G, denoted as $S_G(A, R)$, is a meta-structure of G which is defined over the set of node types A and the set of link types R to represent an overview of all relationships in G.

An example of a heterogeneous network and its network schema is illustrated in Figure 2.6. Figure 2.6(a) shows a small subnetwork of a bibliographic network in the field of computer science. In the given network, there are three node types (i.e. venue, paper, and author). A set of link types of this network contains three link types which include links between venue and paper nodes, between paper and author nodes, and between paper nodes. Each link type in a particular direction has its semantic annotation as shown in Figure 2.6(b). For instance, a link from a paper node to an author node means that the paper was written by the author. Conversely, the meaning of a reverse direction of that link is that the author wrote the paper. Although every links between different node types in a heterogeneous network are directed links, they can be simplified as undirected links if there is no loss of their semantic information. Therefore, sometimes a heterogeneous network can be considered as an undirected network. However, for links connecting between nodes of the same type (e.g. links between paper nodes), it should be aware of semantic annotation loss if their directions are totally discarded. For example, if the link between NetClus and GenClus is treated as an undirected link, then we cannot know which one is cited by another or cites another. In this case, a heterogeneous network still retains its directed links.



(a) A bibliographic heterogeneous network (b) A network schema

Figure 2.6 An example of a heterogeneous network and its network schema [63]

Because this research uses a tripartite network as a representation model of the collected data, the definition of a tripartite network is also given below. According to the definition, a

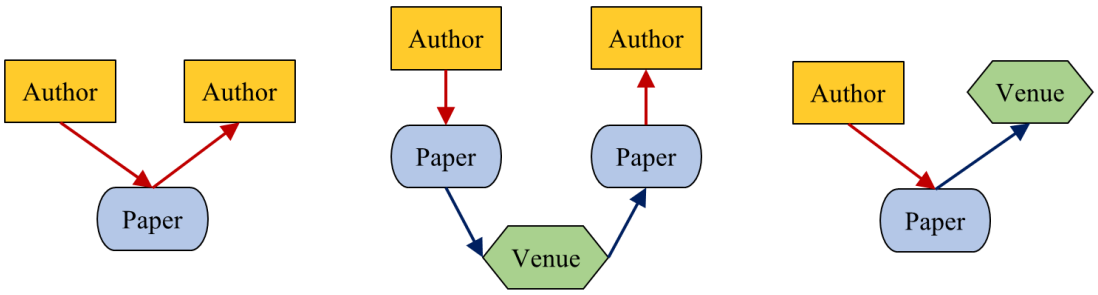
tripartite network can be considered as a special case of heterogeneous networks. Different from other heterogeneous networks, a tripartite network is drawn with three node types and three link types which connect only between nodes of the different types (i.e. no intra-links between nodes of the same types).

Definition 2.3 (Tripartite network). Let $G = (V, E)$ represent a network, where V and E are the set of nodes and links of the network G, respectively. Suppose that $\rho : V \rightarrow A$ is a node type mapping function such that $A = \{A_1, A_2, \ldots, A_n\}$ and $V_i = \{v \in V \mid \rho(v) = A_i\}$ for all $i = 1, 2, \ldots, n$. The network G is called a tripartite network if $V = V_1 \cup V_2 \cup V_3$ and $V_i \cap V_j = \varnothing$ for all $i \neq j$, and $E = \{(s, t) \mid s \in V_i, t \in V_j \text{ and } i \neq j\}$, where $i, j \in \{1, 2, 3\}$.

In a heterogeneous network, any two nodes can be connected together via different patterns of paths, and these patterns contribute different meanings [63]. These path patterns are known as meta-paths, which can be mathematically defined as follows:

Definition 2.4 (Meta-path). Based on a network schema $S_G(A, R)$, a meta-path M with the length l is denoted as $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$, where $A_i \in A$ for all $i = 1, 2, \ldots, l + 1$ and $R_j \in R$ for all $j = 1, 2, \ldots, l$, and $A_1, A_2, \ldots, A_{l+1}$ or $R_1, R_2, \ldots, R_l$ are not all the same. The link types R_j can be omitted if there is only one link type between the same pairs of node types. Then, the meta-path M can be simply written as $A_1 \rightarrow A_2 \rightarrow \ldots \rightarrow A_{l+1}$ or $A_1A_2\ldots A_{l+1}$.

Figure 2.7 demonstrates examples of meta-paths in the bibliographic network. Let $A = \{A_a, A_p, A_v\}$ be the set of node types of this network, where A_a is the author node type, A_p is the paper node type, and A_v is the venue node type. Since there are no multiple link types connecting between the same pairs of node types, each meta-path can be simply written as a sequence of node types. For example, the meta-path Author-Paper-Author is written as $A_aA_pA_a$ for short.



(a) $A_aA_pA_a$ (b) $A_aA_pA_vA_pA_a$ (c) $A_aA_pA_v$

Figure 2.7 Examples of meta-paths in a bibliographic network [63]

Each meta-path has its own meaning. For example, the meta-path $A_aA_pA_a$ means that two authors work together on a paper. The meta-path $A_aA_pA_vA_pA_a$ is about authors who publish their papers in the same venue. Finally, the meta-path $A_aA_pA_v$ means which venue an author publishes his or her paper. A meta-path is considered as a symmetric meta-path if a reverse order of a meta-path sequence is the same as the ordinary meta-path sequence, such as $A_aA_pA_a$ and $A_aA_pA_vA_pA_a$.

Meta-paths are path structures defined for extracting paths from a heterogeneous network. All paths corresponding to a meta-path is called path instances, which can be defined below. Since a network represents a collection of relationships between nodes, a meta-path can be considered as a composite relationship of a pair of node types [63]. Path instances under a meta-path for a particular pair of nodes can serve as supporting evidence of the relationship between those nodes with semantic information provided by that meta-path. Thus, a count of path instances or a path count can be used as a relatedness or similarity measure between two nodes in a heterogeneous network.

Definition 2.5 (Path instance). Suppose that $M = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$ defined over a network schema $S_G(A, R)$ of a heterogeneous network G with a node type mapping function $\rho: V \rightarrow A$ and a link type mapping function $\varphi: E \rightarrow R$, where V is the set of nodes, E is the set of links, A is the set of node types, and R is the set of link types in G. A path $v_1 \xrightarrow{e_1} v_2 \xrightarrow{e_2} \ldots \xrightarrow{e_l} v_{l+1}$ is a path instance under meta-path M if $\rho(v_i) = A_i$ for all $i = 1, 2, \ldots, l + 1$ and $\varphi(e_j) = R_j$ for all $j = 1, 2, \ldots, l$. The number of all path instances extracted from G for a particular pair of node v_1 and v_{l+1} is called a path count under meta-path M.

Based on the bibliographic network and the given meta-paths, path instances under each meta-path can be extracted from the network as shown in Table 2.3. For example, the path Sun-NetClus-Han is a path instance corresponding to the meta-path $A_aA_pA_a$. Paths which start and terminate at the same nodes are excluded from measuring similarity or relatedness between two different nodes. A path count is the simplest measure for evaluating relatedness between nodes. Based on the meta-path $A_aA_pA_a$, two authors who show the greatest number of common papers are San, Y. and Han, J. Three path instances which demonstrate their collaborations are Sun-NetClus-Han, Sun-PathSim-Han, and Sun-GenClus-Han. Based on the meta-path $A_aA_pA_vA_pA_a$, two authors who publish their papers in the same venue are Sun, Y. and Han, J. In addition, meta-paths can also be used to identify relationships between different node types, such as between author and venue nodes. The meta-path $A_aA_pA_v$ observes venues that each author prefers to publish their works. For example, Sun, Y. published two papers in VLDB (i.e. Sun-PathSim-VLDB and Sun-GenClus-VLDB) and one paper in KDD (i.e. Sun-NetClus-KDD). The high value of a path count between an author and a venue may imply to the author's preference or research interests.

Table 2.3 Path instances corresponding to the given meta-paths (revised from [63])

Meta-path	Path instance
Author-Paper-Author ($A_aA_pA_a$)	Sun-NetClus-Han, Sun-PathSim-Han, Sun-GenClus-Han, Sun-NetClus-Yu, Sun-PathSim-Yan, Yu-NetClus-Yan, Han-PathSim-Yan
Author-Paper-Venue-Paper-Author ($A_aA_pA_vA_pA_a$)	Sun-PathSim-VLDB-GenClus-Han
Author-Paper-Venue ($A_aA_pA_v$)	Sun-NetClus-KDD, Sun-PathSim-VLDB, Sun-GenClus-VLDB, Yu-NetClus-KDD, Han-NetClus-KDD, Han-PathSim-VLDB, Han-GenClus-VLDB, Yan-PathSim-VLDB

Path counts from different meta-paths contribute relatedness measures from different point of views. For example, the meta-path $A_aA_pA_a$ measures the numbers of common papers that both authors work on, but the meta-path $A_aA_pA_vA_pA_a$ measures the number of common venues

where both authors publish their works. The former suggests authors who prefer to work together whereas the latter could search for authors who are in the same research fields or share their interests. Therefore, multiple meta-paths are usually employed to gain and combine diverse semantic information for identifying node relatedness and similarity in a heterogeneous network. Moreover, meta-paths with the length more than one or two can be used to identify more sophisticated relationships which cannot be detected by simple meta-paths [64].

2.4.2 Applications of meta-paths in drug repositioning

Most networks in real world are heterogeneous networks where integrate diverse node and relation types, such as a social network fused across platforms (i.e. Facebook, Twitter, etc.). This heterogeneity brings riches of semantic information which needs a powerful approach to mine its. A meta-path based approach is an effective method that can extract semantic information from a heterogeneous network. During many recent years, this method has been gained attention and widely adopted in numerous applications, such as decision making, information retrieval, product recommendation, drug-target interaction predictions, and drug repositioning.

To the best of my knowledge, only few meta-path based methods were proposed for predicting drug-disease associations due to their recent introducing in drug repositioning. Tian et al. [45] proposed HeteSim_DrugDisease (HSDD) for identifying new drug-disease associations. They employed a meta-path based measure called HeteSim [65] for predicting new links between drugs and diseases in a heterogeneous network of drug and disease similarity scores. For each drug-disease pair, multiple meta-paths with lengths less than five were utilized, and then multiple HeteSim scores were combined with penalization of longer meta-paths due to their less semantic contributions. As a result, HSDD outperforms existing methods that use the same drug-disease network such as HGBI (a network propagation method) and MBiRW (a bi-random walk method).

With machine learning techniques, Wu et al. [11] developed Ensemble Meta-Paths and Singular Value Decomposition (EMP-SVD) for predicting drug-disease associations. From the drug-protein-disease heterogeneous network, they utilized five meta-paths to generate five feature matrices containing path counts of all drug-disease pairs. Each of them was reduced its dimension

by using SVD to obtain a latent feature matrix, and then it was used to build a random forest classifier. They combined five classifiers to obtain an ensemble model for classifying drug-disease pairs. Despite less data used (i.e. drug-disease, drug-protein, and disease-protein associations), EMP-SVD significantly outperforms other methods that require more drug and disease data, such as PREDICT, LRSSL, and SCMFDD. Moreover, EMP-SVD does not precompute drug and disease similarity scores but takes advantages of meta-paths to extract more meaningful information about drugs and diseases from the heterogeneous network.

In addition to SVD, network embedding with meta-paths is also applied to find low-dimensional representation of drugs and diseases from a heterogeneous network. Yang et al. [66] proposed Heterogeneous network Embedding for Drug-disease association (HED) to uncover new associations between drugs and diseases. This method adopts a meta-path based network embedding method called metapath2vec [67]. In brief, metapath2vec employs meta-paths in guiding a random walk algorithm to create neighborhood information of a node and then leverages the skip-gram model to obtain a low-dimensional representation vector of each node. This vector of a drug and a disease node were concatenated to obtain a feature vector of a drug-disease pair. Then, Yang et al. [66] built an SVM classifier to predict drug-disease associations based on the network embedding features. By network embedding and meta-paths, HED outperforms the Random Walk with Restart on the Heterogeneous network (RWRH) method. Recently, Zhou et al. [8] also proposed another network embedding method with meta-paths called NEDD or Network Embedding for predicting Drug-Disease associations. They employed a two-layer heterogeneous network, integrating drug and disease similarity scores with known drug-disease associations, and adopted a new network embedding method with meta-paths called HIN2vec [44]. By using HIN2vec, they obtained low-dimensional representations of drug and disease nodes, and then they performed an element wise product between a latent feature vector of a drug and that of a disease to obtain a feature vector of a drug-disease pair. Based on the network embedding features, a random forest classifier was used to predict drug-disease associations. They found that NEDD outperforms other methods that use the same heterogeneous network (e.g. HGBI and MBiRW). Furthermore, they claimed that the superior performance of NEDD results from utilizing meta-paths which can extract high-order relationships between drugs

and diseases, especially when the first-order relationships disappear in unknown drug-disease associations.

These meta-path based methods can demonstrate the capability of meta-paths in capturing valuable semantic information for predicting drug-disease associations. However, some limitations of the existing methods should be mentioned for further improvements. First, most methods (e.g. HSDD and EMP-SVD) discard information of intermediate nodes along the meta-path and focus on only two ending nodes in the path. For any two ending nodes, path instances under the meta-path are treated the same way although they include different intermediate nodes in the paths. Generally, in drug repositioning, intermediate nodes are important indicators that link between drugs and diseases and may be greatly valuable for discovering drug-disease associations such as proteins and GO functions associated with drugs and diseases. Second, it is difficult to integrate information from multiple meta-paths in the network embedding methods (i.e. HED and NEDD). Generally, node representation features from a single meta-path, which was manually selected, were used to predict drug-disease associations, leading to loss of information from other meta-paths [68].

2.5 Positive-Unlabeled (PU) learning

In this section, the definition of PU learning and the reasons why it is important are described. Some examples of PU learning methods are also provided to illustrate how they manipulate positive and unlabeled data. Furthermore, existing PU learning methods developed for drug repositioning are reviewed to demonstrate how PU learning can be applied for predicting drug-disease associations.

2.5.1 Introduction to PU learning

To create a binary classification model, both positive-labeled and negative-labeled samples are required. Nonetheless, in many real situations, only positive-labeled data are available. For example, in medical records of patients, only diseases that have been diagnosed are recorded, but there is no information about diseases that have not been diagnosed for the patients. Visited pages of users are recorded and labeled as user interests, but it cannot explicitly determine in which pages users are not interested. Some positive-labeled samples are obtainable in such

situations, and the remaining samples are unlabeled, where each of them can be either positive or negative. With these positive and unlabeled data, a learning approach that is different from a traditional binary classifier is required.

A solution used to deal with positive and unlabeled data is learning from only positive samples, or one-class learning. A well-known method of this approach is one-class support vector machine (SVM) [69]. The aim of this method is to estimate the smallest boundary with the hyperparameters of a radius and a center which covers all available positive data points, as shown in Figure 2.8(a). Nevertheless, the drawback of this method is that it may lead to overfitting models which cannot practically recognize positive samples from unlabeled samples, especially when only small amounts of positive samples are labeled [70]. To improve the performance of one-class learning models, both positive and unlabeled samples are introduced into the learning process. This approach is known as Positive-Unlabeled (PU) learning. The assumption of PU learning is that a group of unlabeled samples contains both positive and negative samples. Thus, the major goal of PU learning is to accurately identify positive sample from unlabeled samples. With the distributional information in unlabeled data, the performance of a classifier can be improved as illustrated in Figure 2.8(b).

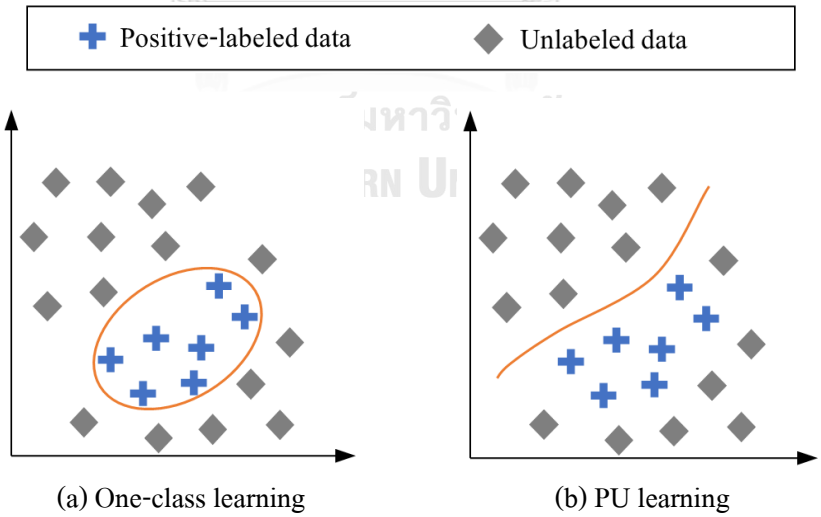


Figure 2.8 One-class learning versus PU learning (adapted from [70])

PU learning has been known to researchers in the field of machine learning since the early 2000s [71], and it has been gained a lot of attention during these recent years. PU learning

has been successfully adopted in various applications, such as remote sensing data analysis, recommendation systems, and bioinformatics. PU learning is simply considered as a special case of semi-supervised learning, which exploits some of labeled samples, including positives and negatives, with unlabeled samples to construct a decision boundary [71].

2.5.2 Categories of PU learning methods

Generally, PU learning methods can be classified into three broader categories: two-step methods, biased learning methods, and bootstrap sampling based methods [71]. In two-step methods, a set of reliable negative samples is primarily specified, and then a traditional binary classifier is employed to learn from positive and those selected negative samples. One commonly used strategy to find reliable negative samples is the spy strategy (Figure 2.9).

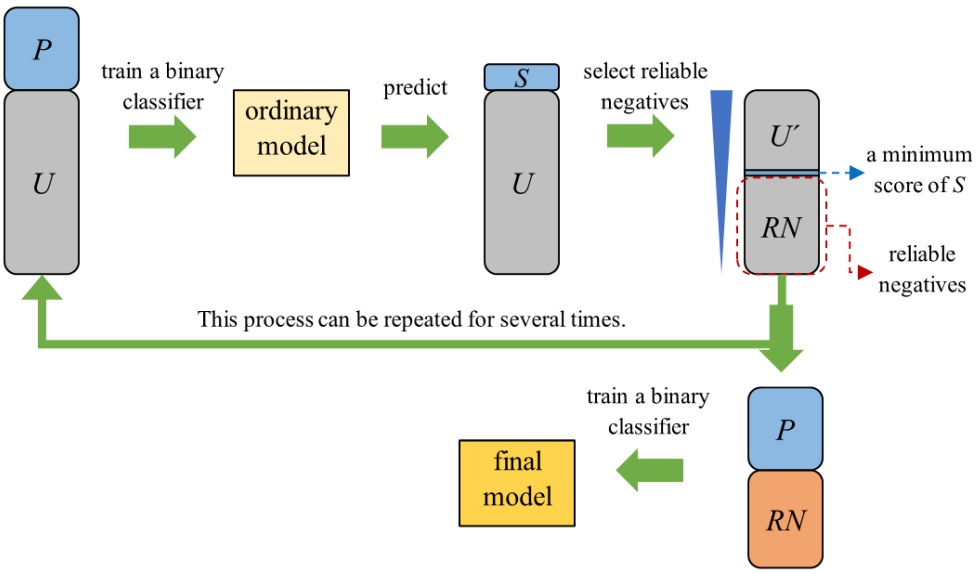


Figure 2.9 Illustration of the spy strategy

In the spy strategy, unlabeled samples (U) are initially used as negative samples to train an ordinary binary classifier with positive samples (P). Then, some of positive samples are randomly selected and act as spy positive samples (S). The trained model is used to predict both spy samples and all unlabeled samples. The minimum predicted score of S serves as a threshold score for identifying a set of reliable negative samples (RN), unlabeled samples with predicted scores less than the threshold score. This whole process can be repeatedly conducted for several

times with different sets of spy samples to obtain more reliable negative samples. Finally, both positive and a final set of reliable negative samples are used to create a final classifier.

Another technique commonly used in most two-step methods is the heuristic strategy. Based on different domain knowledge, assumptions, and techniques, diverse heuristic methods were proposed for identifying their own negative samples. For example, Zheng et al. [72] devised similarity scores between drugs using multiple drug information to identify pairs of drugs which have less similarity scores as credible negative samples in predicting drug-drug interactions. For the same application, DDI-PULearn (Drug-drug interaction prediction based on PU learning) [73] selects reliable negative samples using iterative one-class SVM with a high-recall constraint and drug features different from the previous one.

Although two-step PU methods have been successfully applied in many applications, some disadvantages of this strategy should be mentioned. In a particular domain, diverse techniques were proposed for the reliable negative selection even though there is no standard method to measure the quality of negatives obtained. No one can prove that all reliable negative samples acquired are exactly negative or cover the whole set of negative samples in a data set. Moreover, most two-step heuristic methods use a threshold score (e.g. drug-drug similarity scores for screening out likely potential drug-drug interactions) to include or exclude unlabeled samples into a set of reliable negatives [74], and this set of reliable negatives can directly affect a model performance. However, an optimal threshold score may be data dependent, leading to different sets of negative samples acquired and different model performance achieved.

The second category of PU learning methods is based on biased learning. In these methods, unlabeled samples are considered as negative samples and used to train a traditional classifier with biased weights of samples from different classes. A well-known example of biased learning methods is a biased SVM method [75], which was firstly introduced for text classification. With unlabeled samples that vastly outnumber positives, a biased SVM assumes that most of unlabeled samples are likely negative. In a cost function of this method, a weight of positive errors is typically greater than that of unlabeled errors to minimize the number of unlabeled samples which will be predicted to be positive. There are also other machine learning methods that were proposed with the idea of biased learning for PU data such as weighted logistic regression [76]. An advantage of biased learning is that it does not require a pre-determined set of

reliable negatives like two-step methods. Nevertheless, one of its disadvantages is that the values of biased weights typically deviate relying on data used in the tuning process [74]. These values could result in poor performance of a biased model due to its under-prevention or over-prevention of positive and unlabeled errors.

Another class of PU learning methods is based on bootstrap sampling and ensemble learning. This strategy takes advantages of classifiers' instability occurred when classifiers are trained on positive instances and bootstrap samples of unlabeled instances. Then, an ensemble model is responsible to improve the performance of those unstable classifiers. Mordelet and Vert [77] proposed a bagging SVM for learning PU data. This method is based on the bootstrap aggregating technique, also called the bagging-like strategy (Figure 2.10). In a bagging SVM, unlabeled data are resampled by bootstrap subsampling, randomly selecting with replacement, to generate t bootstrap samples (U_1, U_2, …, U_t). Each of them with the same set of positive samples are used to create each base SVM classifier. Then, multiple predictions of a testing sample are aggregated to obtain an ensemble prediction. According to their results, a bagging SVM outperforms one-class SVM and runs faster than a biased SVM, especially when unlabeled greatly outnumbers positive samples.

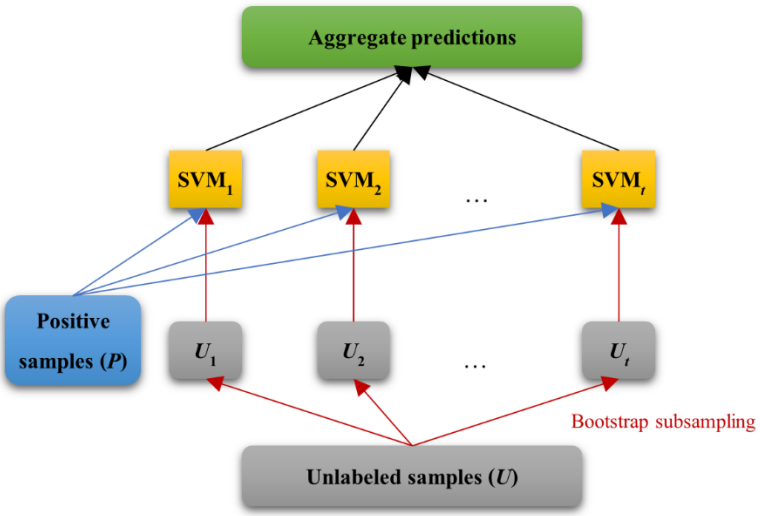


Figure 2.10 A bagging SVM

A bagging SVM has been adopted in various applications. For example, Deepika and Geetha [70] utilized a bagging SVM to predict drug interactions from multiple drug input

features, such as chemical substructures, drug targets, side effects, and drug indications. Recently, Singh et al. [78] compared several deep learning and machine learning methods, including a bagging SVM, for developing an automated COVID-19 detection using lung computerized tomography (CT) scan images. They found that a bagging SVM outperforms other methods with an accuracy of up to 96%. The success of a bagging SVM motivates other researchers to introduce other bootstrap aggregating based methods for PU data, such as a robust ensemble of SVMs [79] and a corrected ensemble of SVMs [80]. One advantage of bootstrap sampling based methods is that they do not require reliable negative samples. However, base classifiers may still suffer from unlabeled positive instances in bootstrap samples, which could propagate classification errors to the ensemble model [74].

2.5.3 PU learning methods for drug repositioning

To the best of my knowledge, only few PU learning methods were proposed for predicting drug-disease associations. Wu et al. [81] developed a PU learning method for drug-disease associations (PUDrDi). This method exploits a biased SVM classifier for learning features of drug-disease pairs derived from drug chemical substructures and disease symptoms. They compared the performance of PUDrDi with some machine learning methods (i.e. *k*-nearest neighbors and a random forest classifier) and one well-known drug repositioning method (i.e. HGBI). The results showed that PUDrDi outperforms all compared methods in most evaluation metrics.

Wu et al. [11] proposed the EMP-SVD (Ensemble Meta-Paths and Singular Value Decomposition) method, which is a two-step PU learning method. They assumed that a drug and a disease which interact with some common proteins are likely associated with each other. Therefore, they selected unlabeled pairs of drugs and diseases which have no common proteins between them as reliable negative samples. With this heuristic strategy, EMP-SVD achieves the highest values in all evaluation metrics when compared to several state-of-the-art methods, such as PREDICT, TL_HGBI, MBiRW, and LRSSL. Another recent method which is also based on the two-step strategy is Topological Similarity and Singular Value Decomposition (TS-SVD) [47]. Although this method is relied on the same heterogeneous network used in EMP-SVD, different techniques to construct features of drug-disease pairs and select reliable negative

samples are employed in TS-SVD. This method assumes that a drug and a disease node which have *k*-step paths (*k* = 1, 2, 3) linking between them in the drug-protein-disease heterogeneous network are more likely associated with each other. Thus, TS-SVD selects unlabeled pairs of drug and disease nodes that have no *k*-step paths between them (*k* = 1, 2, 3) to serve as reliable negatives. When compared to other methods, it was found that TS-SVD performs better than all compared methods, including EMP-SVD, despite less drug and disease information used.

2.6 Extreme Gradient Boosting (XGBoost)

The Extreme Gradient Boosting (XGBoost) [82] method is a state-of-the-art machine learning method that is widely used during these recent years in many applications, such as personal credit assessment, network intrusion detection, financial trading, and drug discovery [83]. This is an improved implementation of the gradient boosting tree (GBT) algorithm to enhance its computational speed and performance. XGBoost, and also GBT, is an ensemble model with a boosting technique, a sequential aggregation of multiple weak learners (e.g. decision trees) to become a strong one. In this section, a mathematical description of XGBoost is provided to gain insights how it is formulated and how it works.

Given a data set $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$ with *n* samples and *m* features. Note that $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$ is a feature vector of the i^{th} sample, and y_i is a class label of the i^{th} sample, where $i = 1, 2, \ldots, n$. The XGBoost model creates *t* base classifiers (i.e. decision trees) providing *t* predictions which are sequentially aggregated as $\hat{y}_i^{(t)}$ for the i^{th} sample, as shown in (2.1). Notice that $\hat{y}_i^{(t)}$ is an aggregate prediction result of the i^{th} sample in the t^{th} iteration, and $f_k(x)$ is a decision tree of the k^{th} iteration, where $k = 1, 2, \ldots, t$.

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\vdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
\tag{2.1}
$$

To build trees in the XGBoost model, it is formulated as an optimization problem as shown in (2.2), where $L(y,\hat{y})$ is the total loss function of the model, and $l(y_i,\hat{y}_i)$ is a loss function of each sample when comparing between its actual and predicted class. $\Omega(f)$ is the regularization term that penalizes a regression tree according to its complexity, as shown in (2.3). In the regularization term, the first term penalizes a tree with more depth and too many leaf nodes, leading to very few examples in each leaf node. The second regularization term is for smoothing the final learnt weights (or the predicted scores of the leaf nodes) to avoid overfitting. Note that γ is the minimum loss reduction, T is the number of leaf nodes in a tree, λ is the regularization parameter to avoid overfitting, and ω_j is a predicted score of the j^{th} leaf node in a tree.

$$\min L^{(t)}(y,\hat{y}^{(t)}) = \min\left(\sum_{i=1}^{n} l(y_i,\hat{y}_i^{(t)}) + \sum_{k=1}^{t}\Omega(f_k)\right) \tag{2.2}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2 \tag{2.3}$$

According to (2.1), substitute $\hat{y}_i^{(t)}$ by $\hat{y}_i^{(t-1)} + f_t(x_i)$ in (2.2) and use the second-order Taylor expansion to approximate the total loss function of the t^{th} iteration. Then, this total loss function can be derived and simplified as shown in (2.4), where g_i and h_i are the first and second partial derivatives of the loss function $l(y_i,\hat{y}_i^{(t)})$ with respect to $\hat{y}^{(t-1)}$.

$$\begin{aligned}\min L^{(t)}(y,\hat{y}^{(t)}) &= \min\left(\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)\right)\\ &= \min\left(\sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + \Omega(f_t)\right)\end{aligned} \tag{2.4}$$

Denote $I_j = \{i \mid q(x_i) = j\}$ represent a set of instances that belong to the j^{th} leaf node in a tree, where $q(x_i)$ is a tree's structure function that map data instance x_i to the j^{th} leaf node, and j = 1, 2, …, T. Thus, a function of a decision tree $f(x)$ can be represented by $\omega_{q(x)}$. Substitute

$f(x) = \omega_{q(x)}$ into (2.4), then the rewritten equation can be shown in (2.5), where $G_j = \sum_{i \in I_j} g_i$ and $H_j = \sum_{i \in I_j} h_i$.

$$\begin{aligned}
\min L^{(t)}(y, \hat{y}^{(t)}) &= \min\left(\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T_t + \frac{1}{2}\lambda \sum_{j=1}^{T_t} \omega_{t,j}^2 \right) \\
&= \min\left(\sum_{j=1}^{T_t} \left[\left(\sum_{i \in I_j} g_i \right) \omega_{t,j} + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_{t,j}^2 \right] + \gamma T_t \right) \qquad (2.5) \\
&= \min\left(\sum_{j=1}^{T_t} \left[G_j \omega_{t,j} + \frac{1}{2} \left(H_j + \lambda \right) \omega_{t,j}^2 \right] + \gamma T_t \right)
\end{aligned}$$

Note that T_t is the total number of leaf nodes in the t^{th} tree, $\omega_{t,j}$ is a predicted score of the j^{th} leaf node in the t^{th} tree. Now the objective function of the t^{th} iteration can be obtained as shown in (2.5). This is a function of $\omega_{t,j}$, because g_i and h_i are values that can be computed from the loss function. To get the optimal $\omega_{t,j}$ at the t^{th} step ($\omega^*_{t,j}$), take the derivative of the variable $\omega_{t,j}$ and solve for the root of the equation. Then, $\omega^*_{t,j}$ that minimizes the objective function is shown in (2.6), and its minimum value of the objective function ($\tilde{L}^{(t)}$) is shown in (2.7).

$$\omega^*_{t,j} = -\frac{G_j}{H_j + \lambda} \qquad (2.6)$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T_t} \frac{G_j^2}{H_j + \lambda} + \gamma T_t \qquad (2.7)$$

The function in (2.7) can be used to evaluate the quality of a tree with the structure $q(x)$, similar to the impurity score. The example for illustrating the method to compute this score is given in Figure 2.11.

Data

Instance index	Name	Age	Sex	Gradient statistics
1	A	13	Male	g_1, h_1
2	B	12	Female	g_2, h_2
3	C	77	Male	g_3, h_3
4	D	25	Female	g_4, h_4
5	E	60	Female	g_5, h_5

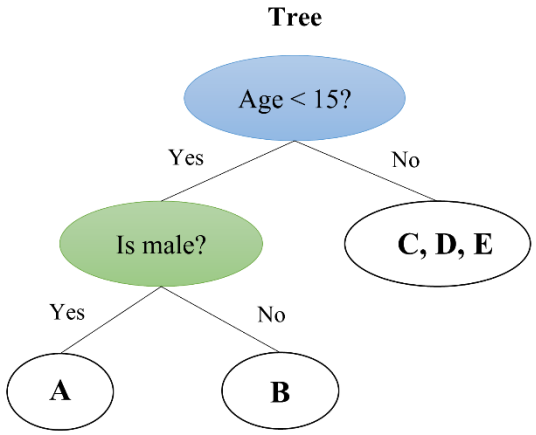


Figure 2.11 An example illustrating the score calculation of a tree structure (adapted from [82])

According to Figure 2.11, $I_1 = \{1\}$, $I_2 = \{2\}$, and $I_3 = \{3, 4, 5\}$. From (2.7), the score of the given tree can be computed as in (2.8). The smaller score indicates the better structure of the decision tree.

$$\begin{aligned} L(q) &= -\frac{1}{2}\sum_{j=1}^{3}\frac{G_j^2}{H_j+\lambda}+3\gamma \\ &= -\frac{1}{2}\left(\frac{(g_1)^2}{h_1+\lambda}+\frac{(g_2)^2}{h_2+\lambda}+\frac{(g_3+g_4+g_5)^2}{(h_3+h_4+h_5)+\lambda}\right)+3\gamma \end{aligned} \tag{2.8}$$

It is impossible to generate all possible decision trees and then compute the structure scores to select the best one at each iteration. Thus, a greedy algorithm that begins with the tree of a single node and then iteratively splits to create the optimal tree is exploited. The formula that is used to select the split candidates is written in (2.9), where I_L and I_R represent the sets of instances in the left and right nodes after the split. Note that $I = I_L \cup I_R$, where I is the instance set before the split. The chosen candidate for the split is the one that gives the maximum loss reduction (L_{split}).

$$L_{split} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{2.9}$$

Because XGBoost is implemented with the aim to enhance GBT, several unique features of XGBoost make it more attractive and popular until now. The advantages of XGBoost are discussed as follows:

- Parallelization - XGBoost can take advantages of multiple cores on CPU, because it is designed by a block structure, where data in each block can be sorted parallelly. This makes XGBoost run faster than many other boosting algorithms. In addition, XGBoost provides an option to construct each tree in parallel.
- Sparsity awareness - Most of large data matrices contain a lot of zero elements, which probably are missing values or generated by some data pre-processing techniques (e.g. one hot encoding features). To handle with missing values, XGBoost assigns them to a default direction that will minimize training loss.
- Regularization - XGBoost provides an option to penalize a complex model using L1 and L2 regularization to prevent overfitting models.
- Effective pruning - Most efficient algorithms for tree pruning cost high time complexity. XGBoost constructs a tree with the depth up to the parameter *max_depth* that is primarily specified, and then it begins pruning backward until the loss reaches below the given threshold.
- Continued training - XGBoost provides an option to retrain the fitted model on new coming data.

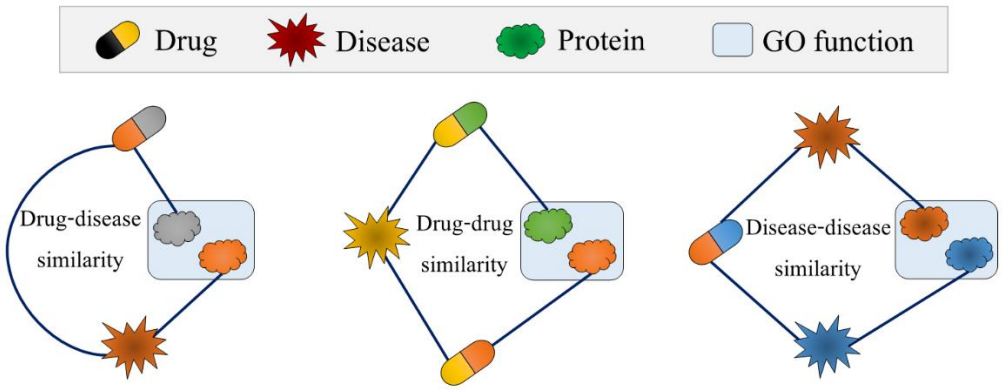
CHAPTER III

PROTEINS VERSUS FUNCTIONAL INFORMATION

In this chapter, utilizing proteins and functional information, or gene ontology (GO) functions, in discovering relationships of drugs and diseases are discussed. The aim of this study is to demonstrate the advantages of using GO functions when compared to a classical technique that directly uses proteins. Herein, the key to identify associations of drugs and diseases is a similarity based on proteins or GO functions that are associated with drugs and diseases.

3.1 An overview of this study

Since proteins play crucial roles in drug actions and disease processes, high similarities based on drug-associated and disease-associated proteins could indicate the relationships of drugs and diseases [49, 52]. Nevertheless, some relationships with lack of some common proteins cannot be detected by protein-based similarities [58]. For example, a drug and a disease are associated with two different proteins, but these proteins work together in the same biological functions. To overcome this limitation, the scope of the consideration should be extended beyond proteins. Herein, functional annotations of proteins or GO functions are used to reveal the relationships of drugs and diseases for drug repositioning. There are three relationships of drugs and diseases investigated in this study, as shown in Figure 3.1.



(a) Drug-disease relationship (b) Drug-drug relationship (c) Disease-disease relationship

Figure 3.1 Three relationships of drugs and diseases under investigation

The first case is the direct relationship between drugs and diseases. Figure 3.1(a) shows that although a drug and a disease are relevant to different proteins, this kind of relationships could be detected by some common GO functions or drug-disease similarity based on GO functions. The second case is the relationship between two drugs which can treat some common diseases. In Figure 3.1(b), two drugs could interact with different proteins, but they could affect the same downstream biological functions resulting in the abilities to treat the same diseases. For this case, drug-drug similarity based on GO functions are measured to identify potential drug-drug associations. The third case is the relationship between two diseases which can be treated by the same drugs. Figure 3.1(c) demonstrates that two diseases could be associated with each other through some common GO functions rather than proteins. For this case, disease-disease similarity based on GO functions is more promising to predict disease-disease associations, when compared to the similarity based on proteins.

An overview of this study is shown in Figure 3.2. Four data sets, functional annotation data of human proteins, drug-protein association data, disease-protein association data, and drug-disease association data, were used. Based on drug-disease association data, all drug-disease, drug-drug, and disease-disease pairs can be generated and prepared for further classifications.

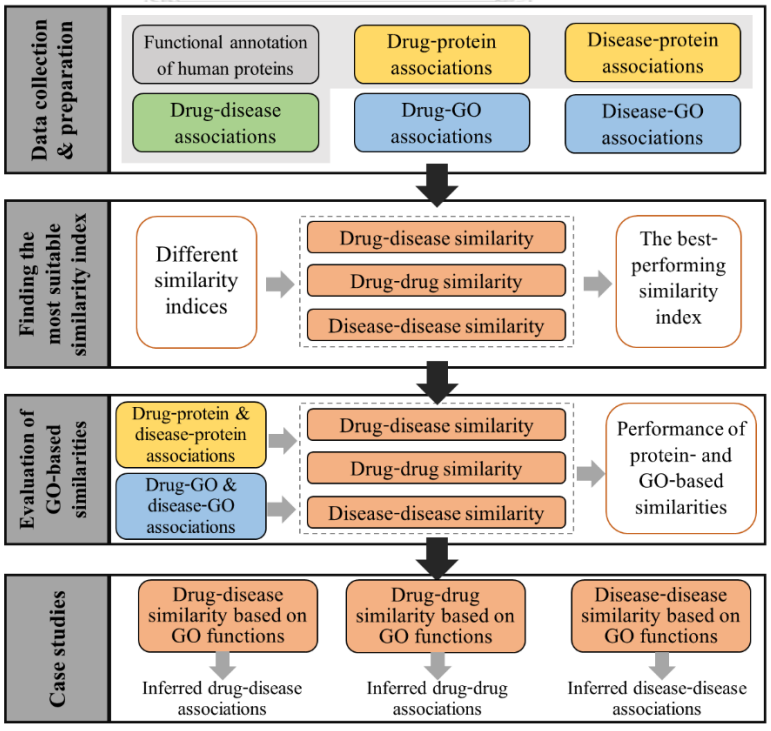


Figure 3.2 A conceptual diagram depicting an overview of the study

Drug-GO and disease-GO associations were derived from the functional annotations, the drug-protein associations, and the disease-protein associations. Based on drug-GO associations, disease-GO associations, and a similarity index, similarity scores between drugs and diseases, between drugs, and between diseases can be computed. These similarities are called as GO-based or functionality-based similarities. Similarly, protein-based similarity scores between drugs and diseases, between drugs, and between diseases can be calculated from drug-protein and disease-protein associations.

Since a number of similarity indices can be used to compute both protein-based and functionality-based similarity scores, the most suitable similarity index for those tasks should be initially determined by comparing among different similarity indices. Based on the selected similarity index, both protein-based and functionality-based similarity scores were computed and then used to classify drug-disease, drug-drug, and disease-disease associations. The performance of protein-based similarities is used as the baseline performance, and it is compared with that of functionality-based similarities to demonstrate the improved performance of functionality-based similarities. To exemplify some predictions using functionality-based similarities, three case studies (an inferred drug-disease, drug-drug, and disease-disease association) are shown with their supporting evidence.

3.2 Data sets

For this research, four main data sets were collected from different sources as shown in. Table 3.1. Drug-disease associations were downloaded from Comparative Toxicogenomics Database (CTD) [84], a version released in August 2019, and the study of Gottlieb et al. [3]. To obtain a reliable data set, only CTD drug-disease relations with supporting literature were used. Gottlieb et al. [3] created a gold-standard data set by manually assembling drug-disease associations found in at least two sources of multiple databases. Both data sets were systematically integrated to obtain a larger one. The collection of drug-disease associations used in this research can be accessible at http://ieeedataport.org/3540. All approved drugs and their target proteins were downloaded from DrugBank (version 5.1.3) [85]. Diseases and their associated proteins were collected from the curated data available on DisGeNET (version 6.0)

[86]. The functional annotation data of all human proteins were downloaded from the Gene Ontology Annotation (GOA) database (version 191) [87].

Table 3.1 The list of data with their sources and versions

Data	Source	Version
Drug-disease associations	CTD [84] and the study of Gottlieb et al. [3]	CTD version released in August 2019
Drug-protein associations	DrugBank [85]	5.1.3
Disease-protein associations	DisGeNET [86]	6.0
Functional annotation of human proteins	GOA [87]	191

3.3 Methods

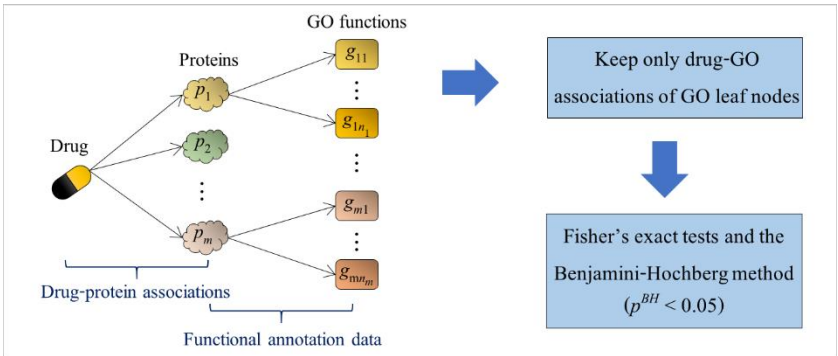
3.3.1 Preparation of drug-disease, drug-drug, and disease-disease pairs

In this step, all drug-disease, drug-drug, and disease-disease pairs are generated and labeled based on the list of known drug-disease associations. All drug-disease pairs were generated by combining all drugs and all diseases. Drug-disease pairs which are known drug-disease associations are considered as positive samples whereas the remaining pairs are unlabeled samples. In case of drug-drug pairs, all possible drug-drug pairs were generated by pairing two distinct drugs together. Drug-drug pairs which share at least one common disease are positive samples, and the remaining pairs are unlabeled samples. Similarly, all possible disease-disease pairs were originated by pairing two different diseases. Disease-disease pairs which have some common diseases are labeled as positive samples, and the remaining pairs are unlabeled samples.

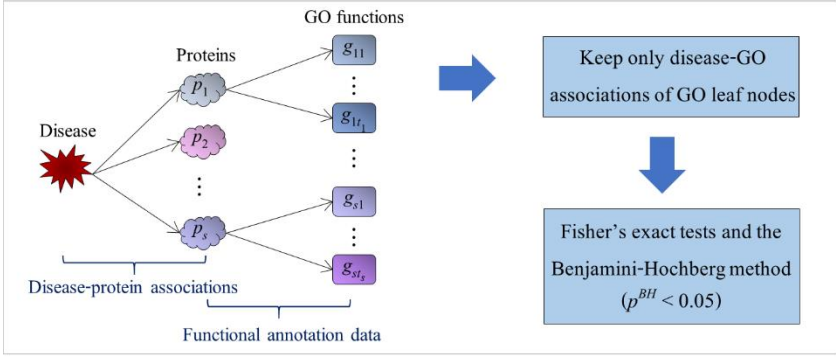
3.3.2 Construction of drug-GO and disease-GO associations

In this work, drug-GO and disease-GO associations are principal components mainly used to connect drugs and diseases. The schematic diagrams that summarize the methods of constructing drug-GO and disease-GO associations are shown in Figure 3.3. In this figure, m and s represent the number of proteins associated with a drug and a disease, respectively. Let n_1, n_2, …, and n_m be the number of GO functions associated with drug's proteins p_1, p_2, …, and p_m,

respectively. Similarly, t_1, t_2, …, and t_s are the number of GO functions associated with disease's proteins p_1, p_2, …, and p_s, respectively.



(a) Construction of drug-GO associations



(b) Construction of disease-GO associations

Figure 3.3 Schematic diagrams summarizing how to construct drug-GO and disease-GO associations

From Figure 3.3(a), the drug-protein association data and the functional annotation data of human proteins were used to map drugs to their associated GO functions. Similarly, all diseases were linked to their associated GO functions by using the disease-protein association data and the GO annotation data, as shown in Figure 3.3(b). GO functions of all GO aspect, including Cellular Component (CC), Molecular Function (MF), and Biological Process (BP), were used to link to drugs and diseases to gain as much as possible information about drugs and diseases. GO terms are represented via a tree-like structure which can describe the relations among different GO terms. GO functions at the higher level provide broader statements about genes and gene products than those at the lower level. By directly mapping a protein to its GO

functions, redundant GO terms from different levels can be linked to that protein. To avoid the redundant annotation, only drug-GO and disease-GO associations of the most detailed GO functions or GO leaf nodes were kept using the R package named multidimensional gene set analysis of genomic data (mdgsa) [88].

After that, an enrichment analysis was performed for selecting GO functions significantly associated with drugs and diseases using one-sided Fisher's exact tests [89]. For a given pair of a drug (a disease) and a GO function, the *p*-value (p) of that pair can be calculated by the hypergeometric distribution, as shown in (3.1). N and M represent the total number of human proteins and the number of proteins annotated by that GO function, respectively. n and m are the number of proteins associated with that drug (disease) and the number of proteins annotated by that GO function and associated with that drug (disease), respectively.

$$p = \sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} \tag{3.1}$$

The multiple testing conducted for all drug-GO or disease-GO associations increases a False Discovery Rate (FDR) resulting in an increased number of false positive associations obtained. To reduce false positives, the Benjamini-Hochberg method [90] was performed. To obtain Benjamini-Hochberg *p*-values (p^{BH}), *p*-values from the Fisher's exact tests (p) were adjusted following to (3.2). Note that i is a rank of a drug-GO or a disease-GO association when all drug-GO or disease-GO associations are sorted in ascending order according to their *p*-values from the Fisher's exact tests, and r represents the total number of tests. Drug-GO and disease-GO associations with p^{BH} values less than 0.05 are significant relations and then preserved for further using.

$$p_i^{BH} = \min\left\{ \min_{j \geq i} \left\{ \frac{r}{j} p_j \right\}, 1 \right\} \tag{3.2}$$

3.3.3 Measurement of protein-based and functionality-based similarities

After getting drug-GO and disease-GO associations, functionality-based similarity scores can be computed for all drug-disease, drug-drug, and disease-disease pairs, as illustrated in Figure 3.4. Then, these similarity scores were used to classify drug-disease, drug-drug, and disease-disease associations. A similarity index can be employed to measure functionality-based similarities from the lists of drug-GO and disease-GO associations. In parallel, protein-based similarities between drugs and diseases, between drugs, and between diseases are also measured from drug-protein and disease-protein associations. Because a number of similarity indices can be applied for the similarity measurement, comparison among several similarity indices to select the most suitable one for the measurement is necessary.

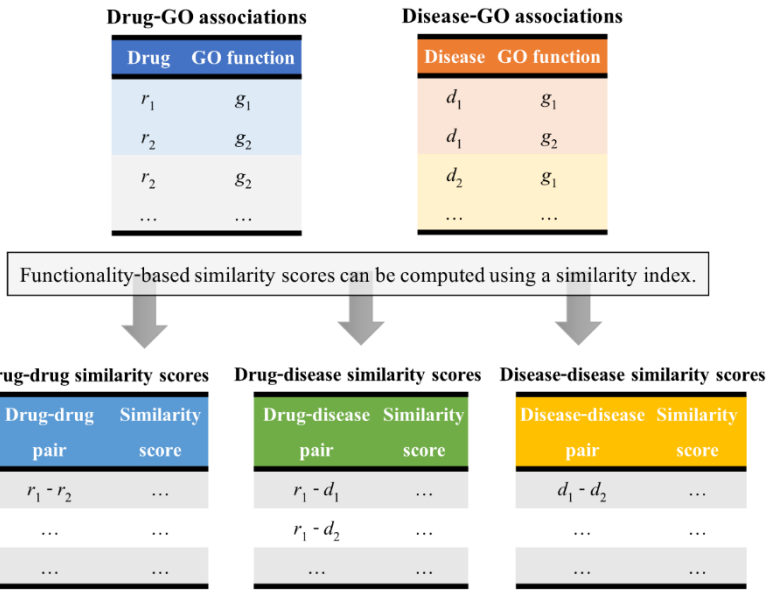


Figure 3.4 An overview of the measurement of functionality-based similarities

In this step, seven commonly used similarity indices, the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, and derived Jaccard similarity index, are applied for the measurement. Let x and y be a drug or a disease, and $S_{SimilarityIndex}(x, y)$ be a function that gives a similarity score between x and y by using a particular similarity index. X and Y are the set of proteins or GO functions associated with x and y, respectively. All similarity indices are

formulated as shown in (3.3) - (3.9), where | · | is the number of all elements in a set, and " \ " is the difference between any two sets.

$$S_{Jaccard}(x,y)=\frac{|X\cap Y|}{|X\cup Y|} \tag{3.3}$$

$$S_{BraunBlanquet}(x,y)=\frac{|X\cap Y|}{\max(|X|,|Y|)} \tag{3.4}$$

$$S_{Simpson}(x,y)=\frac{|X\cap Y|}{\min(|X|,|Y|)} \tag{3.5}$$

$$S_{Cosine}(x,y)=\frac{|X\cap Y|}{\sqrt{|X|\cdot|Y|}} \tag{3.6}$$

$$S_{Sorgenfrei}(x,y)=\frac{(|X\cap Y|)^2}{|X|\cdot|Y|} \tag{3.7}$$

$$S_{McConnaughey}(x,y)=\frac{(|X\cap Y|)^2-(|X\setminus Y|\cdot|Y\setminus X|)}{|X|\cdot|Y|} \tag{3.8}$$

$$S_{DerivedJaccard}(x,y)=\frac{\log(1+|X\cap Y|)}{\log(1+|X\cup Y|)} \tag{3.9}$$

3.3.4 Performance measurement

To compare among seven similarity indices or between protein-based and functionality-based similarities, the performance of all similarity indices, protein-based similarities, and functionality-based similarities is measured. According to the actual and predicted classes of testing samples, a confusion matrix can be generated as shown in Figure 3.5. *TP*, *FP*, *FN*, *TN* represent the number of true positives, false positives, false negatives, and true negatives, respectively. In a confusion matrix, the higher the number of true positives and true negatives of a model are, the better the model performs. Several confusion matrices were constructed and investigated to compare the prediction results of drug-disease, drug-drug, and disease-disease associations between using protein-based and functionality-based similarity scores.

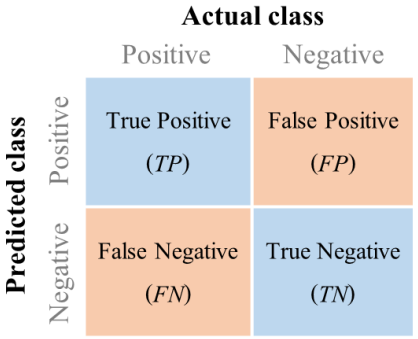


Figure 3.5 A confusion matrix

All information in a confusion matrix can be summarized as a single number of an evaluation metric for easier interpretation of the prediction results. Several well-known evaluation metrics were employed in this study, including precision (*PRE*), recall (*REC*), accuracy (*ACC*), and F_1-score (F_1). They can be calculated following to (3.10) - (3.13). However, only positive and unlabeled samples are identified in this study but not for negative samples. To estimate values of the evaluation metrics, all unlabeled samples are considered as negative ones. Values of all evaluation metrics are ranged from zero to one. Higher values of the metrics indicate better model performance.

$$PRE = \frac{TP}{TP + FP} \tag{3.10}$$

$$REC = \frac{TP}{TP + FN} \tag{3.11}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.12}$$

$$F_1 = \frac{2 \times PRE \times REC}{PRE + REC} \tag{3.13}$$

In addition, a Receiver Operating Characteristic (ROC) and a Precision-Recall (PR) curve are plotted to investigate the overall performance of protein-based and functionality-based similarities. An ROC curve is a plot between true positive rates (*TPR*) and false positive rates (*FPR*) which can be computed following to (3.14) - (3.15). Each value of *TPR* or *FPR* is calculated according to the one threshold score. To create this plot, threshold scores are changed

from low to high values. When negatives largely outnumber positives, values of *FPR* could be flatten by a high value of *TN* resulting in a misleading ROC curve [91]. An additional plot recommended for this situation is a PR curve, which is a plot of precision and recall values according to several threshold scores. To summarize information of an ROC and a PR curve, an Area Under an ROC curve (AUROC) and an Area Under a PR Curve (AUPRC) are computed. An AUROC and an AUPRC value are between zero and one. The higher the value of AUROC or AUPRC of a model is, the better the model performs.

$$TPR = \frac{TP}{TP + FN} \tag{3.14}$$

$$FPR = \frac{FP}{TN + FP} \tag{3.15}$$

3.3.5 Classification of drug-disease, drug-drug, and disease-disease associations

By using a given similarity index, protein-based or functionality-based similarity scores can be computed for all drug-disease, drug-drug, and disease-disease associations. In all association types, these similarity scores are directly used to identify positive associations from unlabeled associations. Each drug-disease, drug-drug, or disease-disease association is classified as either a "1" (positive) or "0" (unlabeled) based on its similarity score (x) and a given threshold score (t) as shown in (3.16), where $f(x)$ is a function assigning a binary class for a sample.

$$f(x) = \begin{cases} 1 \text{ if } x \geq t \\ 0 \text{ if } x < t \end{cases} \tag{3.16}$$

In predicting drug-disease, drug-drug, and disease-disease associations as the binary classes, the Youden's index [92] is used to find an optimal threshold score. Based on this method, an optimal threshold (t^*) is estimated at a point which provides the maximum difference between the values of *TPR* and *FPR* in an ROC curve. By [93], this can be written as shown in (3.17), where t is a threshold score. TPR_t and FPR_t represent the values of true positive rate and false positive rate at t, respectively.

$$t^* = \max_t \left\{ TPR_t - FPR_t \right\} \tag{3.17}$$

3.4 Results

3.4.1 Data summarization

For further investigation drug-associated and disease-associated proteins in a protein-protein interaction (PPI) network, drugs and diseases whose all associated proteins cannot be mapped into the PPI network are excluded from this study. In total, 904 drugs, 524 diseases, 6,782 proteins, and 8,301 GO functions are included in the study (Table 3.2). Out of these GO functions, 901 (10.9%) are Cellular Component (CC) terms, 2,407 (29.0%) are Molecular Function (MF) terms, and 4,993 (60.1%) are Biological Process (BP) terms.

Table 3.2 The total numbers of drugs, diseases, proteins, and GO functions

Drugs	Diseases	Proteins	GO functions
904	524	6,782	8,301

The total numbers and some statistical information of all relations are summarized in Table 3.3. For the drug-related relations, there are 9,427 drug-protein interactions and 52,038 drug-GO associations in total. For disease-related associations, there are 32,659 disease-protein associations and 91,998 disease-GO associations in total. By mapping proteins to their associated GO functions, the greater numbers of drug-GO associations (up to six times) and disease-GO associations (up to three times) are obtained when compared to the total numbers of drug-protein interactions and disease-protein associations, respectively. This is because most proteins are often related to more than one GO function or annotated with terms of more than one GO aspect [94].

Table 3.3 Statistical information of drug-protein, drug-GO, disease-protein, and disease-GO associations

Statistical information	**Relations of drugs**		**Relations of diseases**	
	Drug-protein interactions	**Drug-GO associations**	**Disease-protein associations**	**Disease-GO associations**
Total number	9,427	52,038	32,659	91,998
Mean (per drug or disease)	10.4 proteins	57.6 GO functions	62.3 proteins	175.6 GO functions
Standard deviation (per drug or disease)	13.1 proteins	51.4 GO functions	162.7 proteins	217.6 GO functions
Minimum	1.0 protein	2.0 GO functions	1.0 protein	1.0 GO function
Maximum	188.0 proteins	545.0 GO functions	1,086.0 proteins	944.0 GO functions

Furthermore, the numbers of proteins or GO functions associated with a drug and a disease are also investigated (Table 3.3). Due to the larger number of drug-GO connections, a drug involves with 57.6 ± 51.4 (Mean ± SD) GO functions of any aspects whereas a drug interacts with only 10.4 ± 13.1 proteins in average. The number of GO functions per drug is up to six times greater than that of proteins per drug. For a drug, the numbers of associated proteins and GO functions range from 1 to 188 proteins and from 2 to 545 GO functions, respectively. For a disease, the ranges of the numbers of associated proteins and GO functions are 1 to 1,086 proteins and 1 to 944 GO functions, respectively. In average, a disease is associated with 175.6 ± 217.6 GO functions of any aspects, up to three times greater than that of proteins (62.3 ± 162.7 proteins per disease). Since two or more GO functions of multiple GO aspects can be linked to drugs and diseases, a drug or a disease usually accumulates its associated GO functions with the larger number than that of proteins. With the more extensive and multiple views of GO functions, potential associations of drugs and diseases could be more efficiently discovered when compared with proteins.

The numbers of all drug-disease, drug-drug, and disease-disease pairs are shown in Table 3.4. Based on 904 drugs and 524 diseases, there are 473,696 drug-disease pairs in total. Only 6,144 out of them (1.3%) are known or positive drug-disease associations whereas 467,552 out of them (98.7%) are unknown or unlabeled drug-disease associations. This extremely low proportion of the positive drug-disease pairs could suggest that there is still a room for discovering potential drug-disease associations. By pairing two distinct drugs, 408,156 drug-drug pairs can be generated. Among these pairs, 47,094 drug-drug pairs (11.5%) share at least one common disease and are labeled as positive whereas 361,062 drug-drug pairs (88.5%) are unlabeled. By combining two different diseases, 137,026 possible disease-disease pairs can be created. In these pairs, there are 17,129 disease-disease pairs (12.5%) which share at least one common drug and are labeled as positives. The remaining disease-disease pairs or 119,897 pairs (87.5%) are clustered together in the unlabeled group.

Table 3.4 The numbers of drug-disease, drug-drug, and disease-disease pairs in each class

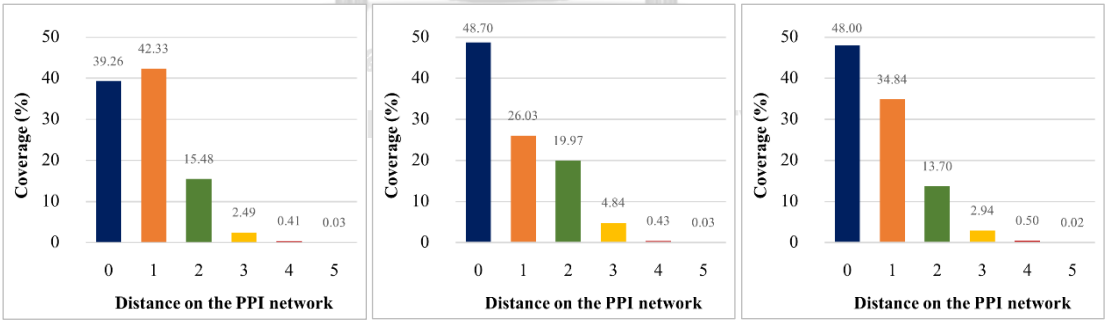
Type of pairs	**Number of pairs in each class (%)**		**Total number of pairs**
	Positive	**Unlabeled**	
Drug-disease pairs	6,144 (1.3%)	467,552 (98.7%)	473,696
Drug-drug pairs	47,094 (11.5%)	361,062 (88.5%)	408,156
Disease-disease pairs	17,129 (12.5%)	119,897 (87.5%)	137,026

3.4.2 Investigation of sharing proteins and GO functions among known associations

To preliminarily provide insights into relationships of drugs and diseases, the investigation of sharing proteins and GO functions among known (positive) drug-disease, drug-drug, and disease-disease associations are conducted. In this experiment, only the positive pairs are mainly observed because they are expected to find some common proteins or GO functions which reflect the existence of their relationships.

Initially, the relationships of drug-associated and disease-associated proteins on a protein-protein (PPI) interaction network are observed for the positive drug-disease, drug-drug, and disease-disease pairs. For each pair, the associated proteins (e.g. drug target proteins or disease-associated proteins) are mapped onto the human PPI network downloaded from the

STRING database (version 11.0) [95]. Drugs and diseases whose no associated proteins can be mapped onto the PPI network are excluded from the experiment. The pairs corresponding to those drugs and diseases are also removed. For each positive pair, to investigate how the associated proteins interact with one another in the PPI network, distances of shortest paths connecting between drug-associated and disease-associated proteins are measured. Because a drug and a disease can interact with one or more proteins, there could be several shortest path distances measured for one positive pair. Herein, a value used to represent a distance for each positive pair is the minimum distance of the shortest paths connecting between two associated proteins of that pair. For example, for a positive drug-disease pair, the distance on the PPI network of each shortest path connecting between a protein associated with the drug and a protein associated with the disease is measured. After getting all shortest path distances, the minimum value of them is used to describe the distance on the PPI network between that drug and that disease. Since two proteins with a small distance on a PPI network are usually expected for their functional relationships, this small distance between drug-associated and disease-associated proteins could point to the relationships between drugs and diseases [96, 97]. A zero distance of a positive pair means that this pair has at least one shared protein. The distances on the PPI network of all positive drug-disease, drug-drug, and disease-disease pairs are summarized in Figure 3.6.



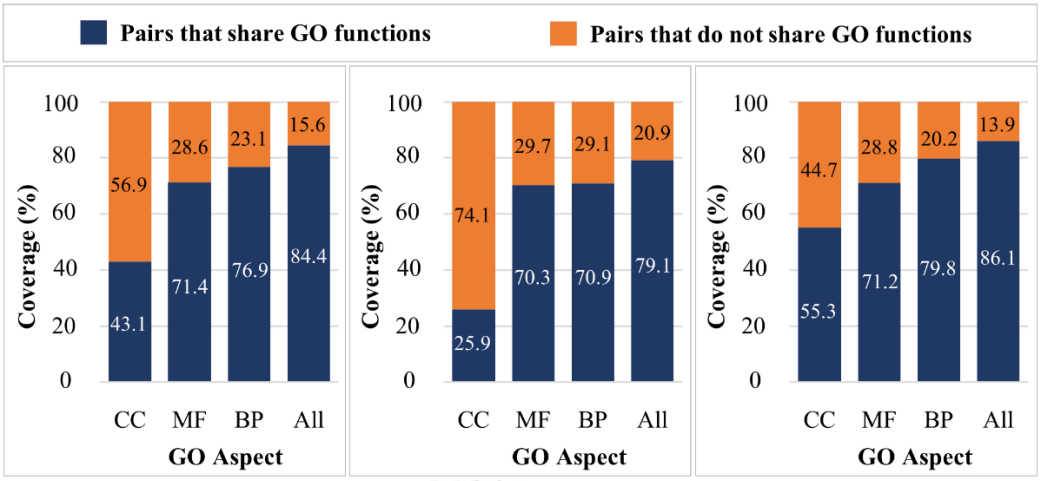
(a) Drug-disease pairs (b) Drug-drug pairs (c) Disease-disease pairs

Figure 3.6 Coverages of positive pairs according to their distances on the PPI network

Out of 6,144 positive drug-disease pairs, only 2,412 pairs (39.26%) have some common proteins between their drug-associated and disease-associated proteins (Figure 3.6(a)). Mostly in 3,732 pairs (60.74%), a drug and a disease connect to each other with the distances up to five on the PPI network. Especially, the direct connections between drug-associated and disease-

associated proteins cover 2,601 positive drug-disease pairs (42.33%). Among 47,094 positive drug-drug pairs, the pairs that share at least one drug-associated protein has the coverage of 22,936 pairs (48.70%), as shown in Figure 3.6(b). In more than a half of the positive drug-drug pairs or 24,158 pairs (51.30%), two drugs link to each other via the paths of up to five steps in the PPI network. To link two associated diseases on the PPI network, only 8,222 positive disease-disease pairs (48%) have some overlapping proteins (Figure 3.6(c)). Conversely, the disease-disease pairs with the distances up to five steps on the PPI network cover 8,907 (52%) of all positive disease-disease pairs.

From the results of the PPI distances of the positive pairs (Figure 3.6), it can be concluded that most drug-disease, drug-drug, and disease-disease associations have more complex relationships than directly interacting with the same proteins. In the PPI network, the connections between drugs and diseases, between two associated drugs, and between two associated diseases mostly occur with the distances of two or more steps. Thus, the association classification solely based on overlapping of drug target proteins and disease-associated proteins can result in low sensitivity of predictions. Rutherford et al. [58] showed that only direct interactions between drug-associated and disease-associated proteins cannot completely discover known drug-disease associations. Indirect interactions, PPI paths of the longer distances, could be considered to increase the sensitivity of predictions, but the question of how long the distance should be is still currently discussed [58, 97]. To resolve this issue, an alternative method is taking advantages of more extensive information (i.e. GO functions) which could be capable of discovering both direct and indirect relationships of drugs and diseases. The coverages of the positive drug-disease, drug-drug, and disease-disease pairs which share at least one common GO function of each GO aspect and all GO aspects are preliminarily investigated and shown in Figure 3.7.



(a) Drug-disease pairs (b) Drug-drug pairs (c) Disease-disease pairs

Figure 3.7 Coverages of positive pairs that share and do not share their GO functions

According to Figure 3.7(a), there are 43.1%, 71.4%, and 76.9% of known drug-disease associations that share at least one common GO function of Cellular Component (CC), Molecular Function (MF), and Biological Process (BP), respectively. Especially when any GO aspect is included, 84.4% of known drug-disease associations share at least one common GO function. In Figure 3.7(b), 25.9%, 70.3%, and 70.9% of known drug-drug associations share at least one common GO function of CC, MF, and BP, respectively. When any GO aspect is considered, up to 80% of the positive drug-drug pairs have at least one overlapping GO function. From Figure 3.7(c), more than a half of known disease-disease associations are found to share their GO functions of CC (55.3%), MF (71.2%), and BP (79.8%) aspects. Greater than these coverages, 86.1% of the positive disease-disease pairs are found to share their GO functions of any GO aspect.

In all three cases, known associations tend to share their BP GO functions greater than GO functions of other aspects. This suggests that among three GO aspects, BP GO functions can be used to achieve higher sensitivity in the predictions of drug-disease, drug-drug, and disease-disease associations. Many previous studies focused on either MF or BP GO functions to uncover the relationships between drugs and diseases [60, 61], between diseases [98, 99], or between drugs [62]. However, different aspects of GO functions contribute functional annotations for drugs and diseases from different points of views. Utilizing all GO aspects could provide the greatest advantages for the predictions of all three associations. This is supported by the results of

the largest coverages of the known associations in all three cases when all GO aspects are used. Hence, all GO aspects are utilized in this study, and sharing GO functions of any GO aspect between drugs and diseases, between drugs, or between diseases is a matter of uncovering relationships of drugs and diseases.

To compare between proteins and GO functions, the numbers of positive pairs that share at least one common protein and GO function are investigated and shown in Table 3.5. It is noticeable that for all pair types, the positive pairs tend to share their GO functions rather than proteins. Lower than 50% of known drug-disease, drug-drug, and disease-disease associations share their proteins. This may be because the relationships of drugs and diseases are complex beyond detecting by overlapping proteins between drugs and diseases, between drugs, or between diseases. By detecting overlapping GO functions, the percent coverages of known drug-disease, drug-drug, and disease-disease associations increase to 84.4%, 79.1%, and 86.1%, respectively. This suggests that GO functions can improve the sensitivity of the predictions of drug-disease, drug-drug, and disease-disease associations, when compared to proteins.

Table 3.5 Comparison of the numbers of pairs that share proteins and GO functions in the positive class

Type of pairs	Number of pairs		Total number
	Sharing protein (s)	Sharing GO function (s)	
Positive drug-disease pairs	2,412 (39.3%)	5,188 (84.4%)	6,144
Positive drug-drug pairs	22,936 (48.7%)	37,259 (79.1%)	47,094
Positive disease-disease pairs	8,222 (48.0%)	14,746 (86.1%)	17,129

The number of pairs that share GO functions in each class is compared and shown in Table 3.6. To verify the greater proportion of GO-sharing pairs in the positive class, Fisher's exact tests were performed. In case of drug-disease pairs, more than 84% of the positive pairs share GO functions, which is significantly greater than the proportion of unknown drug-disease associations that share GO functions (p-value = 1.5×10^{-566}). Moreover, up to 80% of known drug-drug associations have overlapping GO functions whereas about 67% of unknown drug-drug

associations share GO functions. For drug-drug pairs, it can be concluded that the drug-drug pairs that share GO functions are found in the positive class rather than in the unlabeled class (*p*-value = 3.0×10^{-656}). In the positive class, more than 86% of the disease-disease pairs have some overlapping GO functions. When compared to those found in the unlabeled class, it is found that the disease-disease pairs that share GO functions are more frequently found in the positive class than in the unlabeled class (*p*-value = $7.9 \times 10^{-1,234}$). Based on these results, it can be summarized that sharing GO functions among drugs and diseases does not occur by chance and is more commonly found in the positive class than in the unlabeled class. This suggests that it would be promising to use GO functions for identifying the relationships among drugs and diseases.

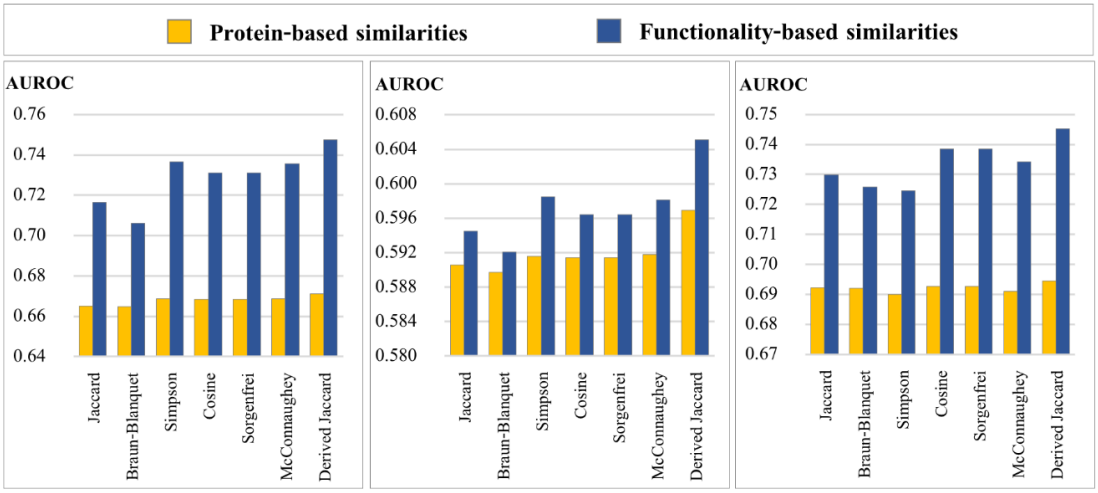
Table 3.6 Comparison of the number of pairs that share GO functions in each class

Type of pairs	**Number of pairs that share GO functions in each class**		***P*-value**
	Positive	**Unlabeled**	
Drug-disease pairs	5,188 (84.4%)	250,564 (53.6%)	1.5×10^{-566}
Drug-drug pairs	37,259 (79.1%)	242,021 (67.0%)	3.0×10^{-656}
Disease-disease pairs	14,746 (86.1%)	69,581 (58.0%)	$7.9 \times 10^{-1,234}$

3.4.3 Selection of the most suitable similarity index

In the previous section, drug-disease, drug-drug, and disease-disease pairs, especially that are positives or known associations, are categorized into a group of pairs that share and do not share their proteins or GO functions. In this section, protein-based or functionality-based similarities between drugs and diseases, between drugs, or between diseases are measured as similarity scores. Since a variety of similarity indices can be used for this task, it is necessary to identify the most suitable one for further using. In this study, seven well-known similarity indices are compared (i.e. the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, derived Jaccard similarity index). By using each similarity index, both protein-based and functionality-based similarity scores are computed for all drug-disease, drug-drug, and disease-disease pairs. Then, these similarity scores were used to classify drug-disease, drug-drug, and

disease-disease associations. From the classification, the AUROC values of all similarity indices for both protein-based and functionality-based similarities are shown in Figure 3.8.



(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 3.8 Areas under the ROC curves (AUROC) of all similarity indices for protein-based and functionality-based similarities

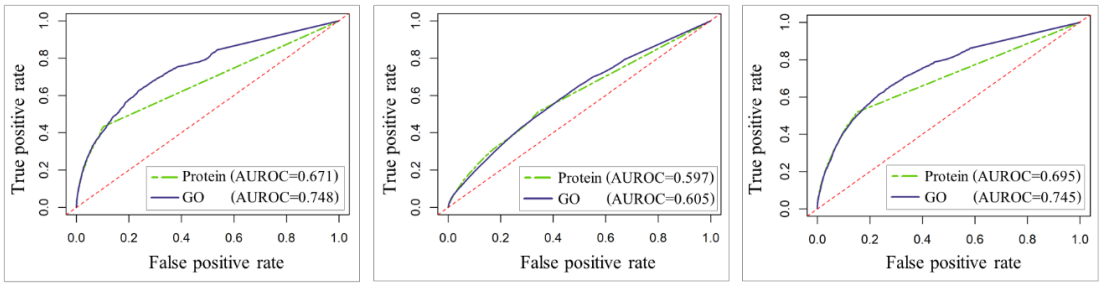
It is noticeable that using functionality-based similarities produces higher AUROC values than those of protein-based similarities in all three cases whatever similarity index are used. When different similarity indices are applied for computing protein-based similarity scores, little improvement in AUROC values is obtained, especially in the classification of drug-disease (Figure 3.8(a)) and disease-disease associations (Figure 3.8(c)). Conversely, significant improvement in AUROC values of functionality-based similarities can be seen, especially in the case of drug-drug associations (Figure 3.8(b)). For protein-based similarities, the similarity index that can produce the highest AUROC values is the derived Jaccard similarity index. Its AUROC values are 0.671, 0.597, and 0.695 in case of drug-disease, drug-drug, and disease-disease associations, respectively. In terms of functionality-based similarities, the derived Jaccard similarity index also performs the best with the AUROC values of 0.748, 0.605, and 0.745 for drug-disease, drug-drug, and disease-disease associations, respectively. Interestingly, the derived Jaccard similarity index produces the highest performance whereas the ordinary Jaccard similarity index performs worse than several similarity indices. It has been revealed that the log-

transformation introduced into the Jaccard similarity index can lead the newly derived similarity index which is uncorrelated with the ordinary one [100]. Moreover, the result showing the best performance of the derived Jaccard similarity index has also found in the study of Wijaya et al. [101], who specified the most appropriate similarity index for classifying matching efficacies of herbal medicine pairs. Therefore, the derived Jaccard similarity index is further used for computing both protein-based and functionality-based similarity scores.

3.4.4 Comparison of protein-based and functionality-based similarities

To assess the advantages of using functional information in uncovering relationships between drugs and diseases, the performance of protein-based similarities serves as the baseline performance and is compared with that of functionality-based similarities. In this experiment, only the derived Jaccard similarity index is applied to compute drug-disease, drug-drug, and disease-disease similarity scores.

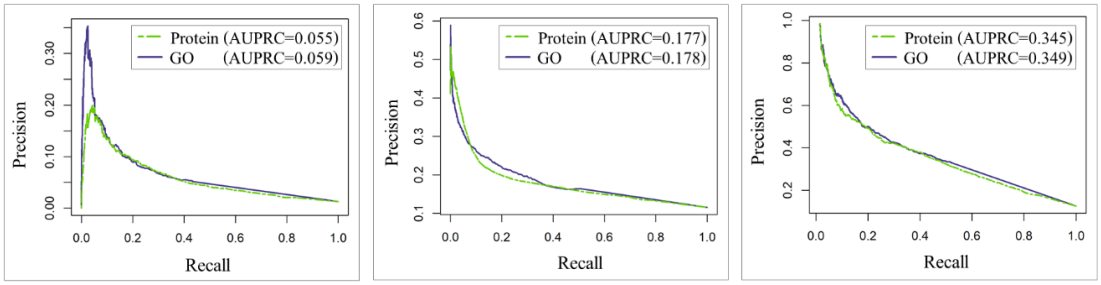
The ROC curves of protein-based and functionality-based similarities are plotted as shown in Figure 3.9. In all three cases, it is noticeable that both protein-based and functionality-based similarities can improve the classification of the completely random model (red-dashed lines). According to Figure 3.9(a), it was found that functionality-based similarity scores of drug-disease pairs can significantly improve the classification of drug-disease associations with an increased AUROC value of 0.748, whereas the AUROC value of protein-based similarities is 0.671. Likewise, the classification of disease-disease associations using functionality-based similarity scores is significantly improved with an AUROC value of 0.745 when compared to that of protein-based similarity scores (AUROC = 0.695), as shown in Figure 3.9(b). In the classification of drug-drug associations (Figure 3.9(c)), using functionality-based similarity scores results in an AUROC value of 0.605 which is slightly greater than that of protein-based similarity scores (AUROC = 0.597).



(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 3.9 ROC curves of protein-based and functionality-based similarities

In addition to the ROC curves, the PR curves of both protein-based and functionality-based similarities are also investigated as shown in Figure 3.10. In the classification of drug-disease associations (Figure 3.10(a)), using functionality-based similarities results in an AUPRC value of 0.059, which is lightly greater than that uses proteins (AUPRC = 0.055). Similarly, the AUPRC value of functionality-based similarities is increased to 0.349 in the classification of disease-disease associations (Figure 3.10(c)), whereas the AUPRC value of proteins is 0.345. Despite insignificant improvement in the classification of drug-drug associations, the AUPRC value of functionality-based similarities (AUPRC = 0.178) is greater than that of protein-based similarities (AUPRC = 0.177), as shown in Figure 3.10(b).

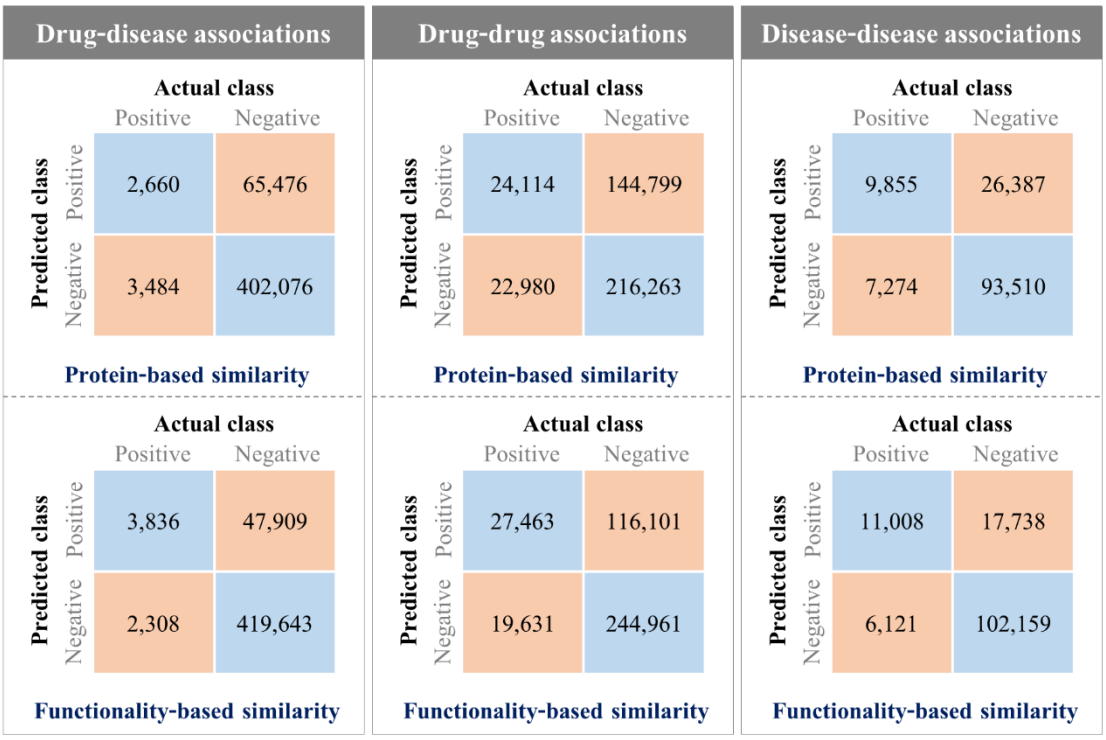


(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 3.10 Precision-recall curves of protein-based and functionality-based similarities

Based on the Youden's index, the optimal threshold scores can be estimated from the ROC curves. When using protein-based similarities, the threshold scores are 0.1, 0.138, 0.099 for classifying drug-disease, drug-drug, and disease-disease associations, respectively. When using functionality-based similarities, the threshold scores are 0.264, 0.369, and 0.297 for predicting

drug-disease, drug-drug, and disease-disease associations, respectively. Based on these threshold scores, the confusion matrices resulting from the predictions of drug-disease, drug-drug, and disease-disease associations can be generated as shown in Figure 3.11.

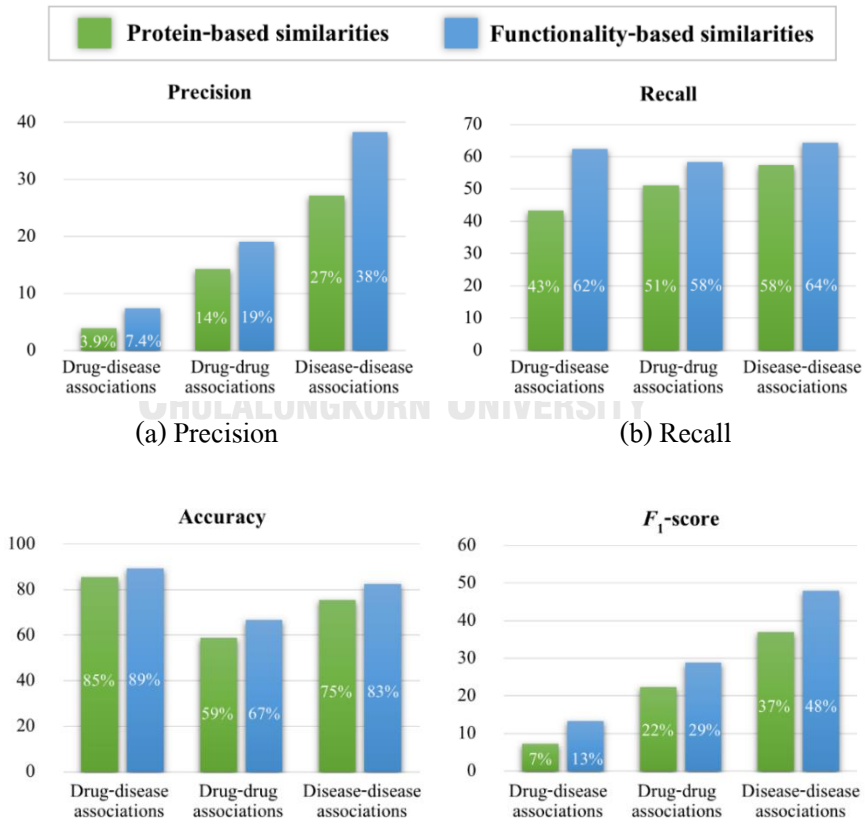


(a) Drug-disease associations (b) Drug-drug associations (c) Disease-disease associations

Figure 3.11 Confusion matrices of protein-based and functionality-based similarities

In each confusion matrix, the total number of each column is a constant, equal to the total number of positive or negative samples. For example, in both confusion matrices of Figure 3.11(a), the summations of the numbers in each column are the total number of positive and negative drug-disease associations (i.e. 6,144 and 467,552, respectively). When comparing between a confusion matrix of a protein-based similarity (top) and that of a functionality-based similarity (bottom), the number of true positives (*TP*) will be increased or decreased by only changing between true positives and false negatives (*FN*). Likewise, the number of true negatives will be increased or decreased by only transferring between true negatives (*TN*) and false positives (*FP*). Therefore, only the numbers of accurate predictions, i.e. *TP* and *TN*, will be discussed here.

According to Figure 3.11(a), the numbers of accurate predictions of drug-disease associations when using protein-based similarities are 2,660 for *TP* and 402,076 for *TN*. It was found that the numbers of accurate predictions increase when using drug-disease similarity scores based on GO functions (*TP* = 3,836 and *TN* = 419,643) for predicting drug-disease associations. From Figure 3.11(b), the number of true positives increases from 24,114 to 27,463 when using functionality-based similarity scores for predicting drug-drug associations. The number of true negatives is also improved from 216,263 to 244,961 by using functionality-based similarities. In the classification of disease-disease associations (Figure 3.11(c)), the greater numbers of accurate predictions are gained when using functionality-based similarities (*TP* = 11,008 and *TN* = 102,159), when compared to those of protein-based similarities (*TP* = 9,855 and *TN* = 93,510). Based on the confusion matrices (Figure 3.11), the values of precision, recall, accuracy, and F_1-score are computed and shown in Figure 3.12.



(a) Precision (b) Recall

(c) Accuracy (d) F_1-score

Figure 3.12 Precision, recall, accuracy, and F_1-score of protein-based and functionality-based similarities

Figure 3.12 clearly shows that using functionality-based similarity scores in the classification of drug-disease, drug-drug, and disease-disease associations can improve values in all evaluation metrics, when compared to those obtained by using protein-based similarity scores. By using GO functions, the percentages of recovered positive samples (recall values) are increased from 43% to 62% in the classification of drug-disease associations, from 51% to 58% in the classification of drug-drug associations, and from 58% to 64% in the classification of disease-disease associations, as shown in Figure 3.12(b). Similarly, the accuracy values are also improved when functionality-based similarity scores are utilized to classify drug-disease associations (89%), drug-drug associations (67%), and disease-disease associations (83%), as illustrated in Figure 3.12(c). Since there are few positive samples relative to samples in another class, the precision values obtained are quite low (Figure 3.12(a)), especially in the classification of drug-disease associations. As mentioned before, less than 2% of drug-disease pairs are positive. Therefore, drug-disease pairs that are predicted to be positive are mostly considered to be *FP*, resulting in a flatten precision value. Nevertheless, the greater numbers of *TP* detected by using functionality-based similarities result in the higher precision values when compared to those obtained by using protein-based similarities. Similarly, despite the flatten F_1 scores, similarities based on GO functions can improve the predictions of drug-disease, drug-drug, and disease-disease associations as shown in Figure 3.12(d).

From these results, it can be concluded that drug-associated and disease-associated GO functions are very useful for identifying relationships between drugs and diseases, between drugs, and between diseases. By using GO functions, the greater numbers of positive drug-disease, drug-drug, and disease-disease associations can be detected resulting in improved performance, when compared to that obtained from using protein information. This may be because broader information provided by GO functions supports drug-disease, drug-drug, and disease-disease similarity detection. Davis et al. [102] investigated BP GO functions and genes overlapping between old and new diseases of three repositioned drugs, which are thalidomide, raloxifene, and sildenafil. They found that relevant diseases in all three cases significantly share their GO functions rather than genes. Moreover, Rutherford et al. [58] observed drug-associated and disease-associated proteins on a PPI network. They revealed that most of known drug-disease associations cannot be easily detected by direct interactions, but they tend to have longer PPI

connections between drug-associated and disease-associated proteins in the PPI network. That is why many positive samples cannot be detected by using protein-based similarities.

3.4.5 Case studies

In this section, a novel drug-disease, drug-drug, and disease-disease association predicted by using functionality-based similarity scores are selected for discussions. Each of them was also searched for supporting evidence from literature and a database of clinical studies (ClinicalTrials.gov).

1) Tolcapone and attention deficit-hyperactivity disorder (ADHD)

This is a discovered drug-disease association with a functionality-based similarity score of 0.484. The drug tolcapone (DB00323) is currently approved for the treatment of Parkinson's disease. ADHD (OMIM: 143465) is a mental health disease which affects several people's behaviors including overactivity, lack of attention, and impulsiveness [103]. In ClinicalTrials.gov, it was found a clinical study (NCT03904498) which has been recently conducted to investigate the use of tolcapone for patients with both alcohol addiction and ADHD. At present, the molecular mechanisms of tolcapone remain unclear, especially those involving with ADHD. However, nine overlapping GO functions between tolcapone and ADHD were found, such as catechol O-methyltransferase (COMT) activity (GO: 00162606), dopamine catabolic process (GO: 0042420), and short-term memory (GO: 0007614). From a literature search, it was found that the COMT enzyme can diminish the level of dopamine in the prefrontal cortex [104], which controls various behaviors involving with ADHD. In addition, tolcapone can inhibit the COMT enzyme to maintain the level of dopamine [105].

2) Glimepiride and nicorandil

This is a potential drug-drug association with a functionality-based similarity score of 0.488. They have four overlapping GO functions which are ion channel binding (GO: 0044325), potassium ion import across plasma membrane (GO: 1990573), inward rectifying potassium channel (GO: 0008282), and sulfonylurea receptor activity (GO: 0008281). Because these two drugs highly involve with similar GO functions, they may be able to treat some common diseases.

Figure 3.13 shows known indications of glimepiride and nicorandil, where some of them could be newly suggested for one of those two drugs. Glimepiride (DB00222) is approved for T2D or type 2 diabetes mellitus (OMIM: 125853) and some other variants of diabetes, including type 1 (OMIM: 125850), type 2 (OMIM: 125851), and type 3 (OMIM: 600496) maturity-onset diabetes in young. Nicorandil (DB09220) is approved for cardiac arrhythmia (OMIM: 115000) and myocardial infraction (OMIM: 608446).

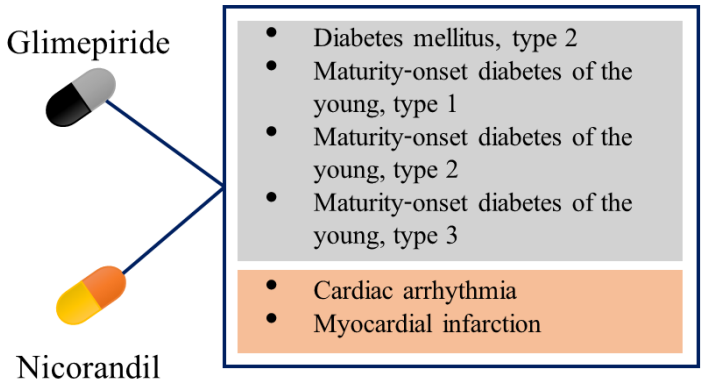


Figure 3.13 An inferred association of glimepiride and nicorandil

By a literature search, it was found that the abnormality in controlling the potassium channel activity involves with the progression of T2D [106], and the dysfunction of several ion channel activities can affect Ca^{2+} controlling and β-cell function in T2D [107]. Additionally, there is a clinical study with an ID of NCT03775902 which was conducted to study the use of nicorandil for patients with T2D. According to all supporting information, an association of nicorandil and T2D can be suggested as a potential one, and some overlapping GO functions may be used to guide how nicorandil and T2D are associated with each other.

3) Myotonia congenita and Gitelman syndrome

This is a proposed disease-disease association with a high functionality-based similarity score of 0.667. They share three GO functions which are voltage-gated chloride channel activity (GO: 0005247), chloride transmembrane transport (GO: 1902476), and chloride channel complex (GO: 0034707). Since these two diseases are highly related to the same GO functions, it would be

possible that they can be treated by some common drugs. The approved drugs of both diseases are listed in Figure 3.14.

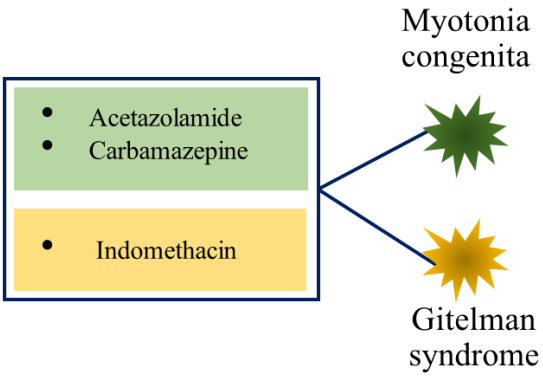


Figure 3.14 An inferred association of Myotonia congenita and Gitelman syndrome

Myotonia congenita (OMIM: 255700) is a rare disease which affects skeletal muscles. The drugs approved for this disease are acetazolamide (DB00819) and carbamazepine (DB00564). Gitelman syndrome (OMIM: 263800) is a rare disease with the abnormality in controlling various ions in the body such as potassium, magnesium, and calcium [108]. One drug that has been approved for this disease is indomethacin (DB00328). Despite no clinical studies found to support a cross relationship between these two diseases, a literature search reveals that carbamazepine can impact the sodium transport in the toad *Pleurodema thaul* [109], and this activity involves with the progression of Gitelman syndrome [110]. According to these results, it could recommend an association between carbamazepine and Gitelman syndrome for further investigation.

3.5 Discussions

Since proteins play crucial roles in drug actions and disease processes, drug-associated and disease-associated proteins are usually utilized to identify the relationships between drugs and diseases, between drugs, and between diseases. An advantage of this approach is that the data of drug-protein and disease-protein associations are uncomplicated and broadly available in many databases, such as DrugBank [85], Therapeutic Target Database (TTD) [111], DisGeNET [86], and OMIM [112]. Moreover, predicting drug-disease, drug-drug, and disease-disease associations

based on protein similarities is manageable and straightforward. However, this approach could not discover some complex relationships with the lack of proteins explicitly shared between drugs and diseases, between drugs, or between diseases.

To improve the predictions of the drug-disease, drug-drug, and disease-disease associations, the proposed approach is based on the functionality-based similarities, which are the similarity measures from drug-associated and disease-associated GO functions. With the broader information of the GO functions, the proposed approach can recover the greater numbers of the positive drug-disease, drug-drug, and disease-disease associations, when compared to those of the protein-based similarities. In addition, the multi-aspects of the GO functions provide more information about drugs and diseases which would be of great advantages for classifying the drug-disease, drug-drug, and disease-disease associations.

However, there are some drawbacks of GO functions that need to be concerned. Firstly, linking GO functions to drugs and diseases is burdensome. Since GO terms are organized in the hierarchical structure, the GO functions of different levels can be mapped to a protein and therefore a drug or a disease. The redundancy of these semantically related GO functions may lead to the inflated similarity scores of the drug-disease, drug-drug, and disease-disease pairs. To reduce the redundancy, only the drug-GO and disease-GO pairs with the most detailed GO functions or leaf nodes were maintained in this study. By this technique, there will be no semantically related GO terms retained for each drug and disease, but those GO functions could be still from the different levels. For example, regulation of translational elongation (GO:0006448) is an ancestor node of regulation of cytoplasmic translational fidelity (GO:0140018), and the former GO function is linked to a drug whereas the latter is mapped to a disease. The similarity score of this drug-disease pair cannot be measured although the drug and disease are functionally related. This leads the functionality-based similarity measures to loss of the similarity information of such GO functions. To recover this kind of the similarity information, a reasonable solution is applying GO semantic similarity measures, such as the Resnik [113] and Wang [114] similarity scores. In addition, the similarity scores should be relative to the levels of the considered GO functions because the deeper GO terms provide the more specific information. For example, a drug-disease pair that possesses term GO:0140018

should have the larger similarity score than another drug-disease pair that shares term GO:0006448.

Secondly, the similarities based on GO functions of the distinct aspects (i.e. BP, MF, and CC) can differently imply to the functional relationships between two proteins. Especially when the similarities are identified solely based on CC GO functions, it could not conclude whether both proteins are functionally related or not. Similarly, the similarities solely based on CC GO functions between drugs and diseases, between drugs, and between diseases could not reliably point to the potential drug-disease, drug-drug, and disease-disease associations, respectively. This can be supported by the results that there are only the small proportions of the positive drug-disease pairs (43%), drug-drug pairs (26%), and disease-disease pairs (55%) that have some common CC GO functions. Therefore, the CC GO functions should be exploited with the BP or MF GO functions to create more reliable functionality-based similarity scores. Additionally, the similarity information provided by the distinct GO aspects should differently contribute to the functionality-based similarity scores.

Finally, some GO terms provide too general information and could not indicate the specific biological activities or pathways in which proteins work, such as protein binding (GO:0005515) and scaffold protein binding (GO:0097110). Such GO functions could inflate the functionality-based similarity measures, resulting in the increased numbers of false positive drug-disease, drug-drug, and disease-disease associations. Therefore, it would be advantages if the trivial GO functions will be removed from the computation of the functionality-based similarity scores.

3.6 Summary

In this study, the feasibility of utilizing GO functions for discovering relationships between drugs and diseases, between drugs, and between diseases is evaluated. Generally, drug-associated and disease-associated proteins are used to identify these relationships via similarity measures based on those proteins. However, the relationships between drugs and diseases can be more complex than interacting with the same proteins, leading to a failure of the protein-based similarity strategy in detecting many of those relationships. This motivates to exploit other

broader information (i.e. GO functions) for predicting drug-disease, drug-drug, and disease-disease associations.

Initially, all drug-disease, drug-drug, and disease-disease pairs were generated. Drug-disease pairs were labeled as positive if they are known drug-disease associations, otherwise they were unlabeled. To connect between drugs and diseases, a pair of two drugs was labeled as positive if they share at least one common disease. If not, it was unlabeled. Similarly, a pair of two diseases was positive if they have at least one common drug. Otherwise, it was unlabeled. All drug-disease, drug-drug, and disease-disease pairs were measured functional similarity levels using drug-GO and disease-GO associations. Seven well-known similarity indices were compared, and the most suitable one (i.e. the derived Jaccard index) was selected to compute the functionality-based similarity scores for all drug-disease, drug-drug, and disease-disease pairs. These scores were directly employed to classify the drug-disease, drug-drug, and disease-disease associations to evaluate how well the functionality-based similarity measures can be used to discover the relationships between drugs and diseases. The performance of the protein-based similarity measures served as the baseline performance and was compared with that of the functionality-based similarity measures.

As a result, the classifications of the drug-disease, drug-drug, and disease-disease associations were improved by using the functionality-based similarity measures with the better values in all evaluation metrics (i.e. precision, recall, accuracy, and F_1), when compared to those of the protein-based similarity measures. Furthermore, the case studies showed that the functionality-based similarity measures can be used to discover new drug-disease associations under investigation or with supporting literature. According to these results, it could suggest that GO information is a promising indicator that can be used to discover the relationships between drugs and diseases, between drugs (probably sharing common diseases), and between diseases (probably sharing common drugs). With these advantages of the GO functions, it would be of great interest to integrate all independent functionality-based similarity information with more sophisticated methods to achieve more reliable and accurate predictions of the drug-disease associations.

CHAPTER IV

META-PATH BASED FUNCTIONAL PROFILES FOR PREDICTING DRUG-DISEASE ASSOCIATIONS

In the previous chapter, GO information is shown as a promising signal in uncovering relationships between drugs and diseases, between drugs, and between diseases. In this chapter, a PU learning method with meta-path based functional profiles is proposed for predicting drug-disease associations. This method takes advantages of GO functions and meta-paths to create novel network-based features of drug-disease pairs, called meta-path based functional profiles.

4.1 An overview of the study

A diagram that provides an overview of this study is shown in Figure 4.1.

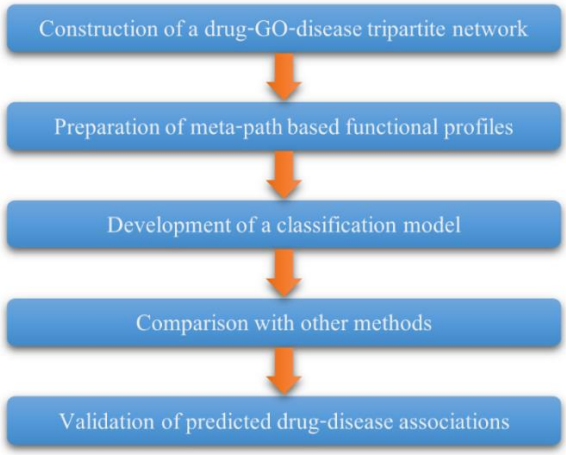


Figure 4.1 A schematic diagram providing an overview of this study

First, a drug-GO-disease tripartite network was constructed by integrating three association data sets, including drug-GO, disease-GO, and drug-disease associations. From the tripartite network, meta-path based functional profiles were generated for all drug-disease pairs by using the proposed algorithms. These meta-path based functional profiles were then prepared to further use in the development of a classification model. To show the superior performance of the proposed method, the researcher compared its performance with those of existing methods.

Finally, the proposed method was employed to predict potential drug-disease associations, and these inferred associations were also searched for supporting evidence in a database of clinical trials and literature to demonstrate the practicality of the proposed method. The source codes and data sets used in this study are freely available at https://github.com/thitipongk/MGPDDA.

4.2 Data sets

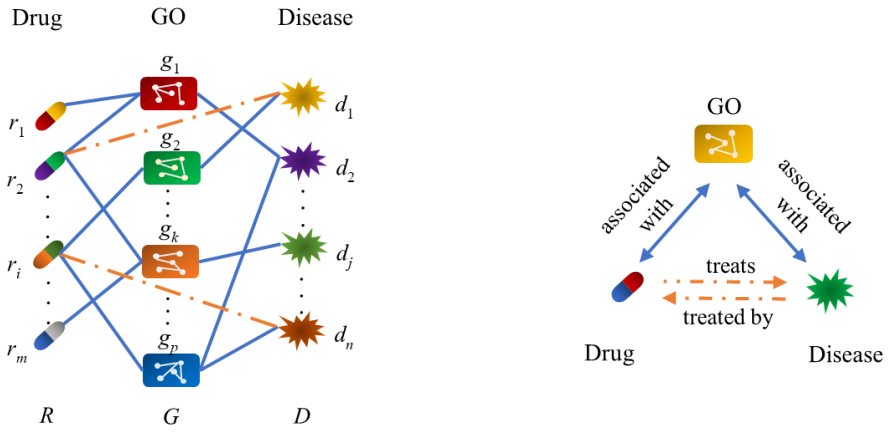
In this method, only three association data sets are used which are drug-GO, disease-GO, and drug-disease associations. These data sets are the same data sets as described in Chapter III (see the section 3.2 for the data sources). In brief, drug-GO associations were generated by using drug-protein associations and GO annotation data of human proteins (see the section 3.3.2 for more details). Similarly, disease-GO associations were constructed by using disease-protein associations and GO annotation data of human proteins (see the section 3.3.2 for more details). A collection of known drug-disease associations was obtained by integrating the gold standard data set of Gottlieb et al. [3] and drug-disease therapeutic relations downloaded from CTD [84].

4.3 Methods

4.3.1 Construction of a drug-GO-disease tripartite network

A drug-GO-disease tripartite network used in this study contains three node types (drug, GO, and disease nodes) as shown in Figure 4.2(a). The set of all drug, GO, and disease nodes in the network are $R = \{r_1, r_2, \ldots, r_m\}$, $G = \{g_1, g_2, \ldots, g_p\}$, and $D = \{d_1, d_2, \ldots, d_n\}$, respectively. m, p, and n represent the total number of all drug, GO, and disease nodes, respectively. In the tripartite network, links are connected only between different node types, and there are three different link types as shown in Figure 4.2(b). The links “associated with” connect between drug and GO nodes or between disease and GO nodes. The links “treats” and “treated by” bridge between drug and disease nodes.

To construct a drug-GO-disease tripartite network, the data of drug-GO, disease-GO, and drug-disease associations are used. Based on these association data, this network is simply represented as an unweighted and undirected network as demonstrated in Figure 4.2(a). However, each link in this network can represent bidirectional relationships with semantic annotations between any two nodes of different types, as shown in Figure 4.2(b).



(a) A drug-GO-disease tripartite network (b) Network schema

Figure 4.2 Illustration of a drug-GO-disease tripartite network and its network schema

4.3.2 Generation of meta-path based functional profiles for drug-disease pairs

A meta-path is a path structure, written by a sequence of node and link types, for extracting paths in a heterogeneous network. Because a heterogeneous network represents a collection of relationships between diverse nodes, a meta-path can be considered as a composite relationship between a starting node type and a terminating node type. Under a meta-path, paths extracted from a network for a pair of nodes can be considered as semantic information describing the relationship between those nodes from a point of view. Therefore, multiple meta-paths are often used to aggregate distinct semantic information from different points of view.

After a drug-GO-disease tripartite network is constructed, novel meta-path based features are generated for each drug-disease pairs. Unlike other existing meta-path based features, the proposed features can retain information of intermediate nodes (i.e. GO functions) by differentiating paths under a meta-path according to their incorporating intermediate nodes and creating as profiles of intermediaries. In this work, functional profiles or GO profiles are generated for each drug-disease pair using meta-paths, that is why these new features is called meta-path based functional profiles.

To generate meta-path based functional profiles for each drug-disease pair, three meta-paths, denoted as M_1, M_2, and M_3, are employed in this study. All meta-paths always start with drug nodes and terminate with disease nodes so that semantic information of drug-disease pairs will be extracted from the network. In a path, a starting drug node is called a target drug, and a

terminating disease node is called a target disease. Only the meta-paths incorporating the GO node type are used so that meta-path based functional profiles can be generated. The lengths of the meta-paths are limited up to three, because too long meta-paths often contribute useless information and are hardly detected in a network [115].

For each drug-disease pair, a meta-path based functional profile is a vector with the length equal to the total number of all GO functions. Each element in that vector is a count of the extracted paths, starting from a target drug node and terminating at a target disease node, under a meta-path considered. Therefore, a meta-path based functional profile can be considered as a functional profile containing association degrees of each GO function to a drug-disease pair based on a count of paths under a meta-path. Each meta-path accumulates GO functions that are involved with a drug-disease pair from different points of view described as follows:

Meta-path 1 (M_1): Drug - GO - Disease

From both drug and disease perspectives, this meta-path accumulates GO functions where both drugs and diseases participate in. For a drug-disease pair, GO functions that overlap between those of a target drug node and a target disease node are accumulated and considered as GO functions associated with both that target drug and disease.

Meta-path 2 (M_2): Drug - GO - Drug - Disease

For each drug-disease pair, this meta-path investigates GO functions where two drugs participate in. From a drug perspective, GO functions that are shared between those of a target drug and each drug known to be associated with a target disease are accumulated and considered as GO functions associated with both that target drug and disease.

Meta-path 3 (M_3): Drug - Disease - GO - Disease

This meta-path observes GO functions which two diseases are involved with for a drug-disease pair. From a disease perspective, GO functions overlapping between those of a target disease and each disease known to be associated with a target drug are collected and considered GO functions that both target drug and disease are involved with.

The pseudocodes of generating M_1-based , M_2-based, and M_3-based functional profiles for all drug-disease pairs (X_{M_1}, X_{M_2}, and X_{M_3}) are shown in Algorithms 4.1 - 4.3. A drug-GO, disease-GO, and drug-disease association matrices are denoted as A_{rg}, A_{dg}, and A_{rd}, respectively. Θ is an element-wise product or the Hadamard product.

Algorithm 1: Generating M_1-based functional profiles (X_{M_1})

Input: $A_{rg} \in \mathbb{R}^{m \times p}$ and $A_{dg} \in \mathbb{R}^{n \times p}$

Output: $X_{M_1} \in \mathbb{R}^{mn \times p}$

1. **For** each drug node $r_i \in R$
2. **For** each disease node $d_j \in D$
3. $X_{M_1}((r_i, d_j), :) = A_{rg}(r_i, :) \Theta A_{dg}(d_j, :)$
4. **EndFor**
5. **EndFor**

Algorithm 2: Generating M_2-based functional profiles (X_{M_2})

Input: $A_{rd} \in \mathbb{R}^{m \times n}$ and $A_{rg} \in \mathbb{R}^{m \times p}$

Output: $X_{M_2} \in \mathbb{R}^{mn \times p}$

1. **For** each drug node $r_i \in R$
2. **For** each disease node $d_j \in D$
3. Find a set of drug nodes associated with d_j from $A_{rd}(:, d_j)$, denoted as $R_{d_j} \subseteq R$.
4. Initialize $X_{M_2}((r_i, d_j), :)$ as a vector whose all elements are zero.
5. **For** each drug node $u \in R_{d_j}$
6. $X_{M_2}((r_i, d_j), :) = X_{M_2}((r_i, d_j), :) + A_{rg}(r_i, :) \Theta A_{rg}(u, :)$
7. **EndFor**
8. **EndFor**
9. **EndFor**

Algorithm 3: Generating M_3-based functional profiles (X_{M_3})

Input: $A_{rd} \in \mathbb{R}^{m \times n}$ and $A_{dg} \in \mathbb{R}^{n \times p}$

Output: $X_{M_3} \in \mathbb{R}^{mn \times p}$

1. **For** each drug node $r_i \in R$
2. **For** each disease node $d_j \in D$
3. Find a set of disease nodes associated with r_i from $A_{rd}(r_i, :)$, denoted as $D_{r_i} \subseteq D$.
4. Initialize $X_{M_3}((r_i, d_j), :)$ as a vector whose all elements are zero.
5. **For** each drug node $v \in D_{r_i}$
6. $X_{M_3}((r_i, d_j), :) = X_{M_3}((r_i, d_j), :) + A_{dg}(d_j, :) \Theta A_{dg}(v, :)$
7. **EndFor**
8. **EndFor**
9. **EndFor**

To clearly demonstrate how to create meta-path based functional profiles, an example with a small drug-GO-disease tripartite network is given in Figure 4.3. This network consists of three drug nodes (i.e. r_1, r_2, and r_3) and two disease nodes (i.e. d_1 and d_2) which are associated with three GO nodes (i.e. g_1, g_2, and g_3). The drug-GO, disease-GO, and drug-disease association matrices that represent this network can be constructed and denoted as $A_{rg} \in \mathbb{R}^{3 \times 3}$, $A_{dg} \in \mathbb{R}^{2 \times 3}$, and $A_{rd} \in \mathbb{R}^{3 \times 2}$, respectively. Each row in A_{rg} can be considered as a functional profile of each drug, a binary vector indicating which GO functions are associated with that drug. Similarly, each row in A_{dg} is a functional profile vector of each disease. A result of an element-wise product performed between two different functional profiles is a profile vector indicating common GO functions between those two profiles.

To create the M_1-based functional profile of drug-disease pair r_1-d_1, an element-wise product is performed on the functional profile vector of drug r_1 and that of disease d_1. Consequently, a vector with the length of three (the total number of GO functions) is obtained, where a value of one appears at the locations of common GO functions shared between those of drug r_1 and disease d_1. In this example, both r_1 and d_1 are associated with all GO functions. Thus, all elements in the M_1-based functional profile of pair r_1-d_1 are ones. Furthermore, each value in

this functional profile represents a count of paths incorporating each GO node under meta-path M_1. In this M_1-based functional profile, the first value of one is from path r_1-g_1-d_1, the second value of one is from path r_1-g_2-d_1, and the third value of one is from path r_1-g_3-d_1, as enumerated in Figure 4.3.

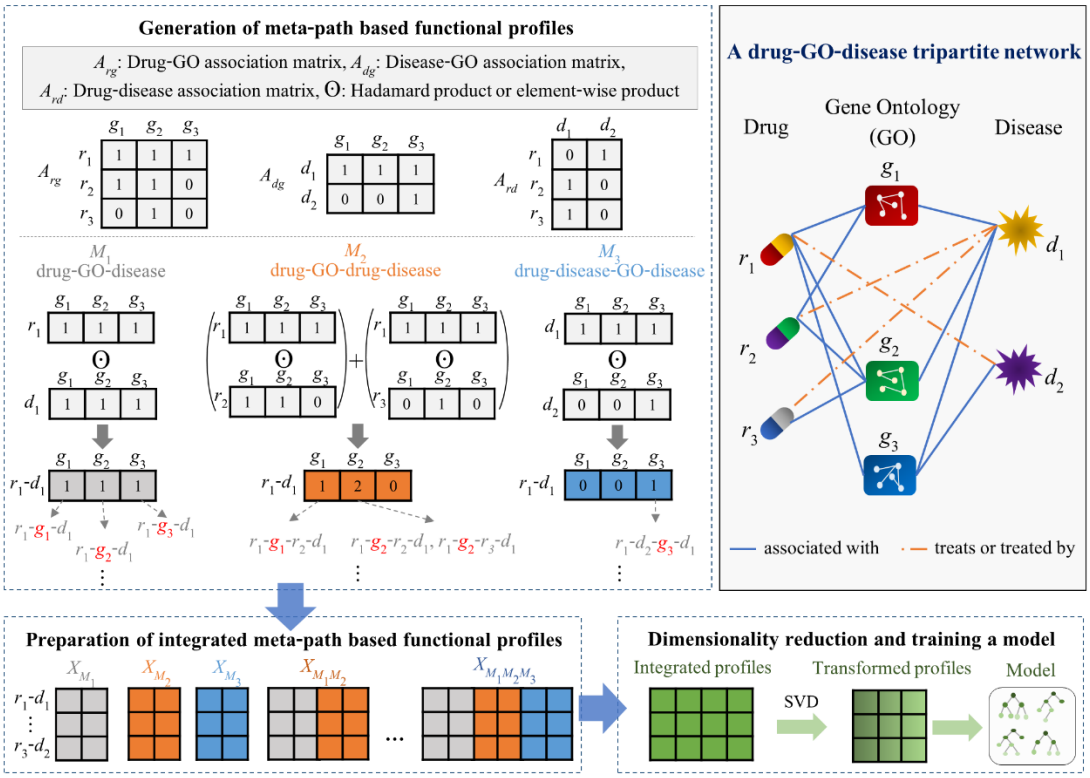


Figure 4.3 A demonstration of generating meta-path based functional profiles and further processes

To investigate GO functions through a view of drugs, drugs associated with target disease d_1 are identified from the drug-disease association matrix A_{rd}. Consequently, there are two drugs (i.e. r_2 and r_3) associated with d_1. Then, an element-wise product is performed between the functional profile vector of target drug r_1 and each drug associated with d_1 separately. To obtain the M_2-based functional profile of r_1-d_1, all profile vectors resulted from the element-wise products are summed up together. Each value in this M_2-based functional profile represents a count of paths that incorporates each GO node under M_2. The value of one at the position of the

GO function g_1 is counted from path r_1-g_1-r_2-d_1. The value of two at the second column of the M_2-based functional profile is a result of paths r_1-g_2-r_2-d_1 and r_1-g_2-r_3-d_1.

To observe GO functions through a disease perspective, all diseases connected to target drug r_1 are primarily identified by searching for values of ones in the second column of A_{rd}. As a result, only disease d_2 is found to be associated with target drug r_1. Then, an element-wise product is performed on the functional profile vector of d_1 and d_2 to obtain the M_3-based functional profile vector of r_1-d_1. In this functional profile, the value of one at the third position is a count of path r_1-d_2-g_3-d_1, which incorporates GO function g_3 in the path.

It is noteworthy that a row summation of X_{M_i}, where i = 1, 2, 3, is equivalent to a total count of paths under a meta-path with regardless of differences of intermediate nodes. This measure is known as a conventional path count and commonly used in many meta-path based studies. The proposed algorithms can differentiate this ordinary path count according to GO functions incorporated in the paths to construct meta-path based functional profiles. These functional profiles could gain more network-based information which are useful for the classification of drug-disease associations, when compared to an ordinary one.

These processes are repeatedly conducted to create meta-path based functional profiles for all drug-disease pairs. Consequently, the matrices of M_1-based, M_2-based, and M_3-based functional profiles or X_{M_1}, X_{M_2}, and X_{M_3} are obtained, where $X_{M_1} \in \mathbb{R}^{6\times3}$, $X_{M_2} \in \mathbb{R}^{6\times3}$, and $X_{M_3} \in \mathbb{R}^{6\times3}$. Because multiple combinations of the functional profile matrices can be produced, an optimal combination should be primarily specified before preparing features of drug-disease pairs. To achieve this task, both independent and combined profile matrices are investigated. To construct a combined functional profile matrix, two different profile matrices are concatenated together and denoted by $X_{M_iM_j}$, where i, j = 1, 2, 3. In case of $X_{M_1M_2M_3}$, it is created by concatenating all three functional profile matrices. In total, seven distinct functional profile matrices are obtained (i.e. X_{M_1}, X_{M_2}, X_{M_3}, $X_{M_1M_2}$, $X_{M_1M_3}$, $X_{M_2M_3}$, and $X_{M_1M_2M_3}$). These functional profiles are high dimensional feature matrices, especially when the total number of GO functions is very high, leading to unsuitability for training and testing a classification model. Therefore, the dimensionality reduction is performed on all seven profile matrices by using the singular value decomposition (SVD). Then, latent features of these different profile matrices are tested to choose the best one for further development of a classification model.

4.3.3 Dimensionality reduction of meta-path based functional profiles

After completing the step of generating meta-path based functional profiles, $X_{M_1} \in \mathbb{R}^{mn \times p}$, $X_{M_2} \in \mathbb{R}^{mn \times p}$, and $X_{M_3} \in \mathbb{R}^{mn \times p}$ are obtained, where m, n, and p are the total number of drugs, diseases, and GO functions. These functional profile matrices are high dimensional features, especially when the total number of GO functions (p) is very high, and all profile matrices are concatenated to produce $X_{M_1M_2M_3}$. In case of $X_{M_1M_2M_3}$, a feature vector of a drug-disease pair has the length as many as $3p$, three times the total number of all GO functions. These high dimensional features could result in the complexity of a model leading to its low generalization. Before using these features in a classification model, dimensionality reduction of meta-path based functional profiles is conducted by using the truncated SVD [116]. This method is commonly used for data dimensional reduction in various applications, such as document clustering [117], medical image classification [118], disease diagnosis models [119], and drug repositioning [11, 47], due to its simplicity and efficiency.

Let $X \in \mathbb{R}^{mn \times q}$ be a functional profile matrix, where $q = p, 2p, 3p$. By SVD, X can be decomposed as shown in (4.1), where U and V are orthonormal matrices, and S is a diagonal matrix containing singular values of matrix X. Singular values in S are sorted in descending order.

$$X_{mn \times q} = U_{mn \times mn} S_{mn \times q} V^T_{q \times q} \tag{4.1}$$

Generally, bottom singular values are near zeros or equal to zeros, which can be truncated for dimensionality reduction. To approximate matrix X as $\hat{X}$, the first r columns of U, the first leading r singular values of S, and the first r rows of V^T are selected, where $r << \min\{mn, q\}$, as shown in (4.2). The r-dimensional latent feature matrix (Y) of functional profile matrix X can be defined in (4.3).

$$X_{mn \times q} \approx \hat{X}_{mn \times q} = \hat{U}_{mn \times r} \hat{S}_{r \times r} \hat{V}^T_{r \times q} \tag{4.2}$$

$$Y_{mn \times r} = \hat{U}_{mn \times r} \hat{S}_{r \times r} \tag{4.3}$$

Note that $\|\cdot\|_F$ is the Frobenius norm which can be defined following to (4.4), where x_j represents the j^{th} column of X, $\|\cdot\|_2$ is the Euclidean norm, and tr(X) is a trace of square matrix X or a summation of all diagonal elements of X. It can be shown in (4.5) that the square of the Frobenius norm of X is the summation of the squares of its singular values.

$$\|X\|_F = \sqrt{\sum_{i=1}^{mn}\sum_{j=1}^{q} x_{ij}^2} = \sqrt{\sum_{j=1}^{q}\left(\|x_j\|_2\right)^2} = \sqrt{\mathrm{tr}(X^T X)} \tag{4.4}$$

$$\left(\|X\|_F\right)^2 = \mathrm{tr}(X^T X) = \mathrm{tr}(VS^T U^T USV^T) = \mathrm{tr}(S^2) = \sum_{i=1}^{\min\{mn,q\}} \sigma_i^2 \tag{4.5}$$

For each functional profile matrix, a latent feature percentage (l) is used to control the number of retained components (r) or the number of latent features obtained from SVD as shown in (4.6), where q is the total number of features in an original functional profile matrix. For a latent feature matrix corresponding to each value of l, the coverage percentage of all singular values (E) can be estimated following to (4.7). The suitable value of l for each functional profile matrix is specified at the minimum value where E reaches at least 95%. This means that the summation of the squares of all singular values should be maintained at least 95% in truncated matrix $\hat{X}$.

$$r = \left\lceil \frac{ql}{100} \right\rceil, \text{ where } \lceil x \rceil = \min\{a \in \mathbb{Z} | a \geq x\} \tag{4.6}$$

$$E = \left(\frac{\left\| \hat{X}_{mn\times q} \right\|_F}{\left\| X_{mn\times q} \right\|_F} \right)^2 \times 100 \tag{4.7}$$

4.3.4 A classification model framework

Normally, only known or positive drug-disease associations can be found whereas no negative drug-disease pair or drug-disease non-association is identified due to lack of its applications [13]. In a set of drug-disease pairs, there are positive and unlabeled data, where a group of unlabeled samples contains both positives and negatives. To deal with this limitation,

many supervised learning based methods consider all unlabeled samples as negatives without awareness of contamination due to undiscovered positives. A classifier trained on positives and this kind of contaminated negatives could gain a deceptive decision boundary, especially when a large number of inaccurate negatives are included [15]. Some of existing methods randomly selected a subset of unlabeled samples to use as negatives with the hope to reduce the number of fallacious negatives obtained. Due to an uncontrollable proportion of noisy negatives, this strategy may lead to an unstable classifier [15] (i.e. different sets of negatives produce seriously distinct decision boundaries of a classifier).

To take advantages of model variances, an ensemble learning with positive-unlabeled (PU) data, so-called a PU bagging classifier, is employed in the proposed method. Bagging or bootstrap aggregate is an ensemble technique known to efficiently reduce model variances and improve model generalization by aggregate predictions from multiple meta-learning models trained on different subsets of data. This technique has been successfully adopted with PU data in many studies such as an ensemble SVMs for predicting kinase substrates [80] and a bagging SVM for identifying drug-drug interactions (DDIs) [70]. A framework of a classification model used in the proposed method is depicted in Figure 4.4.

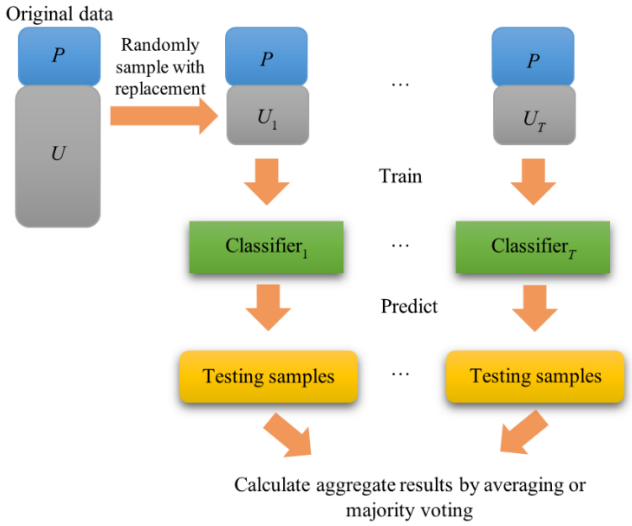


Figure 4.4 A classification model used in the proposed method

In this work, known drug-disease associations are positive samples (P) with the class labels of "1" whereas the remaining drug-disease pairs are unlabeled samples (U) with the class labels of "0". In general, unlabeled drug-disease pairs greatly outnumber positive drug-disease associations. To deal with this, unlabeled drug-disease pairs are randomly sampled with replacement to generate T bootstrap samples (i.e. U_1, U_2, …, U_T) of the sizes equal to P. Then, the same set of positive samples (P) and each bootstrap sample of unlabeled pairs are used to train a base classifier, leading to T base classifiers obtained. Based on latent features of meta-path based functional profiles, these multiple models are used to recognize positive samples from unlabeled samples in the proposed method.

Suppose that x is a latent feature vector of meta-path based functional profiles of a testing drug-disease pair. $h_t(x)$ is a function of the t^{th} base classifier (t = 1, 2, …, T) which gives a predicted probability of being in the positive class of a testing sample. With a given threshold score (z), a testing sample is predicted as either "1" (positive) or "0" (unlabeled) by a function $\sigma(x)$, as shown in (4.8). In this work, a threshold score or z was set as 0.5. To combine multiple predictions obtained from T base classifiers for a testing sample, the averaging and majority voting schemes are employed, which can be defined in (4.9) and (4.10), respectively.

$$\sigma(x) = \begin{cases} 1, & \text{if } h_t(x) \geq z \\ 0, & \text{if } h_t(x) < z \end{cases} \tag{4.8}$$

$$H_{avg}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{4.9}$$

$$H_{mv}(x) = \begin{cases} 1, & \text{if } \frac{1}{T}\sum_{t=1}^{T}\sigma(h_t(x)) \geq \frac{T}{2} \\ 0, & \text{if } \frac{1}{T}\sum_{t=1}^{T}\sigma(h_t(x)) < \frac{T}{2} \end{cases} \tag{4.10}$$

In the proposed method, an Extreme Gradient Boosting model (XGBoost) [82] is used as a base classifier. XGBoost is a boosting ensemble model that has been applied in various fields, such as network intrusion detection, personal credit evaluation, drug discovery, and financial

trading. During these recent years, XGBoost has been a competitive method with deep learning methods due to its beneficial features (e.g. parallelization and regularization) and high performance [83]. Due to its high efficiency, XGBoost is selected to learn PU data with the hope that it can handle the noise in unlabeled data and result in improved performance of a PU bagging classifier.

4.3.5 Experimental settings and performance evaluation

In this section, data manipulation for properly developing a classification model and evaluating performance of the proposed and other compared methods is described. In the model development, which parameter values were tuned and how this process was performed are explained next. Finally, evaluation metrics used in this study are shown.

1) Data manipulation

To properly manage data for each experiment in this study, the data of all drug-disease pairs were manipulated as shown in Figure 4.5, where P and U represent the set of all positive and unlabeled drug-disease pairs, respectively. In the beginning, all samples in P were shuffled and then divided into approximately equal five subsets. Each of them was combined with a set of unlabeled samples resulting from randomly sampling without replacement from U with the size of its positive samples. This set is called an outer testing data set which will be preserved for only performance evaluation of the proposed method. Out of this data set, the remaining data set is called an outer training data set which will be used in the experiments involving with model development. Outer testing data sets cannot be used in the stage of model development to avoid overfitting problems likely occurred when testing sets are already seen. To use in performance evaluation of each model under different settings (i.e. in terms of different meta-path based functional profiles or parameter values used), an inner testing data set was generated by randomly sampling 20% of positive and unlabeled samples in an outer training data set. Out of this data set, a set the remaining drug-disease pairs is called an inner training data set which will be employed for training a classification model. To avoid random bias, the whole process of generating all data sets was repeated for ten times, and the results obtained from all repetitions were used to estimate average performance values of each model.

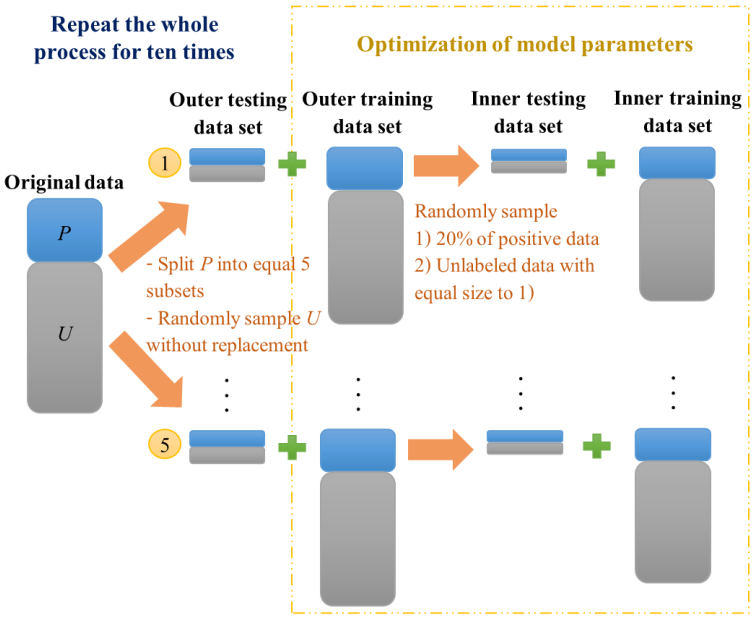


Figure 4.5 Data manipulation for the experiments of this study

During the stage of model development, each model under a particular setting is trained on an inner training data set and then tested on an inner testing data set. A set of parameter values providing the highest performance values over all experimental repetitions will be selected and used in subsequent processes. Under an optimal choice of model settings, a classification model was retrained on the whole set of an outer training data set and then tested on an outer testing data set to measure performance of the proposed method. For other existing methods compared with the proposed method, their parameter values were tuned by using only outer training data sets, similar to those conducted for the proposed method. Each method under its optimal setting will be tested on outer testing data sets. To prevent random bias in performance comparison, both positive and unlabeled samples in all outer testing data sets were maintained as the same for both the proposed methods and other methods.

2) Selection of model inputs and parameter values

There are three parts of a model which can affect the performance of the proposed method, including an input set of meta-path based functional profiles, values of XGBoost parameters, and settings involving in the aggregate process.

- Input sets of meta-path based functional profiles

  Because three different meta-path based functional profiles are generated (i.e. X_{M_1}, X_{M_2}, and X_{M_3}), an optimal choice to use them should be initially specified before tuning other model parameters. Seven different functional profile matrices, including both independent and integrated functional profiles which can be generated by concatenating distinct functional profile matrices together (i.e. X_{M_1}, X_{M_2}, X_{M_3}, $X_{M_1M_2}$, $X_{M_1M_3}$, $X_{M_2M_3}$, and $X_{M_1M_2M_3}$), were investigated. For each profile matrix, the truncated SVD was performed to find its low-dimensional representations of drug-disease pairs. Then, the obtained latent features of drug-disease pairs in inner training data sets were used to build classification models, and then these models were tested with drug-disease pairs in inner testing data sets. A functional profile matrix that provides the greatest average performance values will be selected and used in the next experiments. In this step, the values of all XGBoost parameters were set at the default values, the number of bootstrap samples (T) was fixed at ten, and the aggregate scheme used was the averaging method.

- XGBoost parameters

  After obtaining an optimal input matrix of meta-path based functional profiles, its latent features were used to tune XGBoost parameters, including *learning_rate* (a shrinkage factor of each added tree), *n_estimators* (the number of trees), *max_depth* (a maximum depth of a tree), and *min_child_weight* (a minimum summation of instance weights in a child node). The sets of parameter values under investigation are shown as follows: *learning_rate* = {0.1, 0.3, 0.5}, *n_estimators* = {100, 300, 500}, *max_depth* = {4, 6, 8}, and *min_child_weight* = {3, 5, 7}. A grid search was performed on outer training data sets to find optimal values of these parameters. To avoid laboriousness in independent tuning multiple base classifiers, the set of these optimal values was used for all base classifiers in this study. During tuning XGBoost parameters, the T value was set at ten, the averaging scheme was used to combine multiple predictions from those ten XGBoost classifiers.

- The number of bootstrap samples (T) and aggregate schemes

  After getting optimal values of XGBoost parameters, a suitable value of T and aggregate method were identified. The values of T under investigation are 10, 20, 30, 50, 70, 100, 150, and 200. To combine multiple predictions obtained from multiple XGBoost classifiers, two aggregate schemes were compared, which are the averaging and majority voting scheme. The values of T and aggregate schemes were simultaneously changed to observe which value of T and aggregate scheme can produce the highest performance values. In this experiment, the values of XGBoost parameters were set at the optimal values identified previously.

3) Evaluation metrics

To evaluate performance of the proposed method and other compared methods, well-known performance measures are used, including precision (*PRE*), recall (*REC*), and accuracy (*ACC*). In addition, two useful comprehensive metrics, F_1-score (F_1) and Matthew's correlation coefficients (*MCC*) are also employed in this study. All of these metrics can be computed as shown in (4.11) - (4.15), where *TP*, *FP*, *FN*, *TN* represent the number of true positives, false positives, false negatives, and true negatives, respectively.

$$PRE = \frac{TP}{TP + FP} \tag{4.11}$$

$$REC = \frac{TP}{TP + FN} \tag{4.12}$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.13}$$

$$F_1 = \frac{2 \times PRE \times REC}{PRE + REC} \tag{4.14}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.15}$$

Although a set of truly negative samples is not known in this PU learning, all unlabeled samples are typically treated as negatives to estimate evaluation metrics referred to the negative set, such as precision, accuracy, F_1, and *MCC*. An additional metric that is used in several PU learning based studies [76, 81] is F_1-score for PU learning ($F_{1,\ PU}$), which requires only information from the positive class. This measure can be computed following to (4.16), where N is the total number of testing samples, and UP is the number of unlabeled samples which are tested and predicted as positives.

$$F_{1,PU} = \frac{(REC)^2}{TP + UP/N} = \frac{(REC)^2 \times N}{TP + UP} \tag{4.16}$$

According to the above evaluation metrics, a predefined threshold score is required for predicting testing samples as binary classes and computing their values. To assess performance on all possible threshold scores, an ROC and precision-recall (PR) curve are also plotted. Additionally, an AUROC and AUPRC values are computed to summarize those curves. Furthermore, this PR curve is also exploited to specify an optimal threshold score to predict a binary class for each drug-disease pair. In the PR curve, a selected threshold score is a point where achieves the maximum F_1 value. This strategy was also used in several studies such as [11] and [47].

In many studies, K top ranked drug-disease associations were recommended as potential drug-disease associations. To analyze the capability of recovering positive samples in top-K results, some evaluation metrics are derived and computed based on these top-K predictions as shown in (4.17) - (4.20). These evaluation metrics were also used to evaluate performance of models in many recent studies, such as[45], [120], and [121]. In this study, the values of K were varied from 0 to 1,300 with the step of ten. To make these metrics different from the traditional ones, the researcher denote precision, recall, F_1, and $F_{1,\ PU}$ that are computed at each value of K as *PRE@K*, *REC@K*, F_1*@K*, and $F_{1,\ PU}$*@K*, respectively. Note that K = 0, 10, 20, …, 1,300, and N_p is the number of positive samples in a testing data set.

$$PRE@K = \frac{TP}{K} \tag{4.17}$$

$$REC@K = \frac{TP}{N_p} \tag{4.18}$$

$$F_1@K = \frac{2 \times (PRE@K) \times (REC@K)}{(PRE@K) + (REC@K)} \tag{4.19}$$

$$F_{1,PU}@K = \frac{(REC@K)^2 \times N}{K} \tag{4.20}$$

4.4 Results

4.4.1 The constructed drug-GO-disease tripartite network

Based on drug-disease, drug-GO, and disease-GO association data, the drug-GO-disease tripartite network used in this study was constructed and shown in Figure 4.6. This is an undirected and unweighted tripartite network, where only nodes of different types are linked together. Some basic properties of this network are summarized in Table 4.1.

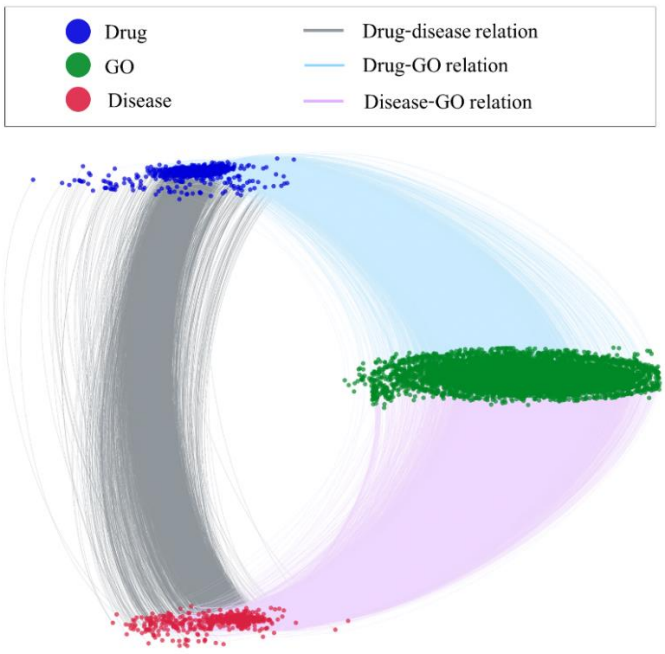


Figure 4.6 The constructed drug-GO-disease tripartite network

In this network, there are 1,022 drug nodes, 585 disease nodes, and 8,320 GO nodes. GO nodes of all aspects were included into this network because they all can provide information about relationships between drugs and diseases. Among all GO nodes, there are 5,009 Biological Process (BP) nodes (60%), 2,408 Molecular Function (MF) nodes (29%), and 903 Cellular Component (CC) nodes (11%).

Table 4.1 Properties of the drug-GO-disease tripartite network

Network compartment	**Type**	**Total number**	**Average node degree or link density**
Nodes	drug	1,022	51.3 GO functions 6.6 diseases
	GO function	8,320	6.3 drugs 11.1 diseases
	disease	585	157.5 GO functions 11.5 drugs
Links	drug-GO	52,463	0.0062
	disease-GO	92,135	0.0189
	drug-disease	6,710	0.0112

According to Table 4.1, there are three link types in the tripartite network which are 52,463 drug-GO links, 92,135 disease-GO links, and 6,710 drug-disease links. Only small number of known drug-disease associations are detected when compared to drug-GO and disease-GO associations. Link densities were calculated from proportions of existing links to all possible links. The link densities of drug-GO, disease-GO, and drug-disease associations are 0.0062, 0.0189, and 0.0112, respectively. The disease-GO links are three times denser than the drug-GO links. This may be because the greater number of disease proteins was included in the study than that of drug target proteins, leading to the larger number of GO functions linked to diseases. In case of drug-disease associations, only 1.12% of all possible drug-disease pairs have already detected in this data set. These light existing links between drugs and diseases suggest that there

is still much room for new drug-disease associations and provide us an opportunity to uncover potential drug-disease associations in this study.

Due to denser existing links between diseases and GO functions, one disease is associated with up to 158 GO functions whereas one drug is associated with about 51 GO functions on average. To link GO functions to their associated drugs or diseases, drug-associated and disease-associated proteins were used. Therefore, the numbers of proteins associated with one drug or one disease can be also investigated. Consequently, the average numbers of GO functions associated with one drug or one disease are more than twice times greater than those of proteins interacting with one drug (9.6 proteins per drug) or one disease (55.9 proteins per disease). When compared to proteins, a drug or a disease is typically relevant to many GO functions, especially when all GO aspects are considered. Conversely, one GO function is also involved with many drugs (6.3 drugs per GO function) and many diseases (11.1 diseases per GO function) on average. According to the results, it suggests that there is much more chance that a drug or a disease would overlap their GO functions with other drugs or diseases when compared to proteins. This enables an opportunity to improve an identification of relationships between drugs and diseases using GO functions. When considered links between drugs and diseases, one drug is associated with 6.6 diseases whereas one disease is associated with 11.5 drugs on average. This result supports the paradigm of using one drug for multiple diseases or vice versa, and also drug repositioning.

4.4.2 Usage of different meta-path based functional profiles

Because three different meta-path based functional profiles are generated (i.e. X_{M_1}, X_{M_2}, and X_{M_3}), which functional profiles should be included in the classification of drug-disease associations needs to be initially specified. Both independent and integrated functional profile matrices are investigated, including X_{M_1}, X_{M_2}, X_{M_3}, $X_{M_1M_2}$, $X_{M_1M_3}$, $X_{M_2M_3}$, and $X_{M_1M_2M_3}$. Notice that $X_{M_iM_j}$ represents an integrated functional profile matrix which concatenates functional profile matrix X_{M_i} and X_{M_j} together, where i, j = 1, 2, 3, and $X_{M_1M_2M_3}$ is the functional profile matrix that concatenates all different functional profile matrices.

Before training classification models on these functional profile matrices, the truncated SVD was performed to find their low-dimensional representations. For each functional profile matrix, the number of latent features acquired is determined by a latent feature percentage, which

is relative to the number of all features in an original functional profile matrix. An optimal number of retained components or the suitable number of latent features is specified at the minimum value where has the coverage percentage of all singular values at least 95%. The number of latent features selected for each functional profile matrix is shown in Table 4.2.

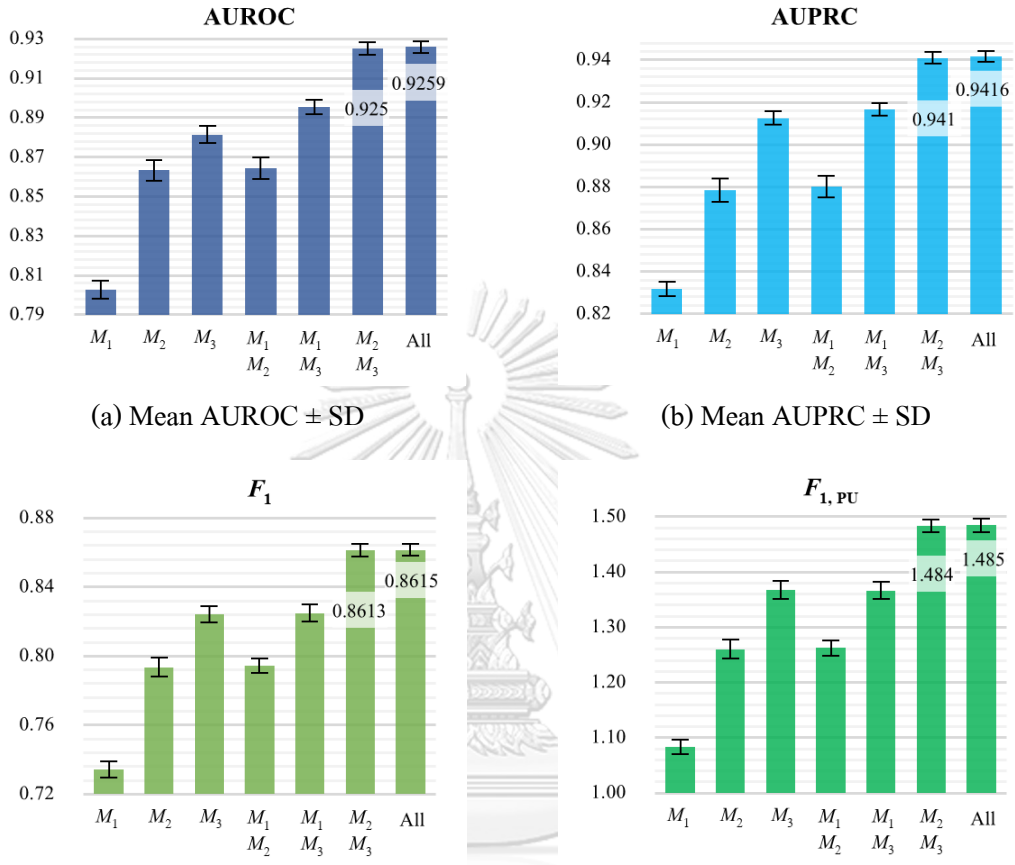
Table 4.2 The selected latent feature percentage for each functional profile matrix

Matrix of meta-path based functional profiles	**Latent feature percentage (*l*)**	**The number of latent features (*r*)**
X_{M_1}	7.992%	665
X_{M_2}	0.432%	36
X_{M_3}	4.699%	391
$X_{M_1M_2}$	0.228%	38
$X_{M_1M_3}$	2.415%	402
$X_{M_2M_3}$	1.868%	311
$X_{M_1M_2M_3}$	1.278%	319

According to Table 4.2, the selected latent feature percentages (l) of all functional profile matrices are less than 10%, and the corresponding numbers of the latent features (r) range from 36 to 665 features. This could indicate that SVD can efficiently reduce the dimensions of all functional profile matrices. Furthermore, the functional profile matrices that originally have large latent feature matrices (i.e. X_{M_1} and X_{M_3}) require smaller numbers of new latent features to retain the same amount of information when they are combined with X_{M_2}. This suggests that combining multiple functional profile matrices contributes much more information than using the independent meta-path based functional profiles. Due to integration of different information provided by multiple meta-paths, the combined functional profile matrices could be more useful for the classification of drug-disease associations.

Next, each latent feature profile matrix obtained from SVD was used to train and test a classification model so that the performance of different functional profile matrices can be

estimated and compared. The mean values of some evaluation metrics and their standard deviations (SD) are shown in Figure 4.7.



(a) Mean AUROC ± SD (b) Mean AUPRC ± SD

(c) Mean F_1 ± SD (d) Mean $F_{1,\ PU}$ ± SD

Figure 4.7 Performance comparison of using different meta-path based functional profiles

When comparing among the independent functional profile matrices, it was found that M_3-based functional profiles can produce the highest values in all performance metrics (AUROC = 0.8815, AUPRC = 0.9126, F_1 = 0.8243, and $F_{1,\ PU}$ = 1.3663). With the data set used in this study, it may suggest that meta-path M_3 is the most effective meta-path which can accumulate the greatest amount of useful information for classifying drug-disease associations. However, the performance of M_3 can still be enhanced by combining X_{M_3} with other functional profile matrices. The performance of a classification model was slightly improved when using the combined functional profile matrix $X_{M_1M_3}$ (AUROC = 0.8955, AUPRC = 0.9167, F_1 = 0.825, and

$F_{1,\,PU}$ = 1.3678), and it was significantly improved when using $X_{M_2M_3}$ (AUROC = 0.925, AUPRC = 0.941, F_1 = 0.8613, and $F_{1,\,PU}$ = 1.484). The highest values in all evaluation metrics are produced by using the functional profile matrix which integrates all meta-path based functional profiles or $X_{M_1M_2M_3}$ (AUROC = 0.9259, AUPRC = 0.9416, F_1 = 0.8615, and $F_{1,\,PU}$ = 1.485). From these results, it can be concluded that integrating multiple meta-path based functional profiles could gain more beneficial information for classifying drug-disease associations, resulting in a performance improvement when combined functional profile matrices are used. By integrating multiple functional profile metrices, diverse information of drug-disease pairs provided by different meta-paths are aggregated, which could be useful for predicting drug-disease associations.

To support why integrating multiple meta-path based functional profiles is more useful, both independent and integrated functional profile matrices are investigated. Because it is difficult to measure how useful information contained in each functional profile matrix is, information loss in each matrix is observed instead. This investigation is limited to only a set of 6,710 positive drug-disease associations because obvious relationships can be found only in this set. For these known drug-disease associations, some GO functions associated with them should be detected by meta-paths and included into their functional profiles so that their relationships are detectable by a classification model. However, it was found that some of them have all-zero functional profiles or all-zero features in some functional profile matrices. To compare information loss between different functional profile matrices, the number of positive samples that have all-zero functional profiles or all-zero features are observed in each functional profile matrix and shown in Table 4.3.

Table 4.3 The number of positive samples that have all-zero functional profiles in each functional profile matrix

Matrix of meta-path based functional profiles	Positive samples that have all-zero functional profiles	
	Number	Percentage
X_{M_1}	1,469	22.30%
X_{M_2}	369	5.50%
X_{M_3}	127	1.89%
$X_{M_1M_2}$	369	5.50%
$X_{M_1M_3}$	127	1.89%
$X_{M_2M_3}$	1	0.01%
$X_{M_1M_2M_3}$	1	0.01%

The numbers of positive samples that have all-zero M_1-based, M_2-based, and M_3-based functional profiles are 1,496 (22.30%), 369 (5.50%), and 127 (1.89%), respectively. This could indicate that there are a greater number of positive samples that M_1 cannot detect their associated GO functions when compared to those exploiting M_2 and M_3. This higher number could lead a classification model to confusion and result in its worse performance, which can be seen in Figure 4.7. Nevertheless, the loss of functional information in a meta-path based functional profile matrix can be resolved by integrating multiple functional profile matrices. For example, by concatenating X_{M_1} and X_{M_2}, the number of positive samples that have no any detected GO functions can be reduced from 1,496 (22.30%) to 369 (5.50%). Especially, those functional profile matrices combining with X_{M_3} (i.e. $X_{M_1M_3}$, $X_{M_2M_3}$, and $X_{M_1M_2M_3}$) can greatly reduce their numbers of all-zero functional profiles in the positive samples. For instance, in $X_{M_2M_3}$ and $X_{M_1M_2M_3}$, there is only one positive samples left (0.01%) that has all-zero functional profiles. These results could partially support why combining multiple meta-path based functional profiles is better than solely using an independent functional profile matrix. Due the greatest values in all evaluation metrics of $X_{M_1M_2M_3}$, this functional profile matrix is used as an input of the proposed method for further classifying drug-disease associations.

4.4.3 Selected values of model parameters

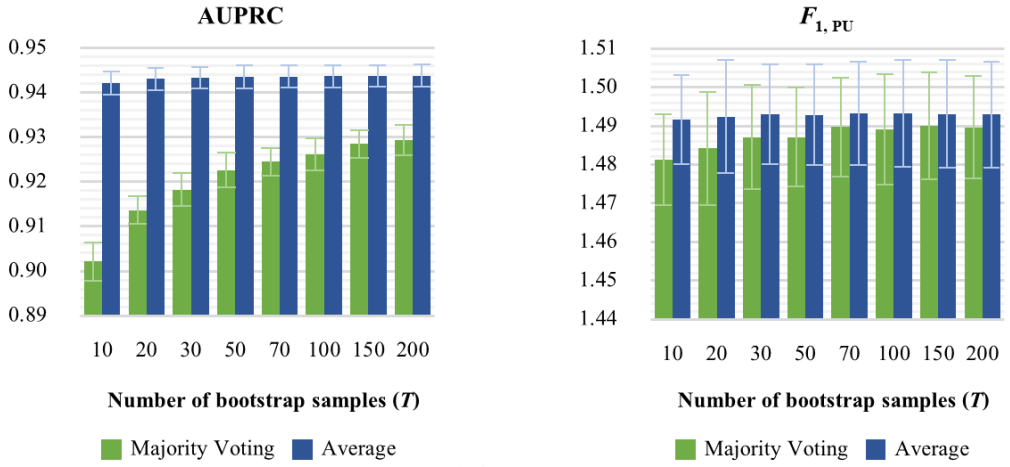
After getting an appropriate input for the proposed method, optimal values of model parameters are then identified. For the proposed method, two main sets of parameters needed to be tuned are XGBoost parameters and those relevant to combining multiple predictions (i.e. the number of bootstrap samples and aggregate schemes).

1) XGBoost parameters

There are four XGBoost parameters were tuned by using outer training data sets. They are *learning_rate* (a shrinkage factor of an added tree), *n_estimators* (the number of tree), *max_depth* (a maximum depth of a tree), and *min_child_weight* (a minimum summation of instance weights in a child node). A grid search was performed to evaluate performance of models with all possible combinations of the predefined values of all parameters. An optimal set of parameter values was manually selected from a set that provides the highest values in most comprehensive evaluation metrics (i.e. AUPRC, AUROC, F_1, and $F_{1,\,PU}$). As a result, the selected set of parameter values are {*learning_rate*, *n_estimators*, *max_depth*, *min_child_weight* } = {0.3, 500, 6, 3}. With this set of parameter values, the model can produce the mean AUPRC of 0.9427, the mean AUROC of 0.9286, the mean F_1 of 0.8633, and the mean $F_{1,\,PU}$ of 1.4921.

2) Aggregate schemes and the number of bootstrap samples (T)

To combine multiple predictions from several base classifiers, a suitable aggregate scheme and the appropriate number of bootstrap samples (T) should be primarily determined. Two aggregate schemes are considered which are the averaging and majority voting scheme. The values of T under investigation are 10, 20, 30, 50, 70, 100, 150, and 200. To find the suitable values of both parameters, they were combined together, and the models with these different settings of both parameters were evaluated their performance. In this experiment, the parameters of all XGBoost classifiers were set at the optimal values previously found. For each T value, the mean AUPRC and $F_{1,\,PU}$ values of both averaging and majority voting schemes with their SDs are computed and shown in Figure 4.8.



(a) Mean AUPRC ± SD (b) Mean $F_{1,\,PU}$ ± SD

Figure 4.8 Performance comparison of using different aggregate schemes

When comparing between the averaging and majority voting scheme, it is noticeable that models with the averaging scheme can produce the higher mean AUPRC and $F_{1,\,PU}$ values, no matter what value of T is used. Furthermore, the majority voting scheme requires a larger value of T when compared to that of the averaging scheme to obtain a stable ensemble model. This may be because the majority voting scheme is directly based on a binary class given for each testing sample (see the section 4.3.4 for more details). With unstable base classifiers, early predicting a testing sample as a binary class could result in too diverse predictions of each testing sample. This leads to the difficulty to create accurate combined predictions by the final ensemble model, especially when the small value of T is used. To resolve this problem, the majority voting scheme requires the larger value of T to make more accurate combined predictions when compared to that used in the averaging scheme. According to these results, the averaging scheme is selected to use in the proposed method.

Next, the suitable value of T is identified when the averaging scheme is applied in the proposed method. When T was increased in each time, the old and new mean AUPRC values of the averaging scheme was compared by one-sided t-tests to examine whether there is a significant improvement in AUPRC values or not. According to Figure 4.9, the mean AUPRC values are significantly improved until T reaches the value of 50. After T is increased above 50, only insignificant improvements in the mean AUPRC values are found. The mean AUPRC value and its SD at T = 50 for the averaging scheme is 0.9436 ± 0.0026. Thus, T is set at 50 and the

averaging scheme is applied to combine multiple predictions obtained from base classifiers in the proposed method.

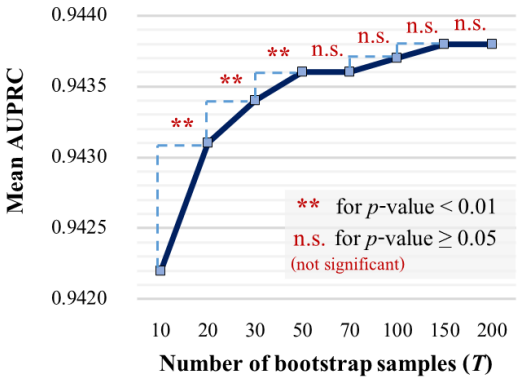


Figure 4.9 Improvements of mean AUPRC values when T was increased and the averaging scheme was used

4.4.4 Comparison with other methods

In this section, the performance of the proposed method with an optimal setting is evaluated by comparing with other existing methods. Four state-of-the-art methods and one baseline method are selected to compare with the proposed method. The first method is a three-layer heterogeneous graph based inference (TL_HGBI) [4] method, which integrates drug target information with drug-drug and disease-disease similarity scores. The second method is MBiRW [5], that exploits a bi-random walk algorithm to predict new links between drugs and diseases in a two-layer heterogeneous network. The third method is the ensemble meta-paths and singular value decomposition (EMP-SVD) [11] method, that uses a meta-path based ensemble model with a heuristic strategy to select reliable negatives for predicting drug-disease associations. The fourth method is the topological similarity and singular value decomposition (TS-SVD) [47] method, an improved version of EMP-SVD that exploits topological similarity features of drugs and diseases. The baseline method directly exploits path counts as features of drug-disease pairs for predicting drug-disease associations. This method is compared with the proposed method to show the superior performance of novel features of drug-disease pairs or meta-path based functional profiles.

Each method was trained on outer training data sets and then tested with outer testing data sets. The same drug-disease pairs were maintained in all outer testing data sets to ensure that all methods are tested on the same data sets. Due to multiple outer testing data sets used, the average values of all evaluation metrics are computed to represent performance of each method. For all methods, the mean values of each evaluation metric and their SDs are shown in Table 4.4.

Table 4.4 Performance comparison of the proposed method and other methods

Metrics	**Methods**					
	Baseline	**TL_HGBI**	**MBiRW**	**EMP-SVD**	**TS-SVD**	**Proposed method**
AUROC	0.864^{**} ± 0.007	0.831^{**} ± 0.007	0.892^{**} ± 0.008	0.928^{**} ± 0.005	0.904^{**} ± 0.006	**0.930** ± 0.006
AUPRC	0.886^{**} ± 0.005	0.829^{**} ± 0.009	0.903^{**} ± 0.008	0.940^{*} ± 0.004	0.919^{**} ± 0.005	**0.944** ± 0.004
PRE	0.782^{**} ± 0.028	0.732^{**} ± 0.018	0.820^{**} ± 0.024	0.846^{**} ± 0.024	0.825^{**} ± 0.026	**0.886** ± 0.021
REC	0.802^{**} ± 0.028	0.839^{**} ± 0.021	0.830^{**} ± 0.022	$\mathbf{0.857}^{\text{n.s.}}$ ± 0.027	0.825^{**} ± 0.027	0.842 ± 0.019
ACC	0.788^{**} ± 0.012	0.765^{**} ± 0.010	0.823^{**} ± 0.010	0.850^{**} ± 0.007	0.824^{**} ± 0.009	**0.867** ± 0.007
MCC	0.578^{**} ± 0.023	0.537^{**} ± 0.017	0.648^{**} ± 0.020	0.700^{**} ± 0.014	0.649^{**} ± 0.017	**0.735** ± 0.015
F_1	0.791^{**} ± 0.008	0.781^{**} ± 0.006	0.825^{**} ± 0.008	0.851^{**} ± 0.006	0.824^{**} ± 0.007	**0.863** ± 0.007
$F_{1,\text{PU}}$	1.253^{**} ± 0.025	1.227^{**} ± 0.019	1.361^{**} ± 0.026	1.449^{**} ± 0.022	1.359^{**} ± 0.022	**1.492** ± 0.024

Note that 1) numbers in the table are average performance values with their standard deviations,

2) a bold number indicates the maximum value of each evaluation metric, and

3) the subscripts denote *p*-values of one-sided *t*-tests, where n.s. (not significant) is marked for *p*-values ≥ 0.05, a single asterisk (*) is marked for 0.01 ≤ *p*-values < 0.05, and a double asterisk (**) is marked for *p*-values < 0.01.

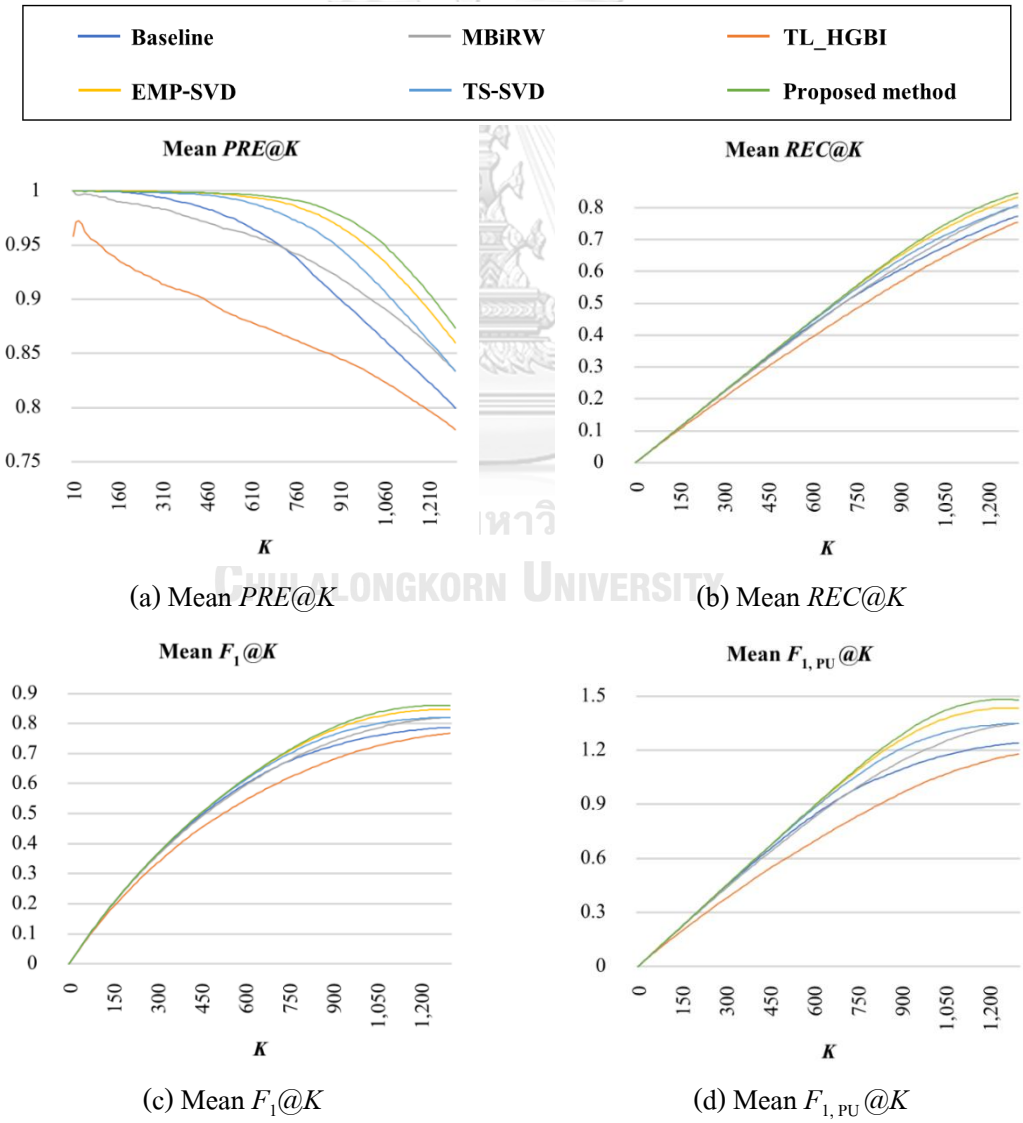
When comparing between the proposed method and the baseline method, it is found that the proposed method outperforms the baseline method at the significant level of 1%, no matter which evaluation metric is considered. This suggests that meta-path based functional profiles are more useful features than ordinary path counts for classifying drug-disease associations. It means that the information of the intermediate nodes (i.e. GO functions) in the tripartite network is also important and very useful in uncovering the relationships between drugs and diseases.

When compared to other state-of-the-art methods, it is noticeable that the proposed method can produce the maximum values in almost evaluation metrics. At the low significance level of 0.01, the mean AUROC, *PRE*, *ACC*, *MCC*, F_1, and $F_{1,\,PU}$ values of the proposed method are significantly greater than those of all compared methods. The mean AUPRC value of the proposed method is significantly higher than those of other methods at the significance level of 0.05 In case of *REC* values, the proposed method has the greater mean *REC* value (0.842 ± 0.019) than those of the others, except EMP-SVD (0.857 ± 0.027). It can be implied that the smaller number of positive samples in a testing data set is recovered by the proposed method when compared to that of EMP-SVD. This may be because the proposed method predicts the fewer number of the positive drug-disease associations than that of EMP-SVD. To support this reason, it is noteworthy that the proposed method has the higher $F_{1,\,PU}$ value than that of EMP-SVD despite the lower *REC* value of the proposed method found. According to the formula of $F_{1,\,PU}$ shown in (4.16), it can suggest that the proposed method has the lower probability of predicting a sample as positives than that of EMP-SVD. Despite this lower probability, those samples that are predicted as positives by the proposed method are more reliable because it has the maximum *PRE* value, the estimated probability that a testing sample is accurately predicted to be positive.

Surprisingly, TS-SVD (an improved version of EMP-SVD) has worse performance than that of EMP-SVD. This is mainly because of different heuristic strategies for selecting reliable negative samples used. In EMP-SVD, negative samples are selected from unlabeled pairs of drugs and diseases that have no common interacting proteins between them. In TS-SVD, unlabeled drug-disease pairs with no k-step neighbors (k = 1, 2, 3) are considered as reliable negative samples. It is noteworthy that the EMP-SVD strategy will be equivalent to the TS-SVD strategy if it uses only k = 2. This means that a set of TS-SVD negative samples is a subset of EMP-SVD negative samples, and the larger number of unlabeled samples are thrown away by the TS-SVD

method. This leads TS-SVD to generate the overfit model with only seen drug-disease pairs (positives and selected negatives) and less generalized for unseen unlabeled drug-disease pairs.

In addition, the performance of all methods was evaluated based on top K ranked predictions. With this strategy, the evaluation metrics (i.e. *PRE*, *REC*, F_1, and $F_{1,\ PU}$) were computed at several values of K and called as *PRE@K*, *REC@K*, $F_1@K$, and $F_{1,\ PU}@K$. The considered K values range from 0 to 1,300 with the step of ten. At a particular value of K, the average value of each evaluation metric is used to represent the estimated value of that metric. The mean *PRE@K*, *REC@K*, $F_1@K$, and $F_{1,\ PU}@K$ values of all methods are shown in Figure 4.10.



(a) Mean *PRE@K* (b) Mean *REC@K*

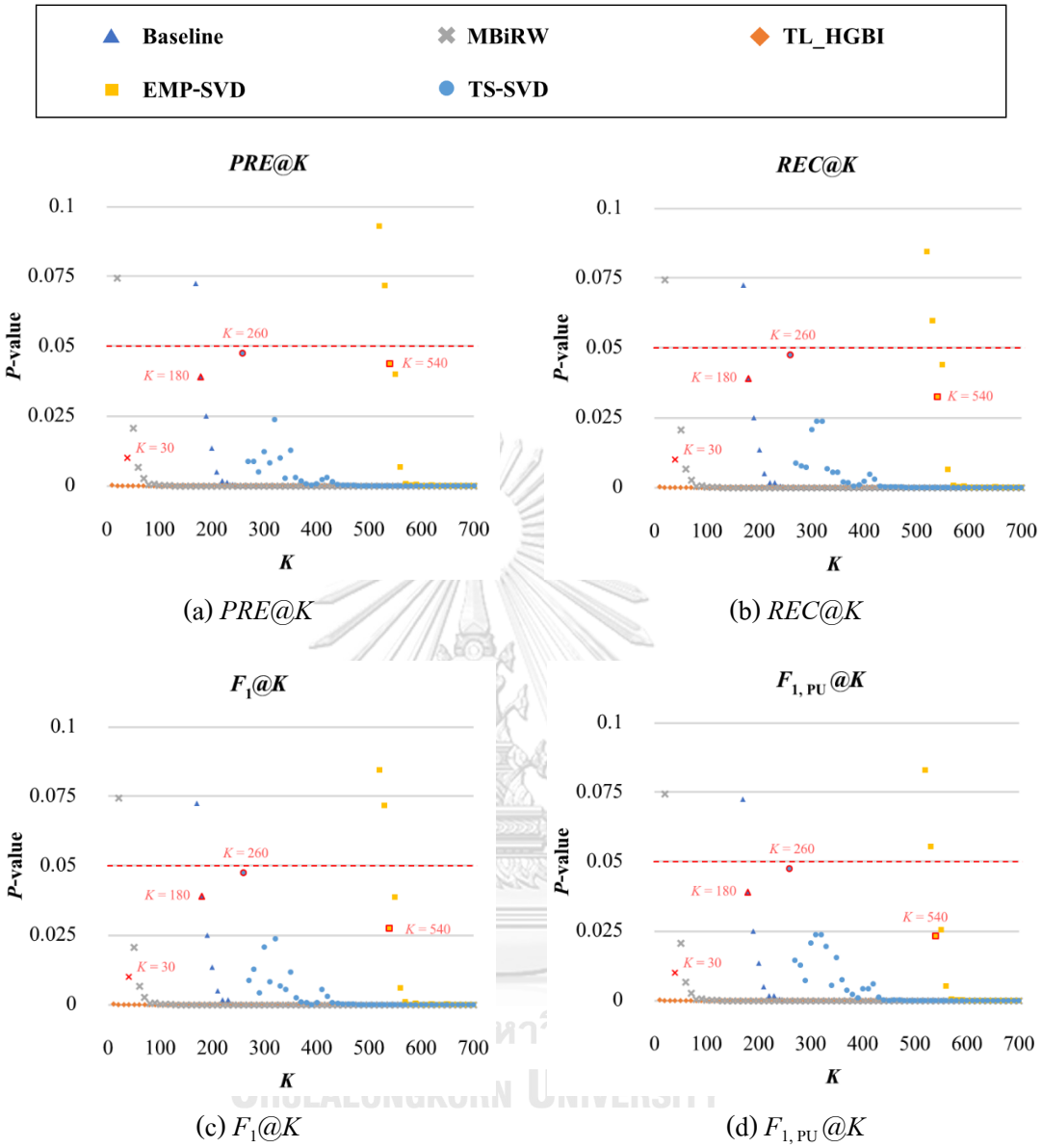
(c) Mean $F_1@K$ (d) Mean $F_{1,\ PU}@K$

Figure 4.10 Comparison of performance based on top-K ranked predictions

According to Figure 4.10, no matter which metric is considered, the proposed method has higher values than those of other methods, especially when *K* is large. When *K* is small, the top-*K* ranked predictions of all methods accurately include only positive drug-disease associations. Nevertheless, the top-*K* predictions of most methods encompass other samples that are not positive, when *K* is increasing. From Figure 4.10(a), it is noticeable that the top *K* predictions of the proposed method covers the greatest number of positive testing samples, especially when *K* is larger than 760. Figure 4.10(b) shows that the proposed method can recover the highest number of positive samples although it is not obviously seen in the plot, especially when compared to EMP-SVD.

To explicitly demonstrate the position of *K* where the proposed method outperforms other methods, the values of each evaluation metric of the proposed method are compared with those of other methods at each value of *K* by using one-sided Wilcoxon signed rank tests. The smaller the value of *K* that initiates the statistical significance is, the better the performance of the proposed method is. *P*-values obtained from the tests for each metric are shown in Figure 4.11. The red dashed lines specify the significance level of 0.05. A marker with a red border indicates the position of *K* where a mean performance measure of the proposed method begins significantly higher than that of a compared method.

In all evaluation metrics, the *K* values that the mean values of the proposed method begin significantly higher than those of TL_HGBI, MBiRW, Baseline, TS-SVD, and EMP-SVD are 10, 30, 180, 260, and 540, respectively. These numbers indicate the values of *K* which top-*K* predictions of the proposed method are more accurate than those of compared methods. For example, when compared to TL_HGBI, only top 10 ranked predictions of the proposed method contain the larger number of accurate predictions than those obtained from TL_HGBI. According to these results, it can conclude that the proposed method can recover the larger number of positive drug-disease associations in top-*K* predictions than those gained from other methods.



(a) *PRE@K* (b) *REC@K*

(c) F_1@*K* (d) $F_{1,\,PU}$@*K*

Figure 4.11 *P*-values of Wilcoxon signed rank tests for comparing performance based on top-*K* ranked predictions

4.4.5 Validation of predicted drug-disease associations

To predict potential drug-disease associations, the proposed method with the optimal settings was used. In this step, all positive drug-disease associations were exploited to train a classification model, and then the trained model was employed to predict all 591,160 unlabeled drug-disease pairs. To predict which unlabeled pairs are probably in the positive class, the threshold score was specified at the maximum $F_{1,\,PU}$ value. By this strategy, almost known drug-

disease associations (98.05%) were predicted as positives. This number is greater than that obtained from using the strategy of the maximum F_1 score, which is commonly used in many studies. As a result, the threshold score used in this study is 0.990385 which provides the maximum $F_{1,\,PU}$ of 76.90. With this threshold score, 895 unlabeled drug-disease pairs (309 drugs and 149 diseases) were recommended as candidate drug-disease associations. Therefore, the remaining 590,265 unlabeled pairs of drugs and diseases can be considered as non-candidate drug-disease associations. The full list of 895 discovered drug-disease associations is provided at http://ieee-dataport.org/3540.

To validate that information supporting the candidate drug-disease associations is not found by chance, both candidate and non-candidate drug-disease associations were searched for their supporting evidence from two databases: ClinicalTrials.gov and CTD. The numbers of the candidate and non-candidate drug-disease associations that their supporting information is found or not found are summarized in Table 4.5.

ClinicalTrials.gov is a database that collects clinical studies conducted in more than 200 countries around the world. This database is maintained by U.S. National Library of Medicine (NLM) and National Institutes of Health (NIH). As a result, 337 out of 895 drug-disease associations (37.7%) were reported in ClinicalTrials.gov whereas only 15,380 non-candidate associations (2.6%) were also found in that database. By the one-sided Fisher's exact test, it can be concluded that this supporting information is significantly found in the candidate drug-disease pairs but not in the non-candidate drug-disease pairs with a p-value of 1.23×10^{-283}.

Table 4.5 Summary of candidate and non-candidate drug-disease associations and their supporting evidence

Source of evidence	**The number of candidate drug-disease associations (%)**		**The number of non-candidate drug-disease associations (%)**	
	Found	**Not found**	**Found**	**Not found**
ClinicalTrials.gov	337 (37.7%)	558 (62.3%)	15,380 (2.6%)	574,885 (97.4%)
CTD with inferred associations	511 (57.1%)	384 (42.9%)	92,408 (15.7%)	497,857 (84.3%)

Comparative Toxicogenomics Database or CTD is a large database that contains a wide variety of relations such as chemical-gene, disease-gene, and chemical-disease relations. From CTD, the newer version of chemical-disease associations (released on March 29, 2020) was downloaded to examine whether the proposed method can identify new associations recently reported in CTD or not. According to those CTD data, 15 therapeutic drug-disease relations have been recently recorded in CTD with supporting literature. Out of those 15 associations, the proposed method can discover four drug-disease associations, which are the benoxaprofen (DB04812) - psoriasis 6 (OMIM: 605364), clonazepam (DB01068) - epilepsy (OMIM: 600131), pioglitazone (DB01132) - anxiety (OMIM: 607834), and resveratrol (DB02709) - autoimmune disease (OMIM: 109100) associations.

In CTD, there also are inferred chemical-disease relations made by using the curated chemical-gene and gene-disease relations. The process to infer chemical-disease relations of CTD has been cautiously conducted, and only chemical-gene and gene-disease relations that have supporting literature have been used. Based on these well-curated relations, a relation between a chemical and a disease can be inferred if they share some common genes. These inferred drug-disease associations in CTD were also exploited to verify the candidate drug-disease associations obtained by the proposed method. Consequently, 511 of 895 candidate drug-disease associations (57.1%) were found in CTD whereas 92,408 non-candidate relations (15.7%) were also inferred by CTD. From the one-sided Fisher's exact test, it can be summarized that the CTD drug-disease relations are significantly found in a group of candidate drug-disease associations but not in a group of non-candidate pairs with a p-value of 2.84×10^{-176}.

4.4.6 Case studies

In this section, some cases of the discovered drug-disease associations are selected for literature investigation to ensure the practicality of the proposed method. Among the discovered drug-disease associations, all novel drugs recommended for esophageal cancer (OMIM: 133239) are selected for further discussion, as shown in Table 4.6.

Esophageal cancer is a type of cancer that occurs in the esophagus, a long tube that connects between the throat and the stomach. In 2020, it has been reported that esophageal cancer is at the sixth rank of cancer causing deaths (544,000 deaths), and one of every 18 cancer deaths

is caused by this cancer [122]. Nevertheless, the causes of this disease still remain unclear, and no effective treatment is provided for this type of cancer [123]. By the proposed method, seven new drugs were discovered for esophageal cancer (Table 4.6). For each drug, it was searched for supporting information from CTD, ClinicalTrials.gov, and literature. In CTD, drug-disease relations can be categorized in two main groups, which are those reported with supporting literature (therapeutic) and those predicted based on overlapping genes between the curated chemical-gene and gene-disease relations (inferred). If a drug is investigated in clinical studies for esophageal cancer, it will be labeled as "Found". A drug with lack of supporting information from CTD or ClinicalTrials.gov is denoted as "NA".

Table 4.6 The list of new drugs proposed for esophageal cancer

DrugBank ID	**Drug name**	**Supporting evidence**	
		CTD (Therapeutic/ Inferred/ NA)	**ClinicalTrials.gov (Found/ NA)**
DB00441	Gemcitabine	Inferred	Found
DB00482	Celecoxib	Inferred	Found
DB01041	Thalidomide	Inferred	Found
DB01234	Dexamethasone	Inferred	Found
DB00635	Prednisone	NA	Found
DB11672	Curcumin	Inferred	NA
DB00541	Vincristine	NA	NA

According to Table 4.6, four drugs were investigated for the treatment of esophageal cancer in clinical studies and inferred by CTD. Gemcitabine (DB00441) is approved for the treatment of various types of cancer, such as ovarian and non-small cell lung cancer. In ClinicalTrials.gov, many clinical studies investigated the combination use of gemcitabine with other drugs and radiation therapy in patients with esophageal cancer, such as the clinical studies ID NCT00759226 and NCT00012363. Recently, Yang et al. [124] conducted a meta-analysis of randomized clinical trials and revealed that gemcitabine-based combination therapy can improve response rate and disease control rate in patients with esophageal cancer. Celecoxib (DB00482) is

a drug that can relieve inflammation and is approved for the treatment of osteoarthritis and rheumatoid arthritis. In ClinicalTrials.gov, many clinical studies focused on using multiple drugs, including celecoxib, for the treatment of esophageal cancer. For example, the clinical study NCT00520091 was conducted to observe the use of celecoxib with irinotecan, cisplatin, and radiation therapy in patients with esophageal cancer. Despite unclear mechanisms of actions of celecoxib, it has been recently found that celecoxib can reduce zinc deficiency, a risk of esophageal cancer, resulting in suppressing tumorigenesis [125]. Thalidomide (DB01041) is currently used for the treatment of erythema nodosum leprosum. The clinical study of NCT01551641 investigated the down-regulated expression of vascular endothelial growth factor (VEGF) genes induced by thalidomide in esophageal cancer patients. Wang et al. [126] also revealed that the use of thalidomide with chemo-radiotherapy significant increases survival rates in esophageal cancer patients with high levels of serum VEGF. Dexamethasone (DB01234) has a wide range of indications, including anti-inflammation and immunosuppression. An example of clinical studies related to this drug is that with an ID of NCT01217060. This clinical trial studied the chemo-radiotherapy, including the use of dexamethasone, in patients with early-stage esophageal cancer before surgery. Many studies were conducted with the attempt to reveal molecular mechanisms of dexamethasone. A study of Yamawaki et al. [127] showed that this drug increases the extracellular secretion of cystatin C in esophageal cancer cells, leading to reduced cancer invasion and metastasis.

Prednisone (DB00635) is another drug that is indicated for anti-inflammation and immunosuppression in various diseases, such as allergic and skin disorders. In CTD, this drug is not inferred to be associated with esophageal cancer due to the lack of common genes between this disease and prednisone. However, this association can be discovered by the proposed method. A clinical study conducted by Shanghai Zhongshan Hospital (NCT03039608) investigated the combination treatment of triamcinolone and prednisone in patients with early esophageal neoplasm. It was shown by [128] that prednisone can prevent esophageal stricture, an abnormal narrowing of the esophageal in patients with esophageal cell carcinoma.

Curcumin (DB11672) is a natural compound which is currently indicated for several uses, such as reducing inflammation, improving cholesterol levels, and maintaining blood sugar levels. Although there is no a clinical study found for this association, it was provided as an

inferred drug-disease association in CTD. From a literature search, it was found that several derivatives of this compound have been developed for the treatment of various types of cancer, including esophageal cancer [129]. Furthermore, Subramaniam et al. [130] also revealed that curcumin can inhibit the growth of esophageal cancer cell lines by reducing Notch-1 activation, which is linked to tumorigenesis.

Vincristine (DB00541) is approved for particular types of cancer, such as acute lymphocytic leukemia and Hodgkin lymphoma. Interestingly, association between vincristine and esophageal cancer is not reported in both CTD and ClinicalTrials.gov, but it was discovered by the proposed method. According to a literature search, a recent finding revealed that the combination treatment by using the drugs vincristine and amlodipine can substantially decrease the viability of neuroblastoma cell lines [131].

4.5 Discussions

The three central compartments of the proposed method are the drug-GO-disease tripartite network, meta-path based functional profiles, and the PU bagging classifier. With these compartments, the proposed method significantly outperforms the existing methods and can efficiently identify the potential drug-disease associations. In the drug-GO-disease tripartite network, GO functions act as central indicators connecting between drugs and diseases, between drugs, and between diseases in the meta-paths. When compared to the methods that exploit proteins as intermediate nodes (i.e. EMP-SVD [11] and TS-SVD [47]), the proposed method produces the superior performance. In the drug-protein-disease network of EMP-SVD and TS-SVD, only a single protein node was employed to link between a drug and a disease node, between two drug nodes, and between two disease nodes. This may lead to loss of many relationships which have no overlapping proteins and loss of the capability of detecting many potential drug-disease associations. The broader information and multi-aspects of GO functions provide more meta-path based information and enable the proposed method to identify the greater numbers of the relationships in the drug-GO-disease tripartite network. However, some relationships between drug and disease nodes, between drug nodes, or between disease nodes which are linked by the semantically related GO functions of the different levels may still be undetectable by the meta-paths. To recover such relationships, the GO semantic similarity scores

(e.g. the Resnik [113] and Wang [114] measures) can be utilized to additionally create the weighted links between GO functions.

In addition to GO functions, the meta-path based functional profiles enable the proposed method to improve the predictions of the drug-disease associations. The meta-path based functional profiles enhance the ordinary path counts by incorporating information of the intermediate nodes (i.e. GO function nodes) along the meta-paths and creating as the functional profiles of each drug-disease pair. When compared to the method that directly uses path counts as the features of the drug-disease pairs (the baseline method), the proposed method significantly outperforms the baseline method. This indicates that the meta-path based functional profiles contain more useful information for classifying the drug-disease associations than the ordinary features. Furthermore, the proposed method with rich of the meta-path based information exploits only three meta-paths to create the superior performance, when compared to EMP-SVD that combines the network-based information from five meta-paths. Nevertheless, the values in the meta-path based functional profiles are derived from the path counts which may overly benefit the GO nodes with high degrees. To improve the meta-path based functional profiles, the meta-path based measures which were proposed to reduce the effect of high-degree nodes can be adopted, such as HeteSim [65] and DPRel [132].

Another benefiting compartment of the proposed method is the PU bagging model. Without reliable negatives, all unlabeled data out of the testing data sets are introduced into the training process of the proposed method to generate multiple bootstrap samples. This approach can advantageously utilize as many as possible data in hand whereas the two-step based methods (e.g. EMP-SVD and TS-SVD) discard a lot of unlabeled data that are not considered as reliable negatives. In EMP-SVD and TS-SVD, only the positive and reliable negative data were used to train the model, which may be overfit and could not practically identify the potential associations from the unlabeled drug-disease pairs.

In addition, each base classifier of the PU bagging model (i.e. XGBoost) can efficiently learn PU data and identify the positive drug-disease associations from the unlabeled pairs. This can be supported by the result of the preliminary study, which the different machine learning algorithms were compared to select the best one to serve as the base classifier of the PU bagging model. There are eight machine learning methods, including the multilayer perceptron (MLP),

logistic regression (LR), *k*-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), adaptive boosting (AdaBoost), gradient boosting tree (GBT), and XGBoost. The drug-disease association data of Gottlieb et al. [3] were used, and the features of each drug-disease pair were generated by integrating all meta-path based functional profiles. The nested 5-fold cross validation was performed to evaluate the performance of each method. The mean $F_{1,\,PU}$ values and their SDs are shown in Table 4.7. The one-sided *t*-tests were conducted to examine whether the mean $F_{1,\,PU}$ value of XGBoost is significantly greater that of each method or not.

Table 4.7 Performance comparison of the machine learning methods

Method	**Mean $F_{1,\,PU}$ and SD**	***P*-value**
MLP	1.091 ± 0.055	1.94×10^{-28}
LR	1.152 ± 0.047	1.22×10^{-26}
KNN	1.185 ± 0.042	1.51×10^{-21}
SVM	1.226 ± 0.041	7.38×10^{-16}
RF	1.275 ± 0.039	9.17×10^{-5}
AdaBoost	1.249 ± 0.040	7.02×10^{-13}
GBT	1.289 ± 0.042	0.0119
XGBoost	1.296 ± 0.044	-

According to Table 4.7, it is noticeably that XGBoost significantly outperforms the others with the mean $F_{1,\,PU}$ value of 1.296 at the significance level of 0.05. With the PU data, the ensemble models with the bagging technique (i.e. RF) and the boosting technique (i.e. AdaBoost, GBT, and XGBoost) perform better than the others. This may be because the ensemble models can reduce the variations of the individual classifiers caused by the unlabeled positive samples. However, too many reduced variations among the base classifiers does not benefit the PU bagging classifier. Thus, the boosting models are selected because they can reduce model bias, but not variances, of each base classifier. Among the boosting algorithms, GBT and XGBoost can produce the high mean values of $F_{1,\,PU}$. Due to many beneficial features of XGBoost (e.g. parallelization) and the best performance, XGBoost is exploited as the base classifiers in the proposed method. XGBoost may increase the variances among the individual classifiers whereas

the regularization of XGBoost could control the prediction errors of each individual classifier. This results in the less propagated errors from each base classifier and then the less total errors of the PU bagging model, when compared to other machine learning methods.

4.6 Summary

In this study, the PU learning method with meta-path based functional profiles is proposed for predicting drug-disease associations. Due to a proof of concept showing the feasibility of using GO functions for drug repositioning, GO functions are utilized as significant indicators for linking drugs to diseases in this method. With useful information of GO functions, only three association data are required for the proposed method (i.e. drug-GO, disease-GO, and drug-disease associations). Then, the drug-GO-disease tripartite network can be constructed using those association data. By taking advantages of meta-paths, the novel features of each drug-disease pair can be generated by differentiating paths from a drug to a disease according to GO functions and creating as the functional profiles. These profile features are called meta-path based functional profiles. Different from ordinary path count features, the proposed features can incorporate information of intermediate nodes (GO functions), which could be very useful in the classification of drug-disease associations. When compared to a method solely depending on path count features (the baseline method), it was found that the proposed method significantly outperforms the baseline method.

Due to the high dimensions of the meta-path based functional profiles, SVD was conducted to find their low-dimensional features. These features were exploited to develop the PU bagging classifier for recovering the positive drug-disease associations from the unlabeled drug-disease pairs. Unlike the existing PU learning methods, the proposed method does not require a set of reliable negatives, generally selected by heuristic strategies. In the proposed method, all unlabeled drug-disease pairs can be introduced into the training process. This enables the classification model to learn from numerous PU data, possibly leading to high generalization performance of the model. However, training the model with a set of unlabeled samples (containing data from both classes) could yield an unstable classifier. To stabilize classification models, the PU ensemble model or the PU bagging classifier is utilized in the proposed method. This method also takes advantages from the boosting technique by using the gradient boosting

method or XGBoost as the base classifier to decrease bias of each base classifier but increase variances among multiple classifiers. By combining both the bagging and boosting technique, the proposed method can significantly outperform several state-of-the-art methods. In addition, a large number of the candidate drug-disease associations with supporting evidence demonstrate an efficiency of the proposed method in discovering new drug-disease associations.

CHAPTER V

CONCLUSIONS AND FUTURE WORKS

In this chapter, overall works conducted in this dissertation are concluded, especially in terms of their contributions and limitations. Furthermore, a perspective of future works is also provided to suggest a direction for further improvement of this research.

5.1 Conclusions

In this research, a new perspective to conduct in silico methods for predicting drug-disease associations is introduced. Typically, drug-associated and disease-associated proteins are the first things that most researchers exploit for uncovering relationships between drugs and diseases. Herein, functional information (i.e. GO functions) about drugs and diseases, beyond the scope of drug-associated and disease-associated proteins, are utilized in a specific way that is different from existing methods.

At the beginning, the feasibility of utilizing GO functions for uncovering relationships between drugs and diseases was assessed. Drug-disease, drug-drug, and disease-disease similarity were investigated using protein and GO information about drugs and diseases. The classification of drug-disease, drug-drug, and disease-disease associations based on the protein-based and functionality-based similarity measures was conducted to examine how well GO information can be used to detect associations between drugs and diseases, between drugs, and between diseases. Drug-disease pairs were labeled as positive if they are known drug-disease associations, otherwise they were unlabeled. To bridge between drugs and diseases, a pair of two drugs was labeled according to how they share their associated diseases. If they share at least one common disease, then that pair is labeled as positive. If not, it is unlabeled. Similarly, a pair of two diseases was labeled as positive if they share at least one common drug, otherwise that pair was unlabeled. To define similarity measures, seven well-known similarity indices (i.e. the Jaccard, Braun-Blanquet, Simpson, Cosine, Sorgenfrei, McConnaughey, and derived Jaccard index) were utilized and compared. The derived Jaccard similarity index was selected as the most suitable one due to its best performance in the classification of drug-disease, drug-drug, and disease-disease associations. With the derived Jaccard similarity index, the performance measures of protein-

based similarity scores in the classification of all pair types were compared with those of functionality-based similarity scores. It was found that functionality-based similarity scores are better measures for identifying drug-disease, drug-drug, and disease-disease associations, because using GO functions can achieve higher values in all evaluation metrics, when compared to those obtained by using proteins.

The first study reveals that GO information about drugs and diseases is very significant and useful for identifying relationships between drugs and diseases, between drugs, and between diseases. With the broader information provided by GO functions, the larger amounts of potential drug-disease, drug-drug, and disease-disease associations could be probably detected, especially drugs and diseases that involve with one another via the more complex relationships than interacting with the same proteins. Although the improvements of classifying the drug-disease, drug-drug, and disease-disease associations are shown by using GO functions, solely using the functionality-based similarity scores is not enough to produce the high values in all evaluation metrics, particularly precision and F_1. In addition, the drug-disease associations are not directly inferred in the predictions of drug-drug and disease-disease associations. Therefore, it would be of great advantages if the functionality-based similarity information of drug-disease, drug-drug, and disease-disease pairs is integrated in a more complex model to identify more credible and accurate drug-disease associations.

According to the feasibility of functionality-based similarity information in the use for drug repositioning, another work of this research is the development of a novel computational method that advantageously exploits GO information for predicting drug-disease associations. Initially, the drug-GO-disease tripartite network was constructed by using only three data sets, which are drug-GO, disease-GO, and drug-disease associations. From the tripartite network, three meta-paths (drug-GO-disease, drug-GO-drug-disease, and drug-disease-GO-disease) were utilized to extract functionality-based similarity information between drugs and diseases, between drugs, and between diseases for each drug-disease pair. Based on each meta-path, this information was created as profiles of GO functions or novel meta-path based features, termed as meta-path based functional profiles. Unlike other existing meta-path based features, the proposed features can incorporate information of intermediate nodes along meta-paths (i.e. GO functions) by differentiating path instances under a meta-path according to GO function nodes included in the

paths. GO information provided by each meta-path is from different perspectives. Thus, an integration of all meta-path based functional profiles helps gaining more complete information about the drug-disease pairs. Before training a classification model, the low-dimensional representation features of the integrated meta-path based functional profiles were obtained by using SVD.

Due to the unavailability of negative samples or non-associated pairs of drugs and diseases, the positive-unlabeled (PU) learning approach was adopted in this study. Positive samples are known drug-disease associations whereas the remaining drug-disease pairs are unlabeled samples, which can be either positive or negative. Due to unlabeled positive samples, a binary classifier trained on positive samples and unlabeled subsamples may gain an unstable decision boundary. To take advantages of these classifiers' variances, a PU bagging classifier, an ensemble model where each base classifier trained on positive data and bootstrap samples of unlabeled data, is utilized in the proposed method. In addition to the bagging technique, the proposed method employs the boosting machine learning method, XGBoost, as a base classifier to learn training samples with noise in the unlabeled subsamples. With the PU ensemble model of the boosting base classifiers, the proposed method significantly outperforms state-of-the-art methods, including TL_HGBI, MBiRW, EMP-SVD, and TS-SVD. Moreover, meta-path based functional profiles are shown to be more useful features for predicting the drug-disease associations than ordinary path counts, which summarize all path instances under a meta-path by discarding information of the intermediate nodes.

With less required but important drug and disease information, the proposed method can produce the superior performance, when compared to other existing methods. By directly integrating GO information of drugs and diseases into the tripartite network, drug-drug and disease-disease similarity measures are not pre-computed for the proposed method, but it utilizes meta-paths for extracting relatedness information of each drug-disease pair from the network. The proposed method does not require a pre-determined set of reliable negative samples, which no one can guarantee that they are truly negative. A particular strategy for selecting negative samples from unlabeled samples could lead to sample selection bias. A selected set of reliable negative samples could be unrepresentative of all negative samples, resulting in low generalization ability of models.

Although many contributions of this research can be shown, there are also some limitations that should be noticed for further improvements. First, the current version of GO annotation data are still not complete. Because new knowledge about genes, gene products, and their functions are discovered every day, GO annotation data are regularly updated, and their new versions are then released. Since the proposed method is solely based on GO information about drugs and diseases, the incomplete GO annotation data could result that some potential drug-disease associations cannot be detected by the proposed method. Second, the proposed method is based on an ensemble model where each base classifier is also an ensemble model (i.e. XGBoost). This costs high computational time in the training process, especially when the number of bootstrap samples (T) is large. However, a multithreading implementation of the proposed method can help reducing its running time.

5.2 Future works

In this research, only GO information about drugs and diseases are used. In the future, other functional information (e.g. pathway information) about drugs and diseases can be utilized to bridge between drugs and diseases. Other valuable drug information (e.g. chemical structures and side effects) and disease information (e.g. phenotypic terms) can be integrated with GO information to improve the predictions of drug-disease associations. To enhance meta-path based functional profiles, other meta-path based similarity measures such as HeteSim can be adopted. In addition, network embedding and deep learning methods are also interesting to be utilized in the PU learning settings for predicting drug-disease associations.

REFERENCES

[1] B. S. Nelson, D. M. Kremer, and C. A. Lyssiotis, "New tricks for an old drug," *Nature Chemical Biology,* vol. 14, no. 11, pp. 990-991, 2018.

[2] K. J. Yella, S. Yaddanapudi, Y. Wang, and G. A. Jegga, "Changing trends in computational drug repositioning," *Pharmaceuticals,* vol. 11, no. 2, pp. 1-21, 2018.

[3] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Molecular Systems Biology,* vol. 7, no. 1, pp. 1-9, 2011.

[4] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinformatics,* vol. 30, no. 20, pp. 2923-2930, 2014.

[5] H. Luo *et al.*, "Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm," *Bioinformatics,* vol. 32, no. 17, pp. 2664-2671, 2016.

[6] X. Liang *et al.*, "LRSSL: Predict and interpret drug–disease associations based on data integration using sparse subspace learning," *Bioinformatics,* vol. 33, no. 8, pp. 1187-1196, 2017.

[7] W. Zhang *et al.*, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinformatics,* vol. 19, no. 1, pp. 1-12, 2018.

[8] R. Zhou, Z. Lu, H. Luo, J. Xiang, M. Zeng, and M. Li, "NEDD: a network embedding based method for predicting drug-disease associations," *BMC Bioinformatics,* vol. 21, no. 13, pp. 1-12, 2020.

[9] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "deepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics,* vol. 35, no. 24, pp. 5191-5198, 2019.

[10] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug–disease associations through layer attention graph convolutional network," *Briefings in Bioinformatics*, to be published. DOI: 10.1093/bib/bbaa243.2020.

[11] G. Wu, J. Liu, and X. Yue, "Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition," *BMC Bioinformatics,* vol. 20, pp. 2-13, 2019.

[12] W. Zhang *et al.*, "Predicting drug-disease associations based on the known association bipartite network," in *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Kansas City, MO, USA, 2017, pp. 503-509.

[13] Z. Li *et al.*, "Identification of drug-disease associations using information of molecular structures and clinical symptoms via deep convolutional neural network," *Frontiers in Chemistry,* Original Research vol. 7, no. 924, pp. 1-14, 2020.

[14] M. Lotfi Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, and J. R. Green, "A review of network-based approaches to drug repositioning," *Briefings in Bioinformatics,* vol. 19, no. 5, pp. 878-892, 2017.

[15] E. Sansone, F. G. B. De Natale, and Z. H. Zhou, "Efficient training for positive unlabeled learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 11, pp. 2584-2598, 2019.

[16] N. Gooneratne, "Overview of drug development," *Academic Entrepreneurship for Medical and Health Scientists,* vol. 1, no. 3, pp. 1-9, 2019.

[17] S. Sinha and D. Vohora, "Drug discovery and development: An overview," in *Pharmaceutical Medicine and Translational Clinical Research*, Boston, MA, USA: Academic Press, 2018, ch. 2, pp. 19-32.

[18] A. Deore, J. Dhumane, R. Wagh, and R. Sonawane, "The stages of drug discovery and development process," *Asian Journal of Pharmaceutical Research and Development,* vol. 7, no. 6, pp. 62-67, 2019.

[19] H. Xue, J. Li, H. Xie, and Y. Wang, "Review of drug repositioning approaches and resources," *International Journal of Biological Sciences,* vol. 14, no. 10, pp. 1232-1244, 2018.

[20] V. Parvathaneni and V. Gupta, "Utilizing drug repurposing against COVID-19 – Efficacy, limitations, and challenges," *Life Sciences,* vol. 259, p. 118275, 2020.

[21] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington, "Diagnosing the decline in pharmaceutical R&D efficiency," *Nature Reviews Drug Discovery,* vol. 11, no. 3, pp. 191-200, 2012.

[22] R. A. Hodos, B. A. Kidd, K. Shameer, B. P. Readhead, and J. T. Dudley, "In silico methods for drug repurposing and pharmacology," *WIREs Systems Biology and Medicine,* vol. 8, no. 3,

pp. 186-210, 2016.

[23] Y. Mei and B. Yang, "Rational application of drug promiscuity in medicinal chemistry," *Future Medicinal Chemistry,* vol. 10, no. 15, pp. 1835-1851, 2018.

[24] J. Langedijk, A. K. Mantel-Teeuwisse, D. S. Slijkerman, and M.-H. D. B. Schutjens, "Drug repositioning and repurposing: Terminology and definitions in literature," *Drug Discovery Today,* vol. 20, no. 8, pp. 1027-1034, 2015.

[25] H. S. Gns, S. Gr, M. Murahari, and M. Krishnamurthy, "An update on drug repurposing: Re-written saga of the drug's fate," *Biomedicine & Pharmacotherapy,* vol. 110, pp. 700-716, 2019.

[26] S. Pushpakom *et al.*, "Drug repurposing: progress, challenges and recommendations," *Nature Reviews Drug Discovery,* vol. 18, no. 1, pp. 41-58, 2019.

[27] J. S. Shim and J. O. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *International Journal of Biological Sciences,* vol. 10, no. 7, pp. 654-663, 2014.

[28] C. Abels and M. Soeberdt, "Can we teach old drugs new tricks?—Repurposing of neuropharmacological drugs for inflammatory skin diseases," *Experimental Dermatology,* vol. 28, no. 9, pp. 1002-1009, 2019.

[29] W. A. Chow, C. Jiang, and M. Guan, "Anti-HIV drugs for cancer therapeutics: back to the future?," *Lancet Oncology,* vol. 10, no. 1, pp. 67-71, 2009.

[30] 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, 2019.

[31] P. Sanseau *et al.*, "Use of genome-wide association studies for drug repositioning," *Nature Biotechnology,* vol. 30, no. 4, pp. 317-320, 2012.

[32] Y. Okada *et al.*, "Genetics of rheumatoid arthritis contributes to biology and drug discovery," *Nature,* vol. 506, no. 7488, pp. 376-381, 2014.

[33] J. T. Dudley *et al.*, "Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease," *Science Translational Medicine,* vol. 3, no. 96, p. 96ra76, 2011.

[34] M. Sirota *et al.*, "Discovery and preclinical validation of drug indications using compendia of public gene expression data," *Science Translational Medicine,* vol. 3, no. 96, p. 96ra77, 2011.

[35] S. Dakshanamurthy *et al.*, "Predicting new indications for approved drugs using a proteochemometric method," *Journal of Medicinal Chemistry,* vol. 55, no. 15, pp. 6832-6848, 2012.

[36] H. Xu *et al.*, "Validating drug repurposing signals using electronic health records: A case study of metformin associated with reduced cancer mortality," *Journal of the American Medical Informatics Association,* vol. 22, no. 1, pp. 179-191, 2014.

[37] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, X. Gao, and J. Li, "Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces," *BMC Bioinformatics,* vol. 20, no. 23, pp. 1-9, 2019.

[38] A. Chiang and A. Butte, "Systematic evaluation of drug–disease relationships to identify leads for novel drug uses," *Clinical Pharmacology & Therapeutics,* vol. 86, no. 5, pp. 507-510, 2009.

[39] W. Wang, S. Yang, and L. Jing, "Drug target predictions based on heterogeneous graph inference," *Pacific Symposium on Biocomputing,* vol. 1, no. 1, pp. 53-64, 2013.

[40] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data," *Artificial Intelligence in Medicine,* vol. 63, no. 1, pp. 41-49, 2015.

[41] F. Napolitano *et al.*, "Drug repositioning: A machine-learning approach through data integration," *Journal of Cheminformatics,* vol. 5, no. 1, pp. 1-9, 2013.

[42] S. Alaimo and A. Pulvirenti, "Network-based drug repositioning: Approaches, resources, and research directions," in *Computational Methods for Drug Repurposing*, New York, NY: Springer New York, 2019, pp. 97-113.

[43] P. Zhang, F. Wang, and J. Hu, "Towards drug repositioning: A unified computational framework for integrating multiple aspects of drug similarity and disease similarity," *AMIA Annual Symposium proceedings,* vol. 2014, pp. 1258-1267, 2014.

[44] T.-y. Fu, W.-C. Lee, and Z. Lei, "HIN2Vec: Explore meta-paths in heterogeneous information networks for representation learning," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, Singapore, 2017, pp. 1797–1806.

[45] Z. Tian, Z. Teng, S. Cheng, and M. Guo, "Computational drug repositioning using meta-path-based semantic network analysis," *BMC Systems Biology,* vol. 12, no. 9, pp. 123-134, 2018.

[46] E. Sansone, F. G. B. D. Natale, and Z. Zhou, "Efficient training for positive unlabeled learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 11, pp. 2584-2598, 2019.

[47] J. Liu, Z. Zuo, and G. Wu, "Link prediction only with interaction data and its application on drug repositioning," *IEEE Transactions on NanoBioscience,* vol. 19, no. 3, pp. 547-555, 2020.

[48] B. J. Samelson-Jones and V. R. Arruda, "Protein-engineered coagulation factors for hemophilia gene therapy," *Molecular Therapy - Methods & Clinical Development,* vol. 12, pp. 184-201, 2019.

[49] M. B. Hamaneh and Y.-K. Yu, "Mechanism-based disease similarity," *Journal of Rare Diseases Research & Treatment,* vol. 1, no. 3, pp. 1-4, 2016.

[50] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences,* vol. 104, no. 21, pp. 8685-8690, 2007.

[51] X. Zhang *et al.*, "The expanded human disease network combining protein–protein interaction information," *European Journal of Human Genetics,* vol. 19, no. 7, pp. 783-788, 2011.

[52] L. Huang, H. Luo, S. Li, F.-X. Wu, and J. Wang, "Drug-drug similarity measure and its applications," *Briefings in Bioinformatics*, to be published. DOI: 10.1093/bib/bbaa265.2020.

[53] M. A. Yildirim, K.-I. Goh, M. E. Cusick, A.-L. Barabási, and M. Vidal, "Drug-target network," *Nature Biotechnology,* vol. 25, no. 10, pp. 1119-1126, 2007.

[54] H. Huang, T. Nguyen, S. Ibrahim, S. Shantharam, Z. Yue, and J. Y. Chen, "DMAP: A connectivity map database to enable identification of novel drug repositioning candidates," *BMC Bioinformatics,* vol. 16, no. 13, pp. 1-11, 2015.

[55] A. P. Davis, C. G. Murphy, C. A. Saraceni-Richards, M. C. Rosenstein, T. C. Wiegers, and C. J. Mattingly, "Comparative Toxicogenomics Database: A knowledgebase and discovery tool for chemical–gene–disease networks," *Nucleic Acids Research,* vol. 37, no. D1, pp. D786-D792, 2008.

[56] L. Yu, B. Wang, X. Ma, and L. Gao, "The extraction of drug-disease correlations based on module distance in incomplete human interactome," *BMC Systems Biology,* vol. 10, no. 4, pp.

531-548, 2016.

[57] J. Sun, K. Zhu, W. J. Zheng, and H. Xu, "A comparative study of disease genes and drug targets in the human protein interactome," *BMC Bioinformatics,* vol. 16, no. 5, pp. 1-9, 2015.

[58] K. D. Rutherford, G. K. Mazandu, and N. J. Mulder, "A systems-level analysis of drug-target-disease associations for drug repositioning," *Briefings in Functional Genomics,* vol. 17, no. 1, pp. 34-41, 2018.

[59] D. P. Hill, B. Smith, M. S. McAndrews-Hill, and J. A. Blake, "Gene ontology annotations: What they mean and where they come from," *BMC Bioinformatics,* vol. 9, no. 5, pp. 1-9, 2008.

[60] S. Mathur and D. Dinakarpandian, "Drug repositioning using disease associated biological processes and network analysis of drug targets," *AMIA Annual Symposium proceedings,* vol. 2011, pp. 305-311, 2011.

[61] S. Li *et al.*, "Building the drug-GO function network to screen significant candidate drugs for myasthenia gravis," *PLoS ONE,* vol. 14, no. 4, pp. 1-17, 2019.

[62] A. Passi, N. K. Rajput, D. J. Wild, and A. Bhardwaj, "RepTB: A gene ontology based drug repurposing approach for tuberculosis," *Journal of Cheminformatics,* vol. 10, no. 1, pp. 1-12, 2018.

[63] C. Shi, Y. Li, J. Zhang, Y. Sun, and P. S. Yu, "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering,* vol. 29, no. 1, pp. 17-37, 2017.

[64] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, "Discovering meta-paths in large heterogeneous information networks," in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015, pp. 754–764.

[65] C. Shi, X. Kong, Y. Huang, P. S. Yu, and B. Wu, "HeteSim: A general framework for relevance measure in heterogeneous networks," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, no. 10, pp. 2479-2492, 2014.

[66] K. Yang, X. Zhao, D. Waxman, and X.-M. Zhao, "Predicting drug-disease associations with heterogeneous network embedding," *Chaos: An Interdisciplinary Journal of Nonlinear Science,* vol. 29, no. 12, pp. 1-9, 2019.

[67] Y. Dong, N. V. Chawla, and A. Swami, "metapath2vec: Scalable representation learning for

heterogeneous networks," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, 2017, pp. 135–144.

[68] X. Fu, J. Zhang, Z. Meng, and I. King, "MAGNN: Metapath aggregated graph neural network for heterogeneous graph embedding," in *Proceedings of The Web Conference 2020*, Taipei, Taiwan, 2020, pp. 2331–2341.

[69] Z. Noumir, P. Honeine, and C. Richard, "On simple one-class classification methods," in *Proceedings of IEEE International Symposium on Information Theory*, Cambridge, MA, USA, 2012, pp. 2022-2026.

[70] S. S. Deepika and T. V. Geetha, "A meta-learning framework using representation learning to predict drug-drug interaction," *Journal of Biomedical Informatics,* vol. 84, pp. 136-147, 2018.

[71] J. Bekker and J. Davis, "Learning from positive and unlabeled data: a survey," *Machine Learning,* vol. 109, no. 4, pp. 719-760, 2020.

[72] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, J. Yin, and J. Li, "Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases," *BMC Bioinformatics,* vol. 19, no. 19, pp. 49-59, 2018.

[73] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, X. Gao, and J. Li, "DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions," *BMC Bioinformatics,* vol. 20, no. 19, pp. 1-12, 2019.

[74] P. Yang, W. Liu, and J. Yang, "Positive unlabeled learning via wrapper-based adaptive sampling," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, Melbourne, Australia, 2017, pp. 3273-3279.

[75] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proceedings of the 3rd IEEE International Conference on Data Mining*, Melbourne, FL, USA, 2003, pp. 179-186.

[76] W. S. Lee and B. Liu, "Learning with positive and unlabeled examples using weighted logistic regression," in *Proceedings of the 20th International Conference on International Conference on Machine Learning*, Washington, DC, USA, 2003, pp. 448–455.

[77] F. Mordelet and J. P. Vert, "A bagging SVM to learn from positive and unlabeled examples," *Pattern Recognition Letters,* vol. 37, pp. 201-209, 2014.

[78] M. Singh, S. Bansal, S. Ahuja, R. K. Dubey, B. K. Panigrahi, and N. Dey, "Transfer

learning–based ensemble support vector machine model for automated COVID-19 detection using lung computerized tomography scan data," *Medical & Biological Engineering & Computing*, to be published. DOI: 10.1007/s11517-020-02299-2.2021.

[79] M. Claesen, F. De Smet, J. A. K. Suykens, and B. De Moor, "A robust ensemble approach to learn from positive and unlabeled data using SVM base models," *Neurocomputing,* vol. 160, pp. 73-84, 2015.

[80] P. Yang, S. J. Humphrey, D. E. James, Y. H. Yang, and R. Jothi, "Positive-unlabeled ensemble learning for kinase substrate prediction from dynamic phosphoproteomics data," *Bioinformatics,* vol. 32, no. 2, pp. 252-259, 2015.

[81] G. Wu, J. Liu, and W. Min, "Prediction of drug-disease treatment relations based on positive and unlabeled samples," *Journal of Intelligent and Fuzzy Systems,* vol. 35, no. 2, pp. 1363-1373, 2018.

[82] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 785–794.

[83] H. Li, Y. Cao, S. Li, J. Zhao, and Y. Sun, "XGBoost model and its application to personal credit evaluation," *IEEE Intelligent Systems,* vol. 35, no. 3, pp. 52-61, 2020.

[84] A. P. Davis *et al.*, "The Comparative Toxicogenomics Database: Update 2019," *Nucleic Acids Research,* vol. 47, no. D1, pp. D948-D954, 2019.

[85] D. S. Wishart *et al.*, "DrugBank 5.0: A major update to the DrugBank database for 2018," *Nucleic Acids Research,* vol. 46, no. D1, pp. D1074-D1082, 2018.

[86] J. Piñero *et al.*, "DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research,* vol. 45, no. D1, pp. D833-D839, 2016.

[87] R. P. Huntley *et al.*, "The GOA database: Gene Ontology Annotation updates for 2015," *Nucleic Acids Research,* vol. 43, no. D1, pp. D1057-D1063, 2015.

[88] D. Montaner and J. Dopazo, "Multidimensional gene set analysis of genomic data," *PLoS ONE,* vol. 5, no. 4, pp. 1-13, 2010.

[89] R. A. Fisher, "The logic of inductive inference," *Journal of the Royal Statistical Society,* vol. 98, no. 1, pp. 39-54, 1935.

[90] Y. Benjamini, R. Heller, and D. Yekutieli, "Selective inference in complex research," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 367, no. 1906, pp. 4255-4271, 2009.

[91] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE,* vol. 10, no. 3, pp. 1-21, 2015.

[92] W. J. Youden, "Index for rating diagnostic tests," *Cancer,* vol. 3, no. 1, pp. 32-35, 1950.

[93] M. D. Ruopp, N. J. Perkins, B. W. Whitcomb, and E. F. Schisterman, "Youden index and optimal cut-point estimated from observations affected by a lower limit of detection," *Biometrical Journal,* vol. 50, no. 3, pp. 419-430, 2008.

[94] R. Balakrishnan, M. A. Harris, R. Huntley, K. Van Auken, and J. M. Cherry, "A guide to best practices for Gene Ontology (GO) manual annotation," *Database,* vol. 2013, no. 1, pp. 1-18, 2013.

[95] D. Szklarczyk *et al.*, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Research,* vol. 47, no. D1, pp. D607-D613, 2018.

[96] J. Huang, C. Niu, C. D. Green, L. Yang, H. Mei, and J.-D. J. Han, "Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network," *PLOS Computational Biology,* vol. 9, no. 3, pp. 1-9, 2013.

[97] E. Guney, J. Menche, M. Vidal, and A.-L. Barábasi, "Network-based in silico drug efficacy screening," *Nature Communications,* vol. 7, no. 1, pp. 1-13, 2016.

[98] P. Wirapati *et al.*, "Meta-analysis of gene expression profiles in breast cancer: Toward a unified understanding of breast cancer subtyping and prognosis signatures," *Breast Cancer Research,* vol. 10, no. 4, pp. 1-11, 2008.

[99] P. Denny, M. Feuermann, D. P. Hill, R. C. Lovering, H. Plun-Favreau, and P. Roncaglia, "Exploring autophagy with gene ontology," *Autophagy,* vol. 14, no. 3, pp. 419-436, 2018.

[100] V. Consonni and R. Todeschini, "New similarity coefficients for binary data," *Match,* vol. 68, no. 2, pp. 581-592, 2012.

[101] S. H. Wijaya, F. M. Afendi, I. Batubara, L. K. Darusman, M. Altaf-Ul-Amin, and S. Kanaya, "Finding an appropriate equation to measure similarity between binary vectors: Case studies

on Indonesian and Japanese herbal medicines," *BMC Bioinformatics,* vol. 17, no. 1, pp. 1-19, 2016.

[102] A. P. Davis *et al.*, "Generating gene ontology-disease inferences to explore mechanisms of human disease at the Comparative Toxicogenomics Database," *PLoS ONE,* vol. 11, no. 5, pp. 1-19, 2016.

[103] J. Kuntsi, R. Pinto, T. S. Price, J. J. van der Meere, A. C. Frazier-Wood, and P. Asherson, "The separation of ADHD inattention and hyperactivity-impulsivity symptoms: Pathways from genetic effects to cognitive impairments and symptoms," *Journal of Abnormal Child Psychology,* vol. 42, no. 1, pp. 127-136, 2014.

[104] H. Sun, F. Yuan, X. Shen, G. Xiong, and J. Wu, "Role of COMT in ADHD: A systematic meta-analysis," *Molecular Neurobiology,* vol. 49, no. 1, pp. 251-261, 2014.

[105] M. J. Bonifácio, P. N. Palma, L. Almeida, and P. Soares-da-Silva, "Catechol-O-methyltransferase and its inhibitors in Parkinson's disease," *CNS Drug Reviews,* vol. 13, no. 3, pp. 352-379, 2007.

[106] D. H. Bonfanti *et al.*, "ATP-dependent potassium channels and type 2 diabetes mellitus," *Clinical Biochemistry,* vol. 48, no. 7, pp. 476-482, 2015.

[107] D. A. Jacobson and S. L. Shyng, "Ion channels of the islets in type 2 diabetes," *Journal of Molecular Biology,* vol. 432, no. 5, pp. 1326-1346, 2020.

[108] D. N. Cruz *et al.*, "Mutations in the Na-Cl cotransporter reduce blood pressure in humans," *Hypertension,* vol. 37, no. 6, pp. 1458-1464, 2001.

[109] M. Suwalsky, S. Mennickent, B. Norris, and H. Cardenas, "The antiepileptic drug carbamazepine affects sodium transport in toad epithelium," *Toxicology in Vitro,* vol. 20, no. 6, pp. 891-898, 2006.

[110] G. Graziani, C. Fedeli, L. Moroni, L. Cosmai, S. Badalamenti, and C. Ponticelli, "Gitelman syndrome: Pathophysiological and clinical aspects," *QJM: An International Journal of Medicine,* vol. 103, no. 10, pp. 741-748, 2010.

[111] Y. Wang *et al.*, "Therapeutic target database 2020: Enriched resource for facilitating research and early development of targeted therapeutics," *Nucleic Acids Research,* vol. 48, no. D1, pp. D1031-D1041, 2019.

[112] 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, 2014.

[113] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *Journal of Artificial Intelligence Research,* vol. 11, pp. 95-130, 1999.

[114] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics,* vol. 23, no. 10, pp. 1274-1281, 2007.

[115] C. Meng, R. Cheng, S. Maniu, P. Senellart, and W. Zhang, "Discovering meta-paths in large heterogeneous information networks," in *Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015, pp. 754-764.

[116] P. C. Hansen, "The truncatedSVD as a method for regularization," *BIT Numerical Mathematics,* vol. 27, no. 4, pp. 534-553, 1987.

[117] A. I. Kadhim, Y. Cheah, and N. H. Ahamed, "Text document preprocessing and dimension reduction techniques for text document clustering," in *Proceedings of 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology*, Kota Kinabalu, Malaysia, 2014, pp. 69-73.

[118] Z. A. Al-Saffar and T. Yildirim, "A hybrid approach based on multiple Eigenvalues selection (MES) for the automated grading of a brain tumor using MRI," *Computer Methods and Programs in Biomedicine,* vol. 201, pp. 1-10, 2021.

[119] Y. Gupta *et al.*, "Prediction and classification of Alzheimer's disease based on combined features from apolipoprotein-E genotype, cerebrospinal fluid, MR, and FDG-PET imaging biomarkers," *Frontiers in Computational Neuroscience,* vol. 13, no. 72, pp. 1-18, 2019.

[120] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods," *Briefings in Bioinformatics,* vol. 21, no. 4, pp. 1425-1436, 2019.

[121] P. Xuan, Y. Cao, T. Zhang, X. Wang, S. Pan, and T. Shen, "Drug repositioning through integration of prior knowledge and projections of drugs and diseases," *Bioinformatics,* vol. 35, no. 20, pp. 4108-4119, 2019.

[122] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*,

to be published. DOI: https://doi.org/10.3322/caac.21660.

[123] X. Hou, J. Wen, Z. Ren, and G. Zhang, "Non-coding RNAs: new biomarkers and therapeutic targets for esophageal cancer," *Oncotarget,* vol. 8, no. 26, pp. 43571-43578, 2017.

[124] J. Yang, X. Liang, Y. Zhai, X. Hu, Z. Jia, and X. Cheng, "Considering gemcitabine-based combination chemotherapy as a potential treatment for advanced oesophageal cancer: A meta-analysis of randomised trials," *International Journal of Clinical Practice,* vol. 74, no. 7, p. e13510, 2020.

[125] X. Yin, Y. Zhang, Y. Wen, Y. Yang, and H. Chen, "Celecoxib alleviates zinc deficiency-promoted colon tumorigenesis through suppressing inflammation," *Aging,* vol. 13, no. 6, pp. 8320-8334, 2021.

[126] J. Wang *et al.*, "Thalidomide combined with chemo-radiotherapy for treating esophageal cancer: A randomized controlled study," *Oncology letters,* vol. 18, no. 1, pp. 804-813, 2019.

[127] C. Yamawaki *et al.*, "Effect of dexamethasone on extracellular secretion of cystatin C in cancer cell lines," *Biomedical reports,* vol. 1, no. 1, pp. 115-118, 2013.

[128] G. Zhou *et al.*, "Efficacy of prednisone for prevention of esophageal stricture after endoscopic submucosal dissection for superficial esophageal squamous cell carcinoma," *Thoracic cancer,* vol. 8, no. 5, pp. 489-494, 2017.

[129] K. Komal, S. Chaudhary, P. Yadav, R. Pramanik, and M. Singh, "The therapeutic and preventive efficacy of curcumin and its derivatives in esophageal cancer," *Asian Pacific Journal of Cancer Prevention,* vol. 20, no. 5, pp. 1329-1337, 2019.

[130] D. Subramaniam *et al.*, "Curcumin induces cell death in esophageal cancer cells through modulating notch signaling," *PLoS ONE,* vol. 7, no. 2, p. e30590, 2012.

[131] A. Taghizadehghalehjoughi, S. Sezen, A. Hacimuftuoglu, and M. Güllüce, "Vincristine combination with Ca^{+2} channel blocker increase antitumor effects," *Molecular Biology Reports,* vol. 46, no. 2, pp. 2523-2528, 2019.

[132] M. Gupta, P. Kumar, and B. Bhasker, "DPRel: A meta-path based relevance measure for mining heterogeneous networks," *Information Systems Frontiers,* vol. 21, no. 5, pp. 979-995, 2019.

APPENDIX

This table contains some examples of potential drug-disease associations discovered by the proposed method. They are the first 100 drug-disease associations, out of 895 discovered drug-disease associations. Drugs are represented by DrugBank IDs and diseases are represented by Online Mendelian Inheritance in Man (OMIM) IDs. Supporting evidence of each drug-disease pair is searched from CTD and ClinicalTrials.gov. In CTD, "Therapeutic" drug-disease relations are reported with supporting literature whereas "Inferred" drug-disease relations are predicted from disease-gene and chemical-gene relations in CTD. "Found" refers to drug-disease associations under investigation in registered clinical studies of the database ClinicalTrials.gov. "NA" means that there is no supporting evidence for a drug-disease pair found in CTD or ClinicalTrials.gov. The full list of all discovered drug-disease associations is freely available online at http://ieee-dataport.org/3540.

No.	DrugBank ID	OMIM ID	Supporting evidence	
			CTD	ClinicalTrials.gov
1	DB01229	OMIM:276300	NA	NA
2	DB00531	OMIM:276300	Inferred	Found
3	DB04812	OMIM:605364	Therapeutic	NA
4	DB01234	OMIM:605027	Inferred	Found
5	DB02709	OMIM:155255	Inferred	NA
6	DB00544	OMIM:254500	Inferred	Found
7	DB00773	OMIM:114550	Inferred	Found
8	DB01169	OMIM:605027	Inferred	Found
9	DB13956	OMIM:102500	Inferred	NA
10	DB00959	OMIM:266600	Inferred	Found
11	DB01042	OMIM:276300	NA	NA
12	DB00457	OMIM:115000	Inferred	NA
13	DB01248	OMIM:605027	Inferred	Found
14	DB02709	OMIM:276300	Inferred	NA

No.	DrugBank ID	OMIM ID	Supporting evidence	
			CTD	ClinicalTrials.gov
15	DB00328	OMIM:276300	Inferred	NA
16	DB00264	OMIM:606799	Inferred	Found
17	DB00773	OMIM:137215	Inferred	Found
18	DB00317	OMIM:276300	Inferred	NA
19	DB01050	OMIM:180300	Inferred	Found
20	DB01162	OMIM:115000	NA	Found
21	DB06202	OMIM:102500	NA	NA
22	DB00883	OMIM:108725	Inferred	Found
23	DB00515	OMIM:601626	Inferred	NA
24	DB00571	OMIM:164230	NA	Found
25	DB00482	OMIM:601626	Inferred	Found
26	DB00675	OMIM:155255	Inferred	Found
27	DB00449	OMIM:608622	NA	Found
28	DB00747	OMIM:600116	Inferred	NA
29	DB01068	OMIM:600131	Therapeutic	NA
30	DB00650	OMIM:276300	NA	Found
31	DB00619	OMIM:601626	Inferred	Found
32	DB01181	OMIM:254500	Inferred	Found
33	DB00762	OMIM:114480	Inferred	Found
34	DB00619	OMIM:109800	Inferred	Found
35	DB00381	OMIM:115000	Inferred	Found
36	DB09110	OMIM:276300	NA	NA
37	DB00650	OMIM:607893	Inferred	Found
38	DB00794	OMIM:103780	NA	NA
39	DB00514	OMIM:147530	NA	NA
40	DB01262	OMIM:276300	Inferred	NA
41	DB00960	OMIM:115000	NA	NA

No.	DrugBank ID	OMIM ID	Supporting evidence	
			CTD	ClinicalTrials.gov
42	DB00541	OMIM:137215	Inferred	NA
43	DB00958	OMIM:114550	Inferred	Found
44	DB00313	OMIM:143465	Inferred	Found
45	DB09070	OMIM:102500	NA	NA
46	DB00945	OMIM:266600	Inferred	NA
47	DB01234	OMIM:155255	Inferred	Found
48	DB00905	OMIM:608622	NA	Found
49	DB01185	OMIM:176807	NA	NA
50	DB01204	OMIM:211980	Inferred	NA
51	DB00570	OMIM:605839	NA	NA
52	DB00373	OMIM:606799	NA	Found
53	DB00305	OMIM:276300	Inferred	NA
54	DB01204	OMIM:608935	Inferred	NA
55	DB00820	OMIM:608622	NA	Found
56	DB00188	OMIM:276300	Inferred	NA
57	DB00244	OMIM:180300	Inferred	Found
58	DB01136	OMIM:108725	Inferred	Found
59	DB08804	OMIM:102500	NA	NA
60	DB00445	OMIM:601626	Inferred	NA
61	DB00334	OMIM:115000	NA	NA
62	DB00367	OMIM:102500	NA	NA
63	DB00958	OMIM:254500	Inferred	Found
64	DB14490	OMIM:137215	NA	NA
65	DB00883	OMIM:606799	Inferred	Found
66	DB01234	OMIM:184700	Inferred	Found
67	DB00246	OMIM:143465	Inferred	Found
68	DB00317	OMIM:137215	Inferred	Found

No.	DrugBank ID	OMIM ID	Supporting evidence	
			CTD	ClinicalTrials.gov
69	DB00987	OMIM:109800	Inferred	Found
70	DB01203	OMIM:115000	NA	Found
71	DB00694	OMIM:109800	Inferred	Found
72	DB00287	OMIM:608622	NA	Found
73	DB00661	OMIM:108725	Inferred	Found
74	DB00290	OMIM:109800	Inferred	Found
75	DB00790	OMIM:108725	Inferred	NA
76	DB00290	OMIM:114480	Inferred	Found
77	DB00776	OMIM:607834	Inferred	Found
78	DB00502	OMIM:143465	Inferred	NA
79	DB00489	OMIM:606799	Inferred	Found
80	DB00722	OMIM:108725	Inferred	Found
81	DB13063	OMIM:601518	Inferred	NA
82	DB01065	OMIM:102500	NA	NA
83	DB01202	OMIM:267740	NA	NA
84	DB00309	OMIM:276300	NA	NA
85	DB00262	OMIM:137215	Inferred	NA
86	DB00255	OMIM:184700	Inferred	NA
87	DB01042	OMIM:601518	Inferred	NA
88	DB00328	OMIM:605027	Inferred	NA
89	DB00829	OMIM:608516	Inferred	Found
90	DB00571	OMIM:143465	Inferred	NA
91	DB00408	OMIM:143465	Inferred	NA
92	DB00502	OMIM:164230	Inferred	Found
93	DB00945	OMIM:106300	Inferred	Found
94	DB01169	OMIM:236000	Inferred	Found
95	DB00541	OMIM:114550	Inferred	Found

No.	DrugBank ID	OMIM ID	Supporting evidence	
			CTD	ClinicalTrials.gov
96	DB00694	OMIM:114480	Inferred	Found
97	DB01234	OMIM:109800	Inferred	Found
98	DB00396	OMIM:192000	NA	NA
99	DB00749	OMIM:180300	Inferred	Found
100	DB00958	OMIM:605027	Inferred	Found

VITA

NAME	Capt. Thitipong Kawichai
DATE OF BIRTH	16 April 1986
PLACE OF BIRTH	Lop Buri, Thailand
INSTITUTIONS ATTENDED	1. B.Sc. (Mathematics) with first class honors, Chiang Mai University, 2008 2. M.Sc. (Bioinformatics and Systems Biology), King Mongkut's University of Technology Thonburi, 2011
PUBLICATION	1. T. Kawichai, A. Suratanee, and K. Plaimas, "Functionality-based similarities for uncovering relationships between drugs and diseases," Science, Engineering and Health Studies, to be published. 2. T. Kawichai, A. Suratanee, and K. Plaimas, "Meta-path based gene ontology profiles for predicting drug-disease associations," IEEE Access, vol. 9, pp. 41809 - 41820, 2021. 3. P. Chuchouisuwan, T. Charoenrat, S. Chittapun, T. Kawichai, and N. Amnuaysin, "Prediction of repining stages of Hom Thong banana using partial least square regression," Thai Journal of Science and Technology, vol. 9, no. 2, pp. 287 - 297, 2020.
AWARD RECEIVED	1. Full scholarships under the Development and Promotion of Science and Technology Talents project (DPST) 2. Best session presentation awards of the 2nd Asia Joint Conference on Computing (AJCC2021), February 25th - 26th, 2021, Thailand