# ENHANCED HETEROGENEOUS NETWORK MODEL WITH ENSEMBLE SIMILARITIES FOR IDENTIFYING PROTEIN TARGETS OF DRUGS

Miss Piyanut Tangmanussukum



A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Applied Mathematics and Computational Science
Department of Mathematics and Computer Science
FACULTY OF SCIENCE
Chulalongkorn University
Academic Year 2021

# แบบจำลองโครงข่ายเฮเทโรจีเนียสแบบเพิ่มสมรรถนะด้วยการมารวมภาวะคล้ายหลายตัวสำหรับการระบุเป้าหมายโปรตีนของยา

น.ส.ปิยะนุช ตั้งมนัสสุขุม



วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์และวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

| | |
|---|---|
| Thesis Title | ENHANCED HETEROGENEOUS NETWORK MODEL WITH ENSEMBLE SIMILARITIES FOR IDENTIFYING PROTEIN TARGETS OF DRUGS |
| By | Miss Piyanut Tangmanussukum |
| Field of Study | Applied Mathematics and Computational Science |
| Thesis Advisor | Assistant Professor KITIPORN PLAIMAS, Ph.D. |

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Master of Science

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

THESIS COMMITTEE

..... Chairman
(Associate Professor KRUNG SINAPIROMSARAN, Ph.D.)

..... Thesis Advisor
(Assistant Professor KITIPORN PLAIMAS, Ph.D.)

..... Examiner
(Thap Panitanarak, Ph.D.)

..... External Examiner
(Associate Professor Treenut Saithong, Ph.D.)

ปิยะนุช ตั้งมนัสสุขุม : แบบจำลองโครงข่ายเฮเทโรจีเนียสแบบเพิ่มสมรรถนะด้วยการมารวมภาวะคล้ายหลายตัวสำหรับการระบุเป้าหมายโปรตีนของยา. ( ENHANCED HETEROGENEOUS NETWORK MODEL WITH ENSEMBLE SIMILARITIES FOR IDENTIFYING PROTEIN TARGETS OF DRUGS) อ.ที่ปรึกษาหลัก : ผศ. ดร.กิติพร พลายมาศ

ในปัจจุบัน การระบุเป้าหมายโปรตีนของยาเชิงคำนวณถูกใช้อย่างแพร่หลาย เพื่อช่วยประหยัดต้นทุนและเวลาในการค้นคว้าและพัฒนายา วิธีที่มีประสิทธิภาพมากที่สุดวิธีหนึ่งคือการทำนายเป้าหมายของยาตามคะแนนภาวะคล้ายระหว่างยาและเป้าหมายโปรตีน แม้จะมีข้อมูลต่างๆที่เกี่ยวข้องกับยาและเป้าหมายอย่างแพร่หลาย แต่โดยส่วนใหญ่แล้ว โครงสร้างทางเคมีและลำดับโปรตีนเท่านั้นที่มักจะถูกใช้เพื่อคำนวณคะแนนภาวะคล้ายของยากับยาและเป้าหมายกับเป้าหมายตามลำดับ ในวิทยานิพนธ์นี้ นำเสนอการรวมภาวะคล้ายโดยไปข้างหน้า (Forward similarity integration หรือ FSI) สำหรับการรวมการวัดภาวะคล้ายอย่างเป็นระบบ เพื่อสร้างแบบจำลองโครงข่ายเฮเทโรจีเนียสด้วยการรวมภาวะคล้ายที่เหมาะสมได้ โดยทำการศึกษาการวัดภาวะคล้ายของยาเจ็ดตัววัด การวัดภาวะคล้ายของเป้าหมายเก้าตัววัด และวิธีการรวมภาวะคล้ายทั้งหมดสี่วิธีในการรวมภาวะคล้ายโดยไปข้างหน้า เป็นผลให้ได้แบบจำลองโครงข่ายเฮเทโรจีเนียสที่เหมาะสมที่สุด จากการรวมการวัดภาวะคล้ายระหว่างยากับยาโดยใช้วิธีการหลอมรวมโครงข่ายภาวะคล้าย (Similarity network fusion หรือ SNF) และการวัดภาวะคล้ายของเป้าหมายกับเป้าหมายโดยใช้ลำดับของโปรตีน ทำให้แบบจำลองโดยการรวมภาวะคล้ายโดยไปข้างหน้านี้มีความแม่นยำถึง 99.8% และมีประสิทธิภาพที่มากกว่าแบบจำลองการรวมภาวะคล้ายแบบสุ่ม แบบเต็ม และแบบจำลองที่ไม่มีการรวมภาวะคล้าย นอกจากนี้ กรณีศึกษาของความสัมพันธ์ระหว่างยากับเป้าหมายที่ค้นพบใหม่ยังแสดงให้เห็นถึงการใช้งานจริงของวิธีการที่เสนอสำหรับการทำนายความสัมพันธ์ระหว่างยากับเป้าหมาย

| | | |
|---|---|---|
| สาขาวิชา | คณิตศาสตร์ประยุกต์และวิทยาการคณนา | ลายมือชื่อนิสิต ..... |
| ปีการศึกษา | 2564 | ลายมือชื่อ อ.ที่ปรึกษาหลัก ..... |

# # 6270065223 : MAJOR APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCE
KEYWORD:

Piyanut Tangmanussukum : ENHANCED HETEROGENEOUS NETWORK MODEL WITH ENSEMBLE SIMILARITIES FOR IDENTIFYING PROTEIN TARGETS OF DRUGS. Advisor: Asst. Prof. KITIPORN PLAIMAS, Ph.D.

Currently, computational identification of drug target proteins is widely used to help saving cost and time for drug discovery and development. One of the most efficient approaches is a prediction of drug-target interactions based on similarity scores between drugs and target proteins. Despite various data about drugs and targets extensively available, only chemical structures and protein sequences are mostly used to compute drug-drug and target-target similarity scores, respectively. In this thesis, the Forward similarity integration (FSI) Framework is proposed for systematically integrating multiple similarity measures to construct a heterogeneous network propagation model with a suitable similarity integration. Seven drug-similarity measures, nine target-similarity measures, and four similarity integration methods were formulated and used in the FSI framework. Thus, the suitable heterogeneous network model combines three drug-similarity measures integrated by using similarity network fusion (SNF) and one target-similarity measure based on protein sequences. The model selected by FSI reached an accuracy of 99.8% and significantly outperformed the models with random, full similarity integration, and the models without similarity integration. Also, the case studies of newly discovered drug-target interactions demonstrate the practicality of the proposed method for drug-target interaction prediction.



| | | |
|---|---|---|
| Field of Study: | Applied Mathematics and Computational Science | Student's Signature ..... |
| Academic Year: | 2021 | Advisor's Signature ..... |

# ACKNOWLEDGEMENTS

This thesis would not have been fully completed if there is no kind support and help from many people during the conduction of my thesis.

First, I am very thankful to my thesis advisor, Assistant Professor Dr. Kitiporn Plaimas, for her constant understanding, support, and precious guidance. In addition, I am grateful to Assistant Professor Dr. Apichat Suratanee, for giving helpful advices and comments. Also, I would like to acknowledge all thesis committees consisting of Associate Professor Dr. Krung Sinapiromsaran, Dr. Thap Panitanarak, and Associate Professor Dr. Treenut Saithong, for their useful comments and suggestions on my thesis.

Furthermore, I especially would like to thank Capt. Dr. Thitipong Kawichai for patience, understanding, valuable suggestions, and encouragement from the beginning to the end of my study. Additionally, I would like to thank all members in my research group and friends in the AMCS for their discussions and their kind supports.

I would like to state my special thanks to my parents and family for their constant love, supports, and believing in me. Importantly, I would like to thank to P'Kaew, my special sister for her encouragement and inspiration through the path of the Master's degree. Moreover, I would like to thank all friends for their encouragement and being always there with me.

Finally, I would like to express my special thanks of gratitude to the Development and Promotion of Science and Technology Talents project (DPST) for financial support during my studies.

Piyanut Tangmanussukum

# TABLE OF CONTENTS

Page

# LIST OF TABLES

**Page**

# LIST OF FIGURES

**Page**

# CHAPTER I
# INTRODUCTION

## 1.1 Background and rationales

Identifying new interactions between drugs and target proteins is of great importance for discovering and developing a new drug or a novel target for drugs. There are several wet-lab techniques to infer drug-target interactions (DTIs), such as biochemical affinity purification, genetic modifications [1], and *in vitro* bioassay systems [2]. However, the discovery of new DTIs through wet labs is a complex process expending a lot of time and costs. To increase the potential of identifying new DTIs, computational inference methods were introduced for more efficiently discovering a plenty of new associations between drugs and targets in a shorter time when compared to the experimental labs.

The computational methods for identifying DTIs can be categorized into three main groups [3], which are ligand-based methods, molecular docking methods, and chemical genomic methods. Ligand-based methods identify promising DTIs by calculating structural similarity between the ligands [4]. However, the prediction results from ligand-based methods are rather sensitive, especially when the ratio of known ligands per protein is low [5]. Molecular docking methods require 3D structures of drugs and proteins to simulate and identify structural interactions between drugs and proteins [6-8]. Nevertheless, not every protein have known 3D structures, resulting that molecular docking cannot be implemented for those proteins [9]. Chemical genomic methods apply various data from both drugs and target proteins for discovering DTIs [3]. This kind of methods is more flexible than the first two categories [10], since there are a variety of choices to differently use biological data, which are currently publicly available in many databases.

One of the most widely used technique in chemical genomic methods is the similarity-based technique, which predicts DTIs from the similarity scores between drugs and between targets [11]. For example, Bleakley et al. [12] designed a method known as the Bipartite Local Model (BLM) for predicting new DTIs using similarity information based on chemical and genomic data. Liu et al. [13] created a new matrix factorization approach for the DTI prediction, namely the Neighborhood Regularized Logistic Matrix Factorization (NRLMF), which calculates the drug similarity scores based on chemical structures of drugs and computes the target similarity scores based on amino acid sequences of target proteins. Wang et al. [14] proposed the Heterogeneous Graph Based Inference (HGBI), which constructs a heterogeneous drug-target network applying similarity scores based on drug chemical structures and protein sequences, to predict new DTIs. Liu et al. [15] the presented the Weighted k-Nearest Neighbor with Interaction Recovery (WkNNIR), which employs neighborhood-based similarity functions for the DTI prediction, and measured the drug similarity scores and target similarity scores based on chemical structures and amino acid sequences of proteins, respectively. Wan et al. proposed [16] NEural integration of neighbOr information for the DTI prediction (NeoDTI), a deep learning-based method combining various information from eight different sources (such as drug side effects, chemical structures of drugs, and protein sequences). According to the existing methods, it can be noticed that most of them focus on taking advantages of similarity scores solely based on drug chemical structures and protein sequences.



With the technology advances in generating and storing biological data, massive information about drugs and target proteins is publicly available in many databases, such as DrugBank [17], Comparative Toxicogenomics Database (CTD) [18], Kyoto Encyclopedia of Genes and Genomes (KEGG) [19], and Side Effect Resource (SIDER) [20]. To enhance the efficiency of the DTI prediction, many researchers utilized multiple similarity measures based on various data of drugs and targets. For example, Cheng et al. [21] combined the profiles of drugs, including the drug

similarities based on drug side effects, chemical structures, and drug indications by using the average, geometric mean, and the maximum function to identify new DTIs. Ding et al. [22] applied Hilbert–Schmidt Independence Criterion-based Multiple Kernel Learning (HSIC-MKL), which is a linear integration similarity method for merging different aspects of data. They utilized various data of drugs and drug targets, such as drug chemical structures, the network of drug-side effect associations, gaussian interaction profiles for drugs, sequence information of target proteins, functional information of targets, protein-protein interaction data, and gaussian interaction profiles for target proteins. In those existing methods, distinct sets of drug and target data were differently fused by various integration methods. To the suitable of my literature review, there is still not a framework for reasonably selecting suitable similarity integration by comparing among distinct similarity measures and similarity integration methods.

In this thesis, we propose the Forward Similarity Integration (FSI) framework to integrate multiple similarity measures of drugs and target proteins into a heterogeneous network propagation model for predicting promising links of DTIs. By systematically finding suitable similarity integration, this framework is developed with an aim to enhance the performance of a traditional heterogeneous network propagation model with a single similarity measure of drugs and a single similarity of target proteins. We firstly collected various data of drugs and drug targets to create different seven drug-drug similarity measures and nine target-target similarity measures. For the drug-drug similarity measures, structural, molecular interaction, and phenotypic data of drugs were used. For the target-target similarity measures, genomic, molecular interaction, and functional data of target proteins were utilized. To combine those multiple similarity measures, we considered four integration functions including both linear and non-linear integration functions. Also, different performance measures are investigated for serving as criterion of the forward selection of the drug and target similarity measures to integrate into a heterogeneous

network propagation model. The superior performance of the heterogeneous network model obtained from FSI was demonstrated by comparing with those of other models, including the conventional heterogeneous network propagation model and the models with full and random similarity integration. Finally, the suitable model obtained from FSI was used to predict new promising DTIs, and then the predictions were validated by searching for supporting evidence to demonstrate the practicality of the proposed framework.

### 1.2 Research objectives

1. To propose an enhanced heterogeneous network model by integrating multiple similarity measures of drugs and target proteins for predicting DTIs

2. To introduce the Forward Similarity Integration (FSI) framework for systematically selecting suitable similarity integration

3. To apply the improved model with the proposed framework for discovering potential interactions between drugs and target proteins



### 1.3 Expected outcomes

To predict DTIs, this research proposes the enhanced heterogeneous network model with the framework for systematically selecting suitable similarity integration from various drug-drug and target-target similarity measures and similarity integration methods. The sets of the considered similarity measures and integration methods different from those used in this thesis can be included in the proposed framework. In addition, the proposed framework can be deployed to preliminarily screen for potential drug-target interactions, which would be useful information for further discovering new drug target proteins, developing novel drugs, inferring drug side effects, and exploring drug repositioning.

# CHAPTER II

# BACKGROUND KNOWLEDGE AND RELATED WORKS

To predict DTIs, this research proposes the enhanced heterogeneous network model with the framework for systematically selecting suitable similarity integration from various drug-drug and target-target similarity measures and similarity integration methods. The sets of the considered similarity measures and integration methods different from those used in this thesis can be included in the proposed framework. In addition, the proposed framework can be deployed to preliminarily screen for potential drug-target interactions, which would be useful information for further discovering new drug target proteins, developing novel drugs, inferring drug side effects, and exploring drug repositioning.

## 2.1 Definitions and notations

The mathematical definitions are provided in this subsection. These definitions are the basic concepts mainly from graph theory [23-25].

**Definition 2.1.1** (Graphs). A network or graph is a pair $G = (V, E)$, where $V$ is a set of $n$ nodes or vertices and $E$ is a set of edges linking between nodes. Each edge $e_{ij} = (v_i, v_j)$ is associated with a weight $w_{ij} \geq 0$, which mostly represents the strength of the relationship between $v_i$ and $v_j$.

**Definition 2.1.2** (Bipartite graphs). A bipartite graph is a graph whose vertices can be divided into two independent sets, $V$ and $U$ such that every edge $(u, v)$ either connects a vertex from $V$ to $U$ or a vertex from $U$ to $V$. We can also say that there is no edge that connects vertices from the same set.

**Definition 2.1.3** (Complete Graphs). A complete graph is a graph of which any two distinct vertices are adjacent.

**Definition 2.1.4** (Adjacency matrix). For a graph $G$ with a set of vertices, $V = \{v_1, \dots, v_n\}$, its adjacency matrix $A$ is a square $n \times n$ matrix such that its element $a_{ij}$ is one when there is an edge from vertex $v_i$ to vertex $v_j$, and zero when there is no edge.

**Definition 2.1.5** (Heterogeneous network). A heterogeneous network is defined as $G = (V, E)$ consisting of a set of node objects, $V$, and a set of edges, $E$, connecting the nodes in $V$. A heterogeneous network also has a node type mapping function, $\phi: V \to O$, and an edge type mapping function defined as $\xi: E \to R$ where $O$ and $R$ denote the set of node object types and edge types, respectively. If the total number of node types $|O| > 1$ or the total number of edge types $|R| > 1$, the network is called heterogeneous; otherwise, homogeneous.

A heterogeneous network is a network containing several different types of nodes and links. In general, a heterogeneous network is usually utilized for describing a complicated system. Some common examples of such systems in the real-world are internet and social networks [26], citation networks [27, 28], movie networks [29, 30], economic networks [31], and financial networks [32]. Figure 2.1 shows an example of heterogeneous networks which is a movie network consisting of three types of nodes (i.e., user, movie, and genre) and two types of links (i.e., user-movie links and movie-genre links). Each link type has its semantic annotation. For example, a link between a user node and a movie node means that the movie was watched by the user, and a link between a movie node and a genre node means that the movie is categorized in the genre.

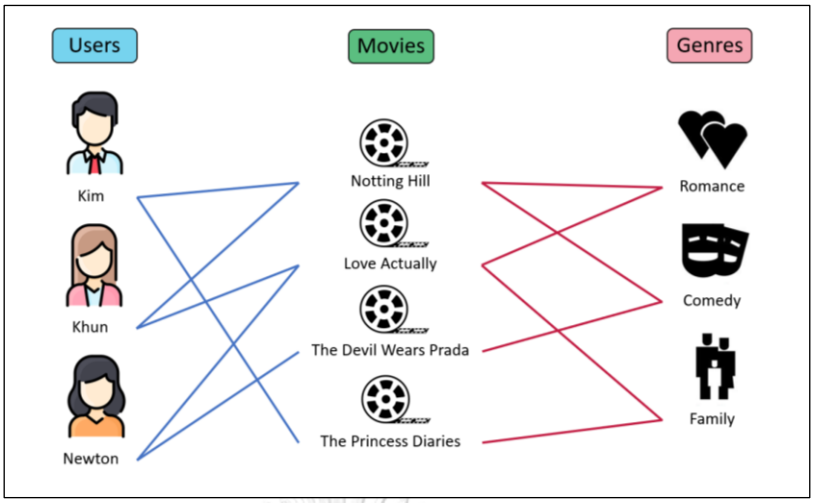



**Figure 2.1** An example of a heterogeneous movie network

## 2.2 Heterogeneous Network Propagation for DTIs

Over the past years, heterogeneous networks are formulated and utilized in many various applications, such as decision making in banking and finance [32], movie recommendation based on user interests [29, 30], product recommendation based on e-commerce search [31], drug repositioning [33, 34], product-rating networks [35], the medical insurance fraud identification [36], and the drug-target interaction (DTI) prediction [34, 37]. For identifying potential DTIs, several computational methods have been proposed mostly based on the simple version of drug-target heterogeneous networks, which combines a drug-target interaction network, a drug-drug similarity network, and a target-target similarity network. For instance, Chen et al. [38] developed Network based Random Walk with Restart on the Heterogeneous network (NRWRH) to discover missing DTI links. This method is based on the heterogenous network of DTIs, drug-drug similarity, and protein-protein similarity.

One of the most promising network-based methods is the Heterogeneous Graph Based Inference (HGBI) [14], which is based on the heterogeneous network consisting of drug-target interaction links, drug-drug similarity links, and target-target similarity links, as shown in Figure 2.2. In this network, there are two types of nodes

(i.e., drug and target nodes) and three types of links (i.e., links between drugs, between targets, and between drugs and targets). The links between drugs and the links between targets represent the degrees of drug-drug similarity based on chemical structures and target-target similarity based on protein sequences, respectively. Links between drug and target nodes represent known drug-target interactions (normal lines) and predicted drug-target interactions (dashed lines).

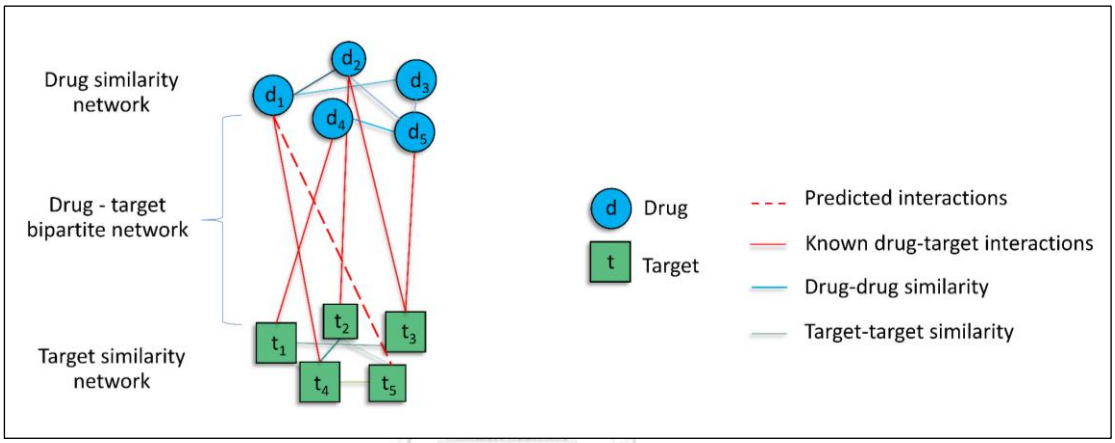



**Figure 2.2** A drug-target heterogeneous network

According to Figure 2.2, the heterogeneous network can be decomposed into three network layers, including the drug similarity network layer, the target similarity network layer, and the drug-target bipartite network. In the drug similarity network, we define the set of $m$ drug nodes as $D = \{d_1, d_2, \dots, d_m\}$ and the set of edges linking between any two drug nodes as $E_{dd}$. Each edge in the drug similarity network is represented with a weight indicating a drug-drug similarity score based on a particular drug property. The matrix containing the weights of all edges of this network is denoted as $W_{dd}$. In the target similarity network, we define the set of $n$ target protein nodes as $T = \{t_1, t_2, \dots, t_n\}$ and the set of edges in the target similarity network as $E_{tt}$. Based on a particular property of proteins, the weights of the edges of the target similarity network can be computed and contained in $W_{tt}$.

In the drug-target bipartite network, we denote $E_{dt}$ as the set of edges linking between drug and target nodes. $W_{dt}$ is defined as the matrix containing the weights of all edges of the drug-target bipartite network. Each element in $W_{dt}$ can be either one or zero, where one represents a known DTI, and zero represent an unknown DTI. Therefore, the drug-target heterogeneous network ($G_{DT}$) can be formulated as shown in Equation (2.1).

$$G_{DT} = \{\{D, T\}, \{E_{dd}, E_{tt}, E_{dt}\}, \{W_{dd}, W_{tt}, W_{dt}\}\} \tag{2.1}$$

The $G_{DT}$ network is an incomplete graph where some edges between drug and target nodes are missing. In this heterogeneous network, the weights of all networks are considered as matrices, i.e., $W_{dd} \in \mathbb{R}^{m \times m}$, $W_{tt} \in \mathbb{R}^{n \times n}$, and $W_{dt} \in \mathbb{R}^{m \times n}$, where $m$ and $n$ are the numbers of drugs and targets, respectively. To predict missing edges of DTIs, the algorithm of heterogeneous network propagation is used to iteratively update the weights in $W_{dt}$ as shown in Equation (2.2).

$$W_{dt}^{i+1} = \alpha W_{dd} \times W_{dt}^{i} \times W_{tt} + (1 - \alpha) W_{dt}^{0} \tag{2.2}$$

In Equation (2.2), $W_{dt}^{0}$ is the matrix of the initial weights of edges linking between drugs and targets. The parameter decay factor ($\alpha$) is in the range of 0 and 1. This parameter is used to determine how much the propagation of the network's weights affects the newly updated $W_{dt}$ when compared to the effects of $W_{dt}^{0}$. The formulation in Equation (2.2) will converge if $W_{dd}$ and $W_{tt}$ are properly normalized [14] as shown in Equation (2.3) and (2.4).

$$w(d_i, d_j) = \frac{w(d_i, d_j)}{\sqrt{\sum_{k=1}^{n} w(d_i, d_k) \sum_{k=1}^{n} w(d_k, d_j)}} \tag{2.3}$$

$$w(t_i, t_j) = \frac{w(t_i, t_j)}{\sqrt{\sum_{k=1}^{n} w(t_i, t_k) \sum_{k=1}^{n} w(t_k, t_j)}} \tag{2.4}$$

where $w(d_i, d_j)$ represents a weight of edge linking between drug $i$ and drug $j$, and $w(t_i, t_j)$ represents a weight of edge linking between target $i$ and target $j$.

**2.3 Similarity integration for the DTI prediction**

In order to identify new DTIs, the various of drug and target related data were considered and used in many new approaches of the DTI prediction. Under the concept that structurally similar drugs or targets are likely to interact with similar proteins or drugs [39, 40], the structures of drugs and targets have been of an interest for the DTI prediction. For example, Neighborhood Regularized Logistic Matrix Factorization (NRLMF) [13], Heterogeneous Graph Based Inference (HGBI) [14], Weighted k-Nearest Neighbor with Interaction Recovery (WkNNIR) [15], and Neural integration of neighbor information for the DTI prediction (NeoDTI) [16] utilized the chemical structure of drugs and the protein sequence of targets to compute similarity scores between drugs and between targets for the DTI prediction, respectively.

Nevertheless, drugs and target are diverse in many aspects and it's not easy to utilize a single similarity measure to accurately explain the relationship among drugs or targets [41]. Thus, several recent studies have attempted to improve the DTI prediction by combining multiple similarity measures from various data sources with chemical structures related drug data and protein sequences related target data. The similarity integration methods can be widely divided into two categories i.e., linear and non-linear integration functions. Table 2.1 and Table 2.2 shows some examples of similarity integration methods based on multiple drug-related data and target-related data, respectively.

**Table 2.1** Examples of the similarity integration method based on drug-related data

| Categories | Drug similarity types | Similarity integration methods | References |
|---|---|---|---|
| Linear integration | Chemical structures, Drug side effects, and Drug indications | Average, Maximum similarity value, and Geometric mean | [21] |
| | Chemical structure and Gaussian interaction profile (GIP) | Linear combination strategy | [42] |
| | Chemical structure, Drug side effects, and Gaussian interaction profile (GIP) | Average, Geometric mean, Maximum similarity value | [43] |
| Nonlinear integration | Chemical structure and drug side effect | The nonlinear fusion formula in Nonlinear integration Section | [45] |
| | Molecular fingerprints, side effect, ATC code, gene expression profile, drug-disease associations, pathways, and Gaussian interaction profile (GIP) | Similarity network fusion (SNF) | [46] |
| | Chemical structure and Gaussian interaction profile (GIP) | Similarity network fusion (SNF) | [47] |
| | Drug-structure, Drug side effect, Drug-protein interaction, Drug-drug interaction, Drug-disease association | Nonlinear end-to-end learning model (NeoDTI) | [16] |

**Table 2.2** Examples of the similarity integration method based on target-related data

| Categories | Target similarity types | Similarity integration methods | References |
|---|---|---|---|
| Linear integration | Amino-acid protein sequence, Gene Ontology annotations, Protein-protein interaction (PPI), Gaussian interaction profile (GIP) | Average, Geometric mean, Maximum similarity value, Similarity network fusion (SNF) | [43] |
| | Genomic sequence similarity (GS), Gene Ontology (GO) similarity, Protein-protein interaction (PPI) network similarity (PPI) | Multiple Similarities Collaborative Matrix Factorization (MSCMF) | [44] |
| Nonlinear integration | Gene Ontology (GO) terms, Protein–protein interaction (PPI) network, and Gaussian interaction profile (GIP) | Similarity network fusion (SNF) | [46] |
| | G protein-coupled receptors (GPCRs), kinase superfamily (Kinases), ion channels (ICs), nuclear receptors (NRs) | Nonlinear end-to-end learning model (NeoDTI) | [16] |



According to Table 2.1 and Table 2.2, it can be noticed that many drug and target data were considered to combine by several integration methods to predict DTIs. For example, chemical structures, side effects, drug indications, drug-drug interactions, drug-disease associations, Gaussian interaction profile (GIP), and Anatomical Therapeutic Chemical (ATC) are the drug-related data. Protein sequence, Gene Ontology annotations, protein-protein interaction (PPI), Gaussian interaction profile (GIP), Genomic sequence similarity (GS), G protein-coupled receptors (GPCRs), kinase superfamily (Kinases), ion channels (ICs), nuclear receptors (NRs) are the target-

related data. In addition, Table 2.1 and Table 2.2 were shown the similarity integration method that applied with those drug and target-related data, such as linear combination strategy, average function, maximum function, geometric mean, Multiple Similarities Collaborative Matrix Factorization (MSCMF), and SNF.

**2.4 Similarity network fusion (SNF)**

Similarity network fusion (SNF) [48] is a computational method for integration of multiple similarity data. SNF utilizes iterative non-linear approach and updates the global similarity network of each layer using a local *k*-nearest neighbors (KNN) approach. A similarity value between nodes is propagated to its *k*-nearest neighbors. In the beginning, SNF was proposed to combine the data of the DNA methylation, the mRNA expression, and the microRNA (miRNA) expression for the identification of cancer subtypes and predicting the survival rates of patients. After that, SNF was widely employed in more various applications, including multi-omics and microbiomes in respiratory diseases [49], the diagnosis of the liver cancer [50], the identification of specific biomolecular disturbances [51], high-risk bronchiectasis identification [52], and the classification of a chronic obstructive pulmonary disease[53].



In SNF, there are three steps to fuse multiple similarity data [49], including creation similarity networks based on particular datasets, fusion of multiple similarity networks, and analysis of the integrated networks (Figure 2.3).

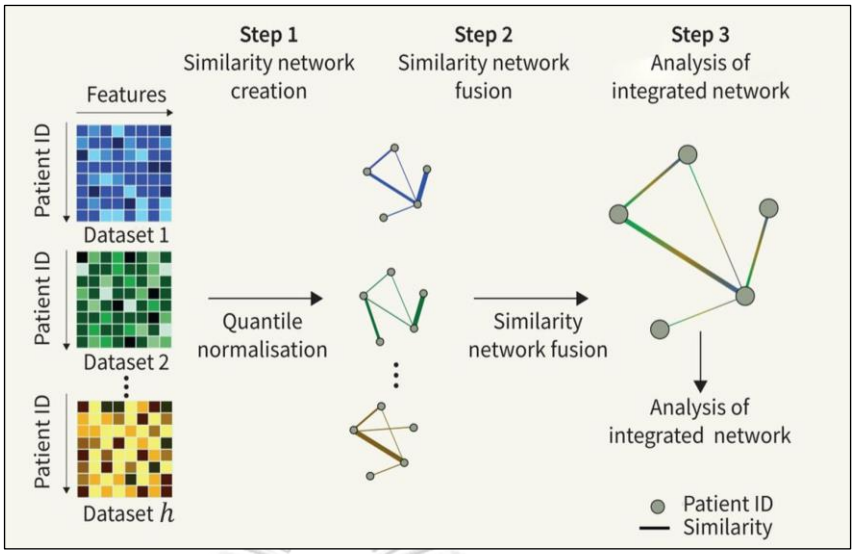



**Figure 2.3** Workflow of similarity network fusion method [49]

To combine $h$ similarity networks, the weights of edges in these similarity networks are considered as the similarity matrices $W^{(h)} \in \mathbb{R}^{m \times m}$, where $m$ be the number of samples or nodes. Initially, two similarity matrices $P_0^{(h)}$ and $S^{(h)}$ are defined. $P_0^{(h)}$ is the normalized version of similarity matrix $W^{(h)}$ which takes the similarity scores of all nodes into consideration, as shown in Equation (2.5). $S^{(h)}$ is another normalized matrix of $W^{(h)}$ where considers only the similarity scores of the *K* most similar samples for each sample, as shown in Equation (2.6), where $N_i$ is a set of nodes $i$'s k-nearest neighbors in $W^{(h)}$ matrices.

$$P_0^{(h)}(i,j) = \begin{cases} \dfrac{W^{(h)}(i,j)}{2\sum_{k \neq i} W^{(h)}(i,k)}, & j \neq i \\ \dfrac{1}{2}, & j = i \end{cases} \tag{2.5}$$

$$S^{(h)}(i,j) = \begin{cases} \dfrac{W^{(h)}(i,j)}{\sum_{k \in N_i} W^{(h)}(i,k)}, & j \in N_i \\ 0, & \text{otherwise} \end{cases} \tag{2.6}$$

Then, $P$ is iteratively updated using the values transferred between the nearest neighbors following to Equation (2.7), where $P_q^{(h)}$ represents the normalized similarity matrix for the $h^{th}$ data type at iteration $q$, and $H$ is number of data types.

Finally, the overall status matrix at iteration $t^{th}$ is calculated as shown in Equation (2.8).

$$P_{q+1}^{(h)} = S^{(h)} \frac{\sum_{k \neq h} P_q^{(h)}}{H-1} S^{(h)q} \tag{2.7}$$

$$P^{(t)} = \frac{\sum_{m \in H} P_q^{(h)}}{H} \tag{2.8}$$

## 2.5 Forward selection technique

Forward selection technique is a type of the stepwise regression [54] that constructs a step-by-step model by adding one variable (e.g., a similarity matrix) to improve the performance and finally obtains the suitable model, as shown example in Figure 2.4.

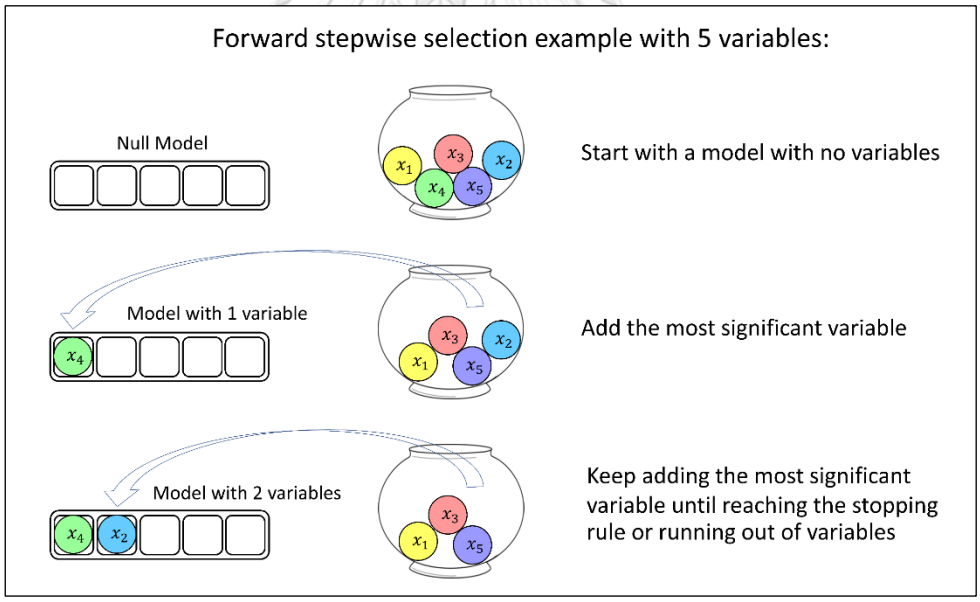



**Figure 2.4** Forward stepwise selection example with 5 variables [55]

According to Figure 2.4, the forward selection technique begins with a null model and add in variables one by one. Then, the model will keep one variable that is most significant variable or give the suitable model. Finally, the model stops adding variables when none of the remaining variables are significant or adding

variables can't improve the performance of the model. Note that once added, a variable is never removed.

To select the most efficient variables for constructing the model, many researchers take advantage of the forward selection technique, such as the financial distress prediction models [56], the bankruptcy prediction models [57], the predictive model for incisional surgical site infections [58], the predictive model for hypertension risk in the Chinese population [59], and selection of the suitable natural fiber for automotive component applications [60].

**2.6 Dependent samples *t*-test**

The *t*-test, known as *t*-statistic or *t*-distribution, is an inferential statistic method used to compare the means of two groups to determine if there is a significant difference. *T*-test can evaluate the difference of a mean value between a sample group and a known value (a one-sample *t*-test), the mean values between two independent sample groups (an independent two-sample *t*-test), and the mean values between two dependent sample groups (a paired or dependent sample *t*-test) [61]. In this section, we review the dependent sample *t*-test that is used to compare the sample means from two related groups, such as the comparison of the effects of a drug from the same patient group before and after taking the drug.

To apply the dependent sample *t*-test, the following assumptions are required to hold [62]:

1) The dependent variable is normally distributed.
2) The observations are sampled independently.
3) The dependent variable is measured on an incremental level, such as ratios or intervals.
4) The independent variables must consist of two related groups or matched pairs.

In the dependent sample *t*-test, there are two possible hypotheses. The first hypothesis is the null hypothesis ($H_0$) which states that there is no significant difference between the means of the two groups ($\mu_1 = \mu_2$). Another hypothesis is the alternative hypothesis ($H_1$) mentioning that the mean of the first group is greater than the mean of the second group ($\mu_1 > \mu_2$), the mean of the first group is less than the mean of the second group ($\mu_1 < \mu_2$), or the means are different ($\mu_1 \neq \mu_2$). The null and alternative hypotheses of those three cases can be mathematically formulated as shown in Equation (2.9) - (2.11). Note that the variances of two groups are not different ($\sigma_1^2 \neq \sigma_2^2$).

Case 1: $H_0 : \mu_1 = \mu_2,\ H_a : \mu_1 \neq \mu_2$ (2.9)

Case 2: $H_0 : \mu_1 = \mu_2,\ H_a : \mu_1 > \mu_2$ (2.10)

Case 3: $H_0 : \mu_1 = \mu_2,\ H_a : \mu_1 < \mu_2$ (2.11)

The test statistic $t$ is computed as $t = \frac{\sum D}{\sqrt{\frac{n\sum D^2 - (\sum D)^2}{n-1}}}$, where $D$ is the mean difference of each sample, $n$ is the sample size, and the degree of freedom ($df$) is equal to $n - 1$. The example data illustrating how to apply the dependent sample *t*-test is shown in Table 2.3. With this data, a teacher wants to know if the two exams are equally difficult [61]. The teacher set two exams from the same content and then has the same group of students to take both exams. The exam scores for all students are shown in Table 2.3.



**Table 2.3** Exam scores for each student

| Student | Scores of Exam 1 | Scores of Exam 2 | Difference Score ($D$) | $D^2$ |
|---|---|---|---|---|
| Bob | 63 | 69 | 6 | 36 |
| Nina | 65 | 65 | 0 | 0 |
| Tim | 56 | 62 | 6 | 36 |
| Kate | 100 | 91 | -9 | 81 |

**Table 2.3** Exam scores for each student (continued)

| Student | Scores of Exam 1 | Scores of Exam 2 | Difference Score ($D$) | $D^2$ |
|---|---|---|---|---|
| Alonzo | 88 | 78 | -10 | 100 |
| Jose | 83 | 87 | 4 | 16 |
| Nikhil | 77 | 79 | 2 | 4 |
| Julia | 92 | 88 | -4 | 16 |
| Tohru | 90 | 85 | -5 | 25 |
| Michael | 84 | 92 | 8 | 64 |
| Jean | 68 | 69 | 1 | 1 |
| Indra | 74 | 81 | 7 | 49 |
| Susan | 87 | 84 | -3 | 9 |
| Allen | 64 | 75 | 11 | 121 |
| Paul | 71 | 84 | 13 | 169 |
| Edwina | 88 | 82 | -6 | 36 |
| **summation** | - | - | 21 | 763 |

According to the teacher's purpose, the statistical hypotheses is preliminarily formulated as $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$ (Case 1). In Table 2.3, the summation of the difference scores between Exam 1 score and Exam 2 score ($\sum D$) and the summation of the squares of the difference scores ($\sum D^2$) are equal to 21 and 763, respectively. So, $t_{cal} = \frac{\sum D}{\sqrt{\frac{n\sum D^2 - (\sum D)^2}{n-1}}} = \frac{21}{\sqrt{\frac{(16)(763)-(21)^2}{16-1}}} = 0.75$, where the sample size is 16, and $df$=16-1=15. Based on the table of *t*-statistic, the value of *t* with $\alpha = 0.05$ and $df = 15$ is 2.131. Thus, the teacher will reject the null hypothesis $H_0$ if $t_{cal}$ > 2.131, otherwise the teacher will accept the null hypothesis. Since $t_{cal}$ = 0.750 < 2.131, the teacher cannot reject the null hypothesis $H_0$. This means that the mean scores are not different ($\mu_1 = \mu_2$), or the two exams are equally difficult.

# CHAPTER III

# MATERIALS AND METHODS

In this chapter, the data sets and the proposed methodology used in this thesis are provided. The methodology of the thesis includes data collection and preparation, the measurement of drug and target similarity, construction of the drug-target heterogeneous network, similarity integration methods, and the Forward Similarity Integration method (FSI) framework. The overview of the methodology is illustrated in Figure 3.1.

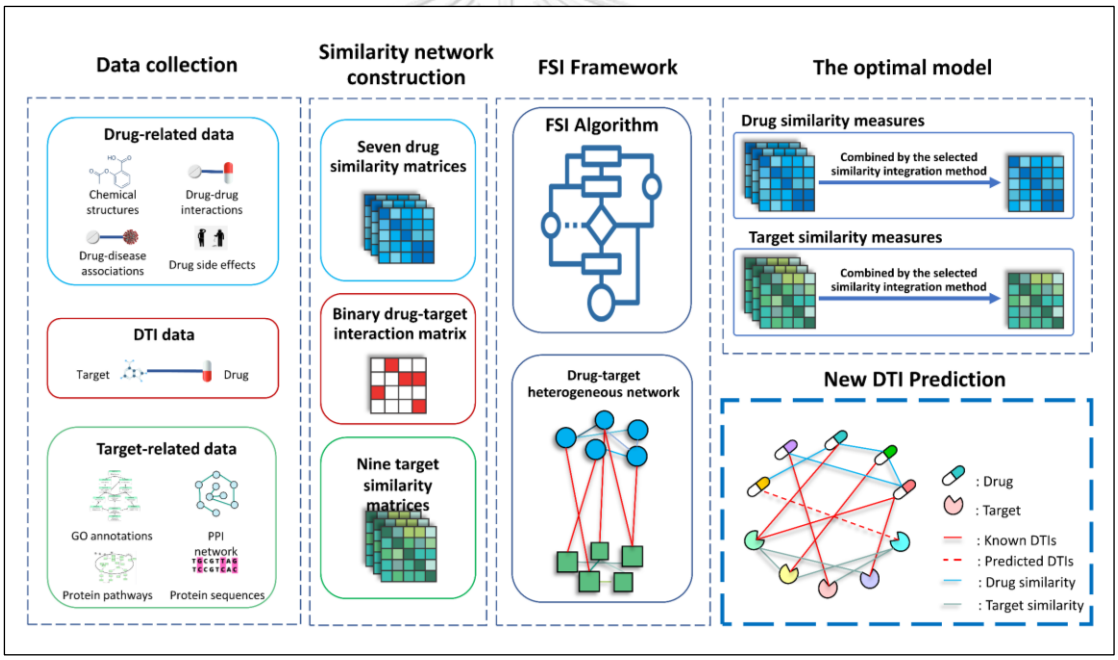



**Figure 3.1** The schematic diagram illustrating an overview of this thesis

According to Figure 3.1, the relevant information including drug related data (i.e., structural, molecular interaction, and phenotypic data of drugs), target related data (i.e., genomic, molecular interaction, and functional data of drug target proteins), and interaction data between drugs and targets were firstly collected from various databases. Next, the drug and target similarity matrices were constructed based on those collected data. Seven drug-drug similarity matrices and nine target-

target similarity matrices are introduced with the drug-target interaction matrix into the framework of Forward Similarity Integration (FSI) to obtain the most suitable heterogeneous network propagation model. In the FSI framework, several integration functions, both linear (i.e., average, maximum, and minimum) and non-linear integration functions (i.e., SNF), were considered. Furthermore, various performance measures (i.e., AUC, AUPR, and F1) were investigated for serving as the FSI criteria to select the similarity matrices fused into the model. After obtaining the FSI model, this model was deployed to predict promising links between drugs and target proteins.

## 3.1 Data collection and preparation

Data collection and preparation include the processes of collecting, cleaning, and manipulating the raw data prior to use for creating the drug-drug and target-target similarity matrices. In this work, there are three groups of the required data, including drug data sets, target protein data sets, and known interactions between drugs and target proteins.

### 3.1.1 Drug data sets

The drug data sets used in this thesis and the databases where these data are collected as shown in Table 3.1. To take advantages of multiple aspects of the drug data, various information about drugs (i.e., chemical, molecular interaction, and phenotypic information) is utilized. However, only drugs that have all those data are included in this work to avoid problems in data integration for creating an integrated version of drug-drug similarity measures.

**Table 3.1** Drug data sets and their sources

| Drug data sets | Data sources |
|---|---|
| Chemical structures | DrugBank [17] |
| Drug-drug interactions | DrugBank [17] |
| Drug-disease associations | CTD [18] |
| Drug side effects | SIDER [20] |

According to Table 3.1, the first data set about drugs is the chemical structures. The chemical structures of drugs in the form of the simplified molecular-input line-entry system (SMILES) [63] were downloaded from the DrugBank database (version 5.0) [17]. In this data set, there are 2,635 drugs in total with different DrugBank IDs. 183 drugs were initially removed due to lack of SMILES data. Furthermore, 1,590 drugs that have not all required drug data been also removed. Finally, there are 862 drugs of DrugBankIDs in total with the structural data available.

The second drug data set is about the molecular interactions about drugs or a data set of drug-drug interactions. Drug-drug interactions (DDIs) are the situations in which one drug affect the activity of another when they are used together. For example, one drug may delay, decrease, or enhance the absorption of other drugs [64]. The data of DDIs were downloaded from the DrugBank database (version 5.0) [17]. There are 4,294 drugs and 2,682,157 drug-drug interactions. However, 3,432 drugs were removed because they did not have all required drug data. Consequently, there are 862 drugs of DrugBank IDs and 924,819 drug-drug interactions remaining.

The third drug data set is about drug-disease associations (DDAs), a phenotypic property of drugs. DDAs are the events in which drugs exert the effect on diseases [65]. The data set of DDAs was extracted from Comparative Toxicogenomics Database (CTD) [18]. This data set contains 3,156 Chemical IDs, 2,425 Disease IDs, and 27,282 chemical-disease associations. To enable linking this data set to other drug data sets, all chemicals in CTD were mapped to the corresponding DrugBank IDs.

After the mapping and removing drugs with lack of some drug data, there are 862 drugs of DrugBank IDs, 2,287 diseases of Disease IDs, and 23,201 drug-disease associations remaining.

The fourth drug data set is about drug side effects (SEs), another phenotypic information of drugs. A drug side effect refers to an undesirable secondary effect which appears in addition to the purposed therapeutic effect of a drug. The data of SEs were received from the SIDE effect Resource (SIDER) database (version 4.1) [20]. In this data set, there are 1,430 drugs of STITCH IDs (Search Tool for Interactions of Chemicals [66]), 5,868 SE terms of UMLS IDs (Unified Medical Language System [67]), and 139,756 drug-SE associations. UMLS is a set of health and biomedical vocabularies compiled to promote and create the interoperable biomedical information systems and services [67]. To enable connecting this data set to other data sets, the drugs of STITCH IDs were mapped to their corresponding DrugBankIDs. As a result, there are 862 drugs of DrugBank IDs, 5,280 drug side effects, and 140,682 drug-SE associations remaining based on side effect related drug data.

3.1.2 Target data sets

The summary of all required data sets about target proteins and the databases where the data sets were downloaded is shown in Table 3.2.



**Table 3.2** Target data sets and their sources

| Target data sets | Data sources |
|---|---|
| Protein sequences | DrugBank [17] |
| GO annotations | GOA [68] |
| Protein-protein interactions | STRING [69] |
| Protein pathways | KEGG [19] |

The first data set of target proteins is about protein sequences, the orders of amino acids in a polypeptide chain [70]. The protein sequences of target proteins

were extracted from the DrugBank database (version 5.0) [17]. In total, there are 2,695 target proteins that have their own protein sequences. After that, 1,178 target proteins were removed due to lack of some required data. Consequently, there are 1,517 target proteins remaining with their available protein sequences.

The second data set of target proteins is about Gene Ontology (GO) annotations, providing functional information of target proteins. GO is a set of the hierarchically structured vocabularies which describe the cellular functions of genes and proteins. GO terms can be split into three domains: Biological Process (BP), Molecular Function (MF), and Cellular Component (CC) [71, 72]. All GO terms are structured in a hierarchical graph, where nodes represent the GO terms, and edges represent the relationships between the GO terms (Figure 3.2). Each edge in the GO graph is assigned a role in the parent-child relationships, such as “is a” and “a part of” relationships. An example of a GO graph is shown in Figure 3.2. In this example, BP term “menaquinone metabolic process” has one parent (i.e., “vitamin K metabolic process”). This means that “menaquinone metabolic process” is a “vitamin K metabolic process”.

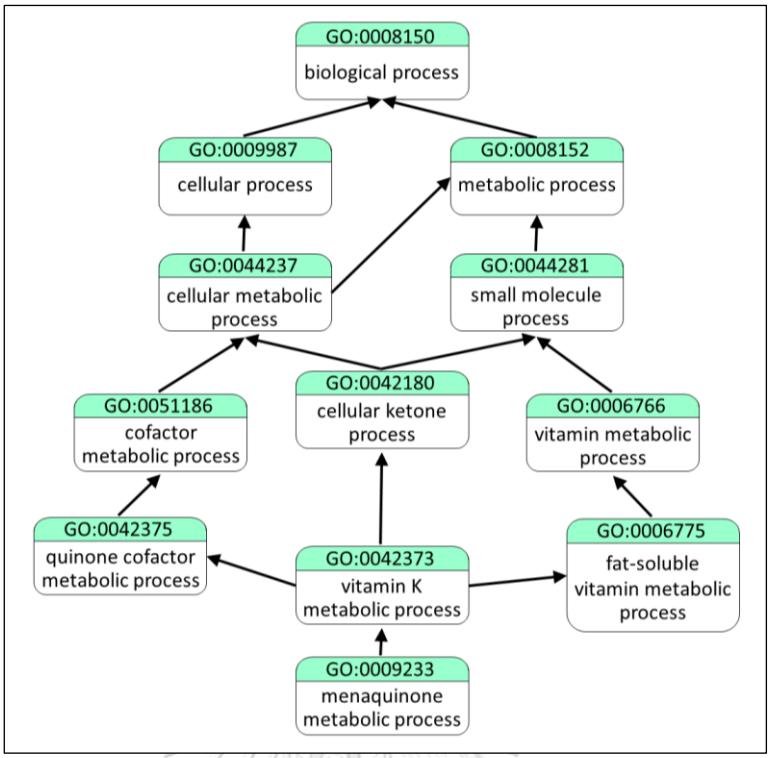



**Figure 3.2** An example of a GO graph

The GO annotation data of human proteins were obtained from Gene Ontology Annotation (GOA) database [68], In total, there are 19,755 UniProt IDs and 18,346 GO IDs. However, 18,238 proteins of UniProt IDs were removed due to lack of some protein data. Finally, there are 1,517 target proteins, 8,924 GO terms and 45,866 interactions remaining with some their annotated GO terms.

The third data set of target proteins is about protein-protein interactions (PPIs) in human. These are the molecular interactions between drug target proteins and other proteins, which can describe the physical or functional interactions between proteins [73]. The PPI network was downloaded from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (version 11.0) [69]. In this network, there are 19,354 proteins of STRING IDs and 11,759,454 PPIs. The STRING database provides the interactions between proteins with their confidence scores ranging from 150 to 999. To avoid false positive interactions, 540 proteins of STRING

IDs and 10,415,577 PPIs were removed because they have the confidence scores lower than 500. The remaining proteins of STRING IDs were mapped to their corresponding UniProt IDs. After the mapping and removing some proteins with lack of some required data, there are 1,517 proteins of UniProt IDs and 218,721 interactions remaining with their PPIs.

The fourth data set of target proteins is about protein pathways, the biological pathways in where proteins involve. This data set was received from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [19] database. Initially, there are 6,471 NCBI-Gene IDs, 283 protein pathways, and 21,008 interactions. To link these proteins of NCBI IDs to other protein data, these proteins were mapped to their corresponding UniProt IDs, and some proteins with lack of some required data of target proteins were removed. Consequently, there are 1,517 proteins, 283 protein pathways, and 7,988 interactions remaining with their protein pathway information.

3.1.3 Drug-target data set

This data set is a list of known drug-target interactions (DTIs), the connections between drugs and target proteins that can lead to some therapeutic effects of drugs [10]. The data of DTIs were downloaded from the DrugBank database (version 5.0) [17]. Initially, there are 2,452 drugs of DrugBank IDs, 2,695 target proteins, and 11,051 DTIs. Some data were removed because these proteins have not all required data. Finally, there are 862 drugs, 1,517 target proteins, and 3,583 drug-target interactions.

**3.2 Measurements of drug-drug and target-target similarities**

After obtaining all required data sets, the different drug-drug and target-target similarities were measured based on those collected data. In this thesis, seven drug-drug similarity measures and nine target-target similarity measures are formulated and investigated. In this section, how each drug-drug and target-target similarity measure is computed based on each data set is described.

3.2.1 Drug-drug similarity measures

Based on four drug data sets, different drug-drug similarity measures can be created. In this subsection, the method to compute each drug-drug similarity measure is described following to the drug data sets used.

- Similarity measurement based on the drug chemical structures

The similarity scores between drugs based on the chemical structures, abbreviated as Structures, are calculated from the structural data of drugs in the simplified molecular-input line-entry system (SMILES) [63]. By using the Chemical Development Kit (CDK) [74], those SMILES data of a drug can be encoded its structural information into a binary string of 2D chemical fingerprints of drugs. After that, the similarity score between two drugs can be calculated based on their binary strings by using the Tanimoto index [75].

- Similarity measurement based on DDIs

The DDI data can be represented in the form of matrix $DDI \in \mathbb{R}^{n_r \times n_r}$, where $n_r$ is the number of all drugs in the DDI network. Each element in this matrix can be either one (if DDI is present) or zero (if DDI is absent). The similarity measures based on DDIs are defined by using the Jaccard and the Cosine index, abbreviated as DDI_Jac and DDI_Cos, respectively. The Jaccard and Cosine index are the common techniques used to measure similarity between two objects and frequently applied to measure drug-drug similarity in many studies [76-79]. The Jaccard and Cosine similarity indices are generally defined in Equation (3.1) and (3.2), where $u$ and $v$ are binary vectors, and $\|\cdot\|$ represents the length of a binary vector. To compute DDI_Jac and DDI_Cos, $u$ is the binary vector of the interactions between drug $u$ to all other drugs, and $v$ is the binary vector of the interactions between drug $v$ to all other drugs. The similarity scores based on DDIs can indicate how much two drugs interact with the same drugs.

$$S_{Jaccard}(u, v) = \frac{u \cdot v}{\|u\|^2 + \|v\|^2 - u \cdot v} \tag{3.1}$$

$$S_{Cosine}(u, v) = \frac{u \cdot v}{\|u\| \cdot \|v\|} \quad (3.2)$$

- Similarity measurement based on DDAs

The DDA data can be represented in the form of matrix $DDA \in \mathbb{R}^{m \times n_d}$, where $m$ is the number of drugs, and $n_d$ is the number of all diseases in the DDA network. In $DDA$, each element can be one (if DDA is present) or zero (if DDA is absent). The Jaccard index (3.1) and Cosine index (3.2) are also used to compute the drug-drug similarity measures based on DDAs. These similarity measures are abbreviated as DDA_Jac for that uses the Jaccard index and DDA_Cos for that uses the Cosine index. For DDA_Jac and DDA_Cos, $u$ and $v$ are binary vectors with the length of $n_d$. The similarity scores based on DDAs can indicate that how much two drugs are associated with the same diseases.

- Similarity measurement based on drug side effects

The data of drug side effects can be represented in the form of matrix $SE \in \mathbb{R}^{m \times n_{se}}$, where $m$ is the number of drugs, and $n_{se}$ is the number of all side effect terms in the drug-side effect network. Similarly, each element in $SE$ can be either one (if drug side effect is present in that drug) or zero (if drug side effect is absent in that drug). Both the Jaccard and the Cosine index, shown in Equation (3.1) and (3.2) are utilized for computing the drug-drug similarity scores based on drug side effects. These similarity measures are defined as SE_Jac for that uses the Jaccard index and SE_Cos for that uses the Cosine index. To compute those similarity measures, $u$ and $v$ are binary vectors or two rows obtained from matrix $SE$. The similarity scores based on SEs can indicate how much two drugs have similar drug side effects.

In terms of drug similarities, there are seven drug-drug similarity measures defined in this thesis. The drug data sets, methods for similarity measurement, and

the abbreviations of the similarity measures are summarized in Table 3.3. The values of all drug-drug similarity measures are in the range of 0 and 1.

**Table 3.3** Seven drug-drug similarity measures and their abbreviations

| Types of data | Methods for similarity measurement | Defined abbreviations |
|---|---|---|
| Chemical structures | Tanimoto index | Structures |
| Drug-drug interactions | Jaccard index | DDI_Jac |
| | Cosine index | DDI_Cos |
| Drug-disease associations | Jaccard index | DDA_Jac |
| | Cosine index | DDA_Cos |
| Drug side effects | Jaccard index | SE_Jac |
| | Cosine index | SE_Cos |

3.2.2 Target-target similarity measures

Based on four data sets of target proteins, different target-target similarity measures can be defined and computed by using various methods for similarity measurement. The methods how to calculate the target-target similarity measures are described according to the data sets used.



- Similarity measurement based on protein sequences

To measure the similarity between two targets based on protein sequences, we applied the Smith-Waterman algorithm [80] and the Needleman-Wunsch algorithm [81]. The Smith-Waterman algorithm performs the local alignment between two amino acid sequences by comparing segments of all possible lengths. The similarity measure based on the local sequence alignments is abbreviated as Seq_Loc. The similarity scores between two strings of protein sequences were calculated and normalized to be in the range of 0 and 1 by the formula suggested in

[12], as shown in Equation (3.3). In this equation, $SW(t_i, t_j)$ is the Smith–Waterman score between amino acid sequences $t_i$ and amino acid sequences $t_j$.

$$S_{SW}(t_i, t_j) = \frac{SW(t_i, t_j)}{\sqrt{SW(t_i, t_i)}\sqrt{SW(t_j, t_j)}} \quad (3.3)$$

The Needleman-Wunsch algorithm serves the global alignment between two amino acid sequences by aligning all sequence segments from beginning to end. The abbreviation of the target-target similarity measure based on the global sequence alignment is Seq_Glo. The similarity between two strings of protein sequences were computed and normalized to be in the range of 0 and 1 by the formula suggested in [82], as shown in Equation (3.4). In this equation, $NW(t_i, t_j)$ is the Needleman-Wunsch score between amino acid sequences $t_i$ and amino acid sequences $t_j$.

$$S_{NW}(t_i, t_j) = \frac{NW(t_i, t_j) - \min(\min_{1 \le i \le n} NW(t_i, t_j), \min_{1 \le j \le m} NW(t_i, t_j))}{\max(\max_{1 \le i \le n} NW(t_i, t_j), \max_{1 \le j \le m} NW(t_i, t_j)) - \min(\max_{1 \le i \le n} NW(t_i, t_j), \max_{1 \le j \le m} NW(t_i, t_j))} \quad (3.4)$$

Both local and global alignment techniques were computed by using an R package named Biostrings version 3.15 [83] based on the BLOSUM62 substitution matrix with a gap opening of 10 and a gap extension of 0.5. These parameters are the same as the default values in the EMBOSS water tool option from the European Bioinformatics Institute.

- Similarity measurement based on GO annotations

To create the target-target similarity matrices based on the semantic similarity of GO annotations, we applied the Wang's method [84] and the Jiang's method [85] by using an R package named GoSemSim version 1.30.2 [86]. The output values of the methods are between 0 and 1. The target-target similarity measures using the Wang's method and Jiang's method are abbreviated as GO_Wang and GO_Jiang, respectively.

As the Jiang's method focuses on the property of nodes on the network, the semantic similarity is represented in the form of the Information Content (IC) [87]. The IC value for term $t$ is define as Equation (3.5), where $p(t)$ is the probability of the presence of GO term $t$ and its descendants in a certain corpus.

$$IC(t) = -\log(p(t)), \tag{3.5}$$

To calculate the semantic similarity between term $t_1$ and $t_2$, we firstly define the most information common ancestor (MICA) of $t_1$ and $t_2$, an ancestor term of both $t_1$ and $t_2$ that has the maximum IC among common ancestors of the terms. Then, the semantic similarity based on the Jiang's method can be calculated as shown in Equation (3.6).

$$sim_{Jiang}(t_1, t_2) = 1 - \min(1, IC(t_1) + IC(t_2) - 2 \times IC(\text{MICA})) \tag{3.6}$$

To measure the semantic similarity by using the Wang's method [84, 87-90], the directed acyclic graph (DAG) of GO term $A$ and its ancestors are represented as $DAG_A = (A, T_A, E_A)$, where $T_A$ is the set of GO terms including $A$ and its ancestors, and $E_A$ is the set of links among nodes of $T_A$ in $DAG_A$. Each GO term $t$ in $DAG_A$ has the semantic contribution (S-value) to target term $A$ or ($S_A(t)$), which was defined in Equation (3.7). In this formula, $w_e$ is the semantic contribution factor for edge $e \in E_A$ linking term $t$ with its child term $t'$. After that, the semantic value of GO term $A$ was defined in Equation (3.8).

$$S_A(t) = 1 \text{ if } t = A \text{ and } S_A(t) = \max_{t' \in \text{children of } t} w_e \times S_A(t') \text{ if } t \neq A, \tag{3.7}$$

$$SV(A) = \sum_{t \in T_A} (S_A(t)) \tag{3.8}$$

Formally, let $DAG_A = (A, T_A, E_A)$ of GO term $A$, and $DAG_B = (B, T_B, E_B)$ of GO term $B$. The semantic similarity between GO term $A$ and GO term $B$ by using the Wang's method is defined in Equation (3.9), where $S_A(t)$ and $S_B(t)$ are the S-values of GO term $t$ related to term $A$ and term $B$, respectively.

$$sim_{Wang}(A,B) = \frac{\sum_{t \in T_A \cap T_B}(S_A(t) + S_B(t))}{SV(A) + SV(B)} \quad (3.9)$$

- Similarity measurement based on PPI information

Based on PPI information, the similarity scores between target proteins are computed by using a PPI network. Three methods are used to define different similarity measures based on PPI information, including the methods of inverse shortest paths, the Jaccard index, and the Cosine Index.

To compute the similarity scores using the inverse shortest paths, abbreviated as PPI_ISP, Dijkstra's algorithm [91] was utilized to find the distances of the shortest paths linking between any two proteins in the PPI network. Then, the distances were transformed to the similarity scores using the formula described in [79], as shown in Equation (3.10). In this equation, $S(p,p')$ is a computed similarity value between two proteins, and $D(p,p')$ is the distance of the shortest path between those proteins in the PPI network. According to [79], $A$ and $b$ were selected to be 0.9 and 1, respectively.

$$S(p,p') = Ae^{-bD(p,p')} \quad (3.10)$$

In addition to PPI_ISP, the PPI information of each target protein is used to compute the similarity scores between target proteins by using the Jaccard and the Cosine index, as shown in Equation (3.1) and (3.2). These similarity measures are defined as PPI_Jac for that uses the Jaccard index and PPI_Cos for that uses the Cosine index. To compute both PPI_Jac and PPI_Cos, the adjacency matrix of the PPI network, $PPI \in \mathbb{R}^{n_p \times n_p}$, is used, where $n_p$ is the number of all proteins in the PPI network. According to Equation (3.1) and (3.2), $u$ and $v$ are binary vectors or two rows obtained from matrix $PPI$. Both PPI_Jac and PPI_Cos can indicate how much two target proteins have similar neighboring proteins or interact with similar proteins.

- Similarity measurement based on protein pathways

The data of protein pathways can be represented in the form of matrix $PW \in \mathbb{R}^{n \times n_{pw}}$, where $n$ is the number of target proteins, and $n_{pw}$ is the number of all pathways. In the matrix, each element can be either one (if a protein involves with a particular pathway) or zero (if a protein does not involve with a pathway). Denote that $u$ and $v$ are binary vectors or two rows obtained from matrix $PW$. The similarity scores between target protein $u$ and $v$ can be computed by using the Jaccard index (PW_Jac) and the Cosing index (PW_Cos). Both similarity scores can indicate how much two target proteins share their pathways or involve in the similar pathways.

In summary, nine target-target similarity measures can be defined based on four data sets of target proteins. The methods to measure similarity between targets for each similarity measure and the abbreviations of all target-target similarity measures are summarized in Table 3.4. The values of all target-target similarity measures are in the range of 0 and 1.

**Table 3.4** Nine target-target similarity measures and their abbreviations

| Types of data | Methods for similarity measurement | Defined abbreviations |
|---|---|---|
| Protein sequences | Smith-Waterman algorithm | Seq_Loc |
| | Needleman-Wunsch algorithm | Seq_Glo |
| GO annotations | GOSemsim: Wang method | GO_Wang |
| | GOSemsim: Jiang Method | GO_Jiang |
| Protein-protein interactions | Inverse shortest path similarity | PPI_ISP |
| | Jaccard index | PPI_Jac |
| | Cosine index | PPI_Cos |
| Protein pathways | Jaccard index | PW_Jac |
| | Cosine index | PW_Cos |

### 3.3 Construction of the drug-target heterogenous network

The drug-target heterogeneous network constructed in this work consists of a drug similarity network layer, a target similarity network layer, and a layer of a bipartite network of drug-target interactions. According to Figure 3.3, seven drug similarity measures and nine target similarity measures were defined based on four drug-related and four target-related data sets. These similarity measures can also be noticed as seven drug-drug similarity matrices and nine target-target similarity matrices, illustrated in Section 3.2. Some or all of them are integrated together to create an integrated drug-drug similarity matrix and an integrated target-target similarity matrix by using a particular similarity integration method (described in Section 3.4). For the selection of drug-drug and target-target similarity measures to be integrated, this research introduces the Forward Similarity Integration (FSI) framework (described in Section 3.5). As a result, the integrated drug similarity matrix, the integrated target similarity matrix, and the drug-target interaction matrix were integrated to construct the drug-target heterogeneous network. To predict undiscovered links between drugs and target proteins, the network propagation algorithm [14] was applied on the constructed heterogeneous network.

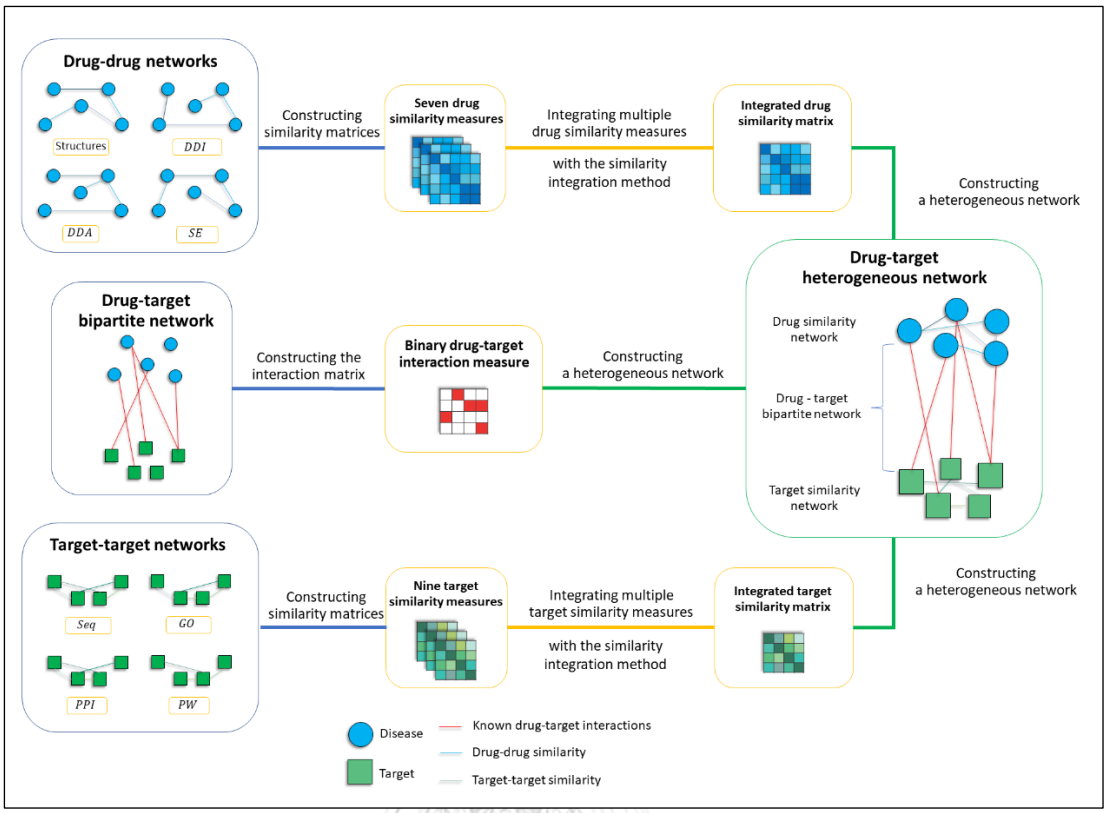



**Figure 3.3** The process of constructing the drug-target heterogeneous network by combining multiple similarity networks

### 3.4 Similarity integration methods

Based on a particular similarity integration method, Algorithm 1 illustrates how to combine multiple drug-drug or target-target similarity measures. Let $SM = \{s_1, s_2, \ldots, s_p\}$ be a subset of given similarity measures which are arrayed according to the order of integration, $Xs_i$ be a matrix containing the similarity scores based on similarity measure $s_i$, and $f$ represents a function to combine two similarity matrices together. The output of Algorithm 1 is an integrated similarity matrix (*IntSimMat*) based on a set of given similarity measures ($SM$) and an integration function $f$. In this algorithm, $\|\cdot\|$ is the number of elements in a set. According to Algorithm 1, $Xs_i$ can be either a drug-drug similarity matrix (i.e., $Xs_i \in \mathbb{R}^{m \times m}$) or a target-target similarity matrix (i.e., $Xs_i \in \mathbb{R}^{n \times n}$).

**Algorithm 1:** Integrating multiple similarity matrices

Input: A set of given similarity measures ($SM$) and a function of an integration method ($f$)

Output: An integrated similarity matrix ($IntSimMat$)

1. for $i = 1$ to $i = |SM| - 1$
2. if $i == 1$ then
3. $IntSimMat = f(Xs_i, Xs_{i+1})$
4. else
5. $IntSimMat = f(IntSimMat, Xs_{i+1})$
6. end if
7. end for

In this thesis, we include both linear and non-linear functions for integrating multiple similarity matrices. We consider four functions which are average ($AVG$), maximum ($MAX$), minimum ($MIN$), and similarity network fusion (SNF) [48]. The SNF method is previously described in Section 2.4. The $AVG$, $MAX$, and $MIN$ are defined in Equation (3.11) to (3.13), where $A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}$ and $B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \dots & b_{mm} \end{bmatrix}$ are $m \times m$ two different similarity matrices based on different similarity measures. According to matrix $A$, $a_{ij}$ is a similarity score between drug (or target) $i$ and drug (or target) $j$ when there are $m$ drugs (or targets). Similar to matrix $B$, $b_{ij}$ is a similarity score between drug (or target) $i$ and drug (or target) $j$ when there are $m$ drugs (or targets).



$$AVG(A,B) = \left(\frac{1}{2}\right)\begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1m}+b_{1m} \\ a_{21}+b_{21} & a_{ij}+b_{ij} & \dots & a_{2m}+b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \dots & a_{mm}+b_{mm} \end{bmatrix} \quad (3.11)$$

$$MIN(A,B) = \begin{bmatrix} min(a_{11}, b_{11}) & min(a_{12}, b_{12}) & \cdots & min(a_{1m}, b_{1m}) \\ min(a_{21}, b_{21}) & min(a_{ij}, b_{ij}) & \dots & min(a_{2m}, b_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ min(a_{m1}, b_{m1}) & min(a_{m2}, b_{m2}) & \dots & min(a_{mm}, b_{mm}) \end{bmatrix} \quad (3.12)$$

$$MAX(A,B) = \begin{bmatrix} max(a_{11}, b_{11}) & max(a_{12}, b_{12}) & \cdots & max(a_{1m}, b_{1m}) \\ max(a_{21}, b_{21}) & max(a_{ij}, b_{ij}) & \dots & max(a_{2m}, b_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ max(a_{m1}, b_{m1}) & max(a_{m2}, b_{m2}) & \dots & max(a_{mm}, b_{mm}) \end{bmatrix} \quad (3.13)$$

### 3.5 Forward similarity method (FSI) Framework

To enhance the heterogeneous network model by integrating multiple similarity measures of drugs and target proteins for predicting DTIs, we proposed the Forward Similarity Integration (FSI) algorithm, which systematically selects suitable similarity integration using the forward selection technique (described in Section 2.5).

Suppose that there are $k$ different drug-drug similarity measures in set $AD = \{dd_1, dd_2, \dots, dd_k\}$ and $l$ different target-target similarity measures in set $AT = \{tt_1, tt_2, \dots, tt_l\}$. To create an integrated similarity measure by function $IntegrateSim$ based on a particular performance measure, the FSI algorithm stepwise finds the suitable subsets of drug-drug and target-target similarity measures, denoted as $OD$ and $OT$, respectively. The pseudocodes explaining the FSI algorithm is shown in Algorithm 2.

**Algorithm 2:** Forward Similarity Integration (FSI)

Input: A set of all drug-drug similarity measures ($AD$) and target-target similarity measures ($AT$)

Output: A suitable subset of drug-drug similarity measures ($OD$) and target-target similarity measures ($OT$), which are orderly integrated to obtain the most suitable combined drug-drug and target-target similarity measures.

1. Initialize $k = 0, OD_0, OT_0 = \varnothing, RD_0 = AD, RT_0 = AT, PRF_0 = 0$.
2. repeat
3. $k = k + 1$
4. $x^*, y^* = \underset{x \in RD_{k-1},\ y \in RT_{k-1}}{\operatorname{argmax}} [EstimatePRF(IntegrateSim(OD_{k-1} \cup \{x\}), IntegrateSim(OT_{k-1} \cup \{y\}))]$ // Add both drug and target similarity
5. Denote the suitable performance as $PRF_{both}, X^* = \{x^*\}, and\ Y^* = \{y^*\}$
6. if $k == 1$ then
7. $PRF_k = PRF_{both}$
8. else // Also consider adding only drug or target similarity
9. $x_d^* = \underset{x \in RD_{k-1}}{\operatorname{argmax}} [EstimatePRF(IntegrateSim(OD_{k-1} \cup \{x\}), IntegrateSim(OT_{k-1}))]$ and denote the suitable performance as $PRF_{drug}$
10. $y_t^* = \underset{y \in RT_{k-1}}{\operatorname{argmax}} [EstimatePRF(IntegrateSim(OD_{k-1}), IntegrateSim(OT_{k-1} \cup \{y\}))]$ and denote the suitable performance as $PRF_{target}$
11. $PRF_k = \max(PRF_{both}, PRF_{drug}, PRF_{target})$
12. if $PRF_k == PRF_{drug}$ then
13. $X^* = \{x_d^*\}$ and $Y^* = \emptyset$
14. else if $PRF_k == PRF_{target}$ then
15. $X^* = \emptyset$ and $Y^* = \{y_t^*\}$
16. end if
17. end if
18. if $PRF_k > PRF_{k-1}$ then
19. Update $OD_k = OD_{k-1} \cup X^*,\ OT_k = OT_{k-1} \cup Y^*,\ RD_k = RD_{k-1} - X^*, RT_k = RT_{k-1} - Y^*$
20. end if
21. until $(PRF_k \leq PRF_{k-1}\ or\ |RD_k| == 0\ or\ |RT_k| == 0)$
22. return the latest $OD$ and $OT$

According to Algorithm 2, the FSI begins with the empty suitable models i.e., $OD = \emptyset$ and $OT = \emptyset$. In each step, a drug-drug and a target-target similarity measure ($dd_i \in AD$ and $tt_j \in AT$) are added into sets $OD$ and $OT$, respectively, to form different heterogeneous network models. These models were compared their performance ($PRF$) to select the suitable one which would be added into sets $OD$ and $OT$. Then, the drug-drug and target-target similarity measure recently added to the suitable subsets would be removed from the remaining sets of drug-drug and target-target similarity measures, denoted as $RD$ and $RT$, respectively. To avoid combining some correlated similarity measures together, the similarity measures derived from the same data set (e.g., DDI_Jac and DDI_Cos) will be removed from $RD$ and $RT$ when one of them is selected to be integrated.

The FSI algorithm iteratively updates $OD$ and $OT$ by adding a drug-drug and target-target similarity measure until no more performance improvement from the adding of similarity measures or no similarity measures remaining in $RD$ or $RT$. Therefore, a heterogeneous network model formed with the integrated similarity measures obtained by the FSI algorithm are expected to show the suitable performance in predicting DTIs. Additionally, the suitable subsets of similarity measures obtained by FSI ($OD$ and $OT$) may depend on the performance measure used to evaluate heterogeneous network models and the integration method applied. In this thesis, we thus consider several performance measures used in FSI and several integration functions to create an integrated similarity matrix from multiple similarity measures.

### 3.6 Performance measurement

To evaluate the performance of all heterogeneous network models, we performed a ten-folds cross-validation technique. All drug-target interactions were classified into the positive class (known DTIs) and negative class (unknown DTIs). Then, we randomly divided both positive and the negative interactions into ten

equal parts which each part treated as the test data in turn, and the remaining nine parts are used as the training data. To compare the predicted and the actual classes, a confusion matrix of binary classes was used, as shown in Table 3.5.

**Table 3.5** A confusion matrix

| | | Actual values: Positive | Actual values: Negative |
|---|---|---|---|
| Predicted values | Positive | True Positive (TP) | False Positive (FP) |
| Predicted values | Negative | False Negative (FN) | True Negative (TN) |

Generally, there are some frequently used metrics in classification, including precision (PRE), recall (REC), Accuracy (ACC), Matthews Correlation Coefficient (MCC), F1-measure (F1), area under a precision-recall curve (AUPR), and area under a receiver operating characteristic curve (AUC), which the values range from 0 to 1. The higher of the values, the better the classifier is. The performance measures used in this work are shown as follows and can be calculated by Equation (3.14) - (3.18).

$$PRE = \frac{TP}{TP + FP} \tag{3.14}$$

$$REC = \frac{TP}{TP + FN} \tag{3.15}$$

$$F_1\ measure = \frac{2 \cdot PRE \cdot REC}{PRE + REC} \tag{3.16}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.17}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.18}$$

# CHAPTER IV
# RESULTS AND DISCUSSION

In this chapter, the results of the preliminary data analysis are shown and discussed. Next, the results of creating the heterogeneous network model with integration of multiple similarity measures (i.e., parameter settings and the selection of similarity integration using the FSI framework) are provided. After obtaining the model with the suitable similarity integration or the FSI model, the results of the performance evaluation of the FSI model are shown and discussed. Finally, the FSI model is utilized to identify new DTIs, and the predictions are then verified by searching for supporting evidence.

## 4.1 Preliminary data analysis

### 4.1.1 Data summarization

In this thesis, four data sets of drugs (i.e., chemical structures, DDIs, DDAs, and SEs) and four data sets of target proteins (i.e., protein sequences, GO annotations, PPIs, and PWs) were used to generate different similarity measures of drugs and target proteins. To avoid problems in integrating multiple similarity measures of drugs and target proteins, only drugs and target proteins that have all required data are included in this thesis. As a result, there are 862 drugs, 1,517 target proteins, and there are 3,583 known DTIs. The summarization of drug and target data are shown in Table 4.1.

All 862 drugs have their own chemical structures. For DDIs, those 862 drugs interact to 4,014 unique drugs with 924,819 drug-drug interactions. In the data set of DDAs, 862 drugs are associated with 2,287 unique diseases, and there are 23,201 drug-disease associations. Moreover, those 862 drugs have 5,280 unique SE terms with 140,682 links between drugs and their SE terms.

In terms of target proteins, 1,517 targets also have their own protein sequences. Those proteins are annotated by 8,924 unique GO terms of any aspects, and there are 45,866 links between target proteins and their annotated GO terms. Moreover, those 1,517 target proteins interact with 14,202 unique proteins in the PPI network of 218,721 PPIs. 1,517 target proteins are also involved in 283 unique pathways, and there are 7,988 links between target proteins and their involved pathways.

**Table 4.1** Summary information of all drug and target data

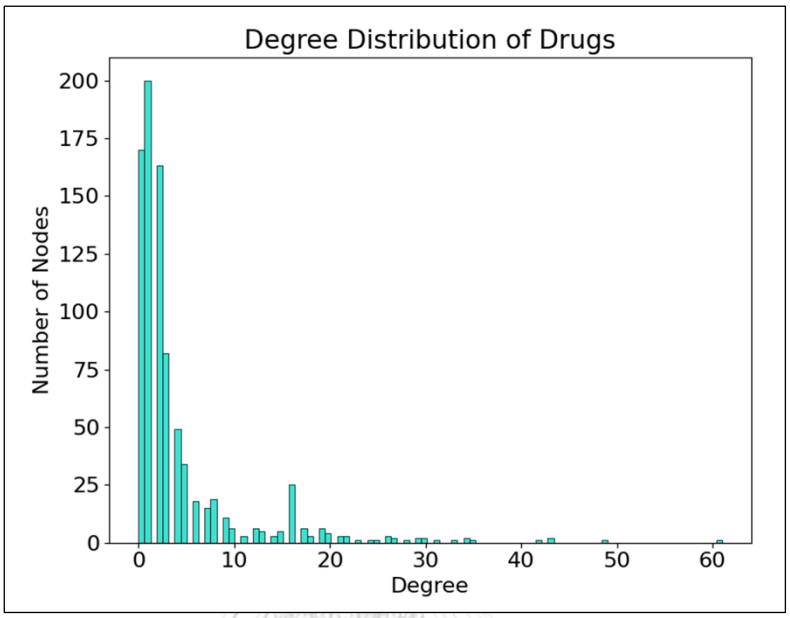
| Data type | Data name | Number of unique entities | Number of Interactions |
|---|---|---|---|
| Drug data (862 drugs) | Structures | 862 structures | - |
| | DDIs | 4,014 drugs | 924,819 |
| | DDAs | 2,287 diseases | 23,201 |
| | SEs | 5,280 SE terms | 140,682 |
| Target data (1,517 target proteins) | Seqs | 1,517 sequences | - |
| | GOs | 8,924 GO terms | 45,866 |
| | PPIs | 14,202 proteins | 218,721 |
| | PWs | 283 pathways | 7,988 |
| DTI data | DTIs | - | 3,583 |



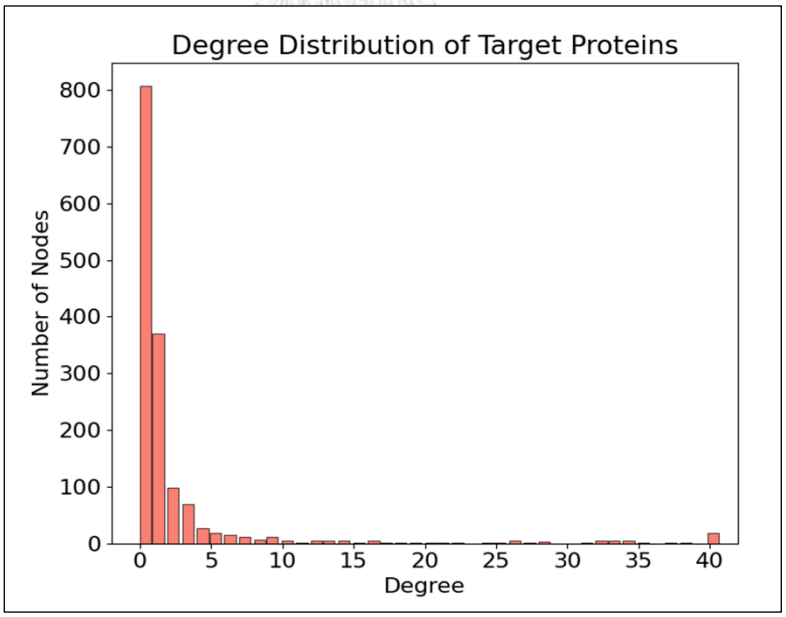
4.1.2 Degree distributions of the DTI network

As the DTI data can be considered as the drug-target bipartite network, the degree distributions of this network are observed to analyze all known DTIs in hands. The degree distributions of the drug-target bipartite network are shown in Figure 4.1. Based on the known drug-target interactions, this figure describes how many target proteins are associated with a drug and how many drugs are associated with a target

protein. In this figure, *x*-axis represents the degree of a node, and *y*-axis represents the fraction of nodes having degree *k*.




(a) The degree distribution of drugs



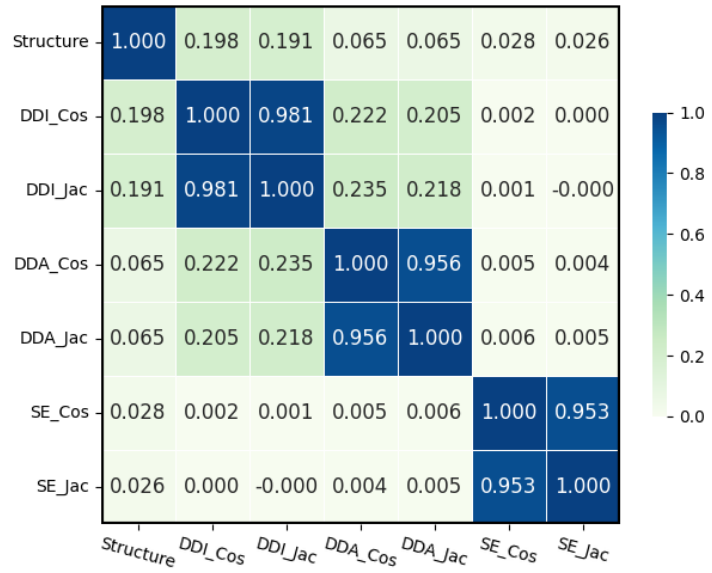

(b) The degree distribution of target proteins

**Figure 4.1** Degree distributions of the drug-target bipartite network

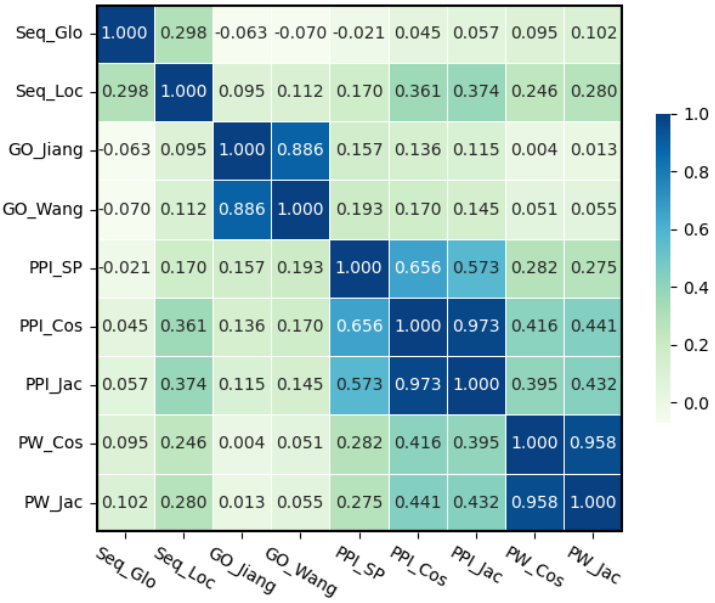
According to Figure 4.1, it's clear that most drugs and target proteins are associated with few target proteins and drugs, respectively. There are 80.97% of the drugs and 91.76% of the target proteins that have the node degrees less than 5. This means that more than 80% of the drugs and more than 90% of the target proteins interact with less than 5 target proteins and drugs, respectively. The maximal degrees of the drug and target nodes are 61 and 40, respectively. In addition, only 0.14% of the drugs have degrees over 40. A drug can bind approximately four target proteins on average, and a target protein can interact with approximately two drugs on average. Additionally, it is observed that the majority of nodes are rarely connected to another while only a few nodes have their own dense links. This suggests that there would be many undiscovered links between drugs and target proteins in the drug-target bipartite network.

4.1.3 Correlation analysis

To preliminary evaluate the relationship among the defined similarity measures, we performed the Pearson correlation analysis between seven drug-drug similarity measures and between nine target-target similarity measures. A Pearson correlation coefficient ($\rho$) is a measure of linear correlation between two data sets that assigns a value between -1 and 1, where 0 is no correlation, 1 is total positive correlation, and -1 is total negative correlation. The Pearson correlation coefficients that we calculate for all pairs of drug-drug and target-target similarity measures are shown in Figure. 4.2, where the darker the color, the greater the correlation.




(a) Drug similarity measures




(b) Target similarity measures

**Figure 4.2** Heatmaps of the Pearson correlation coefficients of drug and target similarity measures.

According to Figure 4.2, most of drug-drug similarity matrices and target-target similarity matrices are slightly positively correlated with one another. This suggests that most drug and target similarity measures tend to provide information

complement that of one another. Nevertheless, the similarity measures based on the same data sets, such as DDI_Cos and DDI_Jac, are highly positive relationship together ($\rho$ > 0.8), as shown in Figure 4.2a. This suggests that these similarity measures are very similar and almost an identical similarity measure. To avoid unfair and useless integrating of multiple drug or target similarity measures, we do not combine the similarity measures based on the same data set in the FSI algorithm.

Interestingly, the target similarity measures based on PPIs, protein sequences, and pathway information are slightly to moderately positively correlated together ($\rho$ < 0.5), as shown in Figure 4.2b. This may be because the target-target similarity based on protein sequences and PPIs could infer to the similarity based on pathways that proteins involve with. Furthermore, the drug similarity measures based on DDIs, structures, and DDAs, and the target similarity measures based on GOs, PPI, and Seq are negligibly positively correlated together ($\rho$ < 0.3). However, the values of the correlation coefficients of all similarity measures, except the similarity measures based on the same data sets, are slightly positive. This could infer that these similarity measures are good complements for being integrated, and combining the different similarity measures in constructing a drug-target heterogeneous network model may improve the DTI prediction.



**4.2 Parameter setting**

In a process of the heterogeneous network propagation, there is a parameter required to be suitably adjusted, i.e., the decay factor. This parameter indicates how much the propagation from the edges' weights affects the weight updates of the drug-target links relative to the initial weights of the drug-target links. Most studies, such as [33, 41], set the value of the decay factor at 0.4 because this value was suggested by [14]. Nevertheless, we used diverse data sets and different integration methods to construct numerous heterogenous network models for finding the suitable model with a suitable similarity integration by FSI. Those models with

different similarity integration may have distinct suitable values of the decay factor. To reduce model variables, this experiment aims to preliminarily specify an estimated value of the decay factor for all heterogeneous network models.

With seven drug-drug similarity measures and nine target-target similarity measures, we can congregate 5,671 possible combinations of drug and target similarity measures. The possible combinations of the drug and target similarity measurements are summarized in Table 4.2, where columns and rows of the table represent the numbers of drug and target similarity measures used for constructing a heterogeneous network propagation model. Among those 5,671 combinations of the similarity measures, there is not any combination that integrates the similar similarity measures generated from the same data sets together.

**Table 4.2** The possible combinations of drug and target similarity measures

| | | Number of drug similarity measures used | | | |
|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** |
| Number of target similarity measures used | **1** | 63 | 210 | 308 | 168 |
| | **2** | 162 | 540 | 792 | 432 |
| | **3** | 180 | 600 | 880 | 480 |
| | **4** | 72 | 240 | 352 | 192 |
| | | **Sum** | | | 5,671 |



In this experiment, we perform a stratified random of 10% of all possible combinations, shown in Table 4.2. This results in 567 combinations of drug and target similarity measures. For each combination with multiple drug and target similarity measures, we integrate the drug and target similarity measures by using all

integration functions (i.e., AVG, MAX, MIN, and SNF) to create the heterogeneous network models. Thus, we have 567 different heterogeneous network models with four integration functions.

To find suitable values of the decay factor, we varied the values of the decay factor from 0.1 to 1 with the step of 0.1 in the network propagation for each model. Then, ten-fold cross validation was conducted to evaluate the model performance when using each value of the decay factor. For each model, we investigated the AUC, AUPR, and F1 values and used each of these metrics to select the suitable value of the decay factor. To compare among all values of the decay factor, we counted numbers of the models that select each value of the decay factor based on AUC, AUPR, and F1, as shown in Figure 4.3.

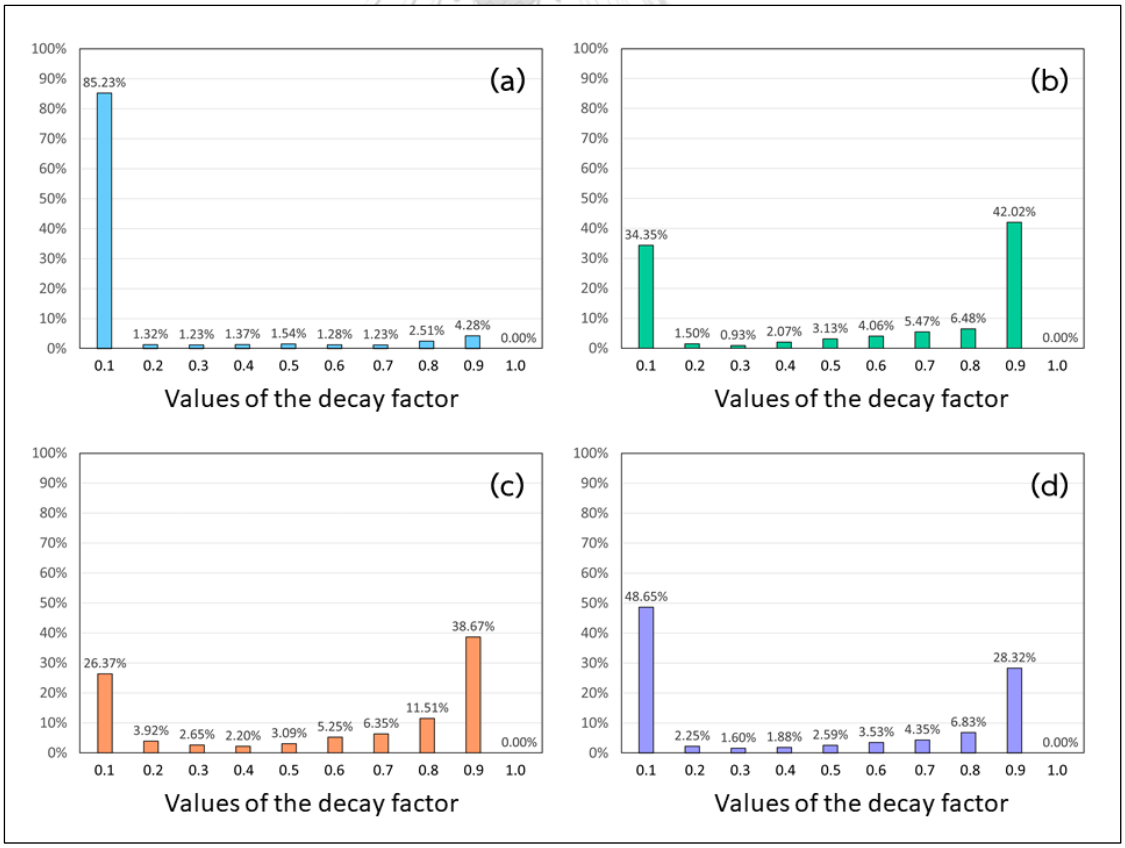



**Figure 4.3** The distributions of the selected values of the decay factor. (a) The number of models with the maximum AUC values (%). (b) The number of models

with the maximum AUPR values (%). (c) The number of models with the maximum F1 values (%). (d) The number of models with the maximum AUC/AUPR/F1 values (%).

According to Figure. 4.3a, most of the models accounted for 85.23% reach the maximum AUC values when the decay factor is 0.1. Meanwhile, Figure 4.3b shows that 42.02% and 34.35% of the sampled models achieve the maximum AUPR values when the decay factor of 0.9 and 0.1 are used, respectively. Moreover, when we exploit the decay factor of 0.9 and 0.1, most of the models, 38.67% and 26.37% of the sampled models, obtain the maximum F1 scores, as shown in Figure 4.3c. To select a value of the decay factor that would provide the approximately suitable AUC, AUPR, and F1 values, we investigate the overall results based on AUC, AUPR, and F1, as shown in Figure 4.3d. As the result, it clearly shows that the decay factor of 0.1 gives the highest coverage of the sampled models with the maximum AUC, AUPR, and F1 values. Therefore, we specified the value of the decay factor as 0.1 in the network propagation for all heterogeneous network models.

**4.3 Selection of the suitable similarity integration using FSI**



Based on a given method of similarity integration, we introduced the FSI framework for building the heterogeneous network model with the suitable similarity integration. The FSI framework systematically discovers the suitable sets of drug and target similarity measures combined by a particular similarity integration method. In the FSI framework, several similarity integration methods (i.e., AVG, MAX, MIN, and SNF) are also compared to select the suitable method. Three comprehensive evaluation metrics, i.e., AUC, AUPR, and F1, are used to serve as the criteria for selecting a similarity measure integrated into a model at each time in FSI. Ten-folds cross validation was executed to evaluate the model performance of different similarity integration and select the suitable one. For each similarity integration

method and each selecting criteria, the suitable models selected by FSI are shown in Table 4.3.

**Table 4.3** The FSI models based on different similarity integration methods and selecting criteria

<table>
<tr><th rowspan="2">Similarity integration method</th><th colspan="3">Performance measure for selecting drug/target similarities</th></tr>
<tr><th>AUC</th><th>AUPR</th><th>F1</th></tr>
<tr><td>AVG</td><td>Model 1<br>{DDA_Jac, PPI_Jac} &<br>{Seq_Loc}</td><td colspan="2" rowspan="2">Model 2<br>{DDA_Jac} & {Seq_Loc}</td></tr>
<tr><td>MAX</td><td>Model 3<br>{DDA_Jac} & {PPI_Jac}</td></tr>
<tr><td>MIN</td><td>Model 4<br>{DDA_Jac, Structures,<br>DDA_Cos} &<br>{PPI_Jac, Seq_Loc}</td><td colspan="2">Model 5<br>{DDA_Jac, SE_Jac, Structures, DDI_Jac} &<br>{Seq_Loc}</td></tr>
<tr><td>SNF</td><td>Model 6<br>{DDA_Jac, DDI_Jac,<br>PPI_Jac} &<br>{Seq_Loc, PW_Jac}</td><td>Model 7<br>{DDA_Jac, DDI_Cos,<br>Structures} &<br>{Seq_Loc}</td><td>Model 8<br>{DDA_Jac, DDI_Jac,<br>Structures} &<br>{Seq_Loc}</td></tr>
</table>



According to Table 4.3, the FSI algorithm selects the models formed by different combinations of drug and target similarities when we used distinct integration methods and performance measures. Nevertheless, when AUPR or F1 are used as the selecting criteria for FSI, the same integrated models are usually obtained (i.e., Model 2 and Model 5). For example, Model 2 uses only a single drug-drug similarity measure (i.e., DDA_Jac) and a single target-target similarity measure (i.e., Seq_loc), when AVG or MAX is specified as the selected similarity integration method. Similarly, the FSI algorithm selects Model 5, which integrates DDA_Jac, SE_Jac, Structures, and DDI_Jac into a single drug similarity measure by MIN and utilizes Seq_Loc as a target similarity measure, whether we used AUPR or F1 as the

selecting criteria. This suggests that AUPR or F1 may be used interchangeably in certain situations.

Moreover, we notice that most models obtained by the FSI algorithm always use DDA_Jac and Seq_Loc because these similarity measures are preliminarily identified as the suitable similarity measures at the initial iteration of the FSI algorithm. This implies that the therapeutic effects of drugs and the local sequence alignments of proteins are relatively important similarity measures for predicting DTIs. This is consistent with several recent studies using therapeutic effects of drugs and the local sequence alignments to predict new DTIs [46, 92-95]. Then, the FSI algorithm additionally selects other drug and target similarities to combine with them in the next iterations. From eight different models obtained, we estimated the performance of each model by performing ten-fold cross validation and calculated four evaluation metrics in addition to AUC, AUPR and F1, i.e., precision, recall, ACC, and MCC. The mean values of all evaluation metrics of each model are shown in Table 4.4. Those values are then compared by performing *t*-tests at a significance level of 0.05.

**Table 4.4** Performance of eight FSI models and results of t-tests

| Model no. | AUPR | AUC | PRE | REC | F1 | ACC | MCC |
|---|---|---|---|---|---|---|---|
| 1 | 0.227 | 0.951 | 0.342 | 0.333 | 0.333 | 0.996 | 0.334 |
| 2 | 0.267[a, b] | 0.935 | 0.389[a, b] | 0.373[a, b] | 0.379[a, b] | 0.997[a, b] | 0.378[a, b] |
| 3 | 0.140 | 0.947 | 0.192 | 0.284 | 0.221 | 0.994 | 0.226 |
| 4 | 0.233 | 0.953 | 0.310 | 0.400 | 0.348 | 0.996 | 0.349 |
| 5 | 0.335[c] | 0.926 | 0.426[c] | 0.443[c] | 0.434[c] | 0.997[c] | 0.433[c] |
| 6 | 0.294 | **0.958** | 0.334 | 0.422 | 0.367 | 0.996 | 0.370 |
| 7 | **0.481**[d, e] | 0.933[e] | **0.578**[d] | 0.508[d] | **0.539**[d, e] | **0.998**[d] | **0.540**[d, e] |
| 8 | 0.481 | 0.933 | 0.564 | **0.515** | 0.538 | 0.998 | 0.537 |

The maximum value is shown in bold.

[a] Significantly greater than a mean value of Model 1 at a significance level of 0.05

[b] Significantly greater than a mean value of Model 3 at a significance level of 0.05

[c] Significantly greater than a mean value of Model 4 at a significance level of 0.05

[d] Significantly greater than a mean value of Model 6 at a significance level of 0.05

[e] Significantly greater than a mean value of Model 8 at a significance level of 0.05

According to Table 4.4, we found that Model 2 significantly performs better than other models based on AVG and MAX (i.e., Model 1 and Model 3) at a significance level of 0.05. When compared among the models based on MIN, Model 5 significantly outperforms Model 4 at a significance level of 0.05. For the models based on SNF, the overall performance of Model 7 is significantly greater than that of Model 6 and 8 in most evaluation metrics, such as AUPR, F1, and MCC, at a significance level of 0.05. After the different models with the similar integration method were compared, we can recommend that using AUPR as a performance measure for selecting drug and target similarities to integrate into a model produces the models with better performance than those of using AUC or F1.

In addition, it is noticeable from Table 4.4 that the maximum performance values mostly are of the models using SNF as a similarity integration method (Model 6, 7, and 8). This may be because the linear integration methods (i.e., AVG, MIN, and MAX) are somewhat sensitive to outliers of some similarity scores whereas SNF is able to reduce noise of weak similarity values and can increase the significance of interactions very confusing values between drugs or targets to the other similarity measures [48]. Currently, SNF is an effective method widely exploited for aggregating multi-omics data in several biological applications, such as DTI prediction [47, 96] and DDA inference [97, 98].

To finally select the suitable model with the most suitable similarity integration, Model 6, 7, and 8 are compared their performance values. By performing *t*-tests, the mean values of almost evaluation metrics, except that of AUC, of Model

7 are significantly greater than those of Model 6 at a significance level of 0.05. When compared between Model 7 and 8, the mean values of AUPR, AUC, F1, and MCC of Model 7 is significantly higher than those of Model 8 at a significance level of 0.05. Therefore, Model 7, which integrates DDA_Jac, DDI_Cos, and Structures into a drug similarity network by SNF and uses Seq_Loc as a target similarity, is selected as the suitable FSI model. This implies that the therapeutic effects of drugs, drug-drug interactions, chemical structures of drugs, and protein sequences are useful similarity measures for predicting DTIs. This corresponds to several recent studies that also utilize those properties of drugs and targets to predict new DTIs [99, 100].

Moreover, it was reported that the therapeutic effects of drugs are associated with the abilities to modulate drug targets in the molecular level and could promote the relationships between drugs and targets [96]. Furthermore, the high similarity scores based on DDIs could infer to highly similar targets or processes that drugs involve. Importantly, two molecules with similar chemical structures can likely relate to same target, and two targets with similar sequence structures are likely to interact with same drugs [4, 39, 101-103]. In addition, SNF which is a non-linear integration method, is suitable to be used to combine the drug-drug similarity measures based on DDA_Jac, DDI_Cos, and Structures because the Pearson correlation coefficient of these similarity measures are slightly positive, according to Figure 4.2.

### 4.4 Performance evaluation of the FSI model

In this section, the heterogeneous network model with the drug and target similarity integration selected by the FSI algorithm outperformances other models. This selected model is termed as the FSI model. By FSI, the selected model combines DDA_Jac, DDI_Cos, and Structures as an integrated drug similarity by using SNF and employs Seq_Loc as a target similarity. In this section, the experiment to demonstrate the FSI efficiency is conducted. In this experiment, the performance of the FSI model is compared with those which fully combines all similarity measures,

randomly integrate similarity measures, and conventionally integrate similarity measures from chemical structures of drugs and protein sequences of targets. Next, the integration of additional drugs' properties (i.e., DDA_Jac and DDI_Cos) are verified and discussed their useful information for predicting DTIs.

4.4.1 Verification of the FSI efficiency

To demonstrate the superior efficiency of the FSI algorithm, we compare the performance of the FSI model with the full integration model, the random integration models, and the conventional model, which uses only Structures and Seq_Loc. Ten-fold cross validation was executed to evaluate the performance of each model.

- Comparing with the full integration model

The full integration model is a model which combines all existing similarity measures of drugs and targets considered in this thesis. The comparison of the performance of the FSI model and the full integration model are shown in Figure 4.4. Noticeably, the FSI model performs better than the full integration model. By *t*-tests, the mean values of all evaluation metrics of the FSI model, except AUC, are significantly greater than those of the full integration model at a significance level of 0.01. This means that a model formed by selecting only some advantageous drug and target similarities by FSI is more efficient than a model integrating all existing drug and target similarities without cautious consideration. Moreover, it can be noticed that the mean values of AUC and ACC of the FSI model is close to full integration model. This could be because the information of DTIs in the thesis is an imbalanced problem, the present DTIs less than the absent DTIs, the mean values of AUC and ACC.

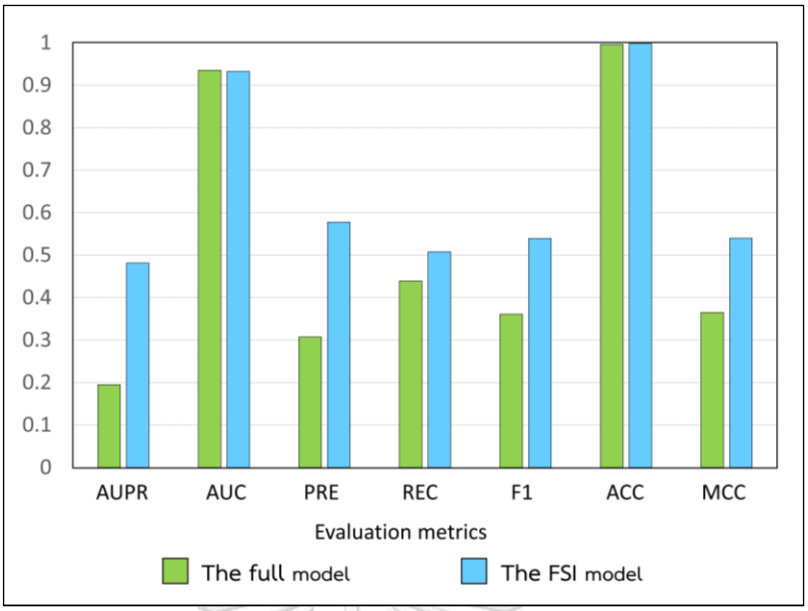



**Figure 4.3** Performance comparison of the full integration model and the FSI model

- Comparing with the random integration models

In the random integration models, drug and target similarities were randomly selected 100 times to combine by using SNF and then construct different 100 models. To cover all models with different numbers of drug and target similarity measures integrated, those 100 integrated models are randomly selected from all possible combinations of the numbers of drug and target similarity measures integrated.



To demonstrate that the performance of the FSI model is better than those of 100 random integration models, the results of the *t*-tests that compare the mean value of each evaluation metric of the FSI model with that of each random integrating model are presented in Figure 4.5.

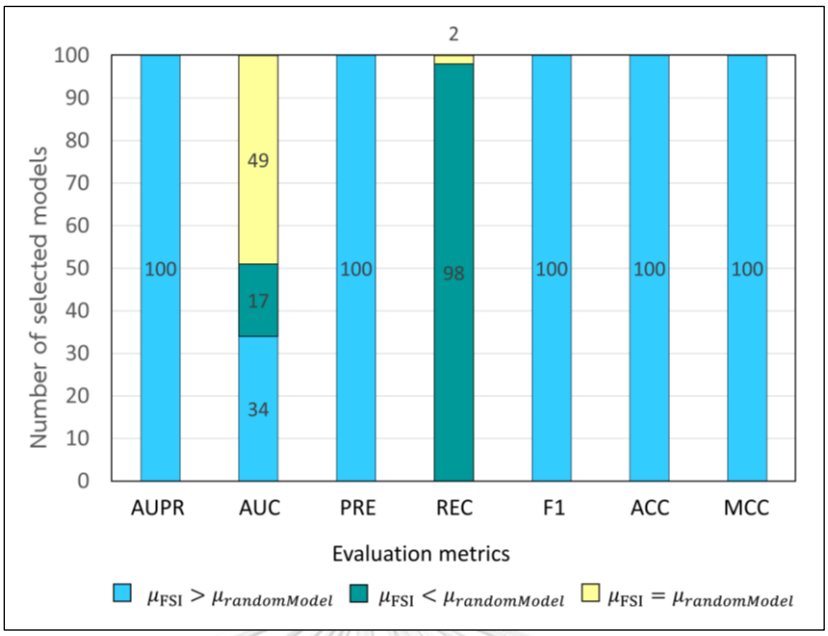



**Figure 4.4** Performance comparison between 100 random integration models and the FSI model

In Figure 4.5, the light blue bars are showed with the numbers of *t*-tests where the mean value of an evaluation metric of the FSI model is greater than that of a random integrating model at a significance level of 0.05. The labeled numbers in the light-yellow bars and light green bars are the numbers of *t*-tests resulting that the mean value of an evaluation metric of the FSI model is equal to or lower than that of a random integration model at a significance level of 0.05. From Figure 4.5, it is showed that the mean values of all evaluation metrics of the FSI model, except AUC and REC, are significantly greater than those of the 100 random integrating models at a significance level of 0.05. This implies that the model formed by systemically selecting drug and target similarities using FSI is more efficient than a model randomly selecting drug and target similarities to integrate into the model.

- Comparing with the conventional model

The conventional model is a model that were commonly used to predict DTIs based on the similarity-based methods (e.g., [12-15]) This model applies only

drug chemical structures to prepare a drug similarity measure and uses only local sequence alignments of target proteins for computing a target similarity measure. To demonstrate the superior performance of the FSI model, we compared the efficiency of the FSI model with that of the conventional model as shown in Figure 4.6.

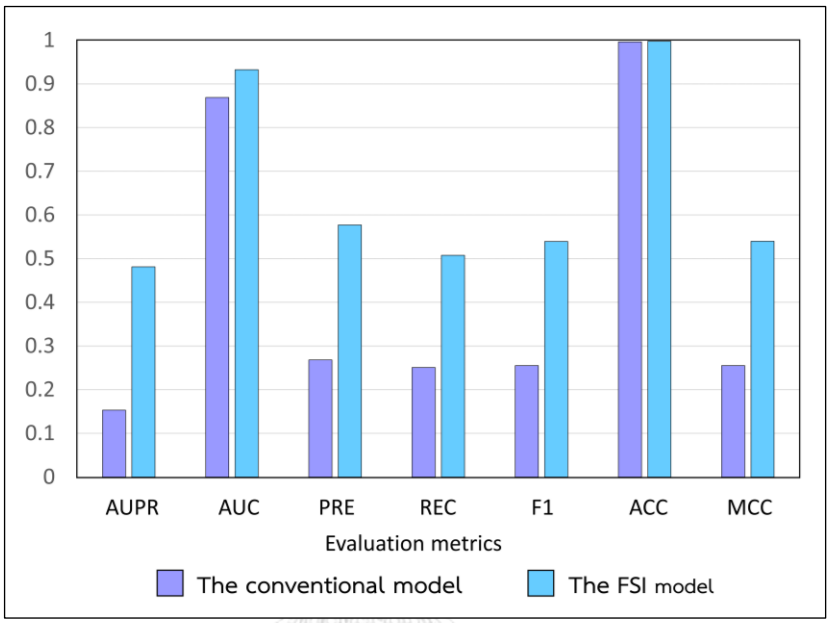



**Figure 4.5** Performance comparison of the conventional model and the FSI model

According to Figure 4.6, we found that the FSI model performs better than the conventional model at a significance level of 0.01. This means that integrating DDA_Jac and DDI_Cos into a drug similarity measure in addition to drug chemical structures can greatly improve the performance of the conventional model in predicting DTIs. Next, both DDA_Jac and DDI_Cos are verified their significance in improving DTI predictions.



4.4.2 Significance of the integrated drug similarities

By using FSI, we obtained the suitable heterogeneous model, which combines DDA_Jac, DDI_Cos, and Structure into a drug similarity measure by SNF and uses Seq_Loc as a target similarity measure. In addition to the conventional model, the FSI model additionally integrates DDA_Jac and DDI_Cos into the model and can

highly improve DTI predictions. To verify the importance of those additional drug similarity measures in predicting DTIs, we compared the performance of the FSI model with those of the models removing DDA_Jac (the DDA_Jac reduced model) and removing DDI_Cos (the DDI_Cos reduced model). Furthermore, we compared the performance of the FSI model with those of the models permuting the DDA matrix (the DDA permuted model) and the DDI matrix (the DDI permuted model). The mean AUPR, AUC, and F1 values of each model and the results of the *t*-tests are shown in Table 4.5.

**Table 4.5** Performance comparison of the FSI model and the reduced models

| Model | AUC | | AUPR | | F1 | |
|---|---|---|---|---|---|---|
| | mean | p-value | mean | p-value | mean | p-value |
| The FSI model | 0.9325 | - | 0.4811 | - | 0.5392 | - |
| DDA_Jac reduced model | $0.9221^{a}$ | $6.46\times10^{-8}$ | $0.4724^{a}$ | $3.65\times10^{-6}$ | $0.5310^{a}$ | $3.91\times10^{-5}$ |
| DDI_Cos reduced model | $0.9241^{a}$ | $2.47\times10^{-7}$ | $0.4711^{a}$ | $1.43\times10^{-6}$ | $0.5278^{a}$ | $1.36\times10^{-6}$ |
| DDA permuted model | $0.9324^{a}$ | 0.0099 | $0.4809^{b}$ | 0.0282 | $0.5384^{ns}$ | 0.0784 |
| DDI permuted model | $0.9137^{a}$ | $5.19\times10^{-9}$ | $0.4693^{a}$ | $1.13\times10^{-6}$ | $0.5240^{a}$ | $1.03\times10^{-5}$ |
| The permuted model of both DDAs and DDIs | $0.9137^{a}$ | $5.20\times10^{-9}$ | $0.4693^{a}$ | $1.11\times10^{-6}$ | $0.5239^{a}$ | $8.76\times10^{-6}$ |

[a] The mean value of the FSI model is significantly greater than the compared model at a significance level of 0.01

[b] The mean value of the FSI model is significantly greater than the compared model at a significance level of 0.05

[ns] The mean value of the FSI model is not significantly different with that of the compared model at a significance level of 0.05

According to Table 4.5, we found that the mean AUC, AUPR, and F1 values of the FSI model, i.e., 0.9325, 0.4811, and 0.5392, respectively, are greater than those of both DDA_Jac reduced model and DDI_Cos reduced model at a significance level of 0.01. This clearly demonstrates the advantages of both DDA_Jac and DDI_Cos for predicting DTIs. To construct the suitable model, we thus cannot remove both DDA_Jac and DDI_Cos from the heterogeneous network model.

Moreover, we further investigated the significance of the DDA_Jac and DDI_Cos by randomly shuffling the existing edges in the DDA and DDI matrices. Then, we reconstructed the heterogeneous network models with the DDA_Jac permuted and DDI_Cos permuted matrices. For both cases, SNF is still used as a similarity integration method. Therefore, there are three possible permuted models, including the DDA permuted model, the DDI permuted model, and the permuted model of both DDAs and DDIs.

According to Table 4.5, it is noticed that the mean AUC, AUPR, and F1 values of the FSI model are greater than those of the permuted model at a significance level of 0.01, except the mean AUPR and F1 values of the DDA permuted model. The mean AUPR value of the FSI model are greater than that of the DDA permuted model at a significance level of 0.05, whereas the mean F1 value of the FSI model is not significantly different from that of the DDA permuted model. Therefore, it cannot be statistically concluded that the FSI model performs the DDA permuted model, especially when considering the mean F1 values. This may be due to the large sparsity of the DDA matrix, resulting that permuting this matrix does not change DDA_Jac from the original one. In summary, the FSI model, which combines DDA_Jac, DDI_Cos, and Structures into a drug similarity network by SNF and takes Seq_Loc as a target similarity, is the model with the suitable similarity integration for predicting new DTIs, based on the available datasets and the similarity integration methods considered in this thesis.

**4.5 Identification of new drug-target interactions**

In this section, the FSI model is demonstrated its efficacy in the new DTI prediction, besides the performance evaluation by the ten-fold cross-validation. In this section, two types of case studies are considered. The first type is the prediction of new drugs for a target protein with only one known drug, and the second type is the prediction of new target proteins for a drug with only one known target.

4.5.1 Predicted drugs for target proteins with one known drug

There are three selected target proteins including tubulin beta-3 chain (Q13509), nicotinic acetylcholine receptor alpha-1 (P02708), and calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1C (Q14123). Their top five predicted drugs are shown in Table 4.6.

**Table 4.6** The top 5 predictions for three selected target proteins.

| Targets | Known Drug | Predicted Drug (DrugBank ID) |
|---|---|---|
| Tubulin beta-3 chain (Q13509) | Lxabepilone (DB04845) | **DB00541**, **DB00570**, **DB11641**, DB00518, DB00643 |
| Nicotinic acetylcholine receptor alpha-1 (P02708) | Lamotrigine (DB00555) | **DB00184**, **DB00657**, DB01273, DB00514, DB00333 |
| Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1C (Q14123) | Caffeine (DB00201) | DB01023, DB01244, DB00622, DB01656, DB00651 |

Note that the drugs found on publications are shown in bold.

The first case of the selected target proteins with one known drug is tubulin beta-3 chain (Q13509). A drug binding to this protein is lxabepilone (DB04845), which can reduce improper cell division caused by several types of cancer, such as breast

cancer, lung cancer, and lymphoma. Based on top five predicted drugs of beta-3 tubulin, three drugs have been reported that they are associated with tubulin, i.e., vincristine (DB00541), vinblastine (DB00570), and vinflunine (DB11641). Moreover, those drugs have been reported in many studies that they interact to tubulins, such as [104, 105] for vincristine, [105, 106] for vinblastine, and [105, 107] for vinflunine. In addition, beta-3 tubulin is also predicted to involve with albendazole (DB00518) and mebendazole (DB00643), drugs for helminth infections. It has been revealed that both drugs interact to the alpha-1A and beta-4B tubulins [108-111]. Therefore, both albendazole and mebendazole could bind to the highly similar protein beta-3 tubulin.

The second case of the selected target proteins is nicotinic acetylcholine receptor alpha-1 or nAChR$\alpha$1 (P02708). It has been reported that this target protein only interacts with lamotrigine (DB00555), an antiepileptic drug approved for the treatment of epilepsy and bipolar disorder [112, 113]. Interestingly, binding lamotrigine to nAChRs results in blocking of voltage-dependent sodium channels on this protein and preventing the release of excitatory neurotransmitters. This causes the prevention of seizures [112, 114]. As the DTI prediction, nAChR$\alpha$1 could be associated to other drugs that play roles in voltage-dependent ion channels of nAChR$\alpha$1, including nicotine (DB00184), and mecamylamine (DB00657), corresponding to several literature, such as [115, 116] for nicotine, and [116, 117] for mecamylamine. In general, nicotine is a stimulant drug that acts as an agonist at nicotinic acetylcholine receptors [118, 119]. The binding of this protein to nicotine activates voltage-gated calcium channels causing the channel to open and allows conductance of sodium, calcium, and potassium [17]. Moreover, nicotine is often used to relieve nicotine withdrawal symptoms and aided to smoke cessation. Meanwhile, mecamylamine is a nicotine antagonist used to treat hypertension and uncomplicated malignant hypertension. By binding to this protein, mecamylamine can act as a nicotinic acetylcholine receptor (nAChR) antagonist, inhibiting all known

nAChR subtypes [120]. Furthermore, nAChR$\alpha$1 is predicted to involve with varenicline (DB01273), dextromethorphan (DB00514) and methadone (DB00333), drugs for the relief of pain [121-123] and treatment of addiction [124-126]. Although there is no clear report that nAChR$\alpha$1 can directly bind to these drugs, there are some publications supporting that these drugs interact to the nAChR$\alpha$4, nAChR$\alpha$7, and nAChβ2 [127-132]. Thus, varenicline, dextromethorphan, and methadone could bind to the highly similar protein nAChR$\alpha$1.

The third case of the selected target proteins is Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1C or PDE1C (Q14123). Ordinarily, the approved drug interacting to target PDE1C is caffeine (DB00201) which is a stimulant present in tea, coffee, and analgesic drugs. By binding to this protein, caffeine can cause vasodilation [133]. According to the DTI predictions obtained from the FSI model, PDE1C could be associated with felodipine (DB01023), bepridil (DB01244), nicardipine (DB00622), and roflumilast (DB01656), the drugs used in chemotherapy for cancer. Despite no clear evidence about those interactions, it has been revealed that those four drugs are related to vasodilation, angina, and ischemic heart disease [134-137]. Furthermore, PDE1C is predicted to interact with dyphylline (DB00651), a drug approved for asthma, bronchospasm, and chronic obstructive pulmonary disease (COPD) [17, 138, 139]. Interestingly, dyphylline and caffeine are in a class of methylxanthines, a purine-derived group of pharmacologic agents used for bronchodilation and stimulation [140, 141].



4.5.2 Predicted target proteins of drugs with one known target protein

Herein, there are two selected drugs that consists of nicorandil (DB09220) and plerixafor (DB06809). Their top five predicted targets are shown in Table 4.7. The first case of the selected drug is nicorandil (DB09220), which is a vasodilatory drug used for patients with angina [142]. ATP-binding cassette sub-family C member 9 or ABCC9 (O60706) is the only one known protein interacting with nicorandil. Binding of this

drug to ABCC9 can activate vasodilation of arterioles and large coronary arteries [17, 143]. According to the DTI prediction, nicorandil could be associated with some target proteins in ABCC subfamily, including ATP-binding cassette sub-family C member 8 or ABCC8 (Q09428), cystic fibrosis transmembrane conductance regulator or ABCC7 (P13569), ATP-binding cassette sub-family C member 5 or ABCC5 (O15440), ATP-binding cassette sub-family C member 2 or ABCC2 (Q92887), and ATP-binding cassette sub-family C member 1 or ABCC1 (P33527). From searching for supporting literatures, we found that nicorandil can reduce an excess of insulin secretion in ABCC8-deficient insulin-producing cells [144]. Interestingly, the predicted ABCCs proteins and ABCC9 are in same subfamily, which are rather structurally conserved. Additionally, the predicted ABCCs proteins overlap some known drugs of ABCC9 [17], i.e., adenosine triphosphate (DB00171) and glyburide (DB01016). Thus, the predicted ABCCs proteins could bind to nicorandil as well [145].

**Table 4.7** The top 5 predictions for two selected target proteins

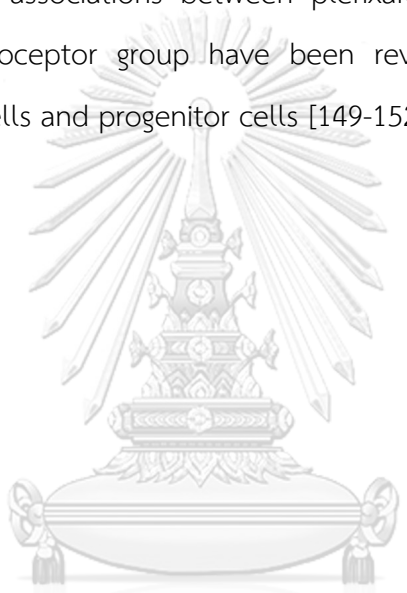
| Drugs | Known Target | Predicted Target (Uniprot ID) |
|---|---|---|
| Nicorandil (DB09220) | ATP-binding cassette sub-family C member 9 (O60706) | **Q09428**, P13569, O15440, Q92887, P33527 |
| Plerixafor (DB06809) | C-X-C chemokine receptor type 4 (P61073) | P35348, P08913, P35368, P18825, P18089 |

Note that the targets found on publications are shown in bold.



The second case of the selected drugs is plerixafor (DB06809), an anti-HIV agent specifically active against the T4-lymphotropic HIV strains [146]. The only one known target protein interacting to this drug is C-X-C chemokine receptor type 4 or CXCR4 (P61073). It has been reported that blocking the interaction between C-X-C

motif chemokine 12 or CXCL12 (P48061) and C-X-C chemokine receptor type 4 or CXCR4 (P61073) by plerixafor stimulation results in mobilize stem cells [17, 147, 148]. According to the DTI predictions, plerixafor could be connected with some target proteins in a group of adrenoceptors, including alpha-1A adrenergic receptor or ADRA1A (P35348), alpha-2A adrenergic receptor or ADRA2A (P08913), alpha-1B adrenergic receptor or ADRA1B (P35368), alpha-2C adrenergic receptor or ADRA2C (P18825), and alpha-2B adrenergic receptor or ADRA2B (P18089). Despite no clear evidence showing the associations between plerixafor and those proteins, some proteins in the adrenoceptor group have been revealed that they can induce mobilization of stem cells and progenitor cells [149-152].



# CHAPTER V
# CONCLUSION AND FUTURE WORK

In this chapter, we provide conclusion and the future direction of this research work for further improvement of the DTI prediction method based on the heterogeneous network propagation and the integration of multiple drug and target similarity measures.

## 5.1 Conclusion

This thesis aims to enhance the heterogeneous network model by integrating multiple drug-drug and target-target similarity measures for identifying protein targets of drugs. The Forward Similarity Integration (FSI) framework is newly introduced heterogeneous network for systematically selecting drug and target similarity measures integrated into a model. Different drug and target similarity measures are generated based on various properties of drugs and target proteins. There are four data sets of drugs (i.e., chemical structures, DDAs, DDIs, and SEs) and four data sets of target proteins (i.e., protein sequences, PPIs, GO annotations, and protein pathways) used to create seven drug-drug similarity matrices and nine target-target similarity matrices, respectively. Moreover, several similarity integration methods and different selecting criteria are also investigated in this thesis. To find the suitable model with the suitable similarity integration, the FSI algorithm is applied.

As the result, the FSI model is derived from combining the Jaccard index of DDAs (DDA_Jac), the Cosine index of DDIs (DDI_Cos), and chemical structures of drugs (Structures) into a drug similarity by using the similarity network fusion (SNF) method. Additionally, the FSI model uses the local alignments of target protein sequences (Seq_Loc) as a target similarity measure. According to the FSI model, AUPR is used as the criteria for selecting the most suitable similarity measure and integration method. To show the superior efficiency of the FSI model, it was compared with the

conventional model and the models with full and random similarity integration. Obviously, the FSI model significantly outperforms those models. According to the case studies, it can be concluded that the FSI framework can be practically used for predicting new DTIs.

According to all results, it can be concluded that the heterogeneous network model with ensemble similarities can improve the DTI predictions. Furthermore, the FSI framework can efficiently construct the heterogeneous network propagation model with the suitable similarity integration. The FSI framework is not limited with only the similarity measures or integration methods used in this thesis. Other data sets of drugs (e.g., gaussian interaction profile and drug-protein interaction) and target proteins (e.g., G protein-coupled receptors, kinase superfamily, and nuclear receptors) can be used as choices of similarity measures in the FSI framework. Furthermore, other similarity integration methods (e.g., Nonlinear end-to-end learning model, and Multiple Similarities Collaborative Matrix Factorization) and selecting criteria (e.g., ACC and MCC) can be introduced into the FSI framework. The FSI framework can also be applied for other applications, apart from the DTI prediction, which require integration of multiple similarity measures.



**5.2 Future work**

The FSI framework proposed in this thesis is based on the heterogeneous network of two layers, a drug-drug similarity layer and a target-target similarity layer connecting together by the links of known drug-target interactions. To extend this work, other useful network layers, such as a protein-protein interaction network and a drug-drug interaction network, can be integrated into the drug-target heterogeneous network. Furthermore, other drug-drug similarity measures and target-target similarity measures can be derived based on new data sets and introduced into the FSI framework. Also, other similarity integration methods can be included in the FSI framework.

# REFERENCES


[1] M. Schenone, V. Dančík, B. K. Wagner, and P. A. Clemons, "Target identification and mechanism of action in chemical biology and drug discovery," *Nature chemical biology,* vol. 9, no. 4, pp. 232-240, 2013.

[2] F. Mohammadipanah and F. Salimi, "Potential biological targets for bioassay development in drug discovery of Sturge–Weber syndrome," *Chemical biology & drug design,* vol. 91, no. 2, pp. 359-369, 2018.

[3] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey," *Briefings in bioinformatics,* vol. 20, no. 4, pp. 1337-1357, 2019.

[4] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," *Nature biotechnology,* vol. 25, no. 2, pp. 197-206, 2007.

[5] L. Jacob and J.-P. Vert, "Protein-ligand interaction prediction: an improved chemogenomics approach," *bioinformatics,* vol. 24, no. 19, pp. 2149-2156, 2008.

[6] G. Pujadas *et al.*, "Protein-ligand docking: A review of recent advances and future perspectives," *Current Pharmaceutical Analysis,* vol. 4, no. 1, pp. 1-19, 2008.

[7] A. C. Cheng *et al.*, "Structure-based maximal affinity model predicts small-molecule druggability," *Nature biotechnology,* vol. 25, no. 1, pp. 71-75, 2007.

[8] H. Li *et al.*, "TarFisDock: a web server for identifying drug targets with docking approach," *Nucleic acids research,* vol. 34, no. suppl_2, pp. W219-W224, 2006.

[9] M. A. Yıldırım, 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.

[10] K. Sachdev and M. K. Gupta, "A comprehensive review of feature based methods for drug target interaction prediction," *Journal of biomedical informatics,* vol. 93, 2019.

[11] Z. Mousavian and A. Masoudi-Nejad, "Drug–target interaction prediction via chemogenomic space: learning-based methods," *Expert opinion on drug*

*metabolism & toxicology,* vol. 10, no. 9, pp. 1273-1287, 2014.

[12] K. Bleakley and Y. Yamanishi, "Supervised prediction of drug–target interactions using bipartite local models," *Bioinformatics,* vol. 25, no. 18, pp. 2397-2403, 2009.

[13] Y. Liu, M. Wu, C. Miao, P. Zhao, and X.-L. Li, "Neighborhood regularized logistic matrix factorization for drug-target interaction prediction," *PLoS computational biology,* vol. 12, no. 2, 2016.

[14] W. Wang, S. Yang, and J. Li, "Drug target predictions based on heterogeneous graph inference," in *Biocomputing 2013*: World Scientific, 2013, pp. 53-64.

[15] B. Liu, K. Pliakos, C. Vens, and G. Tsoumakas, "Drug-target interaction prediction via an ensemble of weighted nearest neighbors with interaction recovery," *Applied Intelligence,* vol. 52, no. 4, pp. 3705-3727, 2022.

[16] F. Wan, L. Hong, A. Xiao, T. Jiang, and J. Zeng, "NeoDTI: neural integration of neighbor information from a heterogeneous network for discovering new drug–target interactions," *Bioinformatics,* vol. 35, no. 1, pp. 104-111, 2019.

[17] 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.

[18] A. P. Davis *et al.*, "Comparative toxicogenomics database (CTD): update 2021," *Nucleic acids research,* vol. 49, no. D1, pp. D1138-D1143, 2021.

[19] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic acids research,* vol. 28, no. 1, pp. 27-30, 2000.

[20] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic acids research,* vol. 44, no. D1, pp. D1075-D1079, 2016.

[21] F. Cheng *et al.*, "Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space," *Journal of chemical information and modeling,* vol. 53, no. 4, pp. 753-762, 2013.

[22] Y. Ding, J. Tang, and F. Guo, "Identification of drug–target interactions via dual laplacian regularized least squares with multiple kernel fusion," *Knowledge-Based Systems,* vol. 204, 2020.

[23] N. Biggs, N. L. Biggs, C. U. Press, and B. Norman, *Algebraic Graph Theory*. Cambridge University Press, 1993.

[24] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of The Web Conference 2020*, 2020, pp. 2704-2710.

[25] X. Wang *et al.*, "Heterogeneous graph attention network," in *The world wide web conference*, 2019, pp. 2022-2032.

[26] C.-T. Li, S.-D. Lin, and M.-K. Shan, "Influence propagation and maximization for heterogeneous social networks," in *Proceedings of the 21st International Conference on World Wide Web*, 2012, pp. 559-560.

[27] X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in *Proceedings of the 2012 SIAM international conference on data mining*, 2012: SIAM, pp. 1119-1130.

[28] Z. Ali, G. Qi, K. Muhammad, A. Khalil, I. Ullah, and A. Khan, "Global citation recommendation employing multi-view heterogeneous network embedding," in *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, 2021: IEEE, pp. 1-6.

[29] W. Yang, X. Cui, J. Liu, Z. Wang, W. Zhu, and L. Wei, "User's interests-based movie recommendation in heterogeneous network," in *2015 International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI)*, 2015: IEEE, pp. 74-77.

[30] C. Shi, B. Hu, W. X. Zhao, and S. Y. Philip, "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering,* vol. 31, no. 2, pp. 357-370, 2018.

[31] Z. Liu *et al.*, "Heterogeneous Network Embedding for Deep Semantic Relevance Match in E-commerce Search," *arXiv preprint arXiv:2101.04850,* 2021.

[32] Y. Guan and M. Pollak, "Contagion in heterogeneous financial networks," *Advances in Complex Systems,* vol. 19, no. 01n02, 2016.

[33] 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, doi: 10.1093/bioinformatics/btw228.

[34] Y. Luo *et al.*, "A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information," *Nature communications,* vol. 8, no. 1, pp. 1-13, 2017.

[35] D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar, "Zoobp: Belief propagation for heterogeneous networks," *Proceedings of the VLDB Endowment,* vol. 10, no. 5, pp. 625-636, 2017.

[36] J. Gong, H. Zhang, and W. Du, "Research on Integrated Learning Fraud Detection Method Based on Combination Classifier Fusion (THBagging): A Case Study on the Foundational Medical Insurance Dataset," *Electronics,* vol. 9, no. 6, 2020.

[37] Q. An and L. Yu, "A heterogeneous network embedding framework for predicting similarity-based drug-target interactions," *Briefings in Bioinformatics,* vol. 22, no. 6, 2021.

[38] X. Chen, M.-X. Liu, and G.-Y. Yan, "Drug–target interaction prediction by random walk on the heterogeneous network," *Molecular BioSystems,* vol. 8, no. 7, pp. 1970-1978, 2012.

[39] A. Masoudi-Nejad, Z. Mousavian, and J. H. Bozorgmehr, "Drug-target and disease networks: polypharmacology in the post-genomic era," *In silico pharmacology,* vol. 1, no. 1, pp. 1-4, 2013.

[40] T. Cheng, M. Hao, T. Takeda, S. H. Bryant, and Y. Wang, "Large-scale prediction of drug-target interaction: a data-centric review," *The AAPS journal,* vol. 19, no. 5, pp. 1264-1275, 2017.

[41] M. Yang, H. Luo, Y. Li, and J. Wang, "Drug repositioning based on bounded nuclear norm regularization," *Bioinformatics,* vol. 35, no. 14, pp. i455-i463, 2019.

[42] X.-Y. Yan, S.-W. Zhang, and S.-Y. Zhang, "Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network," *Molecular BioSystems,* vol. 12, no. 2, pp. 520-531, 2016.

[43] M. A. Thafar *et al.*, "DTiGEMS+: drug–target interaction prediction using graph embedding, graph mining, and similarity-based techniques," *Journal of Cheminformatics,* vol. 12, no. 1, pp. 1-17, 2020.

[44] X. Zheng, H. Ding, H. Mamitsuka, and S. Zhu, "Collaborative matrix factorization with multiple similarities for predicting drug-target interactions," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 1025-1033.

[45] H. Liu, J. Sun, J. Guan, J. Zheng, and S. Zhou, "Improving compound–protein interaction prediction by building up highly credible negative samples," *Bioinformatics,* vol. 31, no. 12, pp. i221-i229, 2015.

[46] R. S. Olayan, H. Ashoor, and V. B. Bajic, "DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches," *Bioinformatics,* vol. 34, no. 7, pp. 1164-1173, 2018.

[47] C. Yan, J. Wang, W. Lan, F.-X. Wu, and Y. Pan, "Sdtrls: Predicting drug-target interactions for complex diseases based on chemical substructures," *Complexity,* vol. 2017, 2017.

[48] B. Wang *et al.*, "Similarity network fusion for aggregating data types on a genomic scale," *Nature methods,* vol. 11, no. 3, pp. 333-337, 2014.

[49] J. K. Narayana, M. Mac Aogáin, N. A. t. B. M. Ali, K. Tsaneva-Atanasova, and S. H. Chotirmall, "Similarity network fusion for the integration of multi-omics and microbiomes in respiratory disease," *European Respiratory Journal,* vol. 58, no. 2, 2021.

[50] G. Zhang, Z. Peng, C. Yan, J. Wang, J. Luo, and H. Luo, "A novel liver cancer diagnosis method based on patient similarity network and DenseGCN," *Scientific Reports,* vol. 12, no. 1, pp. 1-10, 2022.

[51] B. Wen *et al.*, "Systemic inflammation and metabolic disturbances underlie inpatient mortality among ill children with severe malnutrition," *Science advances,* vol. 8, no. 7, 2022.

[52] M. Mac Aogáin *et al.*, "Integrative microbiomics in bronchiectasis exacerbations," *Nature Medicine,* vol. 27, no. 4, pp. 688-699, 2021.

[53] C.-X. Li, C. E. Wheelock, C. M. Sköld, and Å. M. Wheelock, "Integration of multi-omics datasets enables molecular classification of COPD," *European Respiratory Journal,* vol. 51, no. 5, 2018.

[54] N. N. Narisetty, "Bayesian model selection for high-dimensional data," in *Handbook of Statistics*, vol. 43: Elsevier, 2020, pp. 207-248.

[55] G. Choueiry. "Understand Forward and Backward Stepwise Regression." https://quantifyinghealth.com/stepwise-selection/ (accessed 16/052022, 2022).

[56] S. Fallahpour, E. N. Lakvan, and M. H. Zadeh, "Using an ensemble classifier

based on sequential floating forward selection for financial distress prediction problem," *Journal of Retailing and Consumer Services,* vol. 34, pp. 159-167, 2017.

[57] J. Bauweraerts, "Predicting bankruptcy in private firms: Towards a stepwise regression procedure," *International Journal of Financial Research,* vol. 7, no. 2, pp. 147-153, 2016.

[58] J. T. Dang *et al.*, "Predicting surgical site infections following laparoscopic bariatric surgery: development of the BariWound tool using the MBSAQIP database," *Surgical Endoscopy,* vol. 34, no. 4, pp. 1802-1811, 2020.

[59] K. Chien *et al.*, "Prediction models for the risk of new-onset hypertension in ethnic Chinese in Taiwan," *Journal of human hypertension,* vol. 25, no. 5, pp. 294-303, 2011.

[60] M. Noryani, S. M. Sapuan, M. T. Mastura, M. Y. M. Zuhri, and E. S. Zainudin, "Material selection of natural fibre using a stepwise regression model with error analysis," *Journal of Materials Research and Technology,* vol. 8, no. 3, pp. 2865-2879, 2019.

[61] P. Goos and D. Meintrup, *Statistics with JMP: hypothesis tests, ANOVA and regression*. John Wiley & Sons, 2016.

[62] N. Gleichmann, "Paired vs Unpaired T-Test: Differences, Assumptions and Hypotheses," *Technology Networks,* 2020.

[63] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *Journal of chemical information and computer sciences,* vol. 28, no. 1, pp. 31-36, 1988.

[64] U. D. o. Health and H. Services, "AIDS Info Glossary of HIV/AIDS Related Terms," ed: Rockville, MD: HHS, 2018.

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

[66] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork, "STITCH: interaction networks of chemicals and proteins," *Nucleic acids research,* vol. 36, no. suppl_1, pp. D684-D688, 2007.

[67] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research,* vol. 32, no. suppl_1, pp. D267-D270, 2004.

[68] R. Huntley, E. Dimmer, D. Barrell, D. Binns, and R. Apweiler, "The gene ontology annotation (goa) database," *Nature Precedings,* pp. 1-1, 2009.

[69] D. Szklarczyk *et al.*, "The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets," *Nucleic acids research,* vol. 49, no. D1, pp. D605-D612, 2021.

[70] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association,* vol. 88, no. 3, 2000.

[71] G. B. Whitworth, "An introduction to microarray data analysis and visualization," in *Methods in enzymology*, vol. 470: Elsevier, 2010, pp. 19-50.

[72] M. Milano, "Gene Prioritization Tools," 2019.

[73] J. I. Castrillo, P. Pir, and S. G. Oliver, "Yeast Systems Biology: towards a systems understanding of regulation of eukaryotic networks in complex diseases and biotechnology," in *Handbook of systems biology*: Elsevier, 2013, pp. 343-365.

[74] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics," *Journal of chemical information and computer sciences,* vol. 43, no. 2, pp. 493-500, 2003.

[75] T. T. Tanimoto, "Elementary mathematical theory of classification and prediction," 1958.

[76] P. Zhao *et al.*, "Targets preliminary screening for the fresh natural drug molecule based on Cosine-correlation and similarity-comparison of local network," *Journal of Translational Medicine,* vol. 20, no. 1, pp. 1-9, 2022.

[77] H. Öztürk, E. Ozkirimli, and A. Özgür, "A comparative study of SMILES-based compound similarity functions for drug-target interaction prediction," *BMC bioinformatics,* vol. 17, no. 1, pp. 1-11, 2016.

[78] 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, 2011.

[79] L. Perlman, A. Gottlieb, N. Atias, E. Ruppin, and R. Sharan, "Combining drug and gene similarity measures for drug-target elucidation," *Journal of computational biology,* vol. 18, no. 2, pp. 133-145, 2011.

[80] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology,* vol. 147, no. 1, pp. 195-197, 1981.

[81] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology,* vol. 48, no. 3, pp. 443-453, 1970.

[82] T. Güyer, B. Atasoy, and S. Somyürek, "Measuring disorientation based on the Needleman-Wunsch algorithm," *International Review of Research in Open and Distributed Learning,* vol. 16, no. 2, pp. 188-205, 2015.

[83] H. A. Pagès, P. Gentleman, and R. DebRoy, "R package version 2.56. 0," *Biostrings: Efficient manipulation of biological strings,* 2020.

[84] 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.

[85] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *In the Proceedings of ROCLING X,* 1997.

[86] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics,* vol. 26, no. 7, pp. 976-978, 2010.

[87] Y. Shen, S. Zhang, and H.-S. Wong, "A new method for measuring the semantic similarity on gene ontology," in *2010 IEEE international conference on bioinformatics and biomedicine (BIBM)*, 2010: IEEE, pp. 533-538.

[88] K. Ovaska, M. Laakso, and S. Hautaniemi, "Fast gene ontology based clustering for microarray experiments," *BioData mining,* vol. 1, no. 1, pp. 1-8, 2008.

[89] A. Banu, S. S. Fatima, and K. U. R. Khan, "Information Content Based Semantic Similarity Measure for Concepts Subsumed By Multiple Concepts," *Int. J. Web Appl.,* vol. 7, no. 3, pp. 85-94, 2015.

[90] C. Zhao and Z. Wang, "GOGO: An improved algorithm to measure the semantic

similarity between gene ontology terms," *Scientific reports,* vol. 8, no. 1, pp. 1-10, 2018.

[91] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik,* vol. 1, no. 1, pp. 269-271, 1959.

[92] Y. Li, Y.-A. Huang, Z.-H. You, L.-P. Li, and Z. Wang, "Drug-target interaction prediction based on drug fingerprint information and protein sequence," *Molecules,* vol. 24, no. 16, 2019.

[93] W. Lan, J. Wang, M. Li, F.-X. Wu, and Y. Pan, "Predicting drug-target interaction based on sequence and structure information," *IFAC-PapersOnLine,* vol. 48, no. 28, pp. 12-16, 2015.

[94] I. Lee and H. Nam, "Sequence-based prediction of protein binding regions and drug–target interactions," *Journal of cheminformatics,* vol. 14, no. 1, pp. 1-15, 2022.

[95] Z.-H. Chen, Z.-H. You, Z.-H. Guo, H.-C. Yi, G.-X. Luo, and Y.-B. Wang, "Prediction of drug–target interactions from multi-molecular network based on deep walk embedding model," *Frontiers in Bioengineering and Biotechnology,* vol. 8, 2020.

[96] K. Shao, Y. Zhang, Y. Wen, Z. Zhang, S. He, and X. Bo, "DTI-HETA: prediction of drug–target interactions based on GCN and GAT on heterogeneous graph," *Briefings in Bioinformatics,* 2022.

[97] T. N. Jarada, J. G. Rokne, and R. Alhajj, "SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder," *Knowledge-Based Systems,* vol. 212, 2021.

[98] T. N. Jarada, J. G. Rokne, and R. Alhajj, "SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks," *BMC bioinformatics,* vol. 22, no. 1, pp. 1-20, 2021.

[99] G. Wang, H. Wang, F. Guo, M. Du, and C. Cao, "A novel method for drug-target interaction prediction based on graph transformers model," 2022.

[100] L. Jiang, J. Sun, Y. Wang, Q. Ning, N. Luo, and M. Yin, "Identifying drug–target interactions via heterogeneous graph attention networks combined with cross-modal similarities," *Briefings in Bioinformatics,* vol. 23, no. 2, 2022.

[101] M. Whittle, V. J. Gillet, P. Willett, A. Alex, and J. Loesel, "Enhancing the

effectiveness of virtual screening by fusing nearest neighbor lists: a comparison of similarity coefficients," *Journal of chemical information and computer sciences,* vol. 44, no. 5, pp. 1840-1848, 2004.

[102] H. Matter, "Selecting optimally diverse compounds from structure databases: a validation study of two-dimensional and three-dimensional molecular descriptors," *Journal of Medicinal Chemistry,* vol. 40, no. 8, pp. 1219-1229, 1997.

[103] M. A. Johnson and G. M. Maggiora, *Concepts and applications of molecular similarity*. Wiley, 1990.

[104] S. Lobert, B. Vulevic, and J. J. Correia, "Interaction of vinca alkaloids with tubulin: a comparison of vinblastine, vincristine, and vinorelbine," *Biochemistry,* vol. 35, no. 21, pp. 6806-6814, 1996.

[105] R. D. Arora and R. G. Menezes, "Vinca alkaloid toxicity," in *StatPearls [Internet]*: StatPearls Publishing, 2021.

[106] S. Lobert, J. W. Ingram, B. T. Hill, and J. J. Correia, "A comparison of thermodynamic parameters for vinorelbine-and vinflunine-induced tubulin self-association by sedimentation velocity," *Molecular pharmacology,* vol. 53, no. 5, pp. 908-915, 1998.

[107] L. M. A. Aparicio, E. G. Pulido, and G. A. Gallego, "Vinflunine: a new vision that may translate into antiangiogenic and antimetastatic activity," *Anti-cancer drugs,* vol. 23, no. 1, pp. 1-11, 2012.

[108] H. Solana, J. Sallovitz, C. Lanusse, and J. Rodriguez, "Enantioselective binding of albendazole sulphoxide to cytosolic proteins from helminth parasites," *Methods and findings in experimental and clinical pharmacology,* vol. 24, no. 1, pp. 7-14, 2002.

[109] H. M. Berman *et al.*, "The protein data bank," *Nucleic acids research,* vol. 28, no. 1, pp. 235-242, 2000.

[110] T. Ramírez, L. Benítez-Bribiesca, P. Ostrosky-Wegman, and L. A. Herrera, "In vitro effects of albendazole and its metabolites on the cell proliferation kinetics and micronuclei frequency of stimulated human lymphocytes," *Archives of medical research,* vol. 32, no. 2, pp. 119-122, 2001.

[111] S. W. Chu, S. Badar, D. L. Morris, and M. H. Pourgholami, "Potent inhibition of

tubulin polymerisation and proliferation of paclitaxel-resistant 1A9PTX22 human ovarian cancer cells by albendazole," *Anticancer research,* vol. 29, no. 10, pp. 3791-3796, 2009.

[112] A. S. Vallés, I. Garbus, and F. J. Barrantes, "Lamotrigine is an open-channel blocker of the nicotinic acetylcholine receptor," *Neuroreport,* vol. 18, no. 1, pp. 45-50, 2007.

[113] N. Nishimura *et al.*, "Effects of nicorandil on the cAMP-dependent Cl− current in guinea-pig ventricular cells," *Journal of pharmacological sciences,* 2010.

[114] K. S. Prabhavalkar, N. B. Poovanpallil, and L. K. Bhatt, "Management of bipolar depression with lamotrigine: an antiepileptic mood stabilizer," *Frontiers in pharmacology,* vol. 6, 2015.

[115] M. W. Holladay, M. J. Dart, and J. K. Lynch, "Neuronal nicotinic acetylcholine receptors as targets for drug discovery," *Journal of medicinal chemistry,* vol. 40, no. 26, pp. 4169-4194, 1997.

[116] N. D. Cosford *et al.*, "(S)-(−)-5-Ethynyl-3-(1-methyl-2-pyrroli-dinyl) pyridine Maleate (SIB-1508Y): A Novel Anti-Parkinsonian Agent with Selectivity for Neuronal Nicotinic Acetylcholine Receptors," *Journal of medicinal chemistry,* vol. 39, no. 17, pp. 3235-3237, 1996.

[117] S. Kaushal and P. Tadi, "Nicotinic Ganglionic Blocker," 2020.

[118] R. E. Wittenberg, S. L. Wolfman, M. De Biasi, and J. A. Dani, "Nicotinic acetylcholine receptors and nicotine addiction: A brief introduction," *Neuropharmacology,* vol. 177, 2020.

[119] C. Xiao, C.-y. Zhou, J.-h. Jiang, and C. Yin, "Neural circuits and nicotinic acetylcholine receptors mediate the cholinergic regulation of midbrain dopaminergic neurons and nicotine dependence," *Acta Pharmacologica Sinica,* vol. 41, no. 1, pp. 1-9, 2020.

[120] J. R. Nickell, V. P. Grinevich, K. B. Siripurapu, A. M. Smith, and L. P. Dwoskin, "Potential therapeutic uses of mecamylamine and its stereoisomers," *Pharmacology Biochemistry and Behavior,* vol. 108, pp. 28-43, 2013.

[121] W. M. Hooten and D. O. Warner, "Varenicline for opioid withdrawal in patients with chronic pain: a randomized, single-blinded, placebo controlled pilot trial,"

*Addictive Behaviors,* vol. 42, pp. 69-72, 2015.

[122] A. A. Weinbroum, V. Rudick, G. Paret, and R. Ben-Abraham, "The role of dextromethorphan in pain control," *Canadian Journal of Anesthesia,* vol. 47, no. 6, pp. 585-596, 2000.

[123] E. C. Eyler, "Chronic and acute pain and pain management for patients in methadone maintenance treatment," *The American Journal on Addictions,* vol. 22, no. 1, pp. 75-83, 2013.

[124] S. Ali, B. Tahir, S. Jabeen, and M. Malik, "Methadone treatment of opiate addiction: a systematic review of comparative studies," *Innovations in clinical Neuroscience,* vol. 14, no. 7-8, 2017.

[125] H. Koyuncuoğlu and B. Saydam, "The treatment of heroin addicts with dextromethorphan: a double-blind comparison of dextromethorphan with chlorpromazine," *International journal of clinical pharmacology, therapy, and toxicology,* vol. 28, no. 4, pp. 147-152, 1990.

[126] R. Nocente, M. Vitali, G. Balducci, D. Enea, H. R. Kranzler, and M. Ceccanti, "Varenicline and neuronal nicotinic acetylcholine receptors: A new approach to the treatment of co-occurring alcohol and nicotine addiction?," *The American Journal on Addictions,* vol. 22, no. 5, pp. 453-459, 2013.

[127] J.-H. Lee *et al.*, "Effects of dextrorotatory morphinans on $\alpha_3\beta_4$ nicotinic acetylcholine receptors expressed in Xenopus oocytes," *European journal of pharmacology,* vol. 536, no. 1-2, pp. 85-92, 2006.

[128] M. I. Damaj, P. Flood, K. Ho, E. L. May, and B. R. Martin, "Effect of dextrometorphan and dextrorphan on nicotine and neuronal nicotinic receptors: in vitro and in vivo selectivity," *Journal of Pharmacology and Experimental Therapeutics,* vol. 312, no. 2, pp. 780-785, 2005.

[129] P. Steensland, J. A. Simms, J. Holgate, J. K. Richards, and S. E. Bartlett, "Varenicline, an $\alpha_4\beta_2$ nicotinic acetylcholine receptor partial agonist, selectively decreases ethanol consumption and seeking," *Proceedings of the National Academy of Sciences,* vol. 104, no. 30, pp. 12518-12523, 2007.

[130] K. B. Mihalak, F. I. Carroll, and C. W. Luetje, "Varenicline is a partial agonist at

α4β2 and a full agonist at α7 neuronal nicotinic receptors," *Molecular pharmacology,* vol. 70, no. 3, pp. 801-805, 2006.

[131] R. Talka, R. K. Tuominen, and O. Salminen, "Methadone's effect on nAChRs—a link between methadone use and smoking?," *Biochemical Pharmacology,* vol. 97, no. 4, pp. 542-549, 2015.

[132] R. Talka, O. Salminen, and R. K. Tuominen, "Methadone is a Non-Competitive Antagonist at the α4β2 and α3* Nicotinic Acetylcholine Receptors and an Agonist at the α7 Nicotinic Acetylcholine Receptor," *Basic & Clinical Pharmacology & Toxicology,* vol. 116, no. 4, pp. 321-328, 2015.

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

[134] L. E. DeWald *et al.*, "The calcium channel blocker bepridil demonstrates efficacy in the murine model of marburg virus disease," *The Journal of infectious diseases,* vol. 218, no. suppl_5, pp. S588-S591, 2018.

[135] A. B. Bansal and G. Khandelwal, "Felodipine," 2019.

[136] P. Herych and R. Yatsyshyn, "Optimizing Treatment of Patients with Co-Existing Cardiorespiratory Pathology by Administration of Anti-Inflammatory Roflumilast and Cardioprotective Agent Quercetin," *Galician Medical Journal,* vol. 21, no. 3, pp. 65-69, 2014.

[137] S. Ekins, A. C. Puhl, and A. Davidow, "Repurposing the dihydropyridine calcium channel inhibitor nicardipine as a Nav1. 8 inhibitor in vivo for Pitt Hopkins syndrome," *Pharmaceutical research,* vol. 37, no. 7, pp. 1-9, 2020.

[138] H. Y. Zhao, Y. H. Ren, X. B. Ren, and Y. Wang, "Diprophylline inhibits non-small cell lung cancer A549 cell proliferation and migration, and promotes apoptosis, by downregulating PI3K signaling pathway," *Oncology Letters,* vol. 17, no. 1, pp. 857-862, 2019.

[139] R. Ding, J. Shi, K. Pabon, and K. W. Scotto, "Xanthines down-regulate the drug transporter ABCG2 and reverse multidrug resistance," *Molecular pharmacology,* vol. 81, no. 3, pp. 328-337, 2012.

[140] R. M. Basnet, D. Zizioli, M. Guarienti, D. Finazzi, and M. Memo, "Methylxanthines induce structural and functional alterations of the cardiac system in zebrafish embryos," *BMC Pharmacology and Toxicology,* vol. 18, no. 1, pp. 1-12, 2017.

[141] B. Gottwalt and P. Tadi, "Methylxanthines," in *StatPearls [Internet]*: StatPearls Publishing, 2021.

[142] L. A. Ahmed, "Nicorandil: A drug with ongoing benefits and different mechanisms in various diseased conditions," *Indian Journal of Pharmacology,* vol. 51, no. 5, 2019.

[143] H. P. Rang, M. M. Dale, J. M. Ritter, R. J. Flower, and G. Henderson, *Rang & Dale's pharmacology*. Elsevier Health Sciences, 2011.

[144] D. Guo *et al.*, "Modeling congenital hyperinsulinism with ABCC8-deficient human embryonic stem cells generated by CRISPR/Cas9," *Scientific reports,* vol. 7, no. 1, pp. 1-8, 2017.

[145] P. Radivojac *et al.*, "A large-scale evaluation of computational protein function prediction," *Nature methods,* vol. 10, no. 3, pp. 221-227, 2013.

[146] E. De Clercq, "Mozobil®(Plerixafor, AMD3100), 10 years after its approval by the US Food and Drug Administration," *Antiviral Chemistry and Chemotherapy,* vol. 27, 2019.

[147] M. Ilmer *et al.*, "Stories of drug repurposing for pancreatic cancer treatment—Past, present, and future," in *Drug Repurposing in Cancer Therapy*: Elsevier, 2020, pp. 231-272.

[148] R. Innis-Shelton and L. J. Costa, "Sources of Cells for Hematopoietic Cell Transplantation: Practical Aspects of Hematopoietic Cell Collection," in *Hematopoietic Cell Transplantation for Malignant Conditions*: Elsevier, 2019, pp. 73-84.

[149] A. Dar *et al.*, "Rapid mobilization of hematopoietic progenitors by AMD3100 and catecholamines is mediated by CXCR4-dependent SDF-1 release from bone marrow stromal cells," *Leukemia,* vol. 25, no. 8, pp. 1286-1296, 2011.

[150] N. H. Agha *et al.*, "Vigorous exercise mobilizes CD34+ hematopoietic stem cells to peripheral blood via the β2-adrenergic receptor," *Brain, behavior, and*

*immunity,* vol. 68, pp. 66-75, 2018.

[151] N. Asada *et al.*, "Matrix-embedded osteocytes regulate mobilization of hematopoietic stem/progenitor cells," *Cell stem cell,* vol. 12, no. 6, pp. 737-747, 2013.

[152] F. L. Baker *et al.*, "Systemic β-adrenergic receptor activation augments the ex vivo expansion and anti-tumor activity of Vγ9Vδ2 T-cells," *Frontiers in immunology,* 2020.

# VITA

| | |
|---|---|
| **NAME** | Piyanut Tangmanussukum |
| **DATE OF BIRTH** | 18 October 1996 |
| **PLACE OF BIRTH** | Bangkok, Thailand |
| **INSTITUTIONS ATTENDED** | B.Sc. (Computational Science) Walailak University, 2019 |
| **HOME ADDRESS** | 23 Phonrak-utid Road, Huai Yot, Trang |
| **PUBLICATION** | 1. P. Tangmanussukum, T. Kawichai, A. Suratanee, and K. Plaimas, "Heterogeneous Network Propagation with Optimal Similarity Measure for Drug-Target Associations," in 2021 25th International Computer Science and Engineering Conference (ICSEC), 2021: IEEE, pp. 155-160.<br>2. P. Tangmanussukum, T. Kawichai, A. Suratanee, and K. Plaimas, " Link Prediction of Drug-Target Interactions Using Heterogeneous Network Propagation with Forward Similarity Integration", submitted. |
| **AWARD RECEIVED** | Full scholarships from Development and Promotion of Science and Technology Talents Project (DPST) |