Predicting potential drugs and drug-drug interactions for drug repositioning

A dissertation submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Division of Biomedical Engineering
University of Saskatchewan
Saskatoon

By
Fei Wang

Permission to Use

In presenting this dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my dissertation work was done. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my dissertation.

Disclaimer

Reference in this dissertation to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes.

Requests for permission to copy or to make other uses of materials in this dissertation in whole or part should be addressed to:

Head of the Devision of Biomedical Engineering
Engineering Building
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5C9
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

The purpose of drug repositioning is to predict novel treatments for existing drugs. It saves time and reduces cost in drug discovery, especially in preclinical procedures. In drug repositioning, the challenging objective is to identify reasonable drugs with strong evidence. Recently, benefiting from various types of data and computational strategies, many methods have been proposed to predict potential drugs.

Signature-based methods use signatures to describe a specific disease condition and match it with drug-induced transcriptomic profiles. For a disease signature, a list of potential drugs is produced based on matching scores. In many studies, the top drugs on the list are identified as potential drugs and verified in various ways. However, there are a few limitations in existing methods: (1) For many diseases, especially cancers, the tissue samples are often heterogeneous and multiple subtypes are involved. It is challenging to identify a signature from such a group of profiles. (2) Genes are treated as independent elements in many methods, while they may associate with each other in the given condition. (3) The disease signatures cannot identify potential drugs for personalized treatments.

In order to address those limitations, I propose three strategies in this dissertation. (1) I employ clustering methods to identify sub-signatures from the heterogeneous dataset, then use a weighting strategy to concatenate them together. (2) I utilize human protein complex (HPC) information to reflect the dependencies among genes and identify an HPC signature to describe a specific type of cancer. (3) I use an HPC strategy to identify signatures for drugs, then predict a list of potential drugs for each patient.

Besides predicting potential drugs directly, more indications are essential to enhance my understanding in drug repositioning studies. The interactions between biological and biomedical entities, such as drug-drug interactions (DDIs) and drug-target interactions (DTIs), help study mechanisms behind the repurposed drugs. Machine learning (ML), especially deep learning (DL), are frontier methods in predicting those interactions. Network strategies, such as constructing a network from interactions and studying topological properties, are commonly used to combine with other methods to make predictions. However, the interactions may have different functions, and merging them in a single network may cause some biases. In order to solve it, I construct two networks for two types of DDIs and employ a graph convolutional network (GCN) model to concatenate them together.

In this dissertation, the first chapter introduces background information, objectives of studies, and structure of the dissertation. After that, a comprehensive review is provided in Chapter 2. Biological databases, methods and applications in drug repositioning studies, and evaluation metrics are discussed. I summarize three application scenarios in Chapter 2.

The first method proposed in Chapter 3 considers the issue of identifying a cancer gene signature and predicting potential drugs. The k-means clustering method is used to identify highly reliable gene signatures. The identified signature is used to match drug profiles and identify potential drugs for the given disease. The second method proposed in Chapter 4 uses human protein complex (HPC) information to identify a protein complex signature, instead of a gene signature. This strategy improves the prediction accuracy in the experiments of cancers. Chapter 5 introduces the signature-based method in personalized cancer medicine. The profiles of a given drug are used to identify a drug signature, under the HPC strategy. Each patient has a profile, which is matched with the drug signature. Each patient has a different list of potential drugs. Chapter 6 propose a graph convolutional network with multi-kernel to predict DDIs. This method constructs two DDI kernels and concatenates them in the GCN model. It achieves higher performance in predicting DDIs than three state-of-the-art methods.

In summary, this dissertation has proposed several computational algorithms for drug repositioning. Experimental results have shown that the proposed methods can achieve very good performance.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Prof. Fang-Xiang Wu, who offered me the opportunity to pursue a Ph.D. degree in his lab. Throughout my studies, he constantly provides me insightful suggestions, not only for research but also for my daily life. This work would not have been possible without his help.

I would also like to express my gratitude to other members of my advisory committee, Prof. Ian McQuillan, Prof. Debajyoti Mondal, and Prof. Meena Sakharkar. I have received positive support and valuable suggestions from them during the Ph.D. program. Their courses also helped me lay a solid foundation for my study.

In addition, I am lucky to learn and work with a group of great colleagues. I am particularly indebted to Lin Wu, Yichao Shen, Ping Luo, Lingkai Tang, Yulian Ding, Pi-Jing Wei, Wenjing Zhang, Ming Fang, Yuchen Zhang, Hongyu Guo, Yuting Tan, Ali Akbar Hamali Beyrami, Sakib Mostafa, and Rayyan Khan, for their help in both my life and research.

Finally, I would like to thank all my families for their unconditional support and love. I am also grateful for the financial support provided by the Natural Science and Engineering Research Council of Canada (NSERC) and the China Scholarship Council (CSC).

This dissertation is dedicated to my father Jianzhong Wang and my mother Meiqin Liu, who encouraged me all these years.

Contents

Permission to Use **i**

Abstract **ii**

Acknowledgements **iv**

Contents **vi**

List of Tables **viii**

List of Figures **ix**

List of Abbreviations **x**

1 Introduction **1**
- 1.1 Background 1
- 1.2 Motivations and objectives 2
- 1.3 Organization of the dissertation 3

2 Drug repositioning: computational methods, databases and evaluations **4**
- 2.1 Introduction 5
- 2.2 Biological data 8
- 2.3 Computational methods 12
  - 2.3.1 Signature-based methods 12
  - 2.3.2 Basic machine learning strategies 13
  - 2.3.3 Deep learning strategies 19
- 2.4 Applications in drug repositioning 20
  - 2.4.1 Drug-disease association predictions 21
  - 2.4.2 Drug-drug association predictions 24
  - 2.4.3 Drug-target association predictions 27
- 2.5 Evaluation methods 31
- 2.6 Perspectives and conclusions 32

3 Identifying gene signatures for cancer drug repositioning based on sample clustering **34**
- 3.1 Introduction 35
- 3.2 Methods and materials 37
  - 3.2.1 The GS4CDRSC framework 38
  - 3.2.2 The sample clustering 38
  - 3.2.3 The DEG identification for each subset 39
  - 3.2.4 The gene signature determination 42
  - 3.2.5 The connection score calculation 43
  - 3.2.6 Datasets 44
- 3.3 Results and discussion 46
  - 3.3.1 Cluster analysis 47
  - 3.3.2 Statistical analysis 49
  - 3.3.3 Experiments 49
  - 3.3.4 Overlaps of the signatures 50
  - 3.3.5 Annotations of the known drugs 51
- 3.4 Conclusion 55

4 Human protein complex signatures for drug repositioning 57
4.1 Introduction . . . 57
4.2 Methods and materials . . . 59
4.2.1 Datasets . . . 59
4.2.2 HPCs . . . 61
4.2.3 Data pre-processing . . . 62
4.2.4 HPC signatures . . . 63
4.2.5 Matching method . . . 64
4.3 Results and discussion . . . 65
4.3.1 Parameters and performance evaluation . . . 65
4.3.2 Compared with other methods . . . 65
4.3.3 Analysis of predictions . . . 66
4.4 Conclusion . . . 69

5 Human protein complex-based drug signatures for personalized cancer medicine 71
5.1 Introduction . . . 72
5.2 Methods and materials . . . 73
5.2.1 Design of study . . . 74
5.2.2 Datasets and pre-processing . . . 75
5.2.3 HPC-based drug signature procedures . . . 76
5.2.4 Matching procedure . . . 78
5.2.5 Evaluation metrics . . . 78
5.3 Results and discussion . . . 79
5.3.1 Lung cancer . . . 82
5.3.2 Breast cancer . . . 83
5.3.3 Colorectal cancer . . . 84
5.3.4 Prostate cancer . . . 85
5.3.5 Discussion . . . 86
5.4 Conclusion . . . 86

6 Predicting drug-drug interactions by graph convolutional network with multi-kernel 87
6.1 Introduction . . . 88
6.2 Methods and materials . . . 89
6.2.1 DDI graphs and drug feature matrix . . . 90
6.2.2 Feature representations of drugs . . . 91
6.2.3 Predicting DDIs . . . 92
6.2.4 Datasets . . . 92
6.3 Results and discussion . . . 93
6.3.1 Experimental setting . . . 94
6.3.2 Visualization analysis of embedding features . . . 96
6.3.3 Results . . . 96
6.3.4 Case studies . . . 100
6.4 Conclusion . . . 101

7 Summary, limitations, and future work 104
7.1 Summary . . . 104
7.2 Limitations . . . 105
7.3 Future work . . . 105

References 107

Appendix A List of Publications 133

Appendix B Copyright Permissions 134

List of Tables

2.1 Drug-, Disease- and Protein-Centric Databases. 8
2.2 The feature types and similarities of drug-drug, disease-disease, and target-target associations 11
2.3 The introduced strategies 12
2.4 The confusion table. 31

3.1 The number of samples and platforms in each dataset 45
3.2 The average statistical powers and the number of tumor-normal sample pairs in all clusters and datasets. 47
3.3 The prediction rates by two types of gene signatures identified in six cancer cases and three approaches. 48
3.4 The rates of the overlapped genes in the signatures from the approaches with our proposed framework, compared to the approaches without our proposed framework. 49
3.5 The rates of overlapped genes between two or three approaches in all cases. All the approaches are combined with our proposed framework. 50
3.6 The known and potential drugs of three cancers identified by the three approaches with GS4CDRSC 51

4.1 The disease datasets 61
4.2 The number of known drugs identified by our HPCDR method and two gene signature method 66
4.3 The drugs identified by our HPCDR method 67

5.1 The information of datasets 77
5.2 The prediction rates of the five approaches 81
5.3 The drugs predicted for lung cancer 82
5.4 The drugs predicted for Breast cancer 84
5.5 The drugs predicted for colorectal cancer 85
5.6 The drugs predicted for prostate cancer 86

6.1 The types of features and their dimensions 93
6.2 The prediction performances of the competing methods. 97
6.3 The top 20 predicted DDIs. 97
6.4 The top ten predicted DDIs of breast neoplasms-related drugs. 99
6.5 The top ten predicted DDIs of colorectal neoplasms-related drugs. 102
6.6 The top ten predicted DDIs of lung neoplasms-related drugs. 103

List of Figures

2.1 The drug discovery pipeline. 5
2.2 The workflow of data, methods, and applications in drug repositioning. 7
2.3 An example of a drug-disease heterogeneous network. The solid lines denote the known drug-disease associations, and the weights of dotted lines denote the similarities. Six different weight values are exemplified. 10
2.4 Examples of logistic regression (a) and support vector machine (b). 14
2.5 Examples of decision tree (a) and random forest (b). 15
2.6 The structures of bagging (a) and boosting (b). 16
2.7 Examples of k-nearest neighbor (a), k-means clustering (b), and random walk (c). 17
2.8 The structures of basic neural network (a) and basic autoencoder (b). 18
2.9 The structures of convolutional neural network (a) and recurrent neural network (b). 19
2.10 The structures of Deep autoencoder (a) and Generative adversarial network (b). 21
2.11 The structure of CDF. 27

3.1 The flowchart of the GS4CDRSC framework. 37
3.2 The Silhouette values in each dataset. k is ranging from 2 to 10. 46

4.1 The flow chart of our HPCDR method. (A): The drug perturbation profiles are from the LINCS database. (B): HPCs are selected from the CORUM database. The number of satisfied HPCs is 2,064. (C): Microarray data were downloaded from Gene Expression Omnibus (GEO) database. The microarray data is mapped to the Entrez genes profile. (D): Based on the HPCs, drug perturbation profiles in the LINCS database are transformed into drug perturbation-HPC expression profiles. (E): Based on the HPCs, the Entrez gene expression profiles of disease are transformed into disease-HPC expression profiles. (F): An HPC signature is identified from the disease-HPC expression profiles. (G): A connection method is used to calculate 152,290 connection scores between the HPC signature and profiles. (H): The connection scores are sorted in ascending order and the top 20 perturbations are identified as drug candidates. . . . 60
4.2 The details of the conversion. A: From a gene expression profile of a disease to a profile of Entrez genes. B: From a profile of Entrez genes to that of HPCs. 62

5.1 The flowchart of our HPC-based drug signature approach. I: Producing an HPC signature for each drug. II: Transforming patient gene expression profiles to HPC profiles. III: Matching drug HPC signatures to the patient HPC profiles, and producing a list of candidate drugs. p represents a patient and N is the number of patients, u is a drug profile and s is a merged profile for each drug. The number of approved drugs in LINCS is 1,294. 74
5.2 The statistic of drugs and HPCs in the dataset. 75
5.3 The prediction rates between three types of weighting approaches. 79
5.4 The plot of the prediction rate vs. the length of signature. 79
5.5 The iteration steps when identifying a PRL from the profiles of a drug. 80

6.1 The architecture of GCNMK. **I**: Constructing two DDI graphs from increased, decreased interactions, and inputting drug attributes. **II**: Generating the feature representation of drugs by GCN. **III**: Predicting DDIs. 90
6.2 The influence of learning rate lr. 93
6.3 The influence of L_2-regularization coefficient λ. 94
6.4 The influence of embedding size d. 95
6.5 The influence of feature type. 95
6.6 The visualization analysis of embedding features. 96

List of Abbreviations

ALK	Anaplastic Lymphoma Kinase
AMP	Adenosine MonoPhosphate
AI	Artificial Intelligence
ATR	Ataxia Telangiectasia mutated- and Rad3-related
AUC-ROC	Area Under the ROC Curve
AUC-PR	Area Under the Precision-Recall Curve
BC	Breast Cancer
BH	Benjamini-Hochberg
CC	Cervical Cancer
CCR5	C-C chemokine receptor type 5
CDK4	Cyclin-Dependent Kinases 4
CMap	Connectivity Map
CNN	Convolutional Neural Network
CNS	Central Nervous System
CORUM	COmprehensive Resource of Mammalian protein complex
CRC	Colorectal Cancer
CTD	Comparative Toxicogenomics Database
DAE	Deep AutoEncoder
DBN	Deep-Belief Network
DDA	Drug-Disease Association
DDI	Drug-Drug Interaction
DEG	Differentially Expressed Gene
DF	Deep Forest
DL	Deep Learning
DNA	DeoxyriboNucleic Acid
DNN	Deep Neural Network
DTI	Drug-Target Interaction
EGFR	Epidermal Growth Factor Receptor
ELM	Extreme Learning Machine
FC	Fold-Change
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive

GAE	Graph AutoEncoder
GAN	Generative Adversarial Network
GBM	Gradient Boosting Machine
GCN	Graph Convolutional Network
GCNMK	Graph Convolutional Network with Multi-kernel
GEO	Gene Expression Omnibus
GEP	Gene-knockdown Expression Profile
GIP	Gaussian interaction profile
GNN	Graph Neural Network
GO	Gene Ontology
GSEA	Gene Set Enrichment Analysis
HDAC	Histone DeACetylase
HPC	Human Protein Complex
HR	Hormone Receptor
KC	Kindey Cancer
KEGG	Kyoto Encyclopedia of Genes and Genomes
KMC	K-Means Clustering
KNN	K-Nearest Neighbor
LINCS	Library of Integrated Network-Based Cellular Signatures
MBC	Metastatic Breast Cancer
ML	Machine Learning
MoA	Mechanism of Actions
mPTP	mitochondrial Permeability Transition Pore
MRP1	Multidrug Resistance Protein 1
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
NLP	Natural Language Processing
NN	Neural Network
NSCLC	Non-Small Cell Lung Cancer
PARP	Poly-ADP Ribose Polymerase
PC	Prostate Cancer
PCC	Pearson Correlation Coefficient
PI3K	PhosphoInositide 3-Kinase
PNS	Peripheral Nervous System
PPI	Protein Protein Interaction
PPMI	Positive Pointwise Mutual Information

PRL	Prototype Rank List
RBM	Restricted Boltzmann Machine
RCC	Renal Cell Carcinoma
RF	Random Forest
RLS	Regularized Least Squares
RNA	RiboNucleic acid
RNN	Recurrent Neural Network
RNS	reliable negative samples
ROC	Receiver operating Characteristic
RW	Random Walk
RWR	Random Walk with Restart
SCC	Spearman Correlation Coefficient
SCLC	Small Cell Lung Cancer
SMILES	Simplified Molecular-Input Line-Entry System
SNN	Shallow Neural Network
SP	Statistical Power
sscMap	statistically significant connections' Map
SAE	Stacked AutoEncoder
SVM	Support Vector Machine
TCA	TriCyclic Antidepressant
TCGA	The Cancer Genome Atlas
TKI	Tyrosine Kinase Inhibitor
TN	True Negative
TNF	Tumor Necrosis Factor
TP	True Positive
TTD	Therapeutic Target Database
VAE	Variational AutoEncoder
VEGF	Vascular Endothelial Growth Factor
WHO	World Health Organization

1 Introduction

1.1 Background

Drug repositioning is a strategy for drug development which predicts novel treatments for existing drugs. The most fruitful basis for the discovery of a new drug is to start with an old drug [1]. It saves time and reduces cost in drug discovery, especially in preclinical procedures. Unlike traditional drug repositioning approaches that utilize biological experiments, computational approaches can identify potential drugs more effectively. Benefiting from the development of biotechnology and expansion of biological data, many databases are constructed, which is a foundation of computational approaches. Multiple types of datasets about drugs, diseases, targets, *etc*, are employed in those approaches [2].

Another foundation of computational approaches is the algorithm. Commonly used algorithms are signature-based methods, machine learning (ML), and deep learning (DL) methods. Additionally, network strategies are employed as a part of those methods. Signature-based methods identifies a signature to describe a disease condition and matches it with several drug-induced profiles [3, 4, 5]. According to their matching scores, potential drugs are predicted. Therefore, the signature plays an important role in identifying a reliable result.

ML and DL have been employed to solve problems in many fields, such as medical image processing and semantic analysis, while DL is a subset of ML. They can learn from vast datasets effectively, and construct models in different fields. The various types of basic ML methods, such as classified-based methods [6], ensemble methods [7], instance-based methods [8], and neural network methods [9], have been used to predict potential associations between biological and biomedical entities. A DL model is often a neural network with multiple layers. In drug repositioning studies, the DL models are employed to either reduce feature dimensions of drugs, targets, *etc* [10], or predict potential associations between them [6]. Moreover, network strategies are commonly used to combine with prior methods, such as constructing a heterogeneous network [11] and identifying topological properties on the network [4].

In drug repositioning studies, the predictions of associations are mainly focused on three scenarios: drug-disease associations (DDAs) [12, 13, 14, 15, 16], drug-drug interactions (DDIs) [9, 17, 18, 19, 20], and drug-target interactions (DTIs) [21, 22, 23, 24, 25]. The predicted DDAs are the potential associations between drugs and diseases. Besides the signature-based methods, other methods can also be employed to predict potential DDAs. A DDI refers to a novel pharmacological effect of the two drugs, different from the known effects of two drugs when used alone. A DTI reflects that the target is addressed by a drug to produce the desired effect. Although both DDI and DTI cannot give a direct prediction about potential drugs, they help us to understand the mechanism of actions (MoAs) of drugs for drug repositioning.

This dissertation mainly focuses on signature-based methods and DL models to predict potential drugs for cancers and DDIs, respectively. In my studies, I first develop signature-based methods to predict potential drugs for some types of cancers, such as breast cancer, and colorectal cancer. Both disease signatures and drug signatures achieve good performance in prediction. Then I use a DL model to predict DDIs, which imply possible physiological effects of drugs and infer pharmacological functions.

1.2 Motivations and objectives

The overall objectives of my studies are predicting potential drugs for different diseases with multiple types of data and identifying drug-drug interactions. Several issues are addressed in my studies.

First, considering that I use tumor and normal tissue samples of patients to construct a gene signature of specific cancers, the inner-tumor heterogeneity should not be ignored. Therefore, treating all samples as a homogeneous set may average off the differences among the samples. Thus, developing a strategy to solve this problem is useful in my research.

Second, the gene signatures do not take the dependencies between genes into account, as genes work together in terms of protein complexes in the development of diseases. Therefore, a signature strategy involve in protein complex should be proposed to improve its quality in matching disease signatures and drug profiles.

In previous studies, in order to identify a disease signature, the sample size had to be large enough. However, the disease sample is often a single case in practice, especially for personalized medicine. A drug signature strategy for single patient samples is proposed.

Finally, since some drugs have been identified to have potential treatments for a specific disease, their physiological effects and pharmacological functions are unclear. DDIs help us to understand the MoAs of drugs, and propose potential drug combinations. Additionally, the DDIs have different functions, so that constructing a single network may cause biases. Therefore, a DDI prediction method that utilizes multiple networks is proposed.

Based on these motivations, I have the following objectives:

Objective 1: Review existing computational algorithms and databases for drug repositioning.

Objective 2: Develop a new strategy to identify disease gene signatures and predict potential drugs for several types of cancers.

Objective 3: Develop a new form of signature to describe cancer conditions and predict potential drugs.

Objective 4: Develop a strategy to identify drug signatures for personalized cancer medicine.

Objective 5: Develop a graph convolutional network with multi-kernel to identify potential drug-drug interactions.

1.3 Organization of the dissertation

This is a manuscript-style dissertation. The main content is presented in the form of published or submitted manuscripts that I have written during my Ph.D. study. An introduction is given at the beginning of each chapter to describe the connection of the manuscript in the context of the dissertation. All manuscripts have been reformatted to maintain consistency. The reference lists of all publications have been unified, and there is only one bibliography at the end of the dissertation.

The remainder of the dissertation is organized as follows. Chapter 2 reviews the existing computational methods, databases, evaluation metrics, and applications in drug repositioning. Chapter 3 employs a type of machine learning method in identifying disease signatures from patient samples. Chapter 4 proposes a type of protein complex signature of specific diseases and identifies their potential drugs. Chapter 5 proposes a strategy to identify drug signatures and predict potential drugs for a single patient. Chapter 6 proposes a graph convolutional network with multi-kernel to predict potential drug-drug interactions. Chapter 7 summarizes the work presented in this dissertation and discusses several future directions for this research. The list of publications is listed in Appendix A, while the copyright permissions of the manuscripts are included in Appendix B.

2 Drug repositioning: computational methods, databases and evaluations

Prepared as: Fei Wang, Xiujuan Lei, and Fang-Xiang Wu. A review of drug repositioning based chemical-induced cell line expression data. Current Medicinal Chemistry, 2019, 26, 1-10. FW reviewed the existing literature, and FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

Prepared as: Fei Wang, Yulian Ding, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Machine learning and deep learning strategies in drug repositioning. Current Bioinformatics, 2021. FW reviewed the existing literature, and FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

This chapter presents a literature review of computational methods, databases, and evaluation metrics used in drug repositioning. The review classifies current drug repositioning studies into three scenarios: Drug-Disease Association (DDA), Drug-Drug Interaction (DDI), and Drug-Target Interaction (DTI). Three types of methods, including signature-based methods, basic machine learning (ML) methods, and deep learning (DL) methods, are summarized. Furthermore, network strategy is applied as a part of those methods. The pros and cons of different types of methods are discussed, as well as several perspectives to improve them. Commonly used databases and evaluation metrics are also discussed so that researchers can easily develop their algorithms. This chapter fulfills Objective 1 of this dissertation.

Abstract

Drug repositioning is to find novel usages for existing drugs. It plays an important role in drug discovery, especially in the preclinical stages. Compared with the traditional drug discovery approaches, computational approaches can save time and reduce costs significantly. Since drug repositioning relies on existing drug-, disease-, and target-centric data, many methods have been proposed to identify useful information from multiple data resources. Based on transcriptomic data, signature-based methods are proposed to predict the potential connections between drugs and diseases. The disease profiles are used to identify a signature, which is used to match the drug-induced profiles. According to the matching scores, the potential drugs for a given disease are predicted.

When dealing with more types of data, ML approaches can construct models and learn from vast datasets

effectively. Deep learning (DL) is a subset of ML and appears in drug repositioning much later than basic ML. Nevertheless, DL methods have shown great performance in predicting potential drugs in many studies.

In this chapter, I review some commonly used signature-based methods, basic ML and DL approaches in drug repositioning. Firstly, the related databases are introduced, while all of them are publicly available for researchers. Two types of pre-processing steps, calculating similarities and constructing networks based on those data, are discussed. Secondly, the strategies are illustrated separately. Thirdly, I review the latest studies about the applications in three scenarios: DDA, DDI, and DTI. Finally, I discuss the limitations in current studies and suggest several directions of future work to address those limitations.

2.1 Introduction

In traditional pharmaceutical industry, putting a new drug on the market is very costly and time-consuming. About 1 billion US dollars and ten years are common [26]. The related budgets are still increasing rapidly. In traditional drug discovery pipeline, three major procedures are essential: preclinical experiments, clinical trials, and regulatory approval [27], as shown in Figure 2.1. Several thousands of small compound candidates are typically studied to develop one new drug. However, in many projects, no drug can be taken to the market successfully.

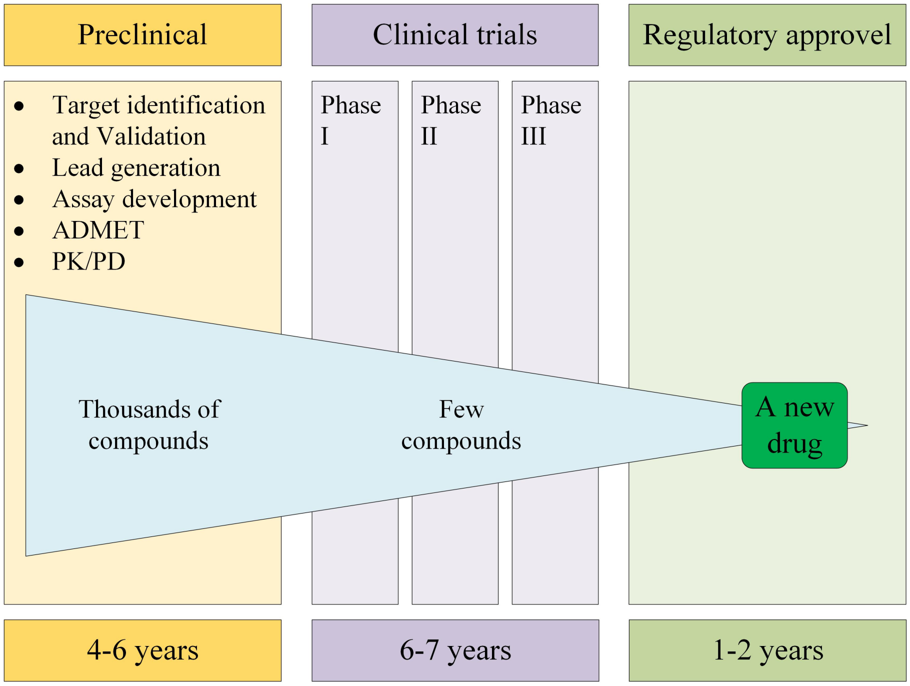


Figure 2.1: The drug discovery pipeline.

Drug repositioning approaches are proposed to identify novel treatments for existing drugs in order to save time, reduce cost, and improve the possibility of success. The safety and other properties of existing drugs have been studied clearly so that preclinical periods can be reduced significantly. Some successful drugs have been identified to have novel treatments for different diseases and approved by the United States Food and Drug Administration (FDA), such as sildenafil, thalidomide, zidovudine, minoxidil, and celecoxib [28]. Those drugs are generated by two types of drug repositioning approaches, which are phenotypic screening and target-based approaches [29]. In the first decade of the 21st century, 45 small compounds were proposed by those two types of approaches, 28 of which were identified by phenotypic screening [30, 31].

However, the traditional drug repositioning approaches still have some limitations. In phenotypic screening, small animal models and cell-based models are necessary. The robustness and relevance of models influence the success of screening [32]. In target-based approaches, the experiments are based on assays, and the number of effective drug targets is limited [33]. Computational drug repositioning approaches are proposed to predict potential drugs without biological experiments. Based on biological data, various algorithms and applications are proposed to identify novel treatments for existing drugs.

Signature-based method identifies a signature of a specific disease (or a specific drug) and predicts a list of potential drugs. The disease signature is a list of genes that characterize a disease condition. Differentially expressed genes (DEGs) between disease and normal conditions are often used to construct a signature of disease. Additionally, some statistical and network centrality methods are proposed to identify a more accurate signature, such as moderated T-test [34], Wilcoxon test [35], and network centrality combination [4]. The identified disease signature is used to query drug perturbation profiles. Gene set enrichment analysis (GSEA) [36] is employed to calculate the connection scores, and a list of potential drugs are identified based on the scores. The signature-based methods have predicted several drugs for diseases in drug repositioning studies [3, 5, 37, 38, 39, 40].

Machine learning (ML) technologies have been applied in many computational fields and achieve good performance in solving regression, classification, and clustering problems. The concept of “machine learning” was proposed by Alan Turing in the 1950s [41]. They are useful tools to identify potential drugs in drug discovery. Deep learning and basic ML are two classes of ML. The basic ML strategies, such as basic neural network (NN) [6, 9, 42, 43, 44, 45, 46, 47, 48, 49, 50], decision tree [7, 8], random forest (RF) [8, 10, 16, 21, 51, 52, 53, 54, 55, 56], k-nearest neighbor (KNN) [8, 19], random walk (RW) [11, 57, 58, 59, 60, 61, 62, 63, 64, 65], support vector machine (SVM) [6, 15, 20, 52, 66, 67, 68], and shallow autoencoder [49, 53, 64, 69, 70, 71], have shown their successful usages in predicting potential drug-disease associations (DDAs), drug-drug interactions (DDIs), and drug-target interactions (DTIs). Those associations and interactions help identify novel treatments for existing drugs. Many researchers apply ML methods to extract drug, disease, and target feature vectors from public databases and make predictions based on those vectors [8, 9, 15, 19, 20, 22]. Other researchers employ ML methods to predict potential missing links on the drug-disease heterogeneous network [11]. The networks are based on known links and similarities. After training ML models on the networks, the missing links are given probability values. The predicted DDAs/DDIs/DTIs are based on those values.

Deep Learning (DL) has also been applied to drug repositioning recently, Wen *et al.* utilized a DL method to predict potential DTIs [25], which was the first DL application for this purpose. After that, many DL methods have been applied to predict potential DTIs, DDAs and DDIs, such as deep neural network (DNN) [6, 52, 65, 70, 72, 73, 74, 75, 76, 77, 78, 79], convolutional network (CNN) [6, 42, 45, 46, 47, 49, 54, 74, 77, 78, 80, 81, 82, 83, 84], recurrent neural network (RNN) [46, 74], and stacked autoencoder (SAE) [10, 55, 65, 85].

In the applications, many methods focus on predicting some novel DDAs, DDIs, and DTIs. DDAs provide essential information for drug repositioning [69]. Novel associations may reveal the treatments of existing diseases with new drugs.

DDIs refer to the pharmacological and clinical responses to a drug combination, different from the known effects of two drugs when used alone. A drug may enhance the therapeutic efficacy of a drug and reduce the toxicity of another drug [86]. The predictions of DDIs help find some drug combinations that have better treatment for a disease, than any of them when given alone. Additionally, based on the "guilt-by-association" principle [87, 88], similar drugs may have the same treatment.

Identifying DTIs is essential as it provides insights into the experimental design of drug discovery [89]. The targets are molecules that have proven associations with particular diseases [90]. Prediction of novel DTIs helps find novel usages of existing drugs.

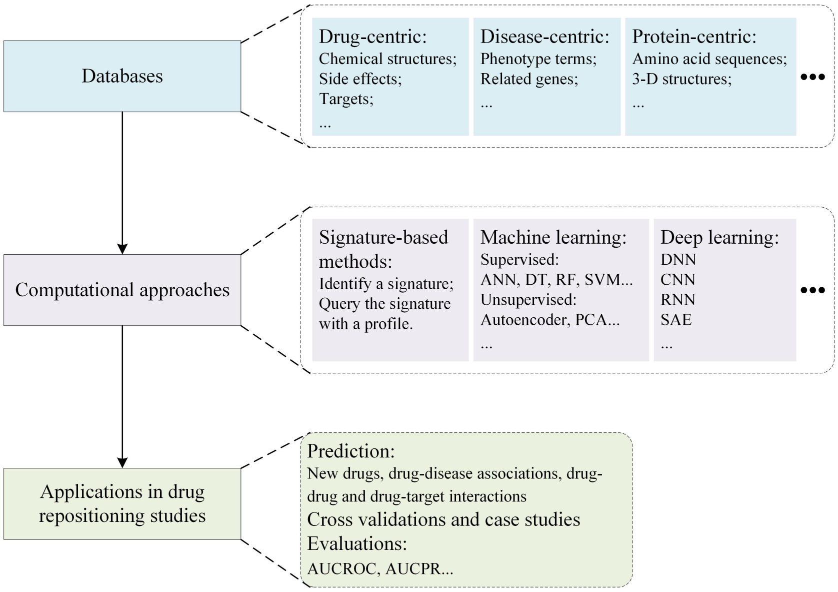


Figure 2.2: The workflow of data, methods, and applications in drug repositioning.

The workflow of this chapter is shown in Figure 2.2. We first summarize some commonly used databases for drug repositioning purposes in Section 2.2. The most commonly used data types are features of drugs, diseases, targets, and associations between them [91]. The signature-based methods are mostly based on drug perturbation profiles and disease tissue samples, which can be treated as drug and disease feature vectors. The basic ML and DL models are based on the feature vectors, associations, and interactions extracted from the databases.

The commonly used methods are introduced in Section 2.3. Then their latest applications in drug repositioning are systematically reviewed in Section 2.4. In order to provide a clear description, we divide them into three scenarios: the predictions of DDAs, DDIs, and DTIs. Finally, we discuss the limitations of those applications and some directions of future work in Section 2.5.

2.2 Biological data

Drugs, diseases, and targets are key components for drug repositioning. Therefore, we first summarize some databases for drug-, disease- and target-centric information in Table 2.1. Those data consist of many feature types, such as drug chemical structures, disease phenotypes, and protein amino acid sequences.

Table 2.1: Drug-, Disease- and Protein-Centric Databases.

Names	Descriptions	URLs
BRaunschweig ENzyme DAtabase (BRENDA)	Drug target sequences and 3-D structures.	https://www.brenda-enzymes.org/
ChEMBL	Physicochemical properties of drugs.	https://www.ebi.ac.uk/chembl
Connectivity Map (CMap)	Drug perturbation profiles.	https://clue.io/cmap
Comparative Toxicogenomics Database (CTD)	Drug-gene, gene-disease, drug-disease and gene-gene associations.	http://ctdbase.org/
DrugBank	Drug-drug interactions, drug substructures, drug-associated enzymes, pathways, and targets.	https://go.drugbank.com/
Drug Gene Interaction DataBase (DGIdb)	Drug related genes, Drug-gene annotations, interactions and potential drug ability database.	https://www.dgidb.org/
Disease-Gene Network (DisGeNet)	Disease related genes.	https://www.disgenet.org/
Drug Target Common (DTC) database	Drug-target interactions.	https://drugtargetcommons.fimm.fi/
Encyclopedia of DNA Elements (ENCODE)	Database of comprehensive parts list of functional elements in human genome.	https://www.encodeproject.org/

FDA Adverse Event Reporting System (FAERS)	Adverse event reports and medication error reports submitted to FDA.	https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers
Gene Expression Omnibus (GEO)	High throughput gene expression datasets.	https://www.ncbi.nlm.nih.gov/geo/
International Union of basic and clinical PHARmacology (IUPHAR) database	Drug-target interactions.	https://www.guidetopharmacology.org/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Databases dealing with genomes, biological pathways, diseases, drugs, and targets.	https://www.genome.jp/kegg/
Library of Integrated Network-based Cellular Signatures (LINCS)	Dataset of transcriptional responses of human cells to chemical and genetic perturbation. 1.3 Million L1000 profiles and tools for their analysis.	https://lincsproject.org/
National Drug File Reference Terminology (NDF-RT)	Drug characteristics, including ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases.	https://bioportal.bioontology.org/ ontologies/NDFRT
National Cancer Institute Developmental Therapeutics Program (NCI-DTP)	Growth inhibition data.	https://dtp.cancer.gov/
Offsides and Twosides	A comprehensive database of drug-drug-effect relationships.	http://tatonettilab.org/offsides/
Online Mendelian Inheritance in Man (OMIM)	Human genes and genetic phenotypes.	https://www.omim.org/
Open Targets Platform	Comprehensive and robust data integration for access to and visualization of potential drug targets associated with disease.	https://www.targetvalidation. org/

PubChem	More than 90 million compounds chemical information along with their bio activities, gene and protein targets.	https://pubchem.ncbi.nlm.nih.gov/
SIDe Effect Resource (SIDER)	Adverse drug reactions, side effects and the indications of marketed medicines, Information on marketed medicines and their recorded adverse drug reactions.	http://sideeffects.embl.de/
Search Tool for the Retrieval of INteracting Genes/proteins (STRING)	Protein-protein interactions, analysis, and networks.	https://string-db.org/
SuperTarget	Drug-target relations.	https://bioinformatics.charite.de/supertarget/
Therapeutic Target Database (TTD)	Dataset of known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs directed at each of these target.	http://db.idrblab.net/ttd/

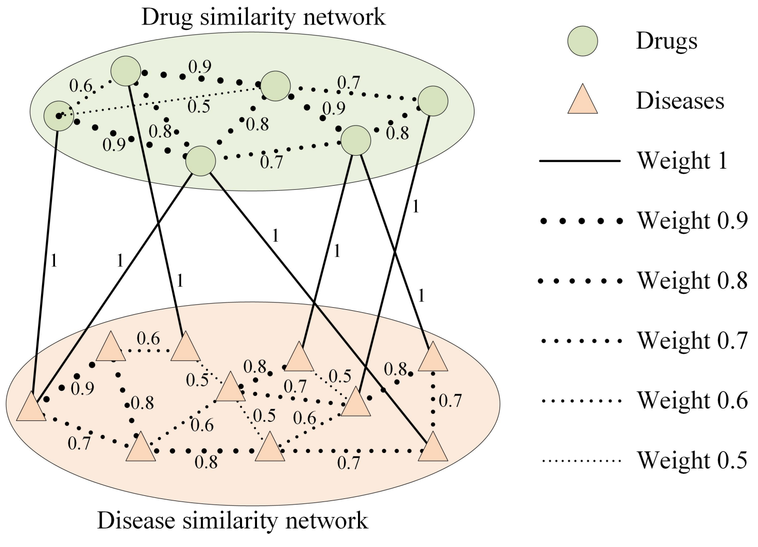


Figure 2.3: An example of a drug-disease heterogeneous network. The solid lines denote the known drug-disease associations, and the weights of dotted lines denote the similarities. Six different weight values are exemplified.

In this dissertation, the most frequently used databases are GEO, CMap, LINCS, and DrugBank. The

Table 2.2: The feature types and similarities of drug-drug, disease-disease, and target-target associations

Association types	Feature types	Similarity methods/tools
Drug-Drug	Chemical structure	CDKSim [92], SIMCOMP [93], Marginalized [94], Tanimoto [95], Spectrum and Lambda-k [96]
	ATC codes	ATCSim [97]
	Associated targets	Tanimoto [95], GIP [19]
	Side effects	Sider2 [98], Aers-bit and Aers-freq [99]
Disease-Disease	Phenotypes	SemFunSim [100], Separation [101]
	Ontologies	DoSim [102]
	Associated genes	GIB and PSB [103], ICod [104]
Target-Target	Amino acid sequences	Smith-Waterman algorithm [105], Spectrum and Mismatch [106]
	Ontologies	Semantic similarity [107]
	Associated drugs	GIP [19]

GEO database consists of a large number of gene expression profiles about different diseases. In my study, I download profiles of several types of cancers from the GEO database. The patient number in each type of cancer varies from tens to hundreds. Both CMap and LINCS are databases of drug perturbation profiles. Gene expression values under different drug perturbations, durations, concentrations, and cell lines are collected. DrugBank is a comprehensive database of drug-related information. In my study, the FDA-approved drugs and drug feature vectors are downloaded from DrugBank.

Many researchers use pre-processing steps when they are introducing those data in their studies. In studying the connections between instances, an association matrix A is constructed. Taking the drug-target associations for example, the rows of A are drugs, while the columns are targets. If there is a known association between drug i and target j, $A(i, j) = 1$; otherwise, $A(i, j) = 0$. Moreover, a row vector can be treated as a feature vector of a drug.

A further step is to calculate a similarity matrix between the same type of instances. As listed in Table 2.2, various methods are proposed to calculate similarities for drug-drug, disease-disease, and target-target pairs.

Another step is to construct a network. It can be either a homogeneous network between the same type of instances, such as a protein-protein interaction network, or a heterogeneous network, such as a drug-disease network. Figure 2.3 [11] contains a drug similarity network, a disease similarity network, and a drug-disease association network. In the similarity networks, the weights of interactions are based on the similarities.

Five different values of weights are used as examples in Figure 2.3. The known drug-disease associations downloaded from databases, such as DrugBank, are used to connect the two similarity networks.

2.3 Computational methods

In this section, we illustrate the commonly used computing strategies, as shown in Table 2.3. For signature-based methods, we describe three methods in identifying a signature. For basic ML, we discuss eleven commonly used methods. For DL, we introduce four types of deep neural networks (DNNs).

Table 2.3: The introduced strategies

Signature-based methods	Moderated T-test
	Wilcoxon test
	Network centrality
Classified-based methods	Logistic regression
	Support vector machine
Ensemble methods	Decision tree
	Bagging
	Boosting
	Random forest
Instance based methods	K-nearest neighbor
	K-means
	Random walk
Neural network methods	Basic neural network
	Basic autoencoder
Deep learning methods	Convolutional neural network
	Recurrent neural network
	Deep autoencoder
	Generative Adversarial network

2.3.1 Signature-based methods

The signature-based methods are identifying a disease signature and calculating similarity scores between the signature and drug profiles, or vice versa. The potential DDAs are predicted based on the similarity scores. The basic strategy in identifying a disease signature is calculating the log 2 fold-change (Log2FC) ratios of genes between disease tissue profiles and normal tissue profiles. Then DEGs are identified as a signature. In order to improve the performance of the signature, some other methods are employed. The moderated T-test and Wilcoxon test are two statistical methods.

Moderated T-test is calculating p-values of genes based on their expression values across all samples. Meanwhile, a Log2FC ratio is assigned to each gene. The genes with small p-values and large Log2FC ratios are identified as a signature. It has been used in predicting potential drugs for many diseases [3, 37, 38, 39].

Different from the moderated T-test that uses the expression values, **Wilcoxon test** is using ranks to calculate z-scores. The ranks are based on the absolute values of differences between normal tissue samples and disease tissue samples. It is also used in predicting novel treatments for existing drugs [3, 5].

Network centrality reflects the topological property of genes in a network. In related studies, the importance of elements on biological networks are correlated with topological centralities [108, 109, 110, 111], such as degree centrality, betweenness centrality, and closeness centrality. The Network centrality strategies are employed to identify gene signatures and predict potential drugs in drug repositioning studies [4, 40].

2.3.2 Basic machine learning strategies

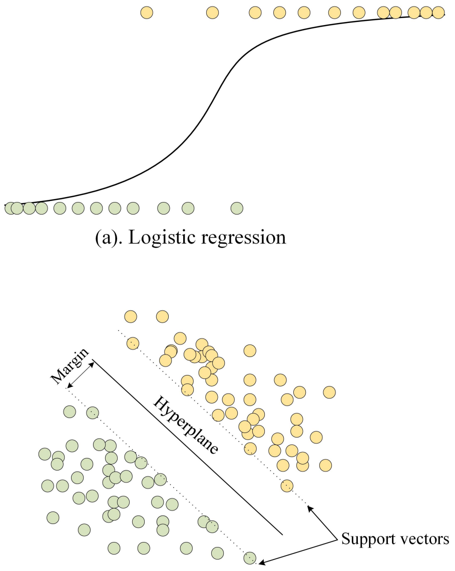
The basic idea of machine learning (ML) is to construct a model based on sample data. The models are used in a variety of applications, such as pattern recognition and drug repositioning. In this section, we introduce eleven widely used basic ML methods in drug repositioning, which are grouped into four categories: regression-based methods, ensemble methods, instance-based methods, and neural network methods.

Classified-based methods

The classified-based methods are based on the linear combination of features to assign samples into two or more classes. The logistic regression and support vector machine are two typical classified-based methods, which are commonly used in binary classification problems of drug repositioning.

Logistic regression (LR) employs a logistic function to model a binary dependent variable. Most of the predictions of DDAs (or DDIs and DTIs) are binary classification problems. Therefore, the binary LR model has a dependent variable with two possible labels: "0" and "1", or "Negative" and "Positive". The log-odds for the value labeled "Positive" is a linear combination of independent variables. The probability of the variable labeled "Positive" varies between 0 and 1, that a logistic function is used to convert log-odds to probability, as shown in Figure 2.4-a. A few researchers employ LR to predict potential drugs. Liu *et al.* utilize several ML models to predict novel DDAs, including LR [52].

Support vector machine (SVM) is one of the most widely used classification algorithms [112]. When dealing with binary classification problems, SVM generates a hyperplane in the sample space. A good separation is achieved by the hyperplane that has the largest distance to the nearest training sample of any class. The larger the distance is, the lower the error of the classifier is. An example of SVM for binary classification is shown in Figure 2.4-b. The SVM can be used to predict potential DDAs, DDIs, and DTIs [6, 8, 15, 52, 66, 67]. Beyond those, Zheng *et al.* employ the SVM algorithm to identify some reliable negative



(a). Logistic regression

(b). Support vector machine

Figure 2.4: Examples of logistic regression (a) and support vector machine (b).

DDIs from unknown DDIs [20]. The known DDIs and reliable negative DDIs are utilized to predict potential DDIs.

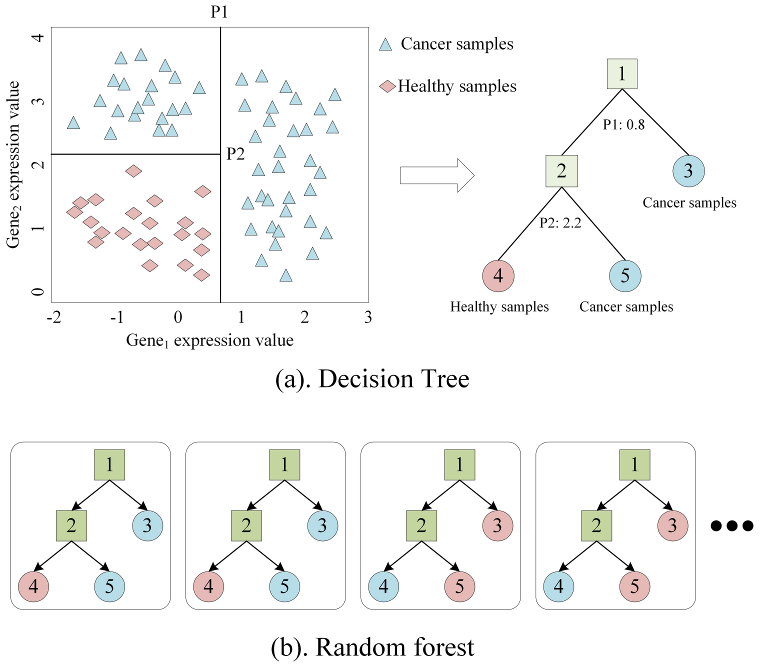
Ensemble methods

The ensemble methods combine multiple models to produce improved results of base models. In drug repositioning, many researchers use decision tree as the base model and apply bagging and boosting methods to improve it. In the following, we mainly review decision tree, bagging, random forest, and boosting methods.

Decision Tree is used in many areas such as radar signal classification, medical diagnosis, and speech recognition [113, 114]. It is a tree structure model. Each internal node is a decision on an attribute, each branch is the outcome of a decision, and each leaf node is a class label. The paths from the root node to leaf nodes are classification rules. An example is shown in Figure 2.5-a, while both cancer samples and healthy samples have two gene values. A decision tree model is constructed to distinguish cancer samples from healthy samples. In this chapter, its employments as a classifier are discussed for predicting potential drugs [7, 8].

Bagging is an abbreviation of "bootstrap aggregating." It is an ensemble algorithm to reduce variance and avoid over-fitting [115]. It is often combined with other ML methods, such as decision trees. An example is shown in Figure 2.6-a. Here, n datasets are generated from the original dataset by sampling with replacement. Each dataset has the same sample size. A classifier is constructed in each subset. The voting of the outputs of all classifiers is the result of the bagging strategy. When processing regression problems, the result is the average of the outputs of all models.

Random forest (RF) is an application of the bagging method in classification. It is a combination



(a). Decision Tree

(b). Random forest

Figure 2.5: Examples of decision tree (a) and random forest (b).

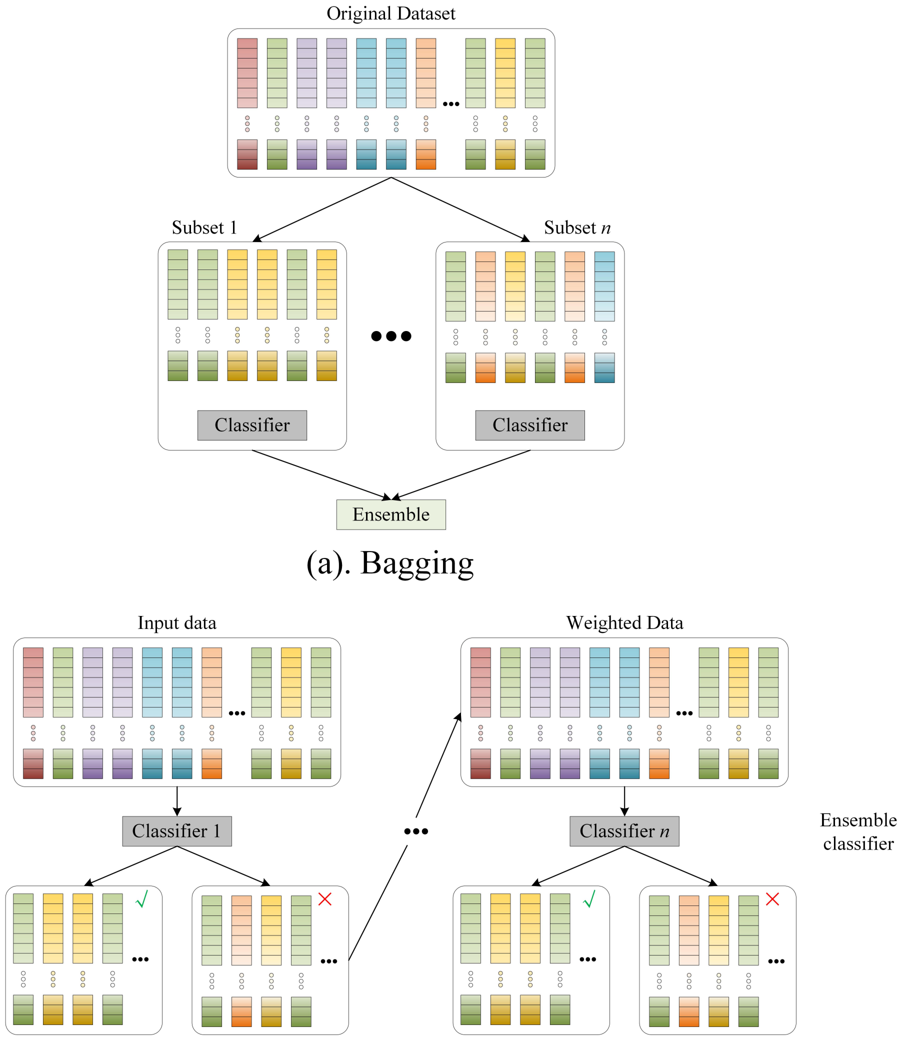
of decision trees that each tree is constructed independently [116], as shown in Figure 2.5-b. It retains the benefits of decision trees while achieving better results by bagging samples [117]. It works well when dealing with biological datasets with a large number of features. Many researchers apply RF to predict potential drugs [7, 8, 10, 16, 21, 51, 52, 53, 54, 55]. In those applications, the RF model is a good classifier when processing vectors with thousands of features.

Boosting is another type of ensemble algorithm [118]. Most boosting algorithms consist of several classifiers in sequence. The first classifier classifies the training data. Then the misclassified data gain a higher weight, and correctly classified data lose weight. The second classifier works on the weighted data and updates the weights, as shown in Figure 2.6-b. The multiple weak classifiers can form a strong classifier via boosting. The Adaptive Boosting (AdaBoost) [119] and Gradient Boosting [120] are two algorithms using boosting method. In AdaBoost, the outputs of the weak classifiers are combined into a weighted sum, while the weights are updated iteratively to adapt to the weak classifiers. In Gradient Boosting, the model is trained based on the residual between the true value and the predicted value of each sample. In predicting potential drugs, those algorithms are often combined with decision tree or RF [21].

Instance-based methods

The instance-based methods are comparing new instances with the training instances. We discussed k-nearest neighbor, k-means clustering, and random walk in this section.

***K*-nearest neighbor (KNN)** is a typical instance-based method, either for classification or for regression problems [121]. Because KNN relies on distances to determine the nearest neighbors, a normalization process



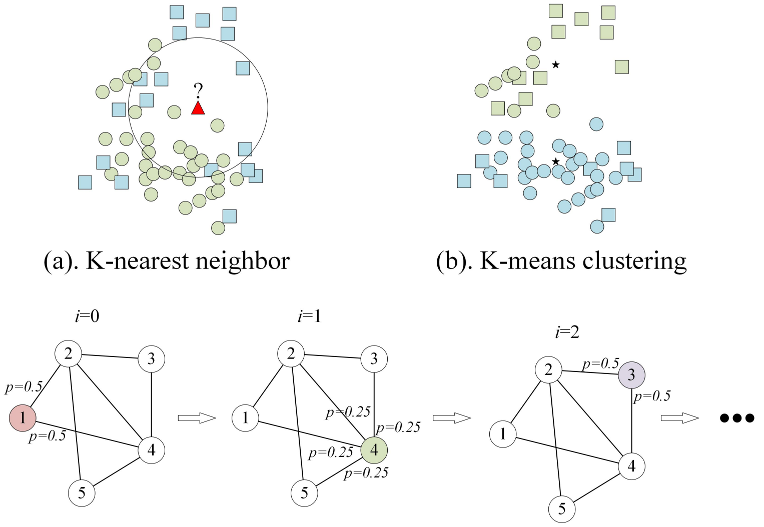
(a). Bagging

(b). Boosting

Figure 2.6: The structures of bagging (a) and boosting (b).

is useful to improve its accuracy, especially when the features vary in different scales. A commonly used distance metric is the Euclidean distance. An example of samples in 2-D space is shown in Figure 2.7-a. In a classification problem, a voting process is employed in the input sample's k nearest neighbors. The input sample is assigned to the class that has more votes among the neighbors. When processing a regression problem, the input sample has an average value of its k nearest neighbors. Both types of problems are applicable in drug repositioning. In [19], each known DDI has an intra-similarity, while the score of an unknown DDI is the average similarity of its k nearest known DDIs. In [8], KNN is applied to predict potential DDIs.

***K*-means clustering (KMC)** aims to cluster samples into k clusters. Each cluster has a center, and each sample belongs to the class whose center is the nearest center to the sample, then each center is updated according to the samples assigned to it, as shown in Figure 2.7-b. k is determined by users, and k samples



(a). K-nearest neighbor

(b). K-means clustering

(c). Random walk

Figure 2.7: Examples of k-nearest neighbor (a), k-means clustering (b), and random walk (c).

are randomly identified as the initial centers of classes. After all the other samples are assigned to the nearest class, the centers are updated. Then the samples are assigned to the nearest classes iteratively. The algorithm is converged when assignments do not change significantly. In drug repositioning, KMC helps find the subsets of a dataset. Wang *et al.* utilize KMC to generate subtypes from cancer samples and identify a gene signature from each subset [122].

Random walk (RW) is a stochastic process that the position of an instance in the (i+1)-th movement is only determined by its position in the i-th movement and a transition probability between those two movements, as shown in Figure 2.7-c. In similarity networks and heterogeneous networks, RW is a useful method to study the topological properties. In drug repositioning, many researchers used RW and its variations to predict potential drugs based on the drug-disease and drug-target heterogeneous networks [11, 57, 58, 59, 60, 61, 62, 63, 64, 65].

Neural network methods

Neural networks are powerful models in machine learning. In the following, we mainly focus on basic neural networks and basic autoencoders while deep networks are discussed in Section 2.3.2.

Basic **neural network (NN)** is a network method that contains three types of layers: input layer, hidden layer, and output layer [123]. The neurons in a layer are fully connected with those in the neighbor layers, as shown in Figure 2.8-a. Taking the neurons in the hidden layer for instance, the information is transformed as follows:

$$H^{Out} = \sigma(W_H H^{In} + B_H) \tag{2.1}$$

where σ is the activation function in the hidden layer, H^{In} and H^{Out} are the inputs and outputs of the

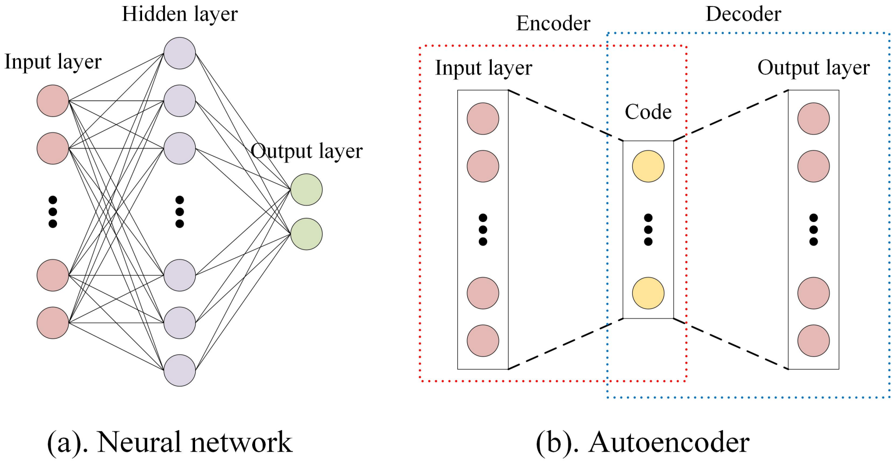


Figure 2.8: The structures of basic neural network (a) and basic autoencoder (b).

hidden layer, respectively. Meanwhile, the inputs of the hidden layer are the outputs of the input layer, and the outputs of the hidden layer are the inputs of the output layer. W_H and B_H are the weight matrix and bias vector of the hidden layer.

There are different activation functions, such as Sigmoid, TanH, eLU, ReLU, Leaky ReLU, and Softmax. The researchers can use any of them according to their requirements.

Many cost functions, which represent the differences between the predicted values and real values, are defined in applications. The cost function is used to optimize the parameters matrices and vectors. One of the frequently used cost functions in processing binary classification problems is the binary cross-entropy cost function as follows:

$$Cost = -\frac{1}{n}\sum_{x}[y\ln(p) + (1-y)\ln(1-p)] \tag{2.2}$$

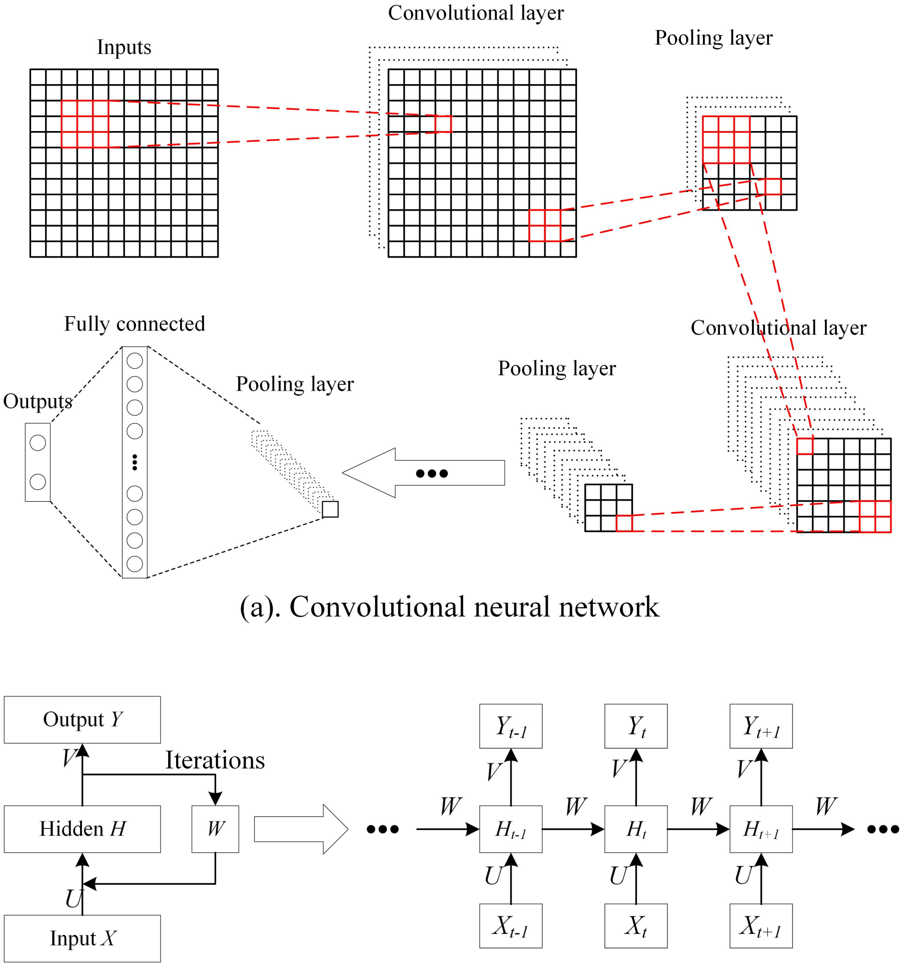
where n is the number of training samples, x is a training sample, and y is the label of x, p is the prediction value. y has two possible values: “0” and “1”.

In this chapter, the NN model is discussed in Section 2.4 for predicting the potential DDAs (or DDIs, DTIs) [6, 9, 42, 43, 44, 45, 46, 47, 48, 49]. The inputs of the NN are feature vectors extracted by different methods, and the outputs are the probabilities of the potential DDAs, DDIs, and DTIs.

Basic **autoencoder** is a type of NN that learns to copy its input to its output. The input layer and the output layer have the same number of neurons. The autoencoder has a code layer that describes a code to represent the input. It consists of two parts: an encoder maps an input to a code, and a decoder maps the code to an output. An example of shallow autoencoder is shown in Figure 2.8-b. In drug repositioning, the autoencoder model is often utilized to reduce the dimensionality of feature vectors [49, 53, 64, 69]. Their dimensions are reduced from thousands to hundreds, and the predictions in the following processes are still satisfying.

2.3.3 Deep learning strategies

The neural network with multiple hidden layers between the input layer and output layer is defined as a "deep neural network (DNN)," which underpins deep learning. The widely used convolutional neural network (CNN) [124], recurrent neural network (RNN) [125], deep autoencoder (DAE) [126], and generative adversarial network (GAN) [127] are different types of DNNs with different structures.



(a). Convolutional neural network

(b). Recurrent neural network

Figure 2.9: The structures of convolutional neural network (a) and recurrent neural network (b).

Convolutional neural network (CNN) utilizes several convolutional layers, pooling layers, and fully connected layers to form the model, as shown in Figure 2.9-a. The convolutional layer uses kernels to encode its input data [128]. In this layer, the widely used activation function is ReLU. The pooling layer aims to reduce the dimensionality of the data by integrating several neighbor neurons of one layer into a single neuron in the next layer. Max-pooling and average-pooling are two common types of pooling. Max-pooling transforms the maximum value among neighbor neurons of the prior layer to the next layer, while the average-pooling layer uses the average value instead. After several convolutional layers and pooling layers, a few fully connected layers are applied to generate the prediction results. CNN models can be employed to predict potential DDAs, DDIs, and DTIs [6, 42, 45, 46, 47, 49, 54, 74, 77, 78, 80].

Recurrent neural network (RNN) is a class of neural networks that the connections between neurons form a directed graph along a temporal sequence, as shown in Figure 2.9-b. The neurons at time t get inputs from other neurons at previous time steps. The calculation processes are as follows:

$$Y_t = g(VH_t + B_Y) \tag{2.3}$$

$$H_t = f(UX_t + WH_{t-1} + B_H) \tag{2.4}$$

where the U, V, and W are weight matrices. B_Y and B_H are bias vectors. X_t, H_t, and Y_t are the matrices of the input layer, hidden layer, and output layer at time t, respectively. g and f are activation functions.

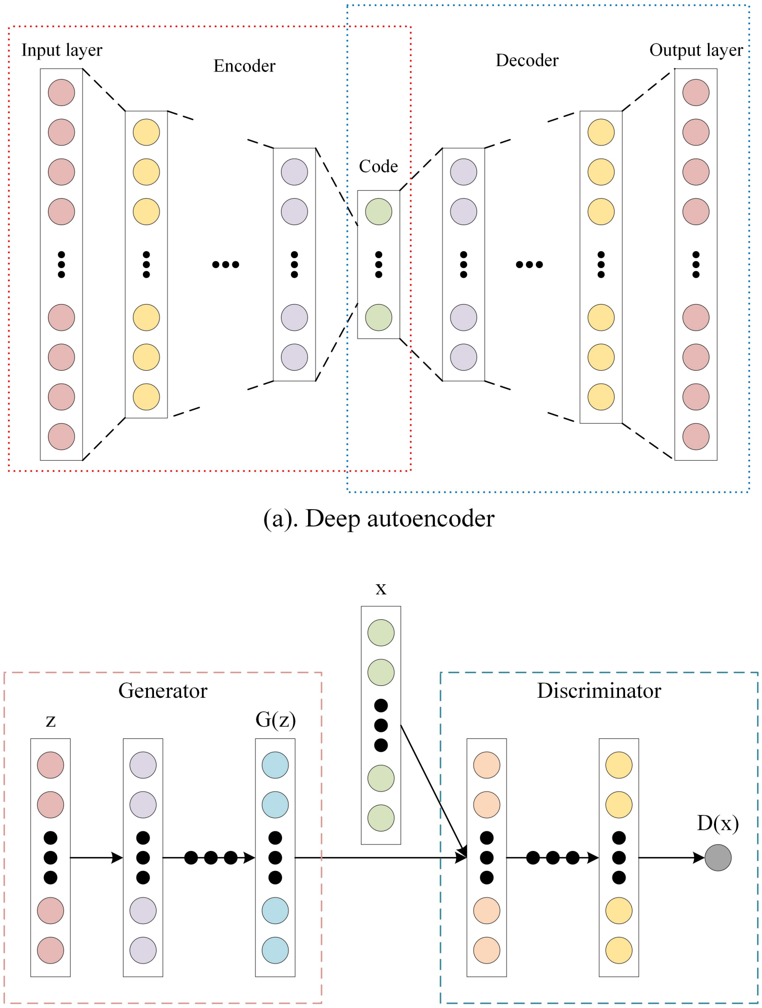
Similar to other types of neural networks, RNN is also used to predict potential drug-target interactions [129].

Deep autoencoder (DAE) is an autoencoder with multiple hidden layers, as shown in Figure 2.10-a. Both the encoder and the decoder consist of some layers with different numbers of neurons, while the code layer often contains a smaller number of neurons than those in the input layer. Similar to the shallow autoencoder, DAE is commonly used to learn the advanced features of drugs/targets in drug repositioning [55, 70, 71], while the advanced features are fed into classifiers to make predictions.

Generative adversarial network (GAN) is based on a game theory that two neural networks contest with each other [127]. The two neural networks are the generator network and discriminator network, as shown in Figure 2.10-b. The generator produces samples, and the discriminator aims to distinguish between the training samples and the samples from the generator [130]. Researchers employed the GAN models to distinguish the known DTIs and the unknown DTIs based on their feature vectors [131].

2.4 Applications in drug repositioning

In the previous two sections, we have discussed the databases and ML/DL methods. In this section, we review some latest applications in drug repositioning. We divide the predictions of novel drugs into three types: drug-disease association (DDA) prediction, drug-drug interaction (DDI) prediction, and drug-target interaction (DTI) prediction. The DDA prediction aims to find some novel drugs directly, based on multiple types of drug features and disease features, such as drug structures, drug side effects, disease phenotypes, and disease genes. The second type is to identify some drug combinations which have better treatment than any of them when given alone. The third type aims to predict some novel DTIs. Mostly, a drug target is a protein, which has essential functions in disease pathways.



(a). Deep autoencoder

(b). Generative adversarial network

Figure 2.10: The structures of Deep autoencoder (a) and Generative adversarial network (b).

2.4.1 Drug-disease association predictions

The signature-based methods are proposed to predict DDAs directly. Benefiting from the drug perturbation databases CMap and LINCS, many researchers have identified gene signatures of multiple diseases and produced lists of potential drugs. Xiao *et al.* generated a Glioblastoma multiforme signature and queried it to CMap [12]. Chandran *et al.* identified two gene signatures for Central Nervous System (CNS) and Peripheral Nervous System (PNS), and three common drugs which appear in both two drug candidate lists were generated [13]. Goss *et al.* [132] and Pessetto *et al.* [133] identified two different signatures for Ewing Sarcoma, while generating the same drug etoposide. Wen *et al.* predicted candidate drugs by integrating a signature from five datasets of colorectal cancer [14].

The transcriptomic data of drugs and diseases used in signature-based methods are variable. A little perturbation in cell culture can make the gene expression values change. Therefore, some more stable features of drugs and diseases, such as drug side effects, chemical structures, target genes, and disease phenotypes, associated genes, were integrated for drug repositioning [134, 135, 136, 137]. Unlike calculating drug-disease similarity scores in signature-based methods, a different strategy is using drug-drug similarity and disease-disease similarity to predict DDAs. Different features may have different methods for calculating similarity, as listed in Table 2.2.

Based on the similarities, some machine learning methods were applied to predict potential drug-disease associations (DDAs), such as random walk [11, 57], SVM [15], and RF [16, 51]. Luo *et al.* applied one type of similarity for each instance and a random walk algorithm to identify new indications for existing drugs [11, 57]. In [11], the drug-drug chemical structure similarity and disease-disease phenotype similarity were proposed to construct a drug similarity network and a disease similarity network. The two networks were connected by known DDAs and form a heterogeneous network. A bi-random walk algorithm was applied in the heterogeneous network, while one random walk was in the drug network and another was in the disease network. Each random walk produced a value, and the average value denoted the probability of the drug-disease association. In [57], the heterogeneous network contained three parts: drug network, disease network, and target network. A random walk with restart (RWR) was applied in the heterogeneous network and produced a probability vector, which contained the probability scores of all drugs associated with a given disease.

The drug-disease association prediction problem is often formulated as a classification problem. Lee-Yoon *et al.* constructed an RF model to predict potential DDAs via genes [16]. The genes were utilized to connect drug target genes and disease genes. Then the drug-disease pairs were represented to gene paths, which were proposed to train an RF model. The known DDAs were assigned as positive samples, and the unknown ones were negative samples. Zhou *et al.* generated a drug-disease heterogeneous network and utilized an RF model to make the prediction [51].

Besides single similarity for drugs and diseases, multiple similarities can be concatenated together to increase the prediction accuracy. Kim *et al.* utilized four types of drug-drug similarity and three types of disease-disease similarity in their work [15]. Furthermore, 1,330 known DDAs were utilized as the basic instances. For a drug-disease association that needed to be predicted, the drug in it had similarities with the drugs in all known DDAs, and the disease in it had similarities with the diseases in all known DDAs. One type of drug similarity and one type of disease similarity were used to construct a classification feature. Twelve types of feature integrations were generated. Finally, an SVM model was constructed, and 10-fold cross-validation was applied to evaluate this model.

Besides basic ML methods, some DL methods are utilized to make the prediction. Liu *et al.* constructed a drug-disease heterogeneous network and applied a DNN model to predict potential DDAs [52]. An adjacent matrix was constructed, while each row or column was treated as the feature vector of an instance. The two feature vectors of a drug-disease pair were integrated and fed into a DNN model, and a probability score was generated. The proposed deep learning method achieved higher scores in multiple measurements than some

ML approaches, including logistic regression, SVM, and RF.

Jarade *et al.* proposed a DNN model [72] and a collective variational autoencoder (cVAE) model [69] to predict DDAs. In their work, several drug similarities and disease similarities were filtered and integrated. The integrated feature vectors were fed into either a DNN model or a cVAE model to finish the prediction. The two models performed better than some machine learning approaches under the measurements of both AUC-ROC and AUC-PR.

Zeng *et al.* proposed a multi-modal deep autoencoder (MDA) model to extract low-dimensional features from multiple networks and a cVAE model to predict potential DDAs [85]. A co-occurrence matrix was generated via random walk on the heterogeneous network. Then the co-occurrence matrix was transformed into a positive pointwise mutual information (PPMI) matrix [138], which was utilized as the input data of MDA [125]. The middle layer of the MDA informative feature, which was part of the input of the cVAE model. Other parts of input data were the known DDAs. The probability score was generated to reflect the potentiality of the drug-disease pairs.

Based on multiple features and similarities, Jiang *et al.* proposed an autoencoder model [53] and a CNN model [54] to predict potential associations. In [53], for a given drug-disease association, the drug chemical structure fingerprint, drug Gaussian interaction profile (GIP) kernel similarity, disease GIP kernel similarity [139], and disease MeSH term similarity were concatenated [140] and fed into an autoencoder. After dimensionality reduction, an RF classifier was applied to finish the prediction. In [54], the autoencoder was replaced by a CNN model, and the RF was also utilized as a classifier.

CNN is another commonly used DL model in drug repositioning. It can effectively extract features from different types of raw data. Li *et al.* proposed a CNN model to conduct a binary classification of DDAs [42]. The drug features were based on the simplified molecular-input line-entry system (SMILES) [141] with a dimensionality of 881. The disease features were retrieved from the human symptoms-disease network [142], and its dimensionality was 322. An 881×322 matrix was constructed and mapped to a gray-scale image. A CNN model was applied to extract feature vectors from the image and generated the prediction results.

Graph neural network (GNN) [143] has several subtypes, including graph convolution network (GCN) and graph autoencoder (GAE). Wang *et al.* proposed a GNN based method to predict potential DDAs [144]. A drug-disease association network was constructed from known associations. Then a GNN model was applied to exploit the high-order features in the network. Yu *et al.* came up with a layer attention GCN (LAGCN) model to predict DDAs after the construction of a drug-disease heterogeneous network [145]. In the embedding process of LAGCN, each layer had a weight parameter to adjust the contribution of different layers. The parameters were determined by NN.

In previous research about DDAs, most of their features were different, for instance, drugs had chemical structures and diseases had phenotype ontologies. However, both of them had associations with genes, which could be measured in microarray platforms. Focusing on the expression values of genes under different drugs in different cell lines could reveal the DDAs directly. In this way, a set of drug perturbation profiles were downloaded from the CMap and LINCS databases, and the disease profiles were downloaded from the GEO database, as listed in Table 2.1.

Wang *et al.* applied a k-means algorithm to cluster the disease profiles into several groups to represent the cancer subtypes [122]. Each group was utilized to identify a list of disease genes. The disease gene signatures were based on the weighted frequencies of genes in the lists, which were mapped with the drug perturbation profiles in the CMap database [146, 147]. The connection score of a disease signature and a drug profile represented the possible association of them, while a negative number meant the drug may have potential treatments for the disease. In comparison with the methods without the k-means algorithm, the proposed framework achieved better prediction accuracy in several types of cancers. Zhao *et al.* used the drug profiles in CMap to train five machine learning classifiers. Based on the drug indications extracted from ATC and MEDI-HPS [148], the positive and negative drug labels were generated. The authors focused their study on three types of diseases and predicted several drugs that have literature evidence.

2.4.2 Drug-drug association predictions

Unlike the drug-disease associations, the drug-drug interactions (DDIs) have the same feature types connecting them, such as chemical structures, targets, enzymes, pathways, transports, indications, and side effects. There are many types of DDIs, which reflect the connections between two drugs, such as the bioavailability/metabolism/serum concentration/therapeutic efficacy of drug a can be decreased/increased by drug b. Therefore, identifying the types of DDIs can help study the drug repositioning potentiality of a drug combination. Additionally, for a single drug, based on the “guilt-by-association” principle, the high similarities with other drugs may reflect their treatment similarities. Those two parts are the main field of DDI prediction for drug repositioning.

Ferdousi *et al.* employed 12 binary features to analyze DDIs [17]. The features were integrated, and the pair similarities were calculated. For the known DDIs, a pre-processing step was added to delete the DDI whose two drugs had no common biological item or had an empty common feature vector. Among the remaining known DDIs, the minimum positive similarity value was set to be the threshold, which was utilized to determine whether an unknown drug-drug pair had the potential to be a DDI.

Yan *et al.* only calculated the similarities of known DDIs and applied a regularized least squares (RLS) classifier to finish the prediction [18, 19]. In [19], eight types of drug features were integrated and made the total dimensionality of the drug vector to be 21,351. Then the similarity of a drug-drug pair was calculated. Based on the known DDIs and similarities, the initial score of an unknown DDI was generated through the KNN method. The drug interaction vector consisted of initial scores between it and all other drugs. The GIP kernel similarity matrix was based on the drug-drug interaction vectors. Finally, an RLS classifier was employed to predict potential DDIs based on the matrix. In [18], the GIP similarity was applied on the adjacent matrix directly, without the initial score procedure. Then the GIP similarity and drug feature cosine similarity were integrated and averaged to construct the similarity matrix.

In many classification methods, the known DDIs were treated as positive samples, and unknown DDIs were negative samples. Some researchers identified reliable negative samples (RNS) from unknown DDIs. Bi *et al.* calculated an average distance between an unknown DDI and all known DDIs, while only the unknown DDIs with large distances were identified as RNS [149]. The residual unknowns were treated as unlabeled samples. The samples with three types of labels were utilized for training an extreme learning machine (ELM) [150] and predicted the potential DDIs. Zheng *et al.* applied an SVM to identify RNSs and another SVM to predict DDIs [20]. Its performance was better than those of Bi's method, based on the measurement of recall and F_1 score.

In many studies, researchers prefer to use multiple types of similarities without any distinction. Rohani *et al.* added a filter procedure and employed a neural network model to predict potential DDIs [9]. In Rohani's method, they first selected several types of similarities with the most information and least redundancy [151], then a nonlinear method was applied to integrate the selected similarity matrices. Each drug had a feature vector in the integrated matrix [152]. A neural network model integrated two drug feature vectors, and the output was a probability value for potential DDI.

Benefiting from the network strategies, a DDI network was constructed based on the known DDIs. Then the DDI prediction problem was transformed into the prediction of missing links in the network. Zhou *et al.* employed a Markov clustering algorithm to identify drug groups from the network, that most of the groups were significantly correlated with certain functions [153]. Munir *et al.* applied the k-means algorithm to generate 12 clusters of drugs and constructed 12 DDI networks [154]. All the drugs were used in the treatment of epidermal growth factor receptor (EGFR) mutations in various cancers. The drugs that link to the nodes with the largest centrality values in each network were selected and combined to construct a final DDI network. Then the same procedure was applied to identify the final drugs with potential interactions. The predicted DDIs had been verified by molecular docking results.

Kastrin *et al.* integrated DDI networks with feature similarities to predict potential DDIs [8]. Their five networks were based on five databases. Five machine learning algorithms, including decision tree, KNN, SVM, RF, and gradient boosting machine (GBM), were applied to finish the prediction based on topological features of the networks and semantic features.

Zhang *et al.* integrated 14 types of similarities to make the DDI prediction [58]. Eight of them were based on drug features, such as chemical structure, targets, and pathways. Six of those were based on the DDI network, which was constructed from the known DDIs. A random walk method was applied on the DDI network with each of the similarity matrices. All the predictions were combined through an ensemble learning procedure [155] to generate an improved final prediction result.

Similar to the drug-disease association predictions, DL methods were utilized to predict potential DDIs. Zhang *et al.* applied multi-modal deep auto-encoders to generate low-dimensional feature vectors of drug pairs and predicted potential DDIs via RF classifier [10]. Ryu *et al.* employed a DNN model to predict potential DDI types [73]. Shukla *et al.* proposed a modified DNN model to make the prediction [74]. In their model, a few CNN and RNN hidden layers were added to process the drug features, while the prediction accuracy of their model was better than either CNN models or RNN models. Lee *et al.* collected three types of data, including drug structures, target genes, and GO terms [70]. For a given drug pair, it had three types of feature vectors. The same types were integrated and fed into an autoencoder. The three code layers of the three autoencoders were integrated again and fed into a DNN model, which was used to predict DDI types. Deng *et al.* utilized four types of similarities and constructed four similarity matrices [75]. The similarity matrices were fed into a DNN model, and the output is the DDI events, which were used to describe the DDI relationships. Feng *et al.* proposed a GCN model to extract the network structure features of drugs from the DDI network and predict DDIs [76]. A 2-layer GCN was utilized to obtain drug features and produce a feature vector matrix. Two drug vectors were integrated and fed into a DNN model, which was used to deduce the potential DDIs.

The previous studies are about drug-drug pairs. In some conditions, a combination of more than two drugs may have potential treatments. Peng *et al.* proposed a novel model to predict the reactions of drug combinations [156]. In the first process, the dimensionalities of drug features were reduced through a neural network model. The new drug vectors were integrated via three approaches: max pooling, mean pooling, and self-attention. The embedding vectors were fed into a second neural network model, and the output value was used to predict the potential reactions of the drug combination.

Some researchers add more entities to the DDI network and construct a new knowledge graph to reflect the new associations. Lin *et al.* utilized drugs, targets, genes, transporters, and enzymes to build a knowledge graph [157]. The drug feature vector of a drug-drug pair was encoded by a 2-layer GNN model. Then the output values were used to predict whether the drug-drug pair had potential interactions.

In many methods, two drugs in a DDI are treated separately. Song *et al.* used a different idea to make the prediction [66]. In their method, the drug-drug pairs were treated as instances. The similarity between two drug-drug pairs was calculated based on the drug similarities as follows:

$$S((d_1, d_2), (d_3, d_4)) = \max(S((d_1, d_3), (d_2, d_4)), S((d_1, d_4), (d_2, d_3))) \tag{2.5}$$

where $S(i, j)$ was the similarity between two instances i and j, (d_1, d_2) was a drug-drug pair. An SVM model was proposed to make the subsequent prediction. A DDI's feature was determined by its similarities with other DDIs. Like the training strategy in other methods, 10-fold cross-validation was applied to evaluate the SVM model. In the results, some DDIs with literature evidence had been predicted, which were not listed in the referenced databases.

Cytochrome P450 enzymes are essential for the metabolism of many medications [158], which are the main reasons for many DDIs. A drug can be a substrate, inhibitor, or inducer of CYP450, which may affect the metabolite of other drugs. Hunta *et al.* predicted potential DDIs via their enzyme actions [67]. Different from other features of drugs, the features in Hunta's study were enzymes and enzyme action types. Machine learning algorithms such as NN and SVM were trained and used to predict the potential DDI.

2.4.3 Drug-target association predictions

A target is a molecule that has a proven association with a particular disease [90]. It is usually a protein. In recent years, many databases and tools have been constructed to reveal interactions between diseases and genes or proteins, which helps researchers predict potential drugs through drug-target interactions (DTI).

The decision tree, RF, and SVM are commonly used classification algorithms in machine learning, that many researchers employ them in drug repositioning. Wang *et al.* applied an RF approach to predict DTIs [21]. In their method, the protein-ligand connection was described by four components: protein sequence, binding pocket, ligand structure, and intermolecular interaction. In general, the total number of features was more than several thousand. A PCA procedure was employed to reduce the dimensionality of features before the RF model. The number of final features was less than a few hundred. After training, their method performed good results in predicting DTIs.

Similar to the drug-disease heterogeneous network, researchers construct a drug-target network to predict potential DTIs. The drug-drug similarities and target-target similarities are calculated from various features, and the known DTIs are downloaded from public databases. Based on the heterogeneous network and similarity matrices, Zeng *et al.* generated feature vectors of drugs and targets separately [22]. A deep forest (DF) classifier was applied to predict potential DTIs from the feature vectors.

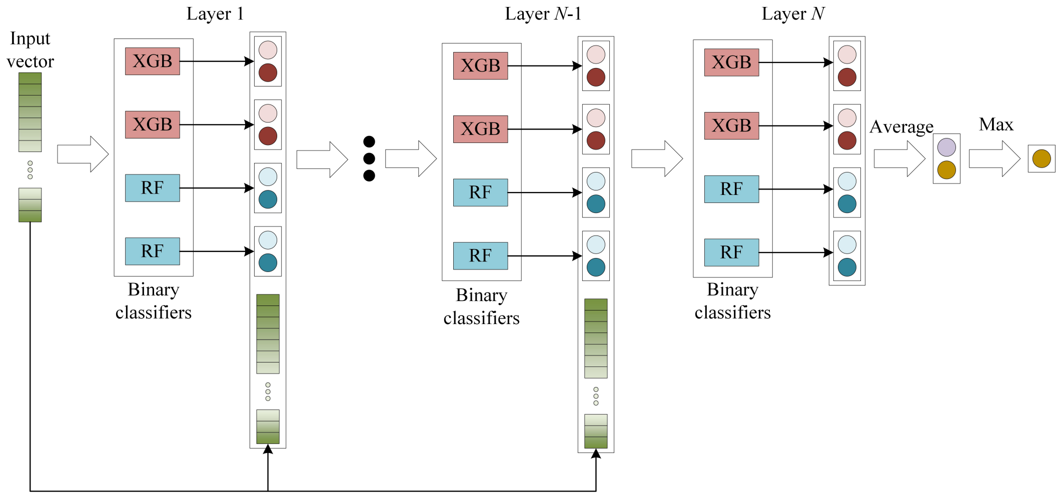


Figure 2.11: The structure of CDF.

Chu *et al.* utilized a cascade deep forest (CDF) model to predict potential DTIs [23]. A few steps were utilized to generate the features, which were fed into the model. Six types of similarities were used to construct the drug-target heterogeneous networks. The networks were merged by a network fusion method [152]. In Chu's work, they used the path nodes between the drug and the target to form the input vector [151]. The path node was either a different drug or a different target, restricted to be the five nearest neighbors of the initial drug and target. As a result, the new form of input vector might be drug-drug-target, drug-drug-drug-target, or four other forms. After fed into the CDF model [159], as shown in Figure 2.11, a final prediction was made from the output. In each layer of CDF, the number of binary classifiers was varied.

Lin *et al.* utilized support vector regression (SVR) to build a model to predict potential DTIs [24]. In their study, the SVR was applied to generate the binding strength of drug-protein pairs. A protein similarity network was constructed, where the similarities were based on the binding strength. The edge betweenness centrality was used to predict shared drugs between proteins, which were the potential DTIs.

Zong *et al.* utilized a DeepWalk method [59], which was a deep model of random walk, to predict DTIs from a network model [160]. The known DDAs, DDIs, and DTIs were downloaded and used to construct a drug-target-disease network. The similarity between two instances was calculated by DeepWalk based on the known edges. After generating the similarities, two approaches were proposed to predict potential DTIs, which were drug-based and target-based similarity inference [161].

Many researchers apply the drug-target heterogeneous network to identify their feature vectors. The dimensionality of each vector is the sum of drug features and disease features. Manoochehri *et al.* proposed a different approach to generate the feature vectors from the drug-target network [43, 44]. For a drug-target pair, the sub-graph was constructed based on their neighbors in the network and themselves, which meant that different interaction has different sub-graphs. An adjacent matrix was identified based on the sub-graph rather than the whole drug-target network. Therefore, the feature vectors also had different dimensionalities. After feeding the features into an NN model, a prediction was made. When training the model, the known DTIs produce known sub-graphs for positive samples, and the negative samples were not selected randomly but built under certain principles [162]. After training, the proposed method achieved higher performance than the baseline methods in terms of AUC-ROC and AUC-PR.

Although the basic ML methods achieve satisfying prediction performance, the DL methods work better in many cases. Wen *et al.* proposed the first deep learning method (DeepDTI) in predicting DTIs [25]. The drug substructure fingerprints were identified as the drug feature vectors, and the target protein sequences were target vectors. Their DeepDTI had a deep-belief network (DBN), which was made by stacking restricted Boltzmann machines (RBMs). In various measurements of predictions, the DeepDTI method achieved better performance than other ML methods, including RF, decision tree, and naive Bayesian.

When applying a DNN model to predict DTIs from drugs and targets feature vectors, some basic ML and DL algorithms are also utilized to generate satisfied feature vectors, such as linear classification [163, 164], random walk with restart (RWR) [60, 61, 62, 63], autoencoder [49, 55, 64, 65, 71], *etc.* Parvizi *et al.* utilized the random walk with restart (RWR) algorithm and skip-gram neural network to generate the feature vectors of drugs and targets [63]. In their method, the drug-target heterogeneous network was replaced by two networks: drug-related network and protein-related network. The drug-related network consisted DDIs and DDAs, while the protein-related network contained protein-protein interactions (PPIs) and protein-disease associations.

Peng *et al.* proposed a similar approach in constructing a drug-related network and a protein-related network [80]. Besides the known interactions and associations, the drug-drug similarities and protein-protein similarities were added in the networks. After integration of the two feature vectors, a deep autoencoder was applied to produce the low dimensional features, which were fed into a CNN model. The prediction performances in terms of AUC-ROC and AUC-PR were increased by adding the similarities, which were also higher than those of other ML methods.

CNN is a commonly used model for deep learning. It can be applied to either make the prediction or produce satisfied feature vectors of drugs and proteins. Hu *et al.* utilized a CNN model to predict DTIs [45]. The drug chemical structure vectors from PaDEL-descriptor [165] and the target amino acid physicochemical property vectors from AAindex [166] were proposed to identify the input matrix of the CNN method. The combination of drug vector and target vector were randomly selected, that the combinations of known DTIs were treated as positive samples, while others were negative samples. With 10-fold cross-validation, the prediction performances were much better than the state-of-the-art methods.

Monteiro *et al.* used CNN models to identify the feature vectors and applied a DNN model to predict DTIs [6]. After generating a drug SMILES vector and a target sequence vector from databases, two CNN models were proposed to process the two types of feature vectors and produced two novel vectors. The two vectors were integrated and fed into a fully connected DNN model. Finally, a prediction of DTI was made. Compared with the method without the CNN pre-processing and the CNN-RF/SVM models, the CNN-DNN architecture yields improved results in the correct classification of both positive and negative interactions.

Similarly, Öztürk *et al.* [77] and Zhao *et al.* [78] applied CNN and DNN models to generate the feature vectors and predict potential DTIs. In Öztürk's method, the drug SMILES features and protein sequence features were processed by two CNNs separately. The two feature vectors of a drug-target pair were generated, integrated, and fed into a DNN model to make a prediction. In Zhao's method, the commonly utilized drug-target heterogeneous network was transformed into a drug-target pair (DTP) network. Different from the heterogeneous network, the nodes in the DTP network were the drug-target pairs. The number of pairs in the DTP network was $n \times m$, where n was the number of drugs and m was the number of targets. A GCN model was processed to extract features from the adjacent matrix of the network. The new features were fed into a DNN model, and the prediction was made.

Huang *et al.* proposed a deep learning library to predict DTIs [46]. In their library, only the drug SMILES vectors and protein amino acid sequence vectors were utilized. Those two vectors were transformed into two new feature vectors through 15 approaches, such as CNN and RNN. Then the two feature vectors were integrated and fed into a multi-layer perceptron to generate the prediction of the drug-target pair.

Lee *et al.* proposed an integrated model to make the prediction [47]. In their method, a convolution layer was applied to process the target sequences, and a fully connected layer was used to process drug fingerprints. Then two vectors were integrated and fed into a CNN model. They compared their method with DeepDTI [25], which had been discussed previously. The DeepConv-DTI achieved higher accuracy and F_1 score.

Similar to the DDA and DDI predictions, GNN is widely used in predicting DTIs. Jiang *et al.* utilized a GNN model to identify the feature vectors, and then an NN model was applied to predict DTIs [48]. The drug's chemical structure was proposed to construct a molecular graph. The nodes were atoms, and the edges were bonds. The protein amino acid graph, which was based on the protein contact map, was produced by PconsC4 [167] based on the amino acid sequences. A new drug vector and target vector were identified by the GNN model, and the integration of them was fed into an NN model to make the prediction.

Lim *et al.* constructed a different graph based on the protein-ligand complex [168]. The structure information of protein and ligand atoms were embedded in two adjacent matrices, A_1 and A_2. A_1 contained covalent interactions only, and A_2 contained both covalent interactions and noncovalent intermolecular interactions. Two node feature vectors were generated from either A_1 or A_2. By subtracting the two feature vectors, their difference was fed into a GNN, and the prediction results were generated.

In many studies, one drug vector is integrated with one target vector. It is crucial to determine which type of target feature is used to identify the integration. In Lee's research, three types of target vectors were proposed to have close relationships with protein functions or drug mechanism of actions (MoAs) [79]. One drug vector, based on differentially expressed genes from the LINCS database [169], was integrated with all three target vectors, including gene knockdown expression profiles (GEPs) from LINCS database, protein-protein interaction (PPI) network from String database [170], and pathway memberships from MSigDB [171]. After integration, the new vector was fed into a DNN model. The concatenation of three types of target vectors showed better performance in terms of AORUC than any single type of them.

Agyeman *et al.* proposed integrated views predictive GAN (IVPGAN) to predict potential DTIs [131]. The model contained two main parts, which were generator and discriminator. The input data of the generator was the integrated vector of drug graph representation, drug SMILES string, and target sequence. The output of the generator, which reflected the binding strength, was combined with the ground truth and fed into the discriminator. Like other DL methods, the authors utilized a 5-fold CV to evaluate the IVPGAN model, and the prediction performance was higher than the parametric models in most of the datasets.

In the previous description, many methods integrate the feature vectors of drugs and targets directly, which fail to learn the low-dimensional features. Autoencoder is an excellent unsupervised approach to reduce dimensionality with high confidence. Wang *et al.* applied a stacked autoencoder to identify protein features from sequence information [55]. An RF classifier was utilized after the integration of protein feature vectors and drug structure vectors. Sun *et al.* proposed a convolutional autoencoder and GAN-based method to predict DTIs [64]. After constructing a drug-target heterogeneous network, the adjacent matrix was fed into a convolutional autoencoder, and a novel feature matrix with lower dimensionality was generated. It was assumed that the new feature vector of a drug or target obeys a Gaussian distribution. After the discriminator, a DTI prediction was made. In the evaluation, the proposed method achieved better performance than some DTI prediction methods, including DTINet by Luo *et al.* [60], Lee's method [61], and DTIGBDT by Xuan *et al.* [62]. The RWR algorithm was applied to capture topological information in the networks of their models.

Torng *et al.* applied a graph autoencoder (GAE) to extract a representation of protein pocket features [49]. Before the final classifier, a fully connected layer was added, taking the joint vector of protein and drug as input, and producing a low-dimensional hidden layer as output. In evaluation, the proposed method outperformed several structure-based and ligand-based methods in AUC-ROC scores.

Wang *et al.* utilized a multi-modal deep autoencoder (MDA) to produce protein and drug feature vectors from several similarities [65]. Each type of similarity had a corresponding network. In each network, the RWR method and PPMI were applied to calculate the topological similarity of drugs and proteins. Then the global structure information was generated. Two MDAs were applied to integrate multiple similarity measures of drugs and targets and learn low-dimensional feature matrices of them. The two features of a drug and a target were merged and fed into a DNN to make a prediction.

Since 2020, COVID-19 has threatened all over the world. Many researchers focus their work on either vaccines or medications to help stop the pandemic. Since SARS-CoV-2's core proteins have been determined, Beck *et al.* used natural language processing (NLP) to identify potential DTI [172]. In NLP, the molecule sequence was analogous to a language. More than 1 million drugs were used to train the models, and several antiviral drugs have been proposed to have potential interactions with SARS-CoV-2 proteins. Remdesivir, which had been reported to be an effective medication for COVID-19 in vitro [173], was among the prediction results.

2.5 Evaluation methods

In related studies in drug repositioning, the predictions of DDA, DDI, and DTI are often treated as binary classifications. Various evaluation metrics are used to measure the prediction performance.

Table 2.4: The confusion table.

		Predicted Condition	
		Predicted Positive	Predicted Negative
Actual Condition	Actual positive	True Positive (TP)	False Negative (FN)
	Actual Negative	False Positive (FP)	True Negative (TN)

Precision, recall, and F_1 score are commonly used to measure the prediction performance. Based on the four basic metrics of true positive (TP), false positive (FP), false negative (FN), and true negative (TN), as shown in Table 2.4, precision is defined as $TP/(TP + FP)$, recall is defined as $TP/(TP + FN)$, and the F_1

score is as follows:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{2.6}$$

A receiver operating characteristic (ROC) curve is created by plotting the TP rate against the FP rate in various thresholds. Similarity, the precision-recall (PR) curve is created by plotting the precision against recall in various thresholds. Furthermore, the area under ROC curve (AUC-ROC) and area under PR curve (AUC-PR) are used to measure the prediction performance, which are the areas under the corresponding curves.

However, in the signature-based methods, the predicted drugs are often focusing on a specific disease, and positive/negative samples are not employed in the methods. Therefore, the above evaluation metrics are not applicable. In those studies, researchers examine the number of known drugs for the given disease among the predicted drug list.

2.6 Perspectives and conclusions

In the former sections, we review some latest studies that employ signature-based and ML/DL methods in drug repositioning. Various methods have been used to predict the potential DDAs, DDIs, and DTIs. Those predictions help find novel treatments for existing drugs. In some cases, researchers also identify some potential drugs for specific diseases by the proposed methods. However, there are still some limitations.

A general issue is about the feature types in databases. As shown in Table 2.1, there are a large number of databases that store the drug-, disease- and target-centric information. Some databases may focus on a single feature type for each category (drug, disease, or target), while others may be comprehensive. In many studies, only one feature type for each category is applied. Although the drug chemical structure, disease phenotype, and protein amino acid sequence are widely used, other types should not be ignored. In [22, 58, 70], the feature vectors are identified from multiple feature types. However, it is still important to select several reliable feature types. In [79], Lee *et al.* propose that three types of target features are closely related to DTIs. The selection of different feature types is attracting attention.

When using multiple feature types, a second issue is how to effectively integrate them. Researchers use many different strategies to perform the integration. In [19, 21], the multiple feature vectors for the same category (drug, disease, or target) are concatenated directly, without any additional processing. The ML models are constructed based on the integrated data. In [8, 9, 23], several approaches are used to integrate the feature vectors or similarities, such as the average similarity of multiple types. In [58], the authors construct 29 models based on the multiple feature types, then merge the results to identify the prediction. Although all different strategies generate satisfied predictions, further ensemble methods need to be proposed.

A third issue which is needed to be improved is the identification of negative samples. A large number of applications are using basic ML and DL models to classify DDAs (or DDIs, DTIs). In many studies, the negative samples are randomly selected from the unknown associations. In order to improve the confidence of samples, a few strategies are proposed to identify reliable negative samples (RNS). In [174], the authors first use the known and unknown associations to construct a classifier, then employ this classifier to classify the unknown associations. Then classified negative samples are identified as RNS. In [20, 149, 175], KNN, RWR, and SVM are applied to extract RNS. Besides calculating distances and similarities, more reliable strategies are needed.

The fourth issue is about the use of ML and DL methods. These methods are just like black boxes, which make the models lack interpretability. Compared with DL models, some basic ML models are more interpretable, such as decision tree and logistic regression. Meanwhile, compared to basic ML models, DL models achieve better performance in predicting potential DDAs, DDIs, and DTIs. Therefore, more interpretable ML and DL models are essential in many application domains, especially in human healthcare-related fields [176], where drug repositioning is applied for. In order to achieve this goal, the improvements of basic ML and DL models with interpretability are necessary.

In this study, we review some latest studies that predict novel treatments for existing drugs. The widely used databases and pre-processing steps are introduced. The six data types in those databases, including drug features, disease features, target features, DDAs, DDIs, and DTIs, are taken into consideration. We then discuss commonly used basic ML methods and DL methods, and their applications to the predictions of DDAs, DDIs, and DTIs. In order to address the limitations of existing methods, we suggest several directions of future work about features, samples, and methods, which could benefit the research community of drug repositioning.

3 Identifying gene signatures for cancer drug repositioning based on sample clustering

Prepared as: Fei Wang, Yulian Ding, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Identifying gene signatures for cancer drug repositioning based on sample clustering. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2020. DOI: 10.1109/TCBB.2020.3019781. FW and FXW discussed the methods. FW implemented the algorithm, designed and performed the experiments. FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

As discussed in Chapters 1 and 2, the signature-based methods identify gene signatures from disease profiles and match them with drug perturbation profiles. A list of potential drugs is produced for a given disease. Most of the methods treat disease profiles, such as the tumor tissue samples, as homogeneous. However, a disease may have some subtypes that the samples are not homogeneous. A strategy should be proposed to identify a signature from heterogeneous samples. In this chapter, a clustering strategy is proposed to identify gene signatures from several gene expression profiles that may consist of a few subtypes. After matching the disease gene signature to the drug perturbation profiles, similarity scores are calculated to represent the connections. Potential drugs for the given disease are identified. The strategy achieves higher performance than other methods in predicting potential drugs. This chapter fulfills Objective 2 of this dissertation.

Abstract

Drug repositioning is an important approach for drug discovery. Computational drug repositioning approaches typically use a gene signature to represent a particular disease and connect the gene signature with drug perturbation profiles. Although disease samples, especially from cancer, may be heterogeneous, most existing methods consider them as a homogeneous set to identify differentially expressed genes (DEGs) for further determining a gene signature. As a result, some genes that should be in a gene signature are averaged off. In this study, we propose a new framework to identify gene signatures for cancer drug repositioning based on sample clustering (GS4CDRSC). GS4CDRSC firstly groups samples into several clusters based on their gene expression profiles. Secondly, an existing method is applied to the samples in each cluster for generating a list of DEGs. Then a weighting approach is used to identify an integrated gene signature from all the lists of DEGs. The integrated gene signature is used to connect with drug perturbation profiles in the

Connectivity Map (CMap) database to generate a list of drug candidates. GS4CDRSC has been tested with several cancer datasets and existing methods. The computational results show that GS4CDRSC outperforms those methods without the sample clustering and weighting approaches in terms of both the numbers and rates of predicted known drugs for specific cancers.

3.1 Introduction

Traditionally, the drug discovery industry is mainly about the screening of chemicals to obtain a small set of potential compounds [177]. However, further studies are needed to identify their therapeutic effects on a particular disease. After that, the screened compounds move forward to animal tests and clinical trials [178]. This whole complex process is so long and expensive that it takes 10-15 years and 0.8-1.5 billion US dollars to bring a drug from theory to product [26]. In order to reduce the time and cost of drug discoveries, researchers propose to find new usages for existing drugs, which have passed the evaluation of human safety [179]. Several successful drug repositioning studies have been published, including sildenafil for erectile dysfunction [180], thalidomide for severe erythema nodosum leprosum and retinoic acid for acute promyelocytic leukemia [181]. However, most of the successful examples of drug repositioning are from phenotypic drug screening and target-based methods [182, 183].

In recent years, the advances of high-throughput technologies, which produce a huge amount of transcriptomic data, provide a great opportunity for studying drug repositioning. Based on transcriptomic data, several databases have been proposed for drug repositioning. Lamb *et al.* constructed a Connectivity Map (CMap) database [146, 147]. In the database, there are 6,100 profiles in CMap build 2, each measuring the expression values of 22,283 genes of a cell line in a particular drug perturbation culture. The total number of drug perturbations is 1,309. In order to increase the scale of perturbations and keep the cost at a low level, the Library of Integrated Network-Based Cellular Signatures (LINCS) was developed [184]. The LINCS database only measures the expression values of 978 genes directly and all other gene expression values are estimated according to the measured values. About 19,811 small compound drug perturbations and 1,319,138 profiles are contained in the LINCS database.

After the construction of CMap and LINCS databases, several computational drug repositioning approaches have been proposed (e.g., [185, 186]). These approaches first identify a gene signature of a particular disease and then calculate the connection scores between the gene signature and the perturbation profiles in CMap database and/or LINCS database. The drugs with a connection score smaller than a threshold are identified as potential drugs for the disease, which are called drug candidates. Usually, among drug candidates, there are some drugs whose treatments for the particular disease are known, which are called known drugs. In general, the number of predicted known drugs can demonstrate the accuracy of the gene signature generated by the prediction method.

Many studies have been proposed to identify DEGs, which are candidates of a gene signature. In order to identify DEGs, gene expression data, which collect gene expression levels in different tissue samples, are needed. The National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [187] is one of the most comprehensive gene expression databases. Based on gene expression data, the fold-change thresholding methods are first used to identify DEGs (e.g., [188, 189]). Each gene has a fold-change ratio between normal tissue samples and disease tissue samples. The genes whose ratios are larger than a threshold are identified as DEGs. In many studies, the threshold is set to be 2.

However, the fold-change thresholding methods do not take variability into account or can not guarantee reproducibility [190]. Then the statistic methods are commonly used to identify DEGs, such as the T-test [14] and Wilcoxon test [191]. Additionally, based on the fact that disease-related proteins tend to have a larger number of interactions and more shared neighbors than non-disease proteins [192], genes can be mapped to protein-protein interaction (PPI) networks and use network methods to identify DEGs (e.g., [4, 193, 194]). In many of these studies, the disease tissue samples are treated as a homogeneous set to identify a gene signature. However, the samples from the same complex disease (e.g., cancer) are still heterogeneous as the complex diseases may have several subtypes. Therefore, treating all disease samples as a homogeneous set may average off the differences among the samples. As a result, DEGs or gene signatures generated by these methods are not good enough and thus their performance for drug repositioning is degraded.

In this study, we propose a new framework to identify gene signatures for cancer drug repositioning based on heterogeneous sample clustering (GS4CDRSC). GS4CDRSC firstly groups cancer samples into some clusters based on their gene expression profiles. Secondly, an existing method is applied to the samples in each cluster for generating a list of DEGs. In the lists of DEGs, a weighting approach is used to give each of the genes a new weight and sort them in descending order. Then the top genes are identified as gene signatures for drug repositioning. Finally, a CMap tool is applied to predict potential drugs from the integrated gene signature.

In order to evaluate its performance for drug repositioning, GS4CDRSC is combined with three existing approaches, while the k-means algorithm is employed to perform sample clustering. All the approaches are used to deal with tissue samples and identify a gene signature of particular cancer. Then the gene signatures are used for drug repositioning and each gene signature obtains a list of drug candidates. In order to evaluate the accuracy of gene signatures, the prediction rate of known drugs on the list of drug candidates has been calculated. Based on the known drugs, other predicted drugs on the list have the potential same treatment. From the experiments we can see that with the proposed GS4CDRSC, higher prediction rates are generated, which means that GS4CDRSC can improve the performance of drug repositioning methods. Finally, we give a discussion about the predicted potential drugs.

3.2 Methods and materials

Typically, the computational drug repositioning approaches contain two main steps [183]: (1) Identifying DEGs based on several tumor tissue samples and normal tissue samples from the GEO database or the like, and further determining a gene signature of specific cancer based on its DEGs; (2) Calculating the connection (or correlation) scores between drugs and gene signatures.

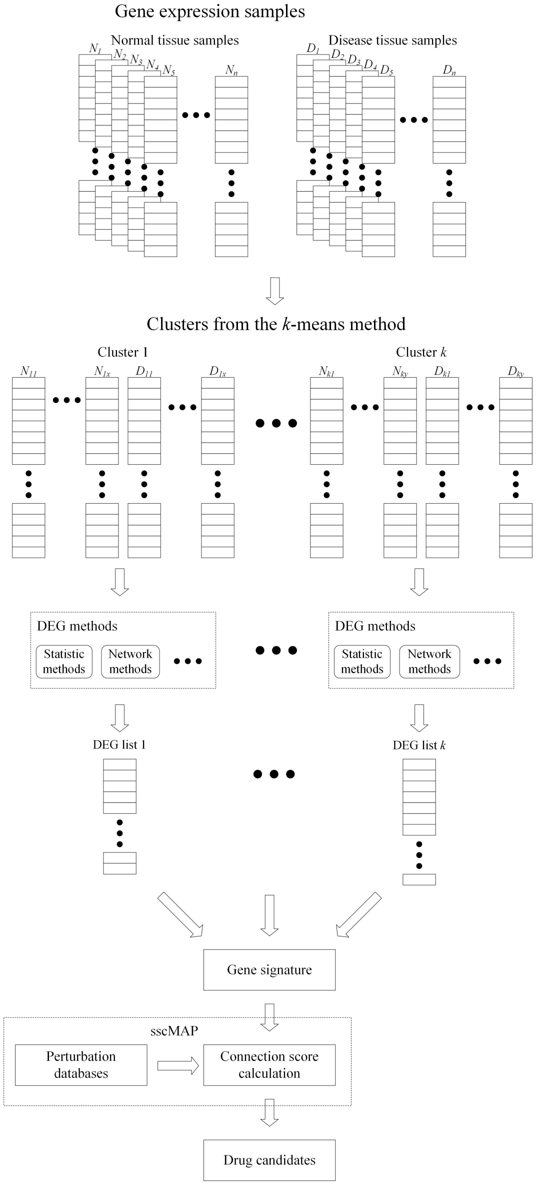


Figure 3.1: The flowchart of the GS4CDRSC framework.

In drug repositioning, the approaches for identifying a gene signature play an important role. In most approaches, such as the fold-change thresholding approaches (e.g., [188, 189, 195]), statistic approaches (e.g., [14, 191]) and network approaches (e.g., [4, 193, 194]), all the samples from patients with the same clinical diagnosed diseases are treated as a homogeneous set. Therefore, the signatures identified by existing methods need to be improved.

One of the reasons is the inner-tumor heterogeneity, where a dataset of a specific cancer may contain several different subtypes. The subtypes of cancer are small groups that cancer can be divided into, based on certain characteristics of the cancer. According to the studies of cancer cells in the past decades, different hierarchies of subtypes are proposed. Taking lung cancer as an example, two main histological subtypes are non-small cell lung cancer (NSCLC, 85% of all lung cancers) and small cell lung cancer (SCLC, 15% of all lung cancers) [196]. There are three subtypes under the NSCLC, which are squamous cell lung carcinoma, adenocarcinomas, and large cell carcinomas. Additionally, when looking into the hierarchy of genes and molecules, some gene mutation-based subtypes are proposed, such as epidermal growth factor receptor (EGFR)-mutation, Kirsten rat sarcoma viral oncogene homolog (KRAS)-mutation, and anaplastic lymphoma kinase (ALK)-mutation [197]. In clinics, some subtypes share similar treatments [197]. In our study, based on the gene expression values of patient samples, we aim to improve the performance of drug repositioning methods based on sample clustering.

3.2.1 The GS4CDRSC framework

In this study, we propose the GS4CDRSC framework for drug repositioning, which focuses on improving the identification of the gene signature of specific cancer. The pipeline of GS4CDRSC is shown in Figure 3.1. Specifically, a clustering algorithm is firstly used to divide the cancer samples into several clusters, each of which is expected to be homogeneous. Then the existing methods are employed to identify DEGs and generate a gene list for each sample cluster. In the list, a weighting approach is proposed to give each of the DEGs a new weight and sort the DEGs in descending order. Then the top M genes are identified as a DEG list. An integrated gene signature is determined over all the DEG lists from different clusters. The genes which appear in most of the DEG lists are utilized to construct the integrated gene signature. Finally, the integrated signature is used to query the CMap database and obtain drug candidates for the given cancer under consideration. The detailed steps are illustrated in the following subsections.

3.2.2 The sample clustering

The sample clustering algorithm is used to produce some clusters that each cluster contains homogeneous samples. In our proposed GS4CDRSC framework, the k-means algorithm is used for this purpose although other clustering algorithms can be used at this step. In the k-means algorithm, the smaller the differences within a cluster, the better the results are [198].

Given a set of samples $s = (s_1, s_2, \ldots, s_n)$, where each sample is a d-dimensional vector and d is the number of genes in a sample. The squared Euclidean distance is used to measure the difference between two samples as follows:

$$dist\left(s_i, s_j\right) = \sum_{t=1}^{d} \left(s_i\left(t\right) - s_j\left(t\right)\right)^2 \tag{3.1}$$

The k-means algorithm is to obtain k clusters $S = (S_1, S_2, \ldots, S_k)$ while the sum of distances within the clusters is the minimum. The objective of the k-means algorithm is to find the optimal S such that for a given k the following sum of squared errors (SSEs) is minimized:

$$J(S) = \sum_{i=1}^{k} \sum_{s \in S_i} dist\left(s, \mu_i\right) \tag{3.2}$$

where μ_i is the mean of the samples in cluster S_i. At the beginning of the algorithm, μ_i can be the profile of any sample. They are iteratively changed until the samples in the clusters are steady. As a result, all cancer samples are divided into k clusters. Then cancer samples in each subset and their corresponding normal samples are paired to make up a subset of samples for identifying DEGs and gene signatures in the following steps.

In GS4CDRSC, the k-means algorithm is based on DEGs and expected to obtain homogeneous subsets from all heterogeneous samples. The value of k is determined in 3.3.1. Additionally, it is expected to reduce the effects of outliers in gene expression profiles. When measuring the values of genes in the microarray platforms, the accuracies of experiments are influenced by many factors, such as the quality of microarray, which produces erroneous values in some samples. When applying the k-means algorithm, the profiles with error values cannot affect all the clusters although they may affect some clusters, which improves its accuracy. Moreover, when considering the samples in a dataset as a whole set, some genes may be averaged and ignored in the gene signature. The clustering algorithm is proposed to help identify such genes. As shown in Table 3.4, most of the genes in the final signatures are new.

3.2.3 The DEG identification for each subset

In GS4CDRSC, a list of DEGs is first generated from each subset of homogeneous samples which is obtained in Section 3.2.2. Then DEGs are used to identify gene signatures for drug repositioning. In this subsection, three DEG identification approaches are briefly described, including the moderated T-test approach, the Wilcoxon test approach, and a network-based approach.

The moderated T-test approach

The T-test is a pioneering approach in identifying DEGs from gene expression profiles. However, the T-test does not take into account the dependencies between genes. In order to address this weakness, the moderated T-test is proposed [34]. Each gene is assigned a p-value based on its gene expression values across all samples. Meanwhile, a fold-change ratio is also assigned to the gene, according to its average expression value in normal tissue samples and that in tumor tissue samples. Then genes with small p-values and large fold-change ratios are identified as DEGs.

Suppose an expression value y_{gij} is from gene $g = (1, \ldots, H)$, array $i = (1, \ldots, n)$ and replicate $j = (1, \ldots, m)$. Let s_g^B be the between-array standard deviation, which is calculated as follows:

$$(s_g^B)^2 = \frac{m}{n-1} \sum_{i=1}^{n} (\overline{y}_{gi} - \overline{y}_g)^2 \tag{3.3}$$

where $\overline{y}_{gi}$ is the mean of the replicates of gene g on array i and $\overline{y}_g$ is the mean of gene g across all arrays.

Let s_g^W be the within-array standard deviation, which is calculated as follows:

$$(s_g^W)^2 = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{gij} - \overline{y}_{gi})^2 \tag{3.4}$$

Then a T score is calculated as follows:

$$T = \frac{\overline{y}_g \times \sqrt{nm[1 + (m-1)\hat{\rho}]}}{s_g} \tag{3.5}$$

where $\hat{\rho}$ is the correlation of gene between replicates and s_g is calculated as follows:

$$s_g^2 = \frac{\{\frac{(n-1)(s_g^B)^2}{1+(m-1)\hat{\rho}} + \frac{n(m-1)(s_g^W)^2}{1-\hat{\rho}}\}}{nm-1} \tag{3.6}$$

A p-value is computed based on the T score. The False Discovery Rate (FDR) α is set to be 0.01 and is controlled by the Benjamini-Hochberg (BH) procedure [199] as follows:

$$p(M) \leq \frac{M}{H} \alpha \tag{3.7}$$

where M is the length of gene signature and H is the number of genes in a sample. The largest M is set to be 100 to make sure that $p(M) \leq 1/H$. So that the maximum number of false genes in the signature is 1.

In order to construct a gene signature with M genes, the fold-change ratio between normal and tumor tissue samples are taken into account. Let μ_1 and μ_2 be the average expression values of gene g in normal and tumor tissue samples, respectively. Then the fold-change ratio of gene g is μ_1/μ_2.

After generating the p-value and fold-change ratio of a gene, if its p-value is smaller than $1/H$ and its fold-change ratio is either larger than R or smaller than $1/R$, the gene is identified as a DEG candidate. R is set to be the threshold of fold-change ratio. Then the satisfied genes are sorted in ascending order based on their p-values. The ith gene in the list is given a weight of $(N - i + 1)/N$, where N is the number of genes in the list. As a result, the top M genes are generated to identify a DEG list of the subset. Finally, k gene lists are obtained from k subsets.

The Wilcoxon test approach

In the Wilcoxon test approach, the p-value of a gene is based on a Z score [35]. Let the vector of differences between normal and tumor tissue samples be $d = (d_1, d_2, \ldots, d_n)$. Then the absolute values of differences are sorted in ascending order $D = (D_1, D_2, \ldots, D_n)$, and a sign vector $q = (q_1, q_2, \ldots, q_n)$ is associated with D, where D_i is the ith smallest absolute value in d. Let d_j be the corresponding value of D_i in d, if d_j is a positive value, then $q_i = 1$, otherwise $q_i = -1$. After that, a rank vector $v = (v_1, v_2, \ldots, v_n)$ is generated, where $v_i = i$. Particularly, if $D_i = D_{i+1} = \cdots = D_{i+j}$, the associated rank value is calculated as follows:

$$v_i = v_{i+1} = \cdots = v_{i+j} = \frac{\sum_{b=0}^{j} (i+b)}{j+1} \tag{3.8}$$

Furthermore the Z score is calculated as follows:

$$Z = \frac{|\sum_{i=1}^{n} q_i v_i|}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \tag{3.9}$$

After that, the p-value is computed based on the Z score. The following steps for obtaining DEGs are similar to those in the previous subsection.

The network-based approach

In the network-based approach, one important step is to identify a DEG network from a PPI network [4]. In this study, we download the PPI data from BioGrid database [200]. Proteins in the PPI network and their corresponding genes in expression datasets are used to construct a gene network [201].

In the PPI network, we have some centrality measures that are appropriate for it. The PPI networks have two properties: small world and scale free [202]. The bridging centrality works well in scale-free networks [203]. Jeong *et al.* propose that proteins with high degree centralities are more likely to be essential proteins [108]. Joy *et al.* conclude that the betweenness centrality is more likely to be essential than the degree centrality [109]. Closeness centrality and clustering coefficient are other commonly used topological parameters in biological network analyses [110, 111].

After obtaining a gene network from the PPI network, DEGs generated from each cluster are mapped into the gene network. In order to obtain a DEG network for each cluster, DEGs and their direct neighbor genes in the gene network are retained. Then all other genes are deleted from the gene network. Finally, the gene network is transformed into a DEG network for each cluster. In the DEG network, the five centralities are used to measure the topological importance of genes, including the degree centrality, betweenness centrality, bridging centrality, closeness centrality, and clustering coefficient.

Let the DEG network be $G = (V, E)$, where $V = (v_1, \ldots, v_{n_1})$ is the set of n_1 vertices and $E = (e_1, e_2, \ldots, e_{n_2})$ is the set of n_2 edges. The degree centrality of a vertex v is calculated as follows:

$$C_D(v) = d(v) = |N(v)| \tag{3.10}$$

where $d(v)$ is the degree of vertex v, and $N(v)$ is the set of all neighbor vertices of v.

The betweenness centrality of a vertex v is calculated as follows:

$$C_B(v) = \sum_{s,v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.11}$$

where σ_{st} is the total number of shortest paths from vertex s to vertex t and $\sigma_{st}(v)$ is the number of those shortest paths that pass through v.

The bridging centrality is calculated as follows:

$$C_R(v) = \frac{1}{d(v) \sum_{i \in N(v)} \frac{1}{d(i)}} \times C_B(v) \tag{3.12}$$

The closeness centrality of vertex v is calculated as follows:

$$C_C(v) = \frac{1}{\sum_{s \in V, s \neq v} dis(v, s)} \tag{3.13}$$

where $dis(v, s)$ is the distance between vertices v and s.

The clustering coefficient is calculated as follows:

$$C_L(v) = \frac{2 \times tri(v)}{|N(v)| \left(|N(v)| - 1\right)} \tag{3.14}$$

where $tri(v)$ is the number of triangles consisting of vertex v and its neighbors in G.

In each type, the centrality values are normalized to [0,1], so that each gene has a new value. Then the five values of a gene are summed up to a new weight, whose range is [0,5]. All genes are sorted in descending order according to the new weight. Then the ranked gene list is used to generate a gene signature.

3.2.4 The gene signature determination

In previous sections, we have generated several gene lists from each of the methods applied in a dataset of the cancers. In this section, we describe a weighting approach to determine the gene signature from those gene lists.

Suppose we have L datasets of a cancer, each has k clusters. In each cluster, several tumor tissue samples and normal tissue samples are contained. One of three previous approaches is used to generate a gene list from a cluster. Then we are handling with $L \times k$ gene lists. Each gene in the list has a sign, either "+" or "-", corresponding to the up-regulation or the down-regulation. In order to identify the up-regulation and the down-regulation, the average expression values of genes in tumor and normal tissue samples are calculated. An up-regulated gene has a larger average expression value in tumor tissue samples than that in normal tissue samples. A down-regulated gene is the opposite, which has a smaller average expression value in tumor tissue samples than that in normal tissue samples.

In addition, a gene on each of the lists has a weight, which is based on three factors, including p-values, statistical powers, and sample size. The p-values are used to describe the Type I error (also known as the false positive), while the statistical powers are used to describe the probability of Type II error (the false negative).

The first factor of weight depends on the rank of p-value in the gene list. The genes are sorted in ascending order based on their p-values. The ith gene in the list has the ith smallest p-value. Then its first factor of weight is $w_1 = (n_l + 1 - i)/n_l$, where n_l is the length of the gene list. If the gene is up-regulated, w_1 has a positive sign, otherwise, it has a negative sign. The larger the statistical power is, the lower probability the Type II error occurs. The statistical power (SP) is the second part of the weight, i.e., $w_2 = SP$. The sizes of the clusters are different. Then we use the size ratio $w_3 = n_c/n_d$ to be the third part of weight, where n_c and n_d are the sizes of a cluster and a dataset, respectively. Finally, the weight of a gene in a cluster is calculated as $w_1 \times w_2 \times w_3$.

In this multiplication procedure, normalization is not an essential step. The ranges of w_1 and w_2 are [0,1] and the w_3 only has 2 possible values, when k is set to be 2. If we apply normalization to the weights, the possible values of w_3 are 0 and 1, that the identified genes are based on the larger clusters. After the multiplication, the final weights of a gene are summed up on all $L \times k$ gene lists and sorted in descending order according to the absolute value. In this procedure, normalization is not essential yet. Suppose the values of w_3 are different, then the ranges of $w_1 \times w_2 \times w_3$ in different clusters are not the same. If we apply a normalization, the w_3 fails to play a role. In addition, we think these three weight factors independently contribute to the final weight. According to the Bayesian rule, the multiplication of independent contribution is more reasonable. So we do not apply normalization after the multiplication.

After generating the final gene list, the top M genes are identified as the gene signature of cancer. The largest value of M is described in 3.2.3. In the BH procedure, it tends to be a strong assumption that there are few signals. In the microarray studies, most genes are not related to the cancers [204]. After the multiplication of $w_1 \times w_2 \times w_3$, these assumptions are also satisfied, that more than 99.7% of the multiplied values in the experiments are 0.

3.2.5 The connection score calculation

In this study, the sscMap platform [205] is used to calculate the connection score between a cancer (represented by its gene signature) and a drug candidate (represented by its induced cell line expression profile in the CMap database).

Given a cancer gene signature $G = (g_1, g_2, \ldots, g_M)$ and a drug-induced profile $P_j (1 \leq j \leq N)$, where M is the number of genes in the integrated gene signature, and N is the number of drug-induced profiles in the CMap database. The genes in P_j are sorted in descending order based on their expression values and $P_j(g_i)$ is denoted as the rank of gene g_i in the profile P_j. Then an intermediate connection score ICS is calculated as follows:

$$ICS(G, P_j) = \sum_{i=1}^{M} s(g_i)(I + 1 - P_j(g_i)) \tag{3.15}$$

where I is the number of genes in the drug-induced profile.

A positive maximum connection score occurs when all the genes in a signature G are up-regulated genes and they are the same as the top s genes in a drug-induced profile P_j. Then a positive maximum connection score $PMCS$ is calculated as follows:

$$PMCS(G, P_j) = \sum_{i=1}^{M} (I + 1 - i) \tag{3.16}$$

Then the connection score between a cancer gene signature G and a drug-induced profile P_j is calculated as follows:

$$CS(g, P_j) = \frac{ICS(G, P_j)}{PMCS(G, P_j)} \tag{3.17}$$

In general, the range of the connection score is [-1,1]. A connection score of -1 indicates that the cancer gene signature and the drug-induced profile are most negatively correlated, which is the best situation that the drug has a potential treatment.

Additionally, a p-value is assigned to the connection score $CS(G, P_j)$. A large number of random gene signatures are identified that the number of genes in a random gene signature is set to be n. Then the connection scores between the random gene signatures and the drug-induced profile P_j are obtained. After that, the p-value is the ratio of the random gene signatures whose connection scores are smaller than $CS(G, P_j)$. The p-value threshold is set to be $1/U$, where U is the number of drugs in the CMap database. Finally, only the drugs whose connection scores are negative and p-values are smaller than $1/U$ are identified as drug candidates.

3.2.6 Datasets

In this study, all gene expression datasets of the cancers are downloaded from the Gene Expression Omnibus (GEO) database [187]. In the GEO database, each cancer has several datasets. However, many of those datasets contain tumor tissue samples only. In our proposed framework, the generated datasets should contain both cancer samples and normal samples. The datasets of cervical cancer (CC), prostate cancer (PC), kidney cancer (KC), breast cancer, (BC) colorectal cancer (CRC), and non-small cell lung cancer (NSCLC) are utilized in the experiments.

BC is the most common cancer in women, the cancer cells are formed in the breast. In order to study its gene signature, three gene expression datasets of breast cancer GSE10780, GSE15852, and GSE50948 are used in this study. PC is the most common cancer in men. It starts in the prostate. The dataset GSE46602 is used in this study. Lung cancer is the second most common cancer in both men and women. About 85%

of lung cancers are NSCLC. Three datasets GSE10072, GSE19804, and GSE27262 are used in this study. CRC is the third leading cause of cancer-related deaths in both men and women in the United States. The cancer cells form in the colon or rectum. The datasets GSE21510, GSE41258, and GSE49355 are used in this study. CC is the fourth most common cancer in women. The dataset GSE63514 is used in this study. KC is a disease that starts in the kidney. The terms “kidney cancer” and “renal cell carcinoma (RCC)” are often used interchangeably. In order to analyze its gene signature, the dataset GSE53757 is downloaded.

Table 3.1: The number of samples and platforms in each dataset

Cases	Datasets	Platforms	Numbers of Samples
Breast	GSE50948	GPL570	80
	GSE15852	GPL96	86
	GSE10780	GPL570	84
Cervical	GSE63514	GPL570	48
Colon	GSE21510	GPL570	70
	GSE41258	GPL96	88
	GSE49355	GPL96	30
Kidney	GSE53757	GPL570	144
Lung	GSE10072	GPL96	48
	GSE19804	GPL570	96
	GSE27262	GPL570	50
Prostate	GSE46602	GPL570	28

All the datasets are listed in Table 3.1 and belong to two platforms: GPL96 and GPL570. The GPL96 platform contains 22,283 probe sets, while the GPL570 platform contains 54,675 probe sets. Although the GPL 570 platform produces more information than the GPL96, the drug repositioning profiles in CMap are based on the GPL96 platform. In order to integrate the datasets from two platforms, we generate datasets with the 22,277 common probe sets among them. All the datasets are normalized using Robust multi-array average (RMA) method [206] and log2-transformed.

In addition, we also study the associations of RNA_Seq datasets with the CMap database. The RNA_Seq datasets are downloaded from the Cancer Genome Atlas (TCGA) program in the National Institutes of Health (NIH) [207]. In order to study the performance of RNA-Seq datasets, we generate 6 datasets from the database, including breast, bronchus and lung, cervix uteri, colon, kidney, and prostate. In addition to the previous approaches, we utilize two new approaches to identify DEGs from RNA_Seq datasets, which are DESeq2 [208] and edgeR [209]. However, the prediction rates of the signatures generated from edgeR are 0 in all cases. Meanwhile, the prediction rates of DESeq2 are 0 in 3 cases, 0.1 in 2 cases, and 0.2 in the colon tumor case. Then we tried to scrutinize the possible reasons. The number of probes in RNA_Seq datasets is 60,483. Mapping the genes in the RNA_Seq dataset to the genes in CMap is an essential process. The Entrez gene IDs are used to be an intermediate to connect these two coding projects. However, only 13,845 probes in RNA_Seq data have their corresponding Entrez genes. Most of the information is lost, which leads to worse results. Thus we do not utilize the RNA_Seq data in the experiment.

3.3 Results and discussion

In this section, we apply our proposed GS4CDRSC framework on six types of cancers, as described above. In the experiments, the gene signatures are generated by the three methods described in Section 3.2 with GS4CDRSC, including the clustering and weighting procedures. In order to make a comparison, the gene signatures are also generated by those methods without GS4CDRSC.

When evaluating the performance of drug repositioning methods, the prediction rate is proposed, which is the rate of the predicted known drugs to all the predicted drugs. The known drugs are the drugs that have shown their therapeutic effects in particular cancer, alone or cooperating with other drugs. The annotations of all known drugs identified by GS4CDRSC are discussed in each case. For the approach to identify a gene signature of cancer, the larger prediction rate with the gene signature can obtain, the better accuracy the gene signature should be. Then the other drug candidates have the potential to achieve the same treatments with the known drugs. We also discuss some annotations about the potential drugs.

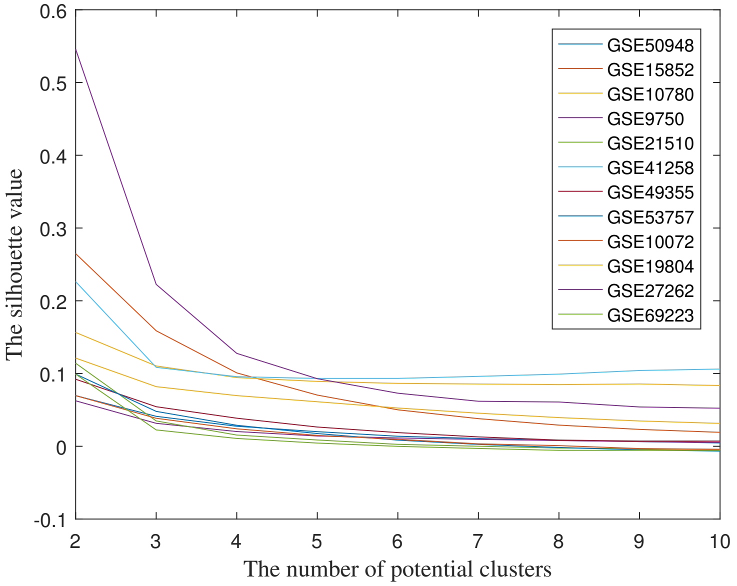


Figure 3.2: The Silhouette values in each dataset. k is ranging from 2 to 10.

3.3.1 Cluster analysis

Before we compare the performance of the approaches with and without GS4CDRSC, we first determine the value of k for the k-means algorithm in GS4CDRSC. Actually, the determination of k for the k-means algorithm is a challenging issue. Although there is no best method for this issue in principle, one of the useful empirical methods is the Silhouette method [210]. In this study, the Silhouette method is utilized to generate validation of consistency within clusters.

As discussed in previous section, given a sample s_i in a cluster S_I, the mean distance between s_i and all other samples in cluster S_I is

$$a(i) = \frac{1}{|S_I| - 1} \sum_{j \in S_I, i \neq j} dist(s_i, s_j) \tag{3.18}$$

Then the smallest distance between s_i and all samples in any other clusters is

$$b(i) = \min_{K \neq I} \frac{1}{|S_K|} \sum_{j \in S_K} dist(s_i, s_j) \tag{3.19}$$

Now we can calculate a silhouette value of a sample s_i:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}, |S_I| > 1 \tag{3.20}$$

Table 3.2: The average statistical powers and the number of tumor-normal sample pairs in all clusters and datasets.

Cancers	Datasets	Cluster 1	Cluster 2	Undivided
NSCLC	GSE10072	0.9873 15	0.9867 9	0.9873 24
	GSE19804	0.9892 23	0.5539 25	0.9878 48
	GSE27262	0.8153 9	0.8134 16	0.8153 25
CRC	GSE21510	0.9954 25	0.9950 10	0.9954 35
	GSE41258	0.9928 34	0.9903 10	0.8282 44
	GSE49355	0.8794 8	0.7011 7	0.7838 15
CC	GSE63514	0.8330 9	0.8199 15	0.3412 24
PC	GSE46602	0.8681 7	0.8624 7	0.8681 14
KC	GSE53757	0.9964 34	0.9967 38	0.9964 72
BC	GSE50948	0.9347 19	0.5966 21	0.8924 40
	GSE15852	0.9858 23	0.5680 20	0.9737 43
	GSE10780	0.9883 24	0.9887 18	0.9480 42

If $|S_I| = 1$, then $s(i) = 0$. Thus the average value of $s(i)$ over all samples is a measure of how appropriately the dataset has been clustered. A larger silhouette value refers to a better cluster result. As shown in Figure 3.2, we have evaluate the 12 datasets in our experiments and the value k is ranging from 2 to 10. When k is set to be 2, the Silhouette values achieve the largest values in all the datasets. Then we utilized k=2 in our experiments.

Table 3.3: The prediction rates by two types of gene signatures identified in six cancer cases and three approaches.

Cancers	Approaches	Without	With
NSCLC	Moderated T-test	0.20	0.20
	Wilcoxon	0.12	0.67*
	Network-based	0.14	0.30
CRC	Moderated T-test	0.20	0.40
	Wilcoxon	0.12	0.30
	Network-based	0.12	0.60
CC	Moderated T-test	0.00	0.00
	Wilcoxon	0.00	0.29*
	Network-based	0.00	0.20
PC	Moderated T-test	0.00	0.40
	Wilcoxon	0.25	0.25*
	Network-based	0.00	0.60
KC	Moderated T-test	0.00	1.00*
	Wilcoxon	0.00	0.25*
	Network-based	0.00	0.00
BC	Moderated T-test	0.13	0.30
	Wilcoxon	0.09	0.30
	Network-based	0.07	0.20

*: The number of drugs in the result is less than 10. Without: The signatures are generated from the datasets without our proposed framework. With: The signatures are generated from the GS4CDRSC framework with the clustering and weighting procedures.

3.3.2 Statistical analysis

In our GS4CDRSC framework, we used the k-means algorithm to identify clusters from gene expression datasets. However, compared with the whole dataset, the size of each cluster is smaller. In this section, we learn the statistical influence of the changes in the sizes. The statistical power (SP) is the probability that it will reject a false null hypothesis.

Suppose we have n_p pairs of tumor-normal tissues. The average expression value of gene g in tumor tissues is m_t, while the standard deviation is sd_t. Then average expression value of gene g in normal tissue is m_n while the standard deviation is sd_n. The confidence level of the test is set to be 0.05, then the critical z score is 1.96 and -1.96. The SP is calculated as follows:

$$\begin{aligned} SP =& \Phi(Z > 1.96 - \frac{(m_t - m_n)\sqrt{n_p}}{sd_n}) + \\ & 1 - \Phi(Z > -1.96 - \frac{(m_t - m_n)\sqrt{n_p}}{sd_n}) \end{aligned} \tag{3.21}$$

where the z score has a corresponding confidence level. Then the SP of gene g is obtained.

The SP is inversely related to the probability of making a Type II error. If a DEG has a large SP, it has a small possibility to be a non-DEG. In order to study the difference of SPs between the undivided dataset and the clusters, we calculate the SPs of all genes. We generate the average SPs of DEGs in each case and list them in Table 3.2. Among the 24 clusters, the average SPs of 11 clusters are larger and those of 5 clusters are equal to those of the datasets. Then we can conclude that although the k-means algorithm decreases the size of profiles in each cluster, the SPs achieve benefits from it in a larger part.

Table 3.4: The rates of the overlapped genes in the signatures from the approaches with our proposed framework, compared to the approaches without our proposed framework.

Cases	Moderated T-test	Wilcoxon test	Network-based
NSCLC	0.49	0.04	0.37
CRC	0.03	0.08	0.44
CC	0.18	0.05	0.18
PC	0.22	0.02	0.11
KC	0.03	0.00	0.52
BC	0.03	0.03	0.23

3.3.3 Experiments

In the experiments, we applied our proposed GS4CDRSC framework to six types of cancers. In order to make a comparison between with and without using the clustering and weighting approaches, we utilized

Table 3.5: The rates of overlapped genes between two or three approaches in all cases. All the approaches are combined with our proposed framework.

Cases	Compare 1	Compare 2	Compare 3	Compare 4
NSCLC	0.22	0.40	0.21	0.11
CRC	0.13	0.23	0.18	0.06
CC	0.17	0.17	0.18	0.05
PC	0.16	0.15	0.05	0.05
KC	0.03	0.17	0.04	0.00
BC	0.16	0.13	0.09	0.02

Compare 1: Between the moderated T-test and network-based approaches. Compare 2: Between the moderated T-test and Wilcoxon test approaches. Compared 3: Between the network-based and Wilcoxon test approaches. Compared 4: Between all the 3 approaches.

three different approaches to identify DEGs in our framework, as shown in Table 3.3. The prediction rates of all the comparisons are listed in Table 3.3. In most cases, we use the gene signature to identify 10 potential drugs. However, the numbers of potential drugs in some cases are less than 10.

In two cases CRC and BC, our GS4CDRSC framework achieves higher prediction rates than without it. In the CC and KC cases, the approaches without our proposed framework cannot identify any known drug. In the PC and NSCLC cases, our GS4CDRSC framework could improve the prediction rates in two out of three approaches. The weakness of our proposed framework is that it cannot help identify any drug of CC with moderated T-test and KC with network-based approach. All the known drugs are discussed in Section 3.3.5.

3.3.4 Overlaps of the signatures

In the experiments, one type of comparison is the gene signatures between with and without the clustering and weighting approaches in our proposed framework. We generate the rates of overlapped DEGs among the signatures, as shown in Table 3.4. The rates are calculated by Jaccard similarity. In general, the rates are small. The most genes identified with our proposed framework are new. We also compared the numbers of overlapped DEGs between two or three approaches with our proposed framework in Table 3.5.

3.3.5 Annotations of the known drugs

In this section, we discuss the treatments of known drugs in six cases. The predicted drugs for the six types of cancers are listed in Table 3.6. Many researchers have done a lot of studies about those drugs and treatments. In all the cases, the histone deacetylase(HDAC) inhibitor is the largest type of drug. Meanwhile, HDAC inhibitors are used in the clinic of many cancers. Some drugs show individual treatment for specific cancer. Some drug combinations are effective in some clinical trials.

Table 3.6: The known and potential drugs of three cancers identified by the three approaches with GS4CDRSC

Cancers	Approaches	Known drugs in the results	Predicted potential drugs
BC	Moderated T-test	Metformin [211], Oligomycin [212], Danazol [213]	Primidone, Rilmenidine, Propidium iodide, Ozagrel, Oxybenzone, Iohexol, Merbromin, Chlorzoxazone
	Wilcoxon	Rosiglitazone [214], MS-275 [215], TTNPB [216]	Monorden, Indomethacin, Lasalocid, Iloprost, Nadolol
	Network-based	Fulvestrant [217], Metformin [211]	Clopamide, Iloprost, Chlorzoxazone, Dicycloverine, Fludrocortisone, Dirithromycin
CC	Moderated T-test	NULL	NULL
	Wilcoxon	Sirolimus [218], LY-294002 [219]	Latamoxef, CP-645525-01, Zuclopenthixol, Picrotoxinin, Zalcitabine
	Network-based	Sirolimus, Valproic acid [220]	0297417-0002B, SC-19220, CP-645525-01, Prochlorperazine, Oxantel, 15(S)-15-Methylprostaglandin E2, Adipiodone, Nortriptyline
CRC	Moderated T-test	Tetrandrine [221], Indomethacin [222], Valproic acid [223], Erastin [224]	CP-320650-01, Mephenytoin, Beclometasone, Mycophenolic acid, Chlorhexidine, Oligomycin
	Wilcoxon	LY-294002 [225], Thioridazine [226], Trichostatin A [227]	Scopolamine, Zalcitabine, Pregnenolone, Fulvestrant, 6-Bromoindirubin-3'-oxime, 0297417-0002B, Maprotiline

	Network-based	Resveratrol [228], Methotrexate [229], Trichostatin A, Trifluridine [230], Etoposide [231], Irinotecan [232]	0173570-0000, Hycanthone, Daunorubicin, PNU-0251126
KC	Moderated T-test	LY-294002 [233]	Irinotecan
	Wilcoxon	Anisomycin [234]	Fulvestrant, CP-690334-01, BCB000039
	Network-based	NULL	Ciclopirox, Estropipate, Ethisterone, Letrozole, Etiocholanolone, Erastin, Benzathine Benzylpenicillin, Metergoline, Selegiline, Rifampicin
NSCLC	Moderated T-test	Clindamycin [235], Glibenclamide [236]	Clopamide, Ajmaline, Lobeline, Azacyclonol, Ampyrone, Danazol, Dirithromycin, Chlorzoxazone
	Wilcoxon	Resveratrol [237, 238], Glibenclamide	Dirithromycin
	Network-based	Indomethacin [239], Glibenclamide, Clindamycin	TTNPB, Anisomycin, Tetraethylenepentamine, Benzathine benzylpenicillin, Pirinixic acid, Lobeline, Ajmaline
PC	Moderated T-test	Pyrvinium [240], Trichostatin A [227]	Prochlorperazine, Diclofenamide, Calmidazolium
	Wilcoxon	Geldanamycin [241]	0225151-0000, Dihydroergocristine, Tanespimycin
	Network-based	Desipramine [242], Sirolimus [243], Withaferin A [244], Menadione [245], Thioridazine [246], Gossypol [247]	Thiostrepton, Isocarboxazid, 6-Benzylaminopurine, 0175029-0000

NULL in the table indicates that there is no result in the experiment.

Breast cancer

There are 7 known drugs in the predicted results. Metformin and MS-275 have shown antitumor effects in a variety of cancers. Metformin is an adenosine monophosphate (AMP) kinase-dependent growth inhibitor for breast cancer cells [211]. MS-275 is an HDAC inhibitor, that inhibits the tumor progression, angiogenesis, and metastasis of breast cancer [215]. Oligomycin is a macrolide created by Streptomyces. It abolishes the growth of human breast cancer cells at remarkably low concentrations [212]. Danazol is a medication used in the treatment of endometriosis. It is an effective treatment for advanced breast cancer [213]. Rosiglitazone is an antidiabetic drug. It sensitizes breast cancer cells to anti-tumor effects of TNF-α, CH11 and CYC202 [214]. Fulvestrant is a medication that is used to treat hormone receptor (HR)-positive metastatic breast cancer [217]. Arotinoid acid (TTNPB) proves to be 100 times more effective than all-trans-retinoic acid (atRA), which also has great growth inhibition of breast cancer cells [216].

Cervical cancer

In the results, only 3 drugs have been studied for their treatment of cervical cancer. LY-294002 is s potent inhibitor of numerous proteins and a strong inhibitor of phosphoinositide 3-kinases (PI3Ks). Its PI3K inhibition produces significant radiosensitization and increases apoptosis in human cervical cancer cell lines [219]. Sirolimus, also known as rapamycin, is a macrolide compound. It can significantly enhance the sensitivity of CaSki cells (a type of human cervical cancer cell lines) to paclitaxel, which is effective against cervical cancer [218]. Valproic acid is used to treat certain types of seizures. It has shown its antitumor effects in NSCLC and CRC. It induces proliferation suppression, cell apoptosis, and cell cycle arrest in cervical cancer cells [220].

Colorectal cancer

Among the predicted drug lists, there are 12 drugs whose treatments have been studied, including tetrandrine, indomethacin, valproic acid, erastin, LY-294002, thioridazine, resveratrol, trichostatin A, methotrexate, trifluridine, etoposide, and irinotecan. Valproic acid and trichostatin A are HDAC inhibitors. Valproic acid has been reported to impair the tumor-cell-induced angiogenesis [248]. It has also been shown to enhance the radiation response in CRC [223]. Trichostatin A reverses epithelial-mesenchymal transition in colorectal cancer and induces apoptosis [227, 249].

Tetrandrine has anti-inflammatory, immunologic, and antiallergenic effects. It inhibits Wnt/β-catenin signaling and suppresses tumor growth of human colorectal cancer [221]. Indomethacin is a nonsteroidal anti-inflammatory drug. It suppresses the growth of colon cancer via inhibition of angiogenesis in vivo [222]. Erastin is a small molecule capable of initiating ferroptosis cell death. It disrupts mitochondrial permeability transition pore (mPTP) and induces apoptotic death of colorectal cancer cells [224]. LY-294002 is a PI3K inhibitor. It has been demonstrated to inhibit cell growth and induce cell apoptosis in colon cancer cell lines [225]. Thioridazine is an antipsychotic drug. It inhibits the proliferation of colorectal cancer stem cells through induction of apoptosis [226]. Resveratrol can depress the growth of colorectal aberrant crypt foci by affecting bax and p21 expression [228]. In further studies, it can inhibit the invasion and metastasis of CRC, in which long non-coding Metastasis Associated Lung Adenocarcinoma Transcript 1 (RNA-MALAT1) plays an important role [250].

Methotrexate is an immune system suppressant that also has anti-tumor treatments in breast cancer and lung cancer. The combination of leucovorin and fluorouracil with it is an active regimen in advanced colorectal cancer [229]. Trifluridine is an anti-herpesvirus antiviral drug. It has recently been approved for the treatment of adult patients with metastatic colorectal cancer [230]. Etoposide is a chemotherapy medication used for the treatment of several types of cancer. It has anti-proliferative effects in colon cancer cells [231]. Irinotecan is a medication used to treat colon cancer and small cell lung cancer. The treatment of it plus fluorouracil and leucovorin is better than a widely used therapeutic regimen of fluorouracil and leucovorin [232]

Kidney cancer

Only 2 drugs have shown their treatment for kidney cancer. LY-294002 is a PI3K inhibitor and PI3K-Akt signaling cascade is, in theory, an ideal therapeutic target for this kidney cancer [251]. The combination of LY-294002 with gefitinib suppresses the viability of gefitinib-resistant kidney cancer cell lines [233]. Anisomycin is an antibiotic that inhibits eukaryotic protein synthesis. It sensitizes human kidney cancer cells to the tumor necrosis factor (TNF)-related apoptosis-inducing ligand (TRAIL)-induced apoptosis [234].

Non-small cell lung cancer

Among the prediction results of our proposed framework, there are 4 drugs, clindamycin, glibenclamide, resveratrol, and indomethacin, whose treatments have been studied. Glibenclamide is predicted by all three approaches, which is a medication used to treat diabetes mellitus type 2. It inhibits multidrug resistance protein 1 (MRP1) activities in human lung cancer cells and enhance their sensitivity to anti-cancer drugs [236]. Clindamycin is predicted by two of the approaches. It is a type of antibiotic. The combination of clindamycin and erlotinib is used for treating NSCLC and reducing the side effect of skin rash [235]. Resveratrol is a stilbenoid, a natural phytoalexin found in many food products, which can down-regulate the expression of survivin and induce apoptosis in multidrug-resistant human NSCLC cells [237]. In addition, resveratrol can enhance the anti-tumor effects of the epidermal growth factor receptor (EGFR) inhibitor erlotinib in NSCLC cells [238]. Indomethacin is a nonsteroidal anti-inflammatory drug. It induces apoptosis in a doxorubicin-resistant lung cancer cell line through an MRP1-dependent mechanism [239].

Prostate cancer

There are 9 known drugs in the predicted results. Trichostatin A is an HDAC inhibitor and has shown antitumor effects in different types of cancers. It reduces cell invasion and migration abilities in prostate cancer cells [227]. Pyrvinium is a known drug for cervical cancer. Androgen receptor (AR) is a type of nuclear receptor. It has a key role in prostate cancer progression [252]. Pyrvinium can suppress prostate cancer cells through endogenous AR in human prostate cancer cell lines [240, 253]. Gossypol is a nature phenol derived from the cotton plant. It is currently in phase II clinical trials as adjuvant therapy for human prostate cancer [247]. Geldanamycin is an antitumor antibiotic that has inhibition of angiogenesis in prostate cancer cells [241].

Desipramine is a tricyclic antidepressant (TCA) used in the treatment of depression. It causes apoptosis via inducing c-Jun NH2-terminal kinase (JNK)-associated caspase-3 activation [242]. Sirolimus shows treatment in both androgen-dependent and independent prostate cancer cells [243]. Withaferin A is a steroidal lactone. It induces mitotic catastrophe and growth arrest in prostate cancer cells [244]. Menadione is an organic compound. The combination of ascorbate and menadione induces cell death in human prostate cancer cells [245]. Thioridazine has shown treatment in colorectal cancer. It significantly inhibited the growth of prostate cancer cells in vitro (including androgen-independent colonies) [246].

Discussions about the predicted drugs

In the experiments, we have identified some small compound drugs that have shown treatments against cancers and some drugs that may have potential treatments. In former subsections, we have talked about the treatments of the known drugs, which are side witnesses of the predicted drugs. In this section, we discuss some of the predicted drugs that have anti-tumor effects in a variety of cancers.

Among the predicted results of NSCLC, danazol and TTNPB have shown some treatments against a variety of cancers, which denotes the potential anti-tumor effects on NSCLC. In the predicted drugs of CRC, chlorhexidine, daunorubicin and oligomycin are known drugs for different cancers. In the third CC case, nortriptyline has shown treatments in many types of cancers. In the predicted drugs of PC, tanespimycin and thiostrepton are identified as anti-tumor agents in a variety of cancers. In the results of KC, irinotecan, fulvestrant, and erastin have some treatments for different cancers. In the predicted drugs of BC, clindamycin, estradiol, gabexate, and altretamine are anti-tumor agents in many cancers. Especially, altretamine is predicted by all three approaches with our proposed framework.

3.4 Conclusion

In this study, we have proposed a GS4CDRSC framework to identify a gene signature of specific cancer for drug repositioning. After sample clustering, the existing DEG approach is performed many times based on the k clusters. At each time, a list of DEGs is identified from each cluster. Then the DEGs from all clusters are used to generate an integrated gene signature. Comprehensive experiments have been conducted to evaluate the performance of the proposed framework. The results demonstrate the effectiveness of GS4CDRSC in identifying a gene signature. With the proposed framework, the gene signatures identified from existing approaches can obtain more known drugs and the prediction rates of known drugs in drug candidates are larger than the approaches without the framework. In the future, we would study more data and expand the applications of the proposed framework for drug repositioning.

Acknowledgments

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC) and by the National Natural Science Foundation of China under Grant No. 61772552 and No. 61428209.

4 Human protein complex signatures for drug repositioning

Prepared as: Fei Wang, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Human protein complex signatures for drug repositioning. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. 2019. FW and FXW discussed the methods. FW implemented the algorithm, designed and performed the experiments. FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

As discussed in Chapter 2 and 3, the signature-based methods identify gene signature from disease tissue samples. Genes are treated as independent elements to represent a disease. However, they may cooperate in disease conditions. In order to reflect the dependencies of genes, I generate the protein complex information in this chapter. The protein complex signature achieves better performance than that of the gene signature. This chapter fulfills Objective 3 of this dissertation.

Abstract

Drug repositioning approaches are attracting more and more attention in the drug discovery field. Benefiting from the high-throughput gene expression data, many computational drug repositioning approaches use gene signatures to represent diseases and drugs, to identify potential drugs for diseases. Then the gene signature is used to identify potential drugs for a disease. However, the gene signatures do not take the dependencies between genes into account in the development of diseases. In this paper, we proposed human protein complex (HPC) signatures to identify potential drugs for diseases. The human protein complex (HPC) features are identified from the comprehensive resource of mammalian protein complexes (CORUM) database.

Based on the gene expression values, the HPC expression values are calculated. All the gene expression profiles of diseases and drug perturbations are transformed to HPC profiles. The HPC signatures are identified from the profiles and a list of drug candidates is generated. The results of 5 cancers indicate that the proposed method identifies more known drugs, compared with gene signature methods.

4.1 Introduction

In the past decades, drug repositioning achieved large progress in drug discovery. In traditional drug discovery approaches, a new drug often costs 8-10 years and 0.8-1.5 billion US dollars before it can be sold in the market

[26]. Reducing such costs is the very first aim of drug repositioning. Drug repositioning has brought some drugs to the market, such as sildenafil for erectile dysfunction [180] and retinoic acid for acute promyelocytic leukemia [181].

The initial drug repositioning approaches are phenotypic drug screening and target-based methods [29]. Between 1999 and 2008, 28 small molecules were identified by phenotypic drug screening, and 17 were proposed by target-based methods [30, 31]. However, the efficiency of both the phenotypic drug screening and target-based methods is limited. As an improvement, the computational approaches can study almost all small compounds in a short time and identify drug candidates in great efficiency [183].

Benefiting from the applications of high-throughput technologies and databases, many computational approaches are used in drug repositioning studies, including pathway-based methods [254, 255], similarity-based methods [256, 257], network-based methods [194, 258, 259, 260], signature-based methods [191, 261, 262], *et al.* The signature-based methods put more attention on the genes whose expression values are significantly changed during disease development. Many gene expression databases are proposed to make those methods more efficient.

In 2006, Lamb *et al.* proposed a drug perturbation database named Connectivity Map (CMap), where a large number of gene expression profiles under specific drug perturbation cultures are encompassed [146, 147]. In their work, a gene signature is used to represent a biological condition and a rank-based matching strategy based on the Kolmogorov-Smirnov statistic [263] is used to calculate the connection score between a gene signature of a disease and a drug perturbation profile. The drug candidates are the drugs that have satisfied the connection scores. In 2008, Zhang *et al.* proposed a simpler and more robust matching method based on the CMap database, named statistically significant connections' map (sscMap), where the statistical significances of all connections were calculated [205, 264]. Wen *et al.* used the sscMap method to study drug candidates for colorectal cancer [14].

However, one significant limitation of CMap is the data coverage. Only 5 cell lines and approximately 1300 small molecules are encompassed. Among them, the number of Food and Drug Administration (FDA)-approved drugs are even smaller. In 2015, the Library of Integrated Network-Based Cellular Signatures (LINCS) program was proposed to create a network-based understanding of biology [169]. The drug perturbation database is an important component of the LINCS program. The LINCS database Phase I, which encompassed 1,319,138 profiles, approximately 70 cell lines and 20,000 small compound perturbations, was published in 2015. Based on the LINCS database, researchers use gene signature-based methods to study drug repositioning [191].

In both CMap and LINCS databases, each drug perturbation expression profile is based on gene features. The disease profile, which is used to identify a gene signature, is also a series of genes expression values. The gene signatures are the connections between drug perturbation profiles and diseases. In those methods, genes are considered as independent elements to represent a disease or a drug. Actually genes work together in terms of protein complexes in the development of diseases [265, 266, 267, 268].

In order to reflect the dependencies of genes in the signature of a disease, we use protein complexes to represent a disease in drug repositioning. A protein complex is a group of proteins that work together in a certain biological process. Proteins in a complex are highly interactive with each other [269, 270]. In our method, we use the human protein complexes (HPCs) to reflect the interactions and co-operations among genes and products. Those HPC signatures are identified from the comprehensive resource of mammalian protein complexes (CORUM) database. Since each HPC has one or more genes, the gene expression profiles of diseases in previous chapter are replaced by disease-HPC expression profiles in this chapter. Then an HPC signature is identified from the HPC profiles. Meanwhile, the drug perturbation profiles in LINCS are also transformed into drug perturbation-HPC profiles. Finally, a connection method is used to calculate the connection scores between an HPC signature and drug perturbation-HPC profiles, and a list of drug candidates is generated.

In order to illustrate the performance of our proposed method, we compare it with two gene signature methods. All three methods are examined in data sets of 5 cancers. In each experiment, the top 20 small compounds in the result are identified as a list of drug candidates. Among them, the drugs whose treatment has been studied are known drugs and other drugs in the list are potential drugs. The number of known drugs in a list is utilized as an evaluation matric. The HPC drug repositioning (HPCDR) method identifies the largest number of known drugs among all three competing methods. Additionally, we study the annotations of the drugs in the DrugBank database. Some known drugs and potential drugs have been identified as antineoplastic agents.

4.2 Methods and materials

In order to identify new potential treatments of old drugs, we propose a novel approach, named HPCDR, to study drug repositioning. The HPCDR method identifies human protein complex (HPC) signatures, instead of gene signatures. Figure 4.1 illustrates the flowchart of the HPCDR method. Figure 4.1-A, -B, and -C describe the databases used in HPCDR. Drug perturbation profiles are from the LINCS database Phase I. Human protein complexes are from the CORUM database. Microarray data are downloaded from the GEO database and mapped to Entrez gene profiles. Figure 4.1-D and -E illustrate the next steps. Both drugs and diseases profiles are mapped to HPC profiles by taking the average values of all genes belonging to an HPC. Figure 4.1-F is the process to identify an HPC signature from the disease profiles. Then the connection scores between the HPC signatures and the drug perturbation-HPC profiles are calculated. All the scores are sorted in ascending order and the top N drugs are identified as drug candidates for that disease.

4.2.1 Datasets

In this paper, the gene expression profiles are downloaded from the Gene Expression Omnibus (GEO) database [187], which is built by the National Center for Biotechnology Information (NCBI). It archives

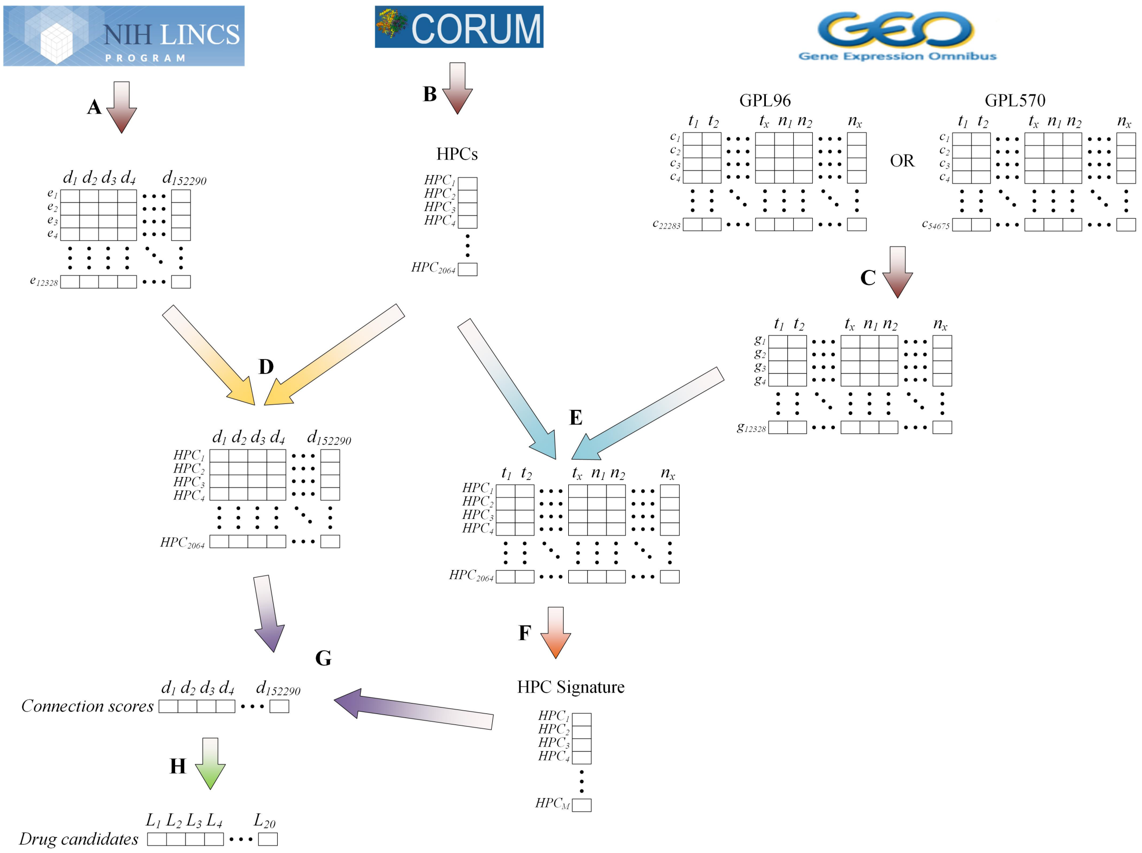


Figure 4.1: The flow chart of our HPCDR method. (A): The drug perturbation profiles are from the LINCS database. (B): HPCs are selected from the CORUM database. The number of satisfied HPCs is 2,064. (C): Microarray data were downloaded from Gene Expression Omnibus (GEO) database. The microarray data is mapped to the Entrez genes profile. (D): Based on the HPCs, drug perturbation profiles in the LINCS database are transformed into drug perturbation-HPC expression profiles. (E): Based on the HPCs, the Entrez gene expression profiles of disease are transformed into disease-HPC expression profiles. (F): An HPC signature is identified from the disease-HPC expression profiles. (G): A connection method is used to calculate 152,290 connection scores between the HPC signature and profiles. (H): The connection scores are sorted in ascending order and the top 20 perturbations are identified as drug candidates.

microarrays and other forms of high-throughput genomic data. In our study, we downloaded the microarray data of 5 cancers, which represent the expression values of genes. In the GEO database, the number of data sets of a specific cancer is very large. However, many of the datasets contain only tumor tissue samples. In order to achieve a meaningful signature from a data set, we utilize the data set which contains both tumor and normal tissue samples. Each tumor tissue sample has a corresponding normal tissue sample. The details of the datasets are listed in Table 4.1.

Besides the gene expression profiles of diseases, we generate the drug perturbation profiles from the LINCS database. Many types of perturbations are compassed in the database, including 19,811 small compound drugs, 18,493 shRNAs, 3,462 cDNAs, and 314 biologics. In order to ensure the small compound drugs are safe, we concentrate on the profiles of FDA-approved drugs in our study. The number of generated small

Table 4.1: The disease datasets

Disease	GEO serial numbers	Platforms	Number of samples
Breast Cancer	GSE10780	GPL570	84
	GSE15852	GPL96	86
	GSE50948	GPL570	80
Cervical Cancer	GSE63514	GPL570	48
Colorectal Cancer	GSE21510	GPL570	70
	GSE41258	GPL96	88
	GSE49355	GPL96	30
Kidney Cancer	GSE66272	GPL570	54
Lung Cancer	GSE10072	GPL96	48
	GSE19804	GPL570	96
	GSE27262	GPL570	50

compounds is 1,273, while that of profiles is 152,290.

4.2.2 HPCs

In previous studies, the drug repositioning methods paid attention to gene signatures, that each gene is considered as an independent unit. However, genes often interacted with each other in complex diseases [271]. In order to reveal the dependencies of genes in cancers, many researchers studied proteins encoded by genes and the roles of protein-protein interactions (PPIs) or protein complexes in cancers. Ivanov *et al.* illustrated that PPIs play an important role in tumor progression, invasion, and/or metastasis [272]. Particularly, Li *et al.* proposed that the Hsp70-Bag3 PPI can be a potential target in cancer [268].

A protein complex is a group of proteins that are highly interactive with each other in a certain biological process [273]. The proteins in a complex play similar roles in a biological process. Sabatini illustrated the roles of mammalian target of rapamycin complexes (mTORCs) in pathways and tumors [266]. Fu *et al.* established essential roles of TWIST/Mi2/NuRD protein complex in cancer metastasis [265].

Furthermore, based on our study of PPIs and protein complexes, we consider human protein complexes (HPCs), instead of individual genes (proteins), to represent a disease in this study. All the HPC information is downloaded from the comprehensive resource of mammalian protein complexes (CORUM) database. It compasses 4,275 protein complexes, among which there are 2,916 HPCs. Because genes are contained in an HPC and the coding scheme in LINCS database is Entrez gene coding, the Entrez genes are used to connect the HPC signature of a disease and LINCS drug perturbation HPC profiles. More importantly, all the genes in an HPC should be measured in LINCS database. The number of satisfied HPCs is 2,064. In the following section, all the profiles are transformed into 2,064-dimensional vectors.

4.2.3 Data pre-processing

In this study, the gene expression data from GEO database are obtained from GPL96 and GPL570 platforms, which compass 22,283 and 54,675 probe sets, respectively. All the data sets are normalized using the robust multi-array average (RMA) method and log2-transformed. Because there are 22,277 common probe sets among the two platforms, we study the differences and similarities of the mapping of them and the other 32,398 probe sets. The 22,277 common probe sets are mapped to 12,315 Entrez genes and the other 32,398 probe sets are mapped to 12,321 Entrez genes. Only 6 Entrez genes are different. Then we choose the common probe sets to do the experiments. All gene expression profiles are transformed to 22,277-dimensional vectors.

The second step is to map probe sets to Entrez genes. The drug perturbation profiles in LINCS database are obtained from the L1000 platform, which contains 12,328 Entrez genes. Among them, there are 978 landmark genes and 11,350 inferred genes. The landmark gene expression values are measured directly from the L1000 platform and the inferred gene expression values are calculated based on the landmark genes. Because an Entrez gene has one or more corresponding probe sets, the average gene expression value of those probe sets is used to be the expression value of the Entrez gene. Then both the profiles of diseases and drug perturbations are 12,328-dimensional vectors. Specifically, the expression values of landmark genes are on the top of the inferred genes, in order to make the experiments more convenient.

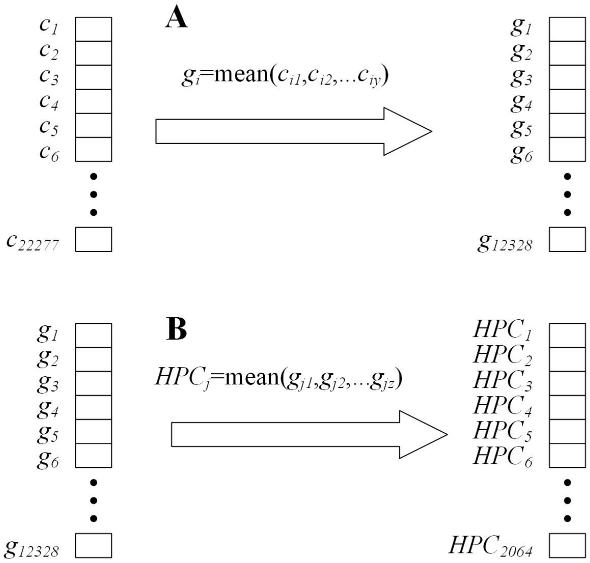


Figure 4.2: The details of the conversion. A: From a gene expression profile of a disease to a profile of Entrez genes. B: From a profile of Entrez genes to that of HPCs.

The third step is to select HPCs from Entrez genes. In CORUM database, most of HPCs contain less than 10 genes. The HPC expression value is the average expression value of genes that belong to it. Then all the Entrez gene expression profiles of diseases and drug perturbations are transformed into 2,064-dimensional

HPC expression profiles, as shown in Figure 4.2.

4.2.4 HPC signatures

In this section, we identify the HPC signatures from the HPC expression profiles of diseases. An HPC can be represented by a 2x-dimensional vector $(t_1, \ldots, t_x, n_1, \ldots, n_x)$, where t_i is the expression value of the HPC in disease tissue profile T_i and n_i is that in normal tissue profile N_i. The fold change ratio r is calculated, based on the average value of *HPC* in disease tissues and normal tissues. Only the HPC whose fold change ratio is larger than 2 is considered as a member of the HPC signature.

Then the paired t-test is used to calculated the statistical significance of the HPCs. The disease-normal difference of HPC_i is denoted as $diff = (t_1 - n_1, t_2 - n_2, \ldots, t_x - n_x)$. Then the *T-score* is calculated as follows:

$$T\text{-}score = \frac{\mu \times \sqrt{x}}{\sigma} \tag{4.1}$$

where μ is the average value of *diff* and σ is the standard deviation of *diff*.

Then a p-value is assigned from the *T-score*, and the HPCs are sorted in ascending ordert according to their p-values. The False Discovery Rate (FDR) α is set to be 0.01 and is controled by the Benjamini-Hochberg procedure [199] as follows:

$$p(M) \leq \frac{M}{H}\alpha \tag{4.2}$$

where H is the number of HPCs in the profile. An HPC whose p-value is smaller than the threshold is identified as a significant HPC. The largest HPC signature length M is 100 to assure that $p(M) \leq 1/H$ so that the maximum number of false HPCs in the signature is 1.

In our experiments, the t-test is calculated in each dataset independently. Each dataset has the same number of normal and tumor profiles to apply the paired t-test.

For the diseases with a single dataset, the HPCs whose fold change ratio is larger than 2 and p-value is smaller than $1/H$ are identified and sorted in ascending order based on their p-values. The top M HPCs are identified as the HPC signature of the disease.

For the diseases with more than one dataset, in each dataset, the HPCs are sorted in ascending order according to their p-values. Each HPC in a dataset has a rank score of $(H + 1 - R)/H$, that R is its rank in the dataset. If the fold change ratio of an HPC is less than 2, then its rank score is set to be 0. The rank scores of an HPC in all datasets of disease are summed up and all features are sorted in descending order according to their total rank scores. The top M HPCs are identified as the HPC signature of the disease.

4.2.5 Matching method

In this section, we use a method to calculate the connection score between an HPC signature and drug perturbation-HPC expression profiles, which is proposed originally to calculate connection scores between a gene signature and CMap profiles [14], and discussed in Section 3.2.5.

Firstly, the drug perturbation-HPC profile $P = (pv_1, pv_2, \ldots, pv_H)$ is replaced by a rank list $PR = (pr_1, pr_2, \ldots, pr_H)$, where pv_i is the expression value of HPC_i and pr_i is its rank in the list. The HPC with the smallest expression value is given a rank of H and the largest one has a rank of 1.

Meanwhile, the HPC signature is divided into two lists, one contains all up-regulated HPCs and another contains all down-regulated HPCs. The up-regulated HPC list indicates that it has a larger expression value in disease tissues than that in normal tissues, while a down-regulated HPC list indicates that it has a smaller expression value in disease tissues than that in normal tissues. Then the *up-score* and *down-score* is calculated as follows:

$$\textit{up-score} = \sum_{i=1}^{H_{up}} (H + 1 - pr(i)) \tag{4.3}$$

$$\textit{down-score} = -\sum_{j=1}^{H_{down}} (H + 1 - pr(j)) \tag{4.4}$$

where H_{up} is the number of HPCs in the up-regulated list and H_{down} is the number of HPCs in the down-regulated list. H is the same variable as mentioned in Section 4.2.4. *pr(i)* is the rank of HPC i in the drug perturbation-HPC list *PR*.

Then a possible maximum connection score is calculated as follows:

$$\textit{poss} = \sum_{i=1}^{M} (H + 1 - i) \tag{4.5}$$

Then a connection score between a HPC signature and a drug perturbation-HPC profile is calculated as follows:

$$\textit{H-score} = \frac{\textit{up-score} + \textit{down-score}}{\textit{poss}} \tag{4.6}$$

In general, its range is [-1,1], a negative score indicates that the drug perturbation reverses the expression of the HPC signature, which means that the drug has a potential treatment for the disease.

All drug perturbations are sorted in ascending order according to their connection scores and the top N drugs are considered as drug candidates for the disease. Since a drug perturbation has more than one profile in the LINCS database, we may have some replicates of a drug among the top N drugs.

4.3 Results and discussion

4.3.1 Parameters and performance evaluation

In order to evaluate the performance of drug repositioning methods, the most commonly used metric is the number of known drugs which are identified by the methods. The known drugs are the drugs whose treatments of a disease have been studied and indicated. In the experiments, given an HPC signature of a disease, we sort the connection scores of all drug perturbation-HPC profiles in descending order and identify the top 20 small compound drugs as the drug candidates for the disease. We compare our proposed HPCDR with two state-of-the-art methods.

In order to analyze the treatments of drugs, we study the annotations in DrugBank database [274]. Some drugs have been identified as antineoplastic agents in DrugBank, that their anti-tumor treatments have been studied. Additionally, the propagation of cancer is a process involving the participation of some enzymes that help develop new drugs [275]. In this study, we also consider the drugs which have been identified as enzyme inhibitors.

4.3.2 Compared with other methods

Entrez gene signatures

In this study, we replace the gene signature with the HPC signature of disease for drug repositioning. In order to illustrate the performance of our proposed method, we use Entrez genes to identify signatures directly and made a comparison with our method.

In this section, all the gene expression profiles of diseases are transformed into profiles of Entrez genes, which are 12,328-dimensional vectors. Similar to our HPCDR method, we use the T-test statistical method [34] to identify DEGs from gene expression profiles of disease and normal tissue samples. Then we calculate the connection scores with drug-perturbation profiles and sort the scores in ascending order. In order to make a comparison, the top 20 small compound drugs are identified as drug candidates.

Landmark gene signatures

Our proposed method use HPC signatures instead of Entrez gene signatures, which can be seen as a feature extraction method. We also compare it with a feature selection method, that we identify landmark gene signatures from Entrez genes. The LINCS drug perturbation profiles contain 12,328 Entrez genes, among which there are 978 landmark genes and 11,350 inferred genes. The expression values of landmark genes are measured directly from the L1000 platform, which can represent approximately 82% information [169]. The expression values of inferred genes are calculated based on the landmark genes.

In this section, the gene expression profiles of diseases are represented by the profiles of landmark genes. The connection scores between disease profiles and drug-perturbation profiles are calculated. The top 20 drug perturbations are identified as drug candidates.

Comparison

In this section, our proposed HPC signature is compared with Entrez gene signature and Landmark gene signature. In order to make a better comparison, we generate the number of known drugs among the top N on the result lists. As the number of connection scores of a disease signature is 152,290, to reduce the scale of drug candidates and focus on the most possible drugs, we set the variable N to be 20. The numbers of known drugs are listed in Table 4.2. One drug has several profiles in LINCS database. They have different concentrations, durations, or cell lines. Therefore, a drug may appear several times among the predicted results. The replicate drugs are deleted in Table 4.3. Based on the results of known drugs, other drugs, which are false positive in the experiments, are lacking in clinical trials. However, that does not mean they are ineffective drugs. They are potential drugs that may have treatment for the given disease.

Table 4.2: The number of known drugs identified by our HPCDR method and two gene signature method

Disease	HPCDR	Entrez gene signatures	Landmark gene signatures
Breast Cancer	12	9	10
Cervical Cancer	10	6	2
Colorectal Cancer	13	8	6
Kidney Cancer	5	2	1
Lung Cancer	10	5	6

The results indicate that our proposed method can identify the most number of known drugs from the five disease data sets. Among 4 out of 5 cancers, the HPCDR method can generate at least 10 known drugs. In kidney cancer, the HPCDR method only identifies 5 known drugs. For the method of Entrez gene signature, the largest number of known drugs is 9. Especially in kidney cancer, only 2 known drugs are obtained. The third method is about landmark gene signature, it only identifies 2 known drugs in cervical cancer and 1 known drug in kidney cancer. In the other three cancers, it generates similar numbers of known drugs with Entrez gene signatures.

4.3.3 Analysis of predictions

In this section, we utilize some literature evidence and annotations in the Drugbank database to analyze the treatments of the drugs which are identified by our method. All the drugs are listed in Table 4.3.

Table 4.3: The drugs identified by our HPCDR method

Disease	Known drugs	Potential drugs
Breast Cancer	aminoglutethimide, atorvastatin, dexamethasone, disulfiram, itraconazole, LY-294002, nitazoxanide, ouabain, resveratrol, vinorelbine, vorinostat	tetracycline, milrinone, nizatidine, clemastine, molsidomine, nimodipine, tolazamide, cefazolin
Cervical Cancer	etoposide, genistein, LY-294002, niclosamide, sirolimus, thioridazine, wortmannin	idarubicin, mitoxantrone, danazol, afatinib, capsaicin, doxepin, tretinoin, digoxin, ABT-751
Colorectal Cancer	atorvastatin, BMS-777607, gefitinib, mitoxantrone, olaparib, saracatinib, vorinostat, zebularine	BMS-777607, mitoxantrone, aliskiren, eplerenone, nifedipine, nimodipine, terconazole
Kidney Cancer	cediranib, panobinostat, tivozanib, vorinostat	brivanib, trimethobenzamide, clofibrate, lorazepam, rivaroxaban, ozagrel, nizatidine, mosapride, ritodrine, exemestane, iniparib, treprostinil, temozolomide, thenoyltrifluoroacetone
Lung Cancer	calcitriol, chlorambucil, entinostat, foretinib, ibuprofen, iloprost, MK-1775, olaparib, pravastatin, tacedinaline, troglitazone, warfarin	fursultiamine, etomidate, fluvoxamine, methantheline, mosapride, trazodone, prazosin

Breast cancer

In the results, 5 of the identified drugs are antineoplastic agents, including aminoglutethimide, dexamethasone, resveratrol, vinorelbine, and vorinostat. Aminoglutethimide has been recognized as a valuable treatment for breast cancer since the 1980s [276]. Dexamethasone is a type of corticosteroid medication, which enhances the effects of ADR on induction of apoptosis and inhibition of cell proliferation [277]. It can also enhance drug efficiency [278]. Resveratrol is a type of natural phenol, which decreases angiogenesis and increases cell apoptosis in vitro and mice experiments [279]. Vinorelbine is an anti-mitotic chemotherapy drug that has been used in the treatment of breast cancer. Vorinostat is a member of histone deacetylases (HDAC) inhibitors. The combination of vorinostat and tamoxifen decreases resistance in breast cancer patients [280].

Besides antineoplastic agents, 5 other drugs are identified as enzyme inhibitors, including atorvastatin, disulfiram. itraconazole, LY-294002 and ouabain. Atorvastatin is a statin medication, that statins increase cell apoptosis, inhibit proliferation and drease metastatic dissemination of breast tumors [281]. The disulfiram-copper complex has the potential to inhibit the proteasomal activity in breast cancer cells [282]. Itraconazole is a member of the triazole medication family, which inhibits breast cancer cell proliferation [283]. LY-294002 is a phosphoinositide 3-kinase (PI3K) inhibitor. The PI3K inhibitor reduces tumor cell proliferation and angiogenesis in a mouse model of breast cancer [284]. Ouabain is a cardiac glycoside and can be used medically in lower doses. The combination of digoxin, proscillaridin A and ouabain induces apoptosis in breast cancer cells [285]. Besides, nitazoxanide induces breast cancer cell apoptosis and suppresses tumor growth [286].

Among the potential drugs whose treatments for breast cancer have not been proposed, there are also two drugs tetracycline and milrinone, identified as enzyme inhibitors. Particularly, tetracycline analogues have shown treatments for prostate cancer [287] and colorectal cancer [288].

Cervical cancer

In the identified drug list, 6 out of 7 drugs are either antineoplastic agents or enzyme inhibitors. Etoposide is a member of the topoisomerase inhibitor family. The combination of etoposide and cisplatin is safe and effective for cervical cancer [289]. Genistein is an angiogenesis inhibitor. It inhibits cell growth in cervical cancer cells [290]. LY-294002 and wortmannin are two PI3K inhibitors, that enhance ratio sensitivity and increase apoptosis [219]. The combination of niclosamide and paclitaxel has been used in the treatment of cervical cancer, where niclosamide sensitizes the responsiveness of cervical cancer cells to paclitaxel [291]. Sirolimus, also known as rapamycin, has a similar treatment of enhancing the sensitivity of cervical cancer cells to paclitaxel [218]. The last drug thioridazine is neither an antineoplastic agent nor an enzyme inhibitor, it induces apoptosis in cervical cancer cells [292].

Among the potential drugs, idarubicin, mitoxantrone, afatinib, tretinoin, and digoxin are either antineoplastic agents or enzyme inhibitors. Particularly, the studies of mitoxantrone [293] and digoxin [294] for prostate cancer, afatinib [295] and tretinoin [296] for lung cancer, have been proposed. The potential treatments of those drugs for cervical cancer should be studied in the future.

Colorectal cancer

In the results of colorectal cancer, atorvastatin and vorinostat have shown treatments for breast cancer in the previous section. Atorvastatin is effective in inhibiting colorectal cancer cells, in combination with celecoxib and aspirin [297]. The combination of vorinostat and bortezomib shows synergistic antiproliferative and proapoptotic effects in colorectal cancer cells [298]. BMS-777607 is a MET tyrosine kinase inhibitor, that has shown promising results in colorectal cancer [299]. Gefitinib is a drug used in the treatment of certain types of cancer [300]. Mitoxantrone is an anthracenedione antineoplastic agent, it shows moderately effective in advanced colorectal cancer cells [301]. Olaparib is a type of poly-ADP ribose polymerase (PARP) inhibitor, which makes colorectal cancer cells sensitive to it [302]. Saracatinib is a dual kinase inhibitor, which has been investigated for the treatments of cancers. It decreases tumor growth in colorectal cancer cells [303].

Zebularine shows anti-tumor activity in colorectal cancer cells [304].

There are 7 potential drugs that their treatments for colorectal cancer can be studied in the future. Particularly, the treatments of BMS-777607 [305] and mitoxantrone for other cancers have been proposed.

Kidney cancer

All of the four identified drugs are both antineoplastic agents and enzyme inhibitors. Cediranib demonstrated significant anti-tumor activity in the treatment of kidney cancer, that its efficacy parameters are comparable to approved drugs [306]. Panobinostat is a non-selective HDAC inhibitor, which inhibits kidney cancer cells [307]. Tivozanib is a vascular endothelial growth factor (VEGF) receptor tyrosine kinase inhibitor and has been recommended in the treatment of advanced kidney cancer [308]. Vorinostat also shows treatment for kidney cancer [309].

15 potential drugs are identified in the results. Among them, the studies of brivanib [310], exemestane [311], iniparib [312] and clofibrate [313] for other cancers have been proposed.

Lung cancer

In the results, 5 out of 12 identified known drugs are either antineoplastic agents or enzyme inhibitors, including chlorambucil, entinostat, ibuprofen, olaparib, and pravastatin. Chlorambucil has been used as an antineoplastic agent for the treatment of various malignant and nonmalignant diseases [314]. The combination of chlortetracycline, nitrogen mustard, and prednisone in lung cancer has been studied [315]. Entinostat is an HDAC inhibitor, which has shown promise in treating lung cancer [316]. Ibuprofen is a medication among nonsteroidal anti-inflammatory drugs, which can enhance the antitumoural activity of cisplatin in lung cancer [317]. Additionally, many drugs show treatment in lung cancer when combined with cisplatin. Calcitriol has shown antiproliferative effects either as a single agent or combined with cisplatin [318]. Olaparib is a PARP inhibitor, the combination of cisplatin with olaparib is more effective than each agent individually [319]. Pravastatin is a statin medication, which reduces progression and limits metastatic diffusion of established hepatocellular carcinoma [320].

Among other known drugs, MK-1775 and tacedinaline have been used in trials studying the treatment of Lung Cancer [321, 322]. Foretinib [323], iloprost [324], troglitazone [325] and warfarin [326] also have treatments in lung cancer.

4.4 Conclusion

Identification of signatures is an important component in computational drug repositioning approaches. In this study, we have proposed a signature identification method, named HPCDR, for drug repositioning. HPCDR generates HPCs from CORUM database. Both the gene expression profiles of diseases and the drug perturbation profiles are transformed into the form of HPCs. The experiments of 5 cancers indicate that our

HPCDR method identifies more known drugs than the other two gene signature methods. The annotations from DrugBank are used to describe the treatments for cancers. In future studies, we would study more applications of HPC signatures in drug repositioning.

Acknowledgments

This work is supported in part by Natural Science and Engineering Research Council of Canada (NSERC), China Scholarship Council (CSC), and by the National Natural Science Foundation of China under Grant No. 61772552 and No. 61428209.

5 Human protein complex-based drug signatures for personalized cancer medicine

Prepared as: Fei Wang, Yulian Ding, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Human protein complex-based drug signatures for personalized cancer medicine. IEEE Journal of Biomedical and Health Informatics. 2021. FW and FXW discussed the methods. FW implemented the algorithm, designed and performed the experiments. FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

As described in Chapter 2, 3 and 4, the disease signatures strategies are identifying signatures from many disease profiles. However, it can not work well when dealing with a single case in practice. In order to address its limitations, I propose a strategy to identify drug signature and employ it in personalized medicine. Our proposed methods could identify a list of potential drugs for even a single patient with high performance. This chapter fulfills Objective 4 of this dissertation.

Abstract

Disease signature-based drug repositioning approaches typically first identify a disease signature from gene expression profiles of disease samples to represent a particular disease. Then such a disease signature is connected with the drug-induced gene expression profiles to find potential drugs for the particular disease. In order to obtain reliable disease signatures, the size of disease samples should be large enough, which is not always a single case in practice, especially for personalized medicine. On the other hand, the sample sizes of drug-induced gene expression profiles are generally large. In this study, we propose a new drug repositioning approach (HDgS), in which the drug signature is first identified from drug-induced gene expression profiles, and then connected to the gene expression profiles of disease samples to find the potential drugs for patients. In order to take the dependencies among genes into account, the human protein complexes (HPC) are used to define the drug signature. The proposed HDgS is applied to the drug-induced gene expression profiles in LINCS and several types of cancer samples. The results indicate that the HPC-based drug signature can effectively find drug candidates for patients and that the proposed HDgS can be applied for personalized medicine with even one patient sample.

5.1 Introduction

In the traditional pharmaceutical industry, putting a new drug in the market is very costly and time-consuming, about ten years and 1 billion US dollars are common in development [3, 26]. Nevertheless, the related budgets are still expanding rapidly. In a traditional drug discovery pipeline, three major procedures are essential: preclinical, clinical trials and regulatory approval [27]. Several thousands of small compound candidates are typically studied to develop one new drug. However, in many projects, no drug can be taken to the market successfully.

In recent decades, drug repositioning has identified some novel treatments for existing drugs, such as sildenafil, thalidomide, zidovudine, minoxidil, and celecoxib [28]. Sildenafil is the most well-known compound in drug repositioning. It was developed for the treatment of coronary artery disease in the 1980s [327], and repurposed to the treatment of erectile dysfunction in the 1990s [180]. Thalidomide was used as a sedative and is now being used to treat multiple myeloma [328].

Two types of approaches have been proposed for drug repositioning initially, which are phenotypic screening and target-based approaches [29]. In the first decade of 21st century, 45 small compounds were proposed by those approaches, 28 of which were identified by phenotypic screening [30, 31]. However, both two types of approaches have some limitations. In phenotypic drug screening, small animal models and cell-based models are necessary. The robustness and relevance of models influence the success of screening [32]. In target-based methods, researchers indicated that only 435 effective drug targets had been proposed [329].

Recently, many high-throughput platforms have been developed to measure the expression values of genes, and some biological databases have been constructed. Many computational approaches have been proposed to use the data more efficiently and identify drug candidates, which are pathway-based methods [254, 255], similarity-based methods [256, 257], network-based methods [71, 76, 259, 260], signature-based methods [3, 122, 191, 261, 262], *etc.* The computational approaches can handle a large number of drug profiles and identify potential drugs for the specific disease in a short period [183].

Lamb *et al.* constructed Connectivity Map (CMap) database which consists of 6,100 profiles under different drug cultures and cell lines [146, 147]. Three main components were utilized in their research. A drug perturbation profile was utilized to describe the differential expression of a drug. A gene signature was a group of significantly expressed genes to represent a disease. A matching strategy was used to connect the drug perturbation profile and the gene signature for producing a connection score [263]. The potential drugs were predicted according to their connection scores.

However, a few cell lines and small compounds were contained in CMap database. Among those small compounds, the Food and Drug Administration (FDA)-approved drugs, which had been studied, were even fewer. Phase I of the Library of Integrated Network-Based Cellular Signatures (LINCS) program was published in 2015, and the sample size of drug perturbation profiles was increased from 6,100 in CMap to 1.3 million [169].

Based on CMap and LINCS databases, many signature-based approaches have been proposed to identify candidates for drug repositioning [14, 205, 264, 330]. In the databases, the expression profiles are based on gene features. In approaches, a signature is a group of genes that are selected independently. Actually, there are some interactions among genes, in the developments of diseases [265, 266, 267, 268, 271]. In order to reflect the dependencies of genes, the associations between genes, proteins, and diseases have been studied [272, 331]. A protein complex is a group of proteins that have strong interactions with each other [332]. The properties of protein complexes and their relationships with diseases have been studied in many studies [333, 334, 335]. Wang *et al.* utilize the human protein complexes (HPCs) to identify new signatures from cancer samples and predict drug candidates for them [3].

In the existing signature-based approaches, a signature is identified from disease samples and compared with the drug perturbation profiles in CMap or LINCS database. In either statistical or network-based approaches, a large number of disease samples are critical in identifying signatures. However, when the disease set has a few samples, it is difficult to identify a reliable signature. These approaches can not especially handle samples from only a few patients.

In this study, instead of creating an HPC-based disease signature from patient samples, we propose an HPC-based drug signature (HDgS) approach to identify drug signatures and predict drug candidates. Based on the HPC information, all drug perturbation profiles and disease samples are transformed into the type of HPCs. An HPC-based drug signature is identified from all the HPC profiles of a specific drug. For disease samples, a differential expression profile is generated. The connection score between an HPC-based drug signature and a patient profile is calculated. Finally, each patient has a list of drugs. After counting the frequencies of drugs that appeared in all lists, ten drugs with the largest frequencies are identified as drug candidates. In the experiments, we compare HDgS with the HPC-based disease signature approach and three other types of drug signatures. Our HDgS approach achieves the highest prediction rates in four types of cancers. The proposed approach can even be used to identify drugs for a single patient, and known drugs are among the prediction results. At the end of the experiments, the annotations, treatments, and literature evidence of the drug candidates are discussed.

5.2 Methods and materials

In this section, we discuss the datasets used in our HDgS approach and the procedures to generate the drug candidates, as shown in Figure 5.1. The human protein complex information in the comprehensive resource of mammalian protein complexes (CORUM) database [336] are utilized in Figure 5.1-I and -II to provide the mapping between genes and protein complexes. The drug perturbation profiles in LINCS are used to produce the drug signatures, which are matched with the patient profiles to generate a list of drug candidates.

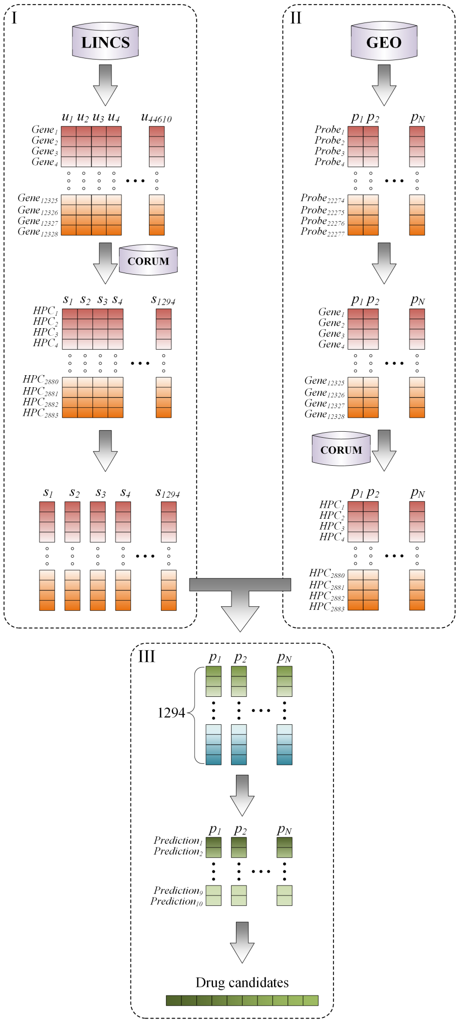


Figure 5.1: The flowchart of our HPC-based drug signature approach. I: Producing an HPC signature for each drug. II: Transforming patient gene expression profiles to HPC profiles. III: Matching drug HPC signatures to the patient HPC profiles, and producing a list of candidate drugs. p represents a patient and N is the number of patients, u is a drug profile and s is a merged profile for each drug. The number of approved drugs in LINCS is 1,294.

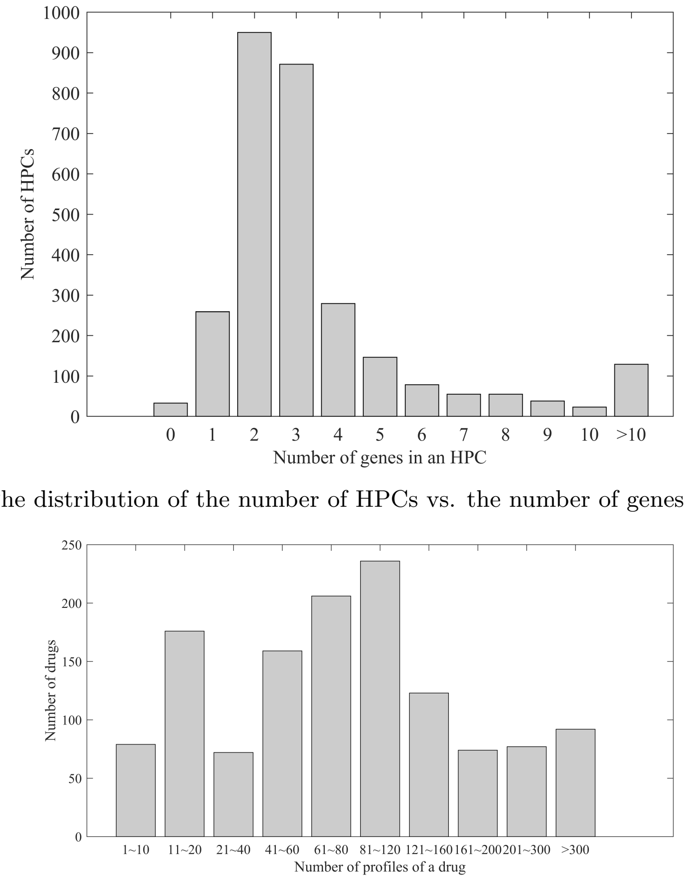
5.2.1 Design of study

The basic idea in our study is to generate a negative connection between a drug signature and a patient profile of a specific disease. The negative connection indicates an opposite effect between a drug and a disease represented by gene expression profiles, which may reflect a potential treatment for the drug to the disease. In Figure 5.1-I, we apply HPC information to describe the drug signatures. In Figure 5.1-II, the patient profiles are transformed from the form of genes to the form of HPCs. A matrix of connection scores is calculated, as shown in Figure 5.1-III. In each patient of a specific disease, the top predictions are generated and merged to produce a list of candidate drugs for the given disease.

5.2.2 Datasets and pre-processing

Three types of data are utilized in our HDgS approach, including the drug perturbation data, patient sample data, and HPC data.

The HPC data is downloaded from the comprehensive resource of mammalian protein complex (CORUM) database [336], which contains 4,273 protein complexes, out of which 2,916 are HPCs. All HPCs cover 4,274 genes. In order to connect with other types of data, those 4,274 genes must match with Entrez gene IDs. After matching, 2,916 HPCs and 3,092 genes remain. Since some HPCs do not encompass any genes that can be matched in Entrez, as shown in Figure 5.2a, we focus on the complexes that contain at least one matched gene. As a result, 2,883 HPCs are used to be the basic features in this study.



(a) The distribution of the number of HPCs vs. the number of genes per HPC.

(b) The distribution of the number of drugs vs. the number of profiles per drug in LINCS.

Figure 5.2: The statistic of drugs and HPCs in the dataset.

The drug perturbation profiles are downloaded from the LINCS database [169]. Phase I of the LINCS database is published in the Gene Expression Omnibus (GEO) database [187]. The expression values of only 978 genes have been measured in LINCS, where these 978 genes are "landmark gene", while the other genes are "inferred genes". The values of inferred genes are calculated based on the values of landmark genes.

These 978 landmark genes are sufficient to recovery 82% information in CMap, where 22,277 gene expression values per profile are measured. In LINCS, there are 12,328 genes in a total of landmark genes and inferred genes, and 1,319,138 profiles produced from 42,080 perturbations and 72 cell lines.

In LINCS database, the types of perturbations are small molecule drugs, shRNAs, cDNAs, and biologics. Since drug repositioning is to find some novel treatment for existing drugs, whose safeties have been studied. In this study, we focus on the drugs in DrugBank that have been approved by FDA [337]. As a result, 1,294 drugs are used in this study. The histogram of the drugs and profiles are shown in Fig 5.2b. The numbers of profiles vary over drugs, while many drugs have a larger number of profiles. In a previous study [3], we have generated the maps between HPCs and LINCS genes. Here drug perturbation profiles are transformed from the type of genes to HPCs, and the value of an HPC is a combination of the gene values in the HPC.

The patient samples are downloaded from the GEO database [187]. In this study, 11 datasets of four common cancers are obtained, as shown in Table 5.1. Among them, two platforms are referred to produce gene expression profiles. One is GPL96, where 22,283 probe sets are utilized to measure gene expression values. Another one is GPL570, where 54,675 probe sets are included. The three datasets of lung cancer come from four different stages. So the lung cancer profiles are divided into four subsets, each representing a cancer stage. In order to analyze data from different platforms, the first step is to select the common probe sets between them, which are 22,777 in CMap. Then the probe sets are transformed into the type of LINCS genes. Since some probe sets may refer to the same gene, the gene expression value is the average of the probe sets which are referred to the same gene. The next mapping step is the same as those LINCS profiles to get the HPC profiles. The cancer datasets from different platforms are transformed into the same type of HPCs. After comparing the tumor tissue samples to the normal tissue samples, a differential HPC profile is generated.

5.2.3 HPC-based drug signature procedures

In this section, we present the procedures to identify our HPC-based drug signature from LINCS database. The profiles are generated from the LINCS Level 5 data, which consists of the differential expressions of the drug perturbations in different concentrations, durations, and cell lines.

In order to reflect the dependencies of genes, HPC is used as the component of signature instead of individual genes in this study. As shown in Figure 5.2a, many complexes contain at least two genes. In our previous study of HPC-based disease signature [3], we chose the average value of the genes in the complex to be the value of the complex. However, the importance of the genes in a complex may not be equal. In this study, we use the Pearson Correlation Coefficient (PCC) to calculate the weights of genes. Additionally, we calculate another type of weight from the Spearman Correlation Coefficient (SCC). We also compare the approach of PCC weights with that of average weights and SCC weights in Section 5.3.

The fingerprinting vector of a gene consists of differential expression values across all profiles. A correlation matrix for a complex is constructed by pair-wise fingerprinting vectors among all genes in the same complex.

Table 5.1: The information of datasets

Cancers	Datasets	Platforms	Disease sample sizes
Lung Cancer	GSE10072	GPL96	32
	-Stage 1		15
	-Stage 2		9
	-Stage 3		6
	-Stage 4		2
	GSE19804	GPL570	59
	-Stage 1		35
	-Stage 2		12
	-Stage 3		12
	GSE27262-Stage 1	GPL570	25
Breast Cancer	GSE10780	GPL570	42
	GSE15852	GPL96	43
	GSE50948	GPL570	40
Colorectal Cancer	GSE21510	GPL570	40
	GSE41258	GPL96	43
	GSE49355	GPL96	12
Prostate Cancer	GSE46602	GPL570	14
	GSE69223	GPL570	15

Stage 1, 2, 3 and 4 are four stages of lung cancer in the datasets.

Then the weight of a gene in the complex is the average correlation to all other genes. All the weights are normalized and summed to 1. The differential expression value of a complex is the linear combination of the gene differential expression values with their weights.

The original drug perturbation profiles are in the form of genes. After mapping genes to HPCs by the weight approach, the novel profiles are in the form of HPCs. If an HPC has a value larger than 1, it is treated as an up-regulated HPC in the profile, while if it has a value smaller than -1, it is a down-regulated HPC. Among all the profiles of a drug, the HPC, which is either up-regulated or down-regulated in at least half of the profiles, is labeled as either an up-regulated HPC or a down-regulated HPC of the drug, respectively. Each HPC has a differential frequency among the profiles, and the up- and down-regulated HPCs are sorted together in descending order according to their frequencies. The length of HPC-based drug signature is determined in Section 5.2.4.

5.2.4 Matching procedure

After generating HPC-based drug signatures and patient differential HPC profiles, the next procedure is matching them together and calculating the matching scores. The matching method is the same which we used with disease signatures [3, 14].

Before matching, a rank list $PR = (pr_1, pr_2, \ldots, pr_H)$ is proposed to replace the patient differential HPC profile $PV = (pv_1, pv_2, \ldots, pv_H)$, where pv_i is the value of HPC_i, pr_i is its rank in the list and H is 2,883. The HPCs are sorted in ascending order according to their values in PV, where $\min(PR) = 1$ and $\max(PR) = H$.

Meanwhile, the signature is divided into two parts, one is the list of down-regulated HPCs, and the other one contains up-regulated HPCs. $score_{up}$ and $score_{down}$ are calculated as follows:

$$score_{up} = \sum_{i=1}^{H_{up}} (H + 1 - pr_{up i}) \tag{5.1}$$

$$score_{down} = -\sum_{j=1}^{H_{down}} (H + 1 - pr_{down j}) \tag{5.2}$$

where H_{up} is the length of the up-regulated list while H_{down} is that of the down-regulated list. up_i is the i^{th} HPC in the up-regulated list while $down_j$ is the j^{th} HPC in the down-regulated list.

A possible maximum score of the connection is calculated as follows:

$$poss = \sum_{i=1}^{M} (H + 1 - i) \tag{5.3}$$

where M is the length of signature. Finally, a connection score between an HPC-based drug signature and a patient profile is calculated as follows:

$$H\text{-}score = \frac{score_{up} + score_{down}}{poss} \tag{5.4}$$

The possible range of *H-score* is [-1,1], where a negative score reflects an inversion of the connection, which means that the drug may reverse the disease condition and have a potential treatment for it.

In our experiments, we use Matlab to implement our algorithm. Its computational time complexity is O(MN), while M is the number of drugs and N is the number of patients.

5.2.5 Evaluation metrics

In previous sections, we have produced a ranked drug list for a patient. The drugs are sorted in ascending order according to their connection scores. The top ten drugs are formed a new list for the following prediction. Therefore, we have N lists for a specific disease, while N is the number of patients. The frequency of each drug that appears on all lists is summarized. Drugs are sorted in descending order according to their frequencies. The ten most frequent drugs are selected as the drug candidates. Their uses as potential treatments for the diseases and literature evidence are discussed in Section 5.3.

In the experiments, the competing methods also produce ten drug candidates. Among the drug candidates, some drugs have been studied about their treatments for the specific disease. Therefore, we use "known drugs" to describe them. The other drugs in the results may have potential treatments for the disease, and we call them "potential drugs". The prediction rate is the rate of known drugs in the results. In the experiments, we use prediction rates to compare various methods.

5.3 Results and discussion

As reported by the World Health Organization (WHO) in 2018, lung, breast, colorectal, and prostate cancers are the four most common cancers in the world [338]. Therefore, in this study, we apply our HPC-based drug signature approach to the four cancers and compare it with other approaches.

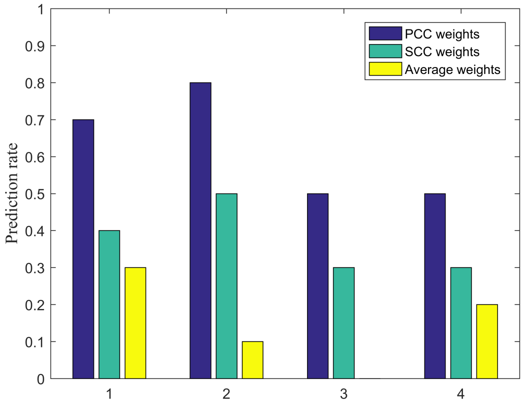


Figure 5.3: The prediction rates between three types of weighting approaches.

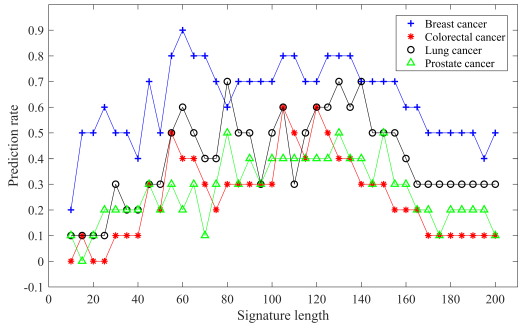


Figure 5.4: The plot of the prediction rate vs. the length of signature.

In Section 5.2.3, we have discussed the PCC weights in calculating HPC values from gene values. In order to ensure the advantage of PCC weights in HPC-based drug signatures, we first compare them with the SCC weights and average weights, as shown in Figure 5.3. The signatures via PCC weights achieve higher prediction rates than the other two types of weights. One possible reason is that genes within an HPC are not equal, while by averaging their values, all genes are treated equally. In SCC, the ranks may weaken the influence of the most differentially expressed genes. In the PCC weighting procedure, the correlations are different, a few genes may have negative correlations with other genes in the HPC. The weighting procedure is the way to enhance the genes with high positive correlations. The length of signature is another parameter that affects the prediction results. Figure 5.4 shows the prediction rates over various signature lengths from 10 to 200 with an increase of 5 for four diseases. From Figure 5.4, the best rate is achieved at the different signature lengths for the different diseases. In this study, we only present the results with its best signature length for a specific disease.

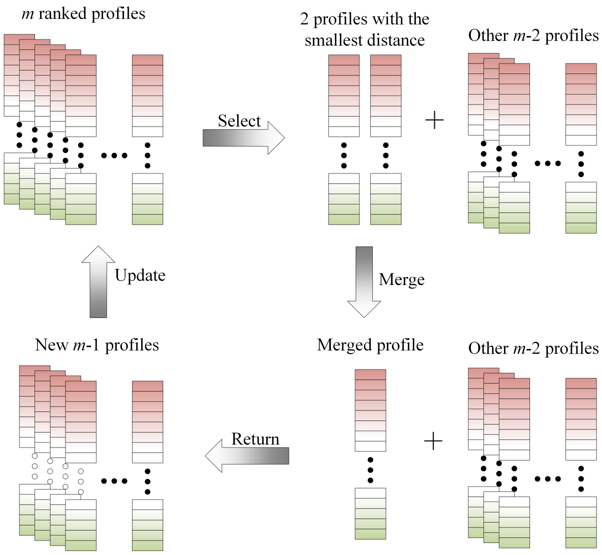


Figure 5.5: The iteration steps when identifying a PRL from the profiles of a drug.

In the previous study [3], we propose an HPC-based disease signature approach. In this study, we apply those two HPC-based signatures to identify drug candidates for four common cancers. Additionally, we utilize three different types of drug signatures in the experiments. The first type is the drug Prototype Ranked List (PRL) signature [339, 340], where the profiles of the same drug are merged hierarchically, as shown in Figure 5.5. A set D is used to reflect a given drug with M ranked profiles. Then the Spearman's Footrule distance is calculated between each pair of them. The two profiles with the smallest distance are deleted from the set D and summed together arithmetically. The new ranked profile is generated and added to the set D. The iteration repeats until there is only one profile in the set. The gene signature contains the same number of top and bottom 50 genes.

The second type is DrugSig, which is an online drug signature resource proposed by Wu *et al.* [341]. 5,913 drug signatures of 1,295 drugs are downloaded from the resource. Each signature contains 500 up-regulated genes and 500 down-regulated genes. The most different aspect is that there's no rank among the drug signature, which means all genes have the same weight. Similarly, genes in disease signatures do not have any ranks. The matching score is the rate of overlap:

$$score_{DrugSig} = \frac{up_{overlap} + down_{overlap}}{length\ of\ the\ disease\ signature} \tag{5.5}$$

where the $up_{overlap}$ is the number of common genes between the disease up-regulated gene list and the drug down-regulated gene list, the $down_{overlap}$ is the number of common genes between the disease down-regulated gene list and the drug up-regulated gene list. The matching score reflects the reverse of the two signatures.

Although several approaches have been proposed to process the LINCS profiles, some researchers prefer to identify drug signatures directly from the LINCS profiles, containing the 978 landmark genes [342]. The third type of compared signature in this study is the landmark signature. One thing that should be noted is that the profiles are not merged into a consensus one, so there may be some replicates in the prediction lists.

In the experiments, each method produces a list of ten drugs. As discussed in Section II.E, we compare the prediction rate of known drugs in the list. The prediction rate indicates the confidence that other drug candidates have the potential for the same treatment. Additionally, we collect the number of publications on PubMed, associated with the candidate drugs for specific cancer.

The prediction rates of five approaches in four types of cancers are listed in Table 5.2. In all four cases, our HDgS approach produces the highest prediction rates. In lung cancer stage 1 and colorectal cancer, there is 1 more approach that can achieve the same prediction rates with HDgS. In the experiment of a single patient, our proposed HDgS approach can achieve a prediction rate of 0.7, the same as in the whole dataset.

Table 5.2: The prediction rates of the five approaches

Cancers	HDgS	HPC-based disease	PRL	DrugSig	Landmark
Lung	**0.7**	0.6	0.3	0.4	0.6
-Stage 1	**0.7**	0.3	0.2	0.4	**0.7**
-Stage 2	**0.6**	0.3	0.2	0.4	0.3
-Stage 3	**0.6**	0.1	0.2	0.2	0.3
-Stage 4 - P1	**0.7**	0.0	0.3	0.2	0.2
-Stage 4 - P2	**0.7**	0.0	0.3	0.2	0.2
Breast	**0.8**	0.6	0.6	0.6	0.4
Colorectal	**0.6**	**0.6**	0.5	0.3	0.1
Prostate	**0.5**	0.4	0.4	0.4	0.3

NULL: Do not have corresponding result.

P1 and P2 are two patients in Stage 4.

Besides the prediction rate, the frequency rate of a given drug is used to reflect the portion of patients for whom the drug has been identified as a drug candidate. The drug candidates for four cancers and their

Table 5.3: The drugs predicted for lung cancer

Labels	Names	The frequency rates in the groups of patients						Num of Ref.
		Whole dataset	Stage 1	Stage 2	Stage 3	Stage 4-P1	Stage 4-P2	
Known drugs	Triptolide	0.948	0.960	0.905	0.944	1	1	58
	Maraviroc	0.871	0.893	0.762	0.889	1	1	6
	Palbociclib	0.629	0.613	0.762	0.556	NULL	1	79
	Crizotinib	0.517	0.493	0.524	0.611	1	NULL	2074
	Neratinib	0.431	0.427	0.222	0.500	1	1	70
	Oxytetracycline	0.414	0.440	NULL	0.556	1	NULL	18
	Caffeine	0.336	0.387	NULL	NULL	1	NULL	143
	Ciglitazone	NULL	NULL	0.333	NULL	NULL	1	26
	Fenretinide	NULL	NULL	NULL	NULL	1	1	46
	Geldanamycin	NULL	NULL	NULL	NULL	NULL	1	64
Potential drugs	Lomerizine	0.500	0.520	0.477	0.389	1	1	0
	Terconazole	0.414	0.360	0.477	0.556	1	NULL	0
	GSK-1059615	0.371	0.373	0.333	0.444	NULL	NULL	0
	Guanadrel	NULL	NULL	0.286	0.333	NULL	NULL	0
	Lofexidine	NULL	NULL	NULL	NULL	1	NULL	0
	Tinidazole	NULL	NULL	NULL	NULL	NULL	1	3
	Oxetacaine	NULL	NULL	NULL	NULL	NULL	1	0

NULL: The drug is not on the prediction list of the corresponding group of patients.

Num of Ref.: The number of publications associated with the predicted drug for lung cancer on PubMed.

frequency rates are listed in Tables 5.3-5.6. The treatments and annotations of drugs are discussed in the following sections.

5.3.1 Lung cancer

In 2.09 million cases of lung cancers [338], about 85% are non-small cell lung cancer (NSCLC), while the others are small cell lung cancer (SCLC). As shown in Table 5.3, ten small compounds are identified by the whole patient group, seven of which are known drugs. Additionally, seven different drugs are identified by the five subsets of patients.

Triptolide is a diterpenoid epoxide that is produced from the Tripterygium Wilfordii plant. It can decrease cell migration and invasion of lung cancer in vitro [343]. Maraviroc is an antiretroviral drug. It reduces lung tumor growth via decreasing the migration of C-C chemokine receptor type 5 (CCR5)+ regulatory T cells

[344]. Palbociclib is an inhibitor of the cyclin-dependent kinases 4 (CDK4) and CDK6. The combination treatment of palbociclib and selumetinib is effective in the models of NSCLC [345].

Crizotinib is an anaplastic lymphoma kinase (ALK) inhibitor that has shown treatments for NSCLC. It is superior to standard chemotherapy in advanced NSCLC patients with ALK rearrangement [346]. Neratinib is a tyrosine kinase inhibitor (TKI) anticancer drug. It has promising activity in NSCLC, according to both preclinical and human studies [347]. Oxytetracycline is a broad-spectrum antibiotic. It displays apparent inhibitions on the proliferation of A549 lung cancer cells [348]. Caffeine is a central nervous system (CNS) stimulant [349]. It increases apoptosis of lung cancer, which is killed by cisplatin, through the inhibition of ataxia telangiectasia mutated- and Rad3-related (ATR) activation [350].

Ciglitazone is a thiazolidinedione. It inhibits growth and induces apoptosis of NSCLC cells through decreased expression of phosphoinositide-dependent protein kinase 1 (PDK1) [351]. Fenretinide is a synthetic retinoid derivative. It induces apoptosis of SCLC cells and inhibits its growth [352]. Geldanamycin is a 1,4-benzoquinone ansamycin antitumor antibiotic. The association of Ad-mda7 gene and geldanamycin inhibits lung cancer cell motility and induces cell death [353].

Seven drugs are predicted to have potential treatments for lung cancer, two of which have been studied for the treatment of tumors and cancers. Lomerizine has the clinical potential to reverse tumor multidrug resistance [354]. GSK-1059615 is a type of kinase inhibitor and has been used in trials studying the treatment for solid tumors and breast cancer [355]. About the other five predictions, more information about the associations with cancers needs to be studied in the future. Terconazole is an antifungal drug. Guanadrel is an antihypertensive agent. Lofexidine is a non-opioid prescription medicine used to treat high blood pressure. Tinidazole is a drug for protozoan infections. Oxetacaine is a potent local anesthetic.

5.3.2 Breast cancer

Breast cancer is the most common cancer (2.09 million cases) in women [338]. As shown in Table 5.4, ten small compounds are predicted by our proposed HPC-based drug signature, nine of which are known drugs for breast cancer.

Palbociclib is a medication for breast cancer that has been sold in the market [356, 357]. Etoposide is a medication for several types of cancers. It is an active and well-tolerated regimen in metastatic breast cancer (MBC) patients [358]. Tretinoin is a medication for leukemia. The tretinoin-loaded lipid core nanocapsules reduce the breast cancer cell viability even at lower concentrations [359]. Teniposide is a chemotherapeutic medication used in the treatment of childhood acute lymphocytic leukemia and several cancers. It suppresses the growth of breast tumor in vivo [360].

Tunicamycin is a mixture of homologous nucleoside antibiotics. The combination of trastuzumab and tunicamycin shows effective treatments for HER2-positive breast cancer cells [361]. Triptolide has shown antitumor effects for lung cancer and predicted in Section 5.3. It inhibits the viability of breast cancer cells and significantly reduces the tumor weight and volume [362]. Idarubicin is an antineoplastic that has shown

Table 5.4: The drugs predicted for Breast cancer

Labels	Names	Frequency rates	Num of Ref.
Known drugs	Palbociclib	0.824	784
	Etoposide	0.560	1195
	Tretinoin	0.432	657
	Teniposide	0.408	39
	Tunicamycin	0.280	95
	Triptolide	0.272	53
	Idarubicin	0.272	107
	Cytarabine	0.264	262
Potential drugs	PHA-793887	0.512	1
	Norethisterone	0.344	255

treatments against breast cancer [363, 364]. Cytarabine is a chemotherapy medication used to treat leukemia. Some cases have suggested that treatment of intrathecal liposomal cytarabine in patients with leptomeningeal metastasis of breast cancer is feasible [365].

In this study, PHA-793887 and norethisterone are predicted to be potential drugs for breast cancer. PHA-793887 is a CDK4 inhibitor, while the CDK4/6 inhibitors could sensitize a subtype of breast cancer to PI3K inhibitors [366]. Norethisterone is a synthetic progestational hormone. It is a very weak inhibitor of CYP2C9 and CYP3A4, which are expressed in breast cancer tissues [367]. Studies about CYP3A4 indicate that it may play a role in breast carcinogenesis [368]. Further studies may concentrate on how to enhance its inhibitions on those genes.

5.3.3 Colorectal cancer

Colorectal cancer is the third most common cancer (1.80 million cases) in the world [338]. Ten small compounds are predicted in the results, as shown in Table 5.5, six out of which are known drugs.

Isosorbide is a bicyclic chemical compound. The combination of aspirin and isosorbide mononitrate shows synergistic apoptosis-inducing effects in human colon cancer cells [369]. Triptolide has been identified in Section 5.3.1 and 5.3.2. It also induces apoptosis of human colon cancer cells and inhibits proliferation [370]. Maraviroc has been used in the treatment of breast cancer. It induces significant apoptotic effects in colorectal cancer cells [371]. Palbociclib has been discussed in Section 5.3.1 and 5.3.2. It promotes colon cancer cell death and induces apoptosis [372].

Tivozanib is a type of kinase inhibitor, and the inhibition is helpful in the treatment of colorectal cancer [373]. In a phase II study, the combination of tivozanib and everolimus shows treatment in 50% of the patients with metastatic colorectal cancer [373]. Trametinib is a MEK inhibitor drug with anti-cancer activities. The combination of dabrafenib and trametinib shows treatment for patients with metastatic colorectal cancer

Table 5.5: The drugs predicted for colorectal cancer

Labels	Names	Frequency rates	Num of Ref.
Known drugs	Isosorbide	0.863	4
	Triptolide	0.726	18
	Maraviroc	0.526	2
	Palbociclib	0.474	5
	Tivozanib	0.347	4
	Trametinib	0.263	29
Potential drugs	Lomerizine	0.884	0
	Alverine	0.589	0
	Oxetacaine	0.558	0
	Tyloxapol	0.495	0

[374, 375].

Four drugs are predicted to have potential treatment for colorectal cancer, two of which have been studied the connections with cancers. Lomerizine is predicted to be a potential drug for both lung cancer and colorectal cancer, that it has the clinical potential to reverse tumor multidrug resistance [354]. Alverine is a medication for gastrointestinal disorders. The combination of MG132 and it shows cytotoxic effects on breast cancer cells [376]. Oxetacaine is a potent local anesthetic. Tyloxapol is a surfactant.

5.3.4 Prostate cancer

Prostate cancer is the second common cancer (1.28 million cases) in men. As shown in Table 5.6, ten small compounds are predicted, five of which are known drugs.

Palbociclib and triptolide are identified in all four cancers. Palbociclib is a novel medication for prostate cancer. A phase II study shows that it may help slow the growth of prostate cancer [377]. Triptolide induces prostate cancer cell death [378]. Maraviroc has been identified in lung and prostate cancers. It reduces prostate tumor bone metastasis in immunocompetent mice [379]. Cisplatin is a chemotherapy medication used to treat several types of cancers, including prostate cancer [380, 381]. Rucaparib is a poly ADP ribose polymerase (PARP) inhibitor, which is used as an anti-cancer medication. It has antitumor activities in prostate cancer patients [382].

Alverine and tyloxapol are predicted to be potential drugs for both colorectal and prostate cancers. Brompheniramine is a histamine H1 antagonist, that histamine has some interactions with cell proliferation and tumor growth [383]. PHA-793887 is a CDK inhibitor, which is used to treat cancers by preventing overproliferation of cancer cells [384]. Disopyramide is an antiarrhythmic medication.

Table 5.6: The drugs predicted for prostate cancer

Labels	Names	Frequency rates	Num of Ref.
Known drugs	Palbociclib	0.897	13
	Triptolide	0.690	28
	Maraviroc	0.517	3
	Cisplatin	0.310	1138
	Rucaparib	0.276	49
Potential drugs	Alverine	0.897	0
	Brompheniramine	0.552	1
	Tyloxapol	0.414	0
	Disopyramide	0.379	1
	PHA-793887	0.276	0

5.3.5 Discussion

In the experiments, we have studied our proposed framework in four types of cancers, including lung cancer, breast cancer, colorectal cancer, and prostate cancer. Among the predicted drug lists for each cancer, some known drugs have been either utilized in the treatment of cancer or studied *in vitro* and *vivo* trials. The lowest rate of the known drugs in the list is 50% in prostate cancer, while even 80% of drugs in the candidate list for breast cancer have shown treatments in previous studies. Those results indicate that our HDgS approach can be used to predict drug candidates for cancers. In this study, we have adopted the HPC-based drug signatures to connect with patient profiles. The datasets used in this study contain only one type of cancer in each sample. However, in principle, if a sample is from a patient with comorbidity, the potential drugs for such a patient should be different from those patients with a single disease. If there are some datasets from patients with comorbidity available, we would like to apply our proposed method to them in the future.

5.4 Conclusion

In this study, we have proposed a novel HPC-based drug signature (HDgS) for drug repositioning. The HPCs are utilized to describe dependencies between genes. Comprehensive experiments have been conducted to evaluate the performance of HDgS and other approaches. In the experiments, each patient is given a list of drug candidates, and the predictions for the cancer are according to the frequency analysis of the lists. The proposed HDgS can identify known drugs for most of the patients. The prediction rates of HDgS are larger than those of the competing approaches. When dealing with two patient samples separately, the proposed HDgS approach can identify seven known drugs, most of which are the same as those from the whole dataset. Based on literature evidence, many of the potential drugs also have anti-cancer properties.

6 Predicting drug-drug interactions by graph convolutional network with multi-kernel

Prepared as: Fei Wang, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Predicting drug-drug interactions by graph convolutional network with multi-kernel. Briefings in Bioinformatics, 2021. FW and FXW discussed the methods. FW implemented the algorithm, designed and performed the experiments. FXW supervised the study. FW and FXW wrote the manuscript. All authors read, revised, and approved the final version of the manuscript.

As described in Chapter 3, 4, and 5, the signature-based methods identify a list of potential drugs. In practice, drug combinations also show treatments for a specific disease. Predicting potential drug combinations, or DDIs, helps us to understand the MoAs of drugs. In many methods, the DDIs are treated as a whole set to construct a DDI network, while there are various types of them. In this chapter, I divided those DDIs into two groups and construct a model to aggregate them together. The model can predict potential DDIs effectively. This chapter fulfills Objective 5 of this dissertation.

Abstract

Drug repositioning is proposed to find novel usages for existing drugs. Among many types of drug repositioning approaches, predicting drug-drug interactions (DDIs) helps explore the pharmacological functions of drugs and achieves potential drugs for novel treatments. Many deep learning methods have been applied to predict DDIs. The DDI network, which is constructed from the known DDIs, is a common part of many of the existing methods. However, the functions of DDIs are different, and thus integrating them in a single DDI graph may overlook some useful information. We propose a graph convolutional network with multi-kernel (GCNMK) to predict potential DDIs. GCNMK adopts two DDI graph kernels for the graph convolutional layers, namely, increased DDI graph consisting of “increase”-related DDIs and decreased DDI graph consisting of “decrease”-related DDIs. The reconstructed drug features are fed into a block with three fully connected layers for the DDI prediction. We compare various types of drug features, while the target feature of drugs outperforms all other types of features and their concatenated features. In comparison with three different DDI prediction methods, our proposed GCNMK achieves the best performance in terms of AUC-ROC and AUC-PR. In case studies, we identify the top 20 potential DDIs from all unknown DDIs, and the top ten potential DDIs from the unknown DDIs among breast, colorectal, and lung neoplasms-related

drugs. Most of them have evidence to support the existence of their interactions.

6.1 Introduction

Drug repositioning is to find novel usages for existing drugs. The safety and other properties of the existing drugs, which have been approved to sell on the market, have been studied clearly. Therefore, drug repositioning helps save time and reduces the cost of drug development greatly. Several successful drugs have been proposed by drug repositioning approaches, such as sildenafil, thalidomide, zidovudine, minoxidil, and celecoxib [28].

In order to increase the prediction efficiency, many computational approaches have been utilized to predict potential drugs for different diseases. A main field is predicting potential links between drugs and related elements, such as drug-disease associations [11, 42, 53, 57, 122, 145], drug-target interactions [6, 25, 45, 48, 60, 65] and drug-drug interactions (DDIs) [10, 17, 19, 20, 58, 76, 153, 385].

When predicting DDAs, Luo *et al.* calculated similarities and constructed a similarity network [11, 57]. Random walk was employed to calculate the probabilities of DDAs. Li *et al.* utilized a convolutional neural network (CNN) model to conduct a binary classification of DDAs, based on the known DDAs and drug/disease feature vectors [42]. In the study of DTI, deep learning (DL) approaches are effective tools to predict potential DTIs. Wen *et al.* constructed a deep-belief network (DBN) to predict potential DTIs [25]. Monteiro *et al.* combined a CNN with a deep neural network (DNN) to make predictions, where the CNN was used to produce novel representations of feature vectors and the DNN was employed to predict DTIs [6].

The DDIs refer to the pharmacological and clinical responses to a drug combination, different from the known effects of two drugs when used alone. The prediction of DDIs helps researchers to have a deep understanding of the mechanisms of actions (MOAs) of drugs. In order to analyze DDIs, various types of drug features have been studied, such as chemical substructures, side effects, targets, pathways, and enzymes, *etc.*

Many approaches have been proposed to predict DDIs based on one or more types of drug features. Ferdousi *et al.* calculated drug-drug similarities based on various types of features and utilized a positive similarity threshold to determine the potential DDIs [17]. However, the similarities of many DDIs are negative, while they cannot be predicted by a constant positive value. Yan *et al.* used a k-nearest neighbor procedure after generating similarities of known DDIs and employed a regularized least squares (RLS) classifier to predict potential DDIs [19]. In the classifier, both positive samples and negative samples are essential. In predicting potential DDIs, the positive samples are those known DDIs, while the negative samples are the unknown DDIs. Zheng *et al.* used an SVM model to produce reliable negative samples (RNS) from the unknown samples and made a further prediction [20]. Zhang *et al.* proposed a multi-modal autoencoder (MDAE) with positive-unlabeled (PU) learning to predict potential DDIs [10].

The DDIs can be utilized to construct a DDI graph, where nodes are drugs and edges are interactions among drugs. Zhou *et al.* used a Markov clustering algorithm on the DDI graph to predict potential drug combinations [153]. Additionally, researchers can combine the drug features with the network structures to predict potential interactions. Zhang *et al.* used a random walk algorithm on the DDI graph [58], while the transition probabilities were based on the drug-drug similarity matrices.

Graph convolutional network (GCN) [143] is a variant of convolutional neural network (CNN) on the graph, while the graph is used as a kernel. Researchers utilize GCN to produce low-dimensional representation vectors of drugs by learning topological structures of drugs in the DDI graph. Feng *et al.* combined GCN with a deep neural network (DNN) to generate feature representation matrix and predict potential DDIs [76]. Huang *et al.* added a skip graph to reflect the indirect connections in the original DDI graph and made predictions based on both the original DDI graph and the skip graph [385].

In many DDI prediction methods, researchers do not distinguish the responses of DDIs. All known DDIs are labeled as positive samples and used to construct the DDI graph. However, there are many types of DDIs relating to various mechanisms. About half of them are "increase"-related, such as "DRUG A may increase the activities of DRUG B," another half of them are "decrease"-related, such as "The metabolism of DRUG A can be decreased when combined with DRUG B."

In this work, we aim to learn novel embeddings from those two types of DDIs. As discussed above, GCN is an effective structure to utilize both DDI graphs and drug feature vectors. We propose a graph convolutional network with multi-kernel (GCNMK) to predict potential increased DDIs. We firstly construct an increased DDI graph and a decreased DDI graph from the "increase"-related and "decrease"-related DDIs, respectively. Two GCN layers are combined to learn low-dimensional representation vectors of drugs with those two graphs and various types of drug features. After generating the node embeddings, two drug vectors are concatenated to be the vector of a DDI. Finally, a block with three fully connected layers is used to make predictions. In the experiments, we investigate the prediction performance of our proposed model on various types of drug features, including chemical substructures, side effects, targets, pathways, and enzymes, *etc.* We compare three state-of-the-art methods with our GCNMK. The results demonstrate that our GCNMK outperforms other competing methods in predicting potential DDIs. In case studies, we predict potential DDIs, and most of them have evidence to support the existence of their interactions.

6.2 Methods and materials

In this section, we introduce the architecture of our GCNMK model, as shown in Figure 6.1. In Figure 6.1-I, an increased DDI graph and a decreased DDI graph are constructed from the "increase"-related and "decrease"-related DDIs, respectively. The two graphs and drug feature matrices are fed into two GCN blocks, respectively. In Figure 6.1-II, these two GCN blocks form the GCN layer L_1, while layer L_2 contains the third block. An additional procedure, whose output is a linear combination of its inputs, is adopted in each block to generate drug embeddings from both increased and decreased DDI graphs. The low-dimensional representation vectors of drugs are produced after the layer L_2. In Figure 6.1-III, the feature vectors of two drugs are concatenated to form a DDI vector. A block with three fully connected layers is employed to predict potential DDIs.

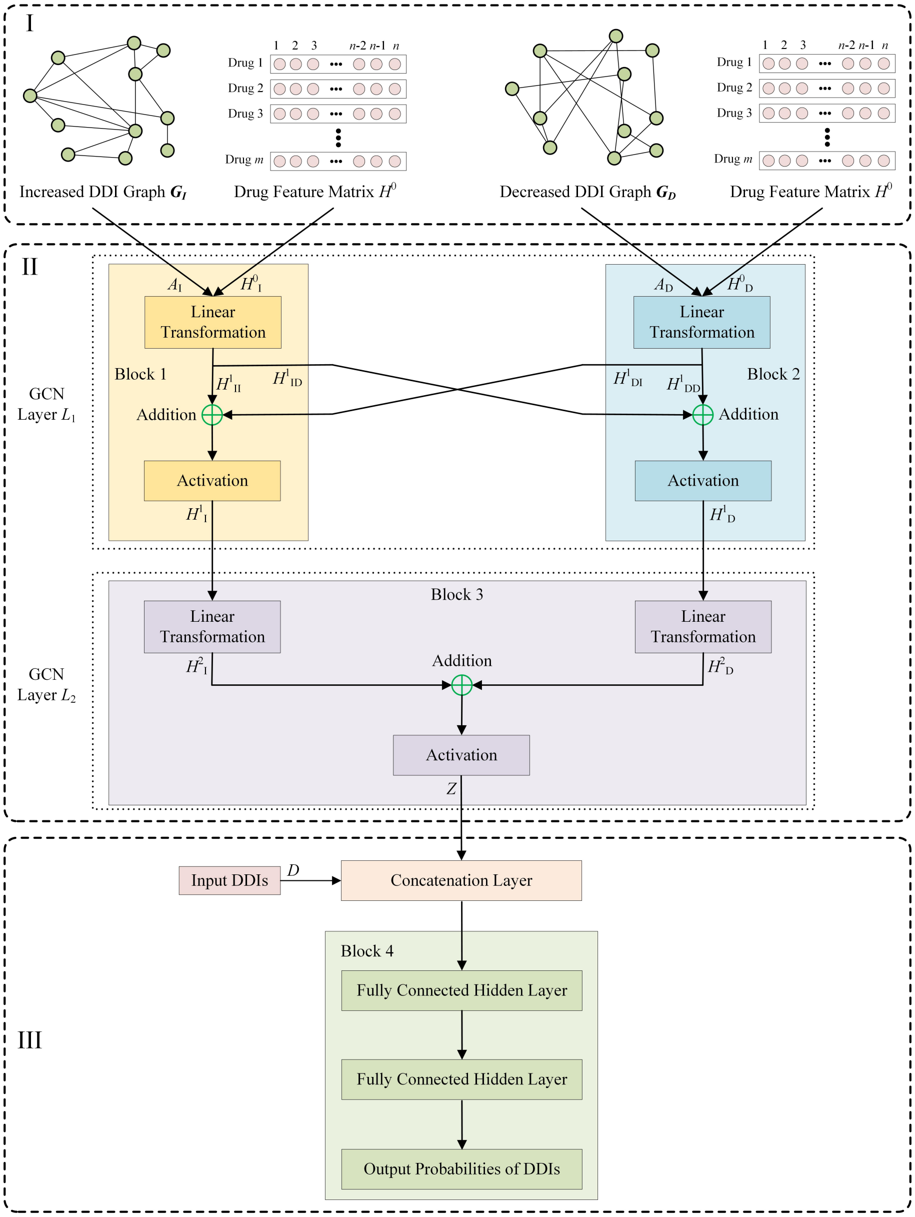


Figure 6.1: The architecture of GCNMK. **I**: Constructing two DDI graphs from increased, decreased interactions, and inputting drug attributes. **II**: Generating the feature representation of drugs by GCN. **III**: Predicting DDIs.

6.2.1 DDI graphs and drug feature matrix

A DDI graph $G = (V, E)$ represents a collection of n nodes and m edges, while nodes are drugs and edges are DDIs, which is described by an association matrix A. The DDI refers to the pharmacological and clinical responses to a drug combination, different from the known effects of two drugs when used alone. If there is a known response between drugs i and j, in the association matrix A, $A(i,j) = 1$. Otherwise, $A(i,j) = 0$. The DDI graph is undirected, that is, $A(i,j) = A(j,i)$.

There are various types of responses between two drugs, including analgesic activity, risk or severity of heart failure, serum concentration, therapeutic efficacy, *etc.* We divide them into two groups. One group contains DDIs that increase one of the responses, while another group contains DDIs that decrease one of the responses. Two DDI graphs G_I and G_D are constructed based on those two groups of DDIs, respectively. Their association matrices are denoted by A_I and A_D.

Another matrix is the drug feature matrix H^0. In order to make a distinction, the feature matrix together with the graph G_I is marked as H_I^i, while the other one is H_D^i, at the i-th layer of GCNs.

6.2.2 Feature representations of drugs

In this study, we construct two DDI graphs G_I and G_D for the increased and decreased DDIs, respectively. Our purpose is to use GCN layers to learn features from both two graphs. In layer L_1, two blocks are adopted, each has an input graph, as shown in Figure 6.1-II. The propagation rules of linear transformation are as follows:

$$H_{II}^1 = F_I H_I^0 W_I^0 \tag{6.1}$$

$$H_{ID}^1 = F_I H_I^0 {W'}_I^0 \tag{6.2}$$

$$H_{DD}^1 = F_D H_D^0 W_D^0 \tag{6.3}$$

$$H_{DI}^1 = F_D H_D^0 {W'}_D^0 \tag{6.4}$$

where H_{II}^1 and H_{DD}^1 are the node embedding matrices transferring within each block, respectively. H_{ID}^1 and H_{DI}^1 transferring between the two blocks in layer L_1. $F_I = \widetilde{D}_I^{-\frac{1}{2}} \widetilde{A}_I \widetilde{D}_I^{-\frac{1}{2}}$, $F_D = \widetilde{D}_D^{-\frac{1}{2}} \widetilde{A}_D \widetilde{D}_D^{-\frac{1}{2}}$. $\widetilde{A}_I = A_I + I$ and $\widetilde{A}_D = A_D + I$ are the association matrices of the graph G_I and G_D, respectively. I is the identity matrix. $\widetilde{D}_I(i,i) = \sum_j \widetilde{A}_I(i,j)$ and $\widetilde{D}_D(i,i) = \sum_j \widetilde{A}_D(i,j)$ are the degree diagonal matrices.W_I^0, ${W'}_I^0$, W_D^0, and ${W'}_D^0$ are the weight matrices.

In each block, an addition procedure is adopted before the activation function as follows:

$$H_I^1 = \sigma(H_{II}^1 + H_{DI}^1) \tag{6.5}$$

$$H_D^1 = \sigma(H_{DD}^1 + H_{ID}^1) \tag{6.6}$$

where H_I^1 and H_D^1 are the outputs. σ is the activation function, which is ReLU in this study.

The GCN layer L_2 contains one block, which is used to integrate the outputs from two blocks in layer L_1 as follows:

$$Z = \sigma(H_I^2 + H_D^2) = \sigma(F_I H_I^1 W_I^1 + F_D H_D^1 W_D^1) \tag{6.7}$$

where Z is the final representation matrix of drugs.

6.2.3 Predicting DDIs

The Block 4 with three fully connected layers is utilized to predict DDIs in our model, as shown in Figure 6.1-III. Before Block 4, a concatenation layer is used to generate the DDI feature matrix. The inputs of concatenation layer are representation matrix Z, and DDI information matrix D. For a pair of drugs i and j in D, its DDI feature vector is the concatenation of Z_i and Z_j, represented as $[Z_i, Z_j]$, where Z_i and Z_j are the feature vectors of drugs i and j in Z, which is fed into Block 4.

In Block 4, the number of neurons in each layer is 64, 16, and 1. The DDI prediction is formulated as a binary classification, that the output values are the probabilities of how likely a drug pair is a true DDI. The activation function is ReLU in hidden layers and Sigmoid in the output layer.

The cross-entropy loss function is used in our GCNMK model:

$$BCE = -\frac{1}{N} \sum_{ij} [y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij})] \tag{6.8}$$

where N is the sample size, $y_{ij} \in [0, 1]$ is the true label for the interaction between drug i and j. “1” represents the label of a positive sample, while “0” represents that of a negative sample. p_{ij} is the predicted probability.

In order to prevent the over-fitting problem, an L_2-regularization is adopted:

$$L_2 = \frac{\lambda}{2N} \sum_{w} w^2 \tag{6.9}$$

where λ is a hyper-parameter, w is an element in the parameter matrices W_I^0, ${W'}_I^0$, W_D^0, ${W'}_D^0$, W_I^1, and W_D^1. As a result, the loss function for training our GCNMK model is $L = BCE + L_2$.

6.2.4 Datasets

In order to make a fair comparison between various types of features and methods, we choose the drugs which have all types of features in both our proposed methods and the competing methods. In our study, we download DDIs from the DrugBank database (Version 5.1.8) [386], while the numbers of “increase”-related and “decrease”-related DDIs are 40,202 and 40,500, respectively, among 613 FDA-approved drugs.

Eight types of features are compared in the experiments, as described in Table 6.1. It should be mentioned that the node2vec feature matrix is generated from the whole DDI graph $G_{all} = G_I \cup G_D$ and that there is an information leak in it. The features about associated drugs, enzymes, side effects, substructures, and targets are generated from the corresponding databases, as listed in Table 6.1. The pathway feature vectors of drugs are based on the drug-related targets and target-pathway associations. The prototype ranked list (PRL) feature vector is generated by merging a group of profiles of a given drug into a single ranked list [339]. The profiles are downloaded from the Library of Integrated Network-based Cellular Signatures (LINCS) database [169].

Table 6.1: The types of features and their dimensions

Feature types	Dimensions	Resources
Associated Drugs	613	DrugBank [386]
Enzymes	454	DrugBank
Pathways	533	DrugBank, CTD [387] and KEGG [388]
Side Effects	4859	SIDER [389]
Substructures	811	DrugBank
Targets	2670	DrugBank and CTD
Node2vec	613	[390] and [385]
PRL	978	LINCS [169] and [339]

6.3 Results and discussion

In this section, we illustrate the performances of our proposed model in various types of data and compare it with three state-of-the-art DDI prediction algorithms. Five aspects are discussed in the following five subsections: datasets in both our proposed model and the competing models; experiment setting; visualization analysis of embedding features; results of competing methods; case studies of our proposed model.

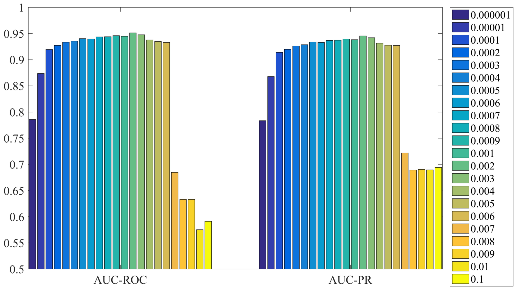


Figure 6.2: The influence of learning rate lr.

6.3.1 Experimental setting

In this study, we use 5-fold cross-validation (5-CV) to evaluate the prediction performance of our GCNMK model and the competing methods. The known DDIs are represented as positive samples, and the unknown DDIs are represented as negative samples. The number of positive samples is 80,702, while that of negative samples is 106,876. In order to make the training data balanced, 80,702 negative samples are randomly selected. Both the positive samples and the selected negative samples are divided into five subsets randomly. At each time, a positive subset and a negative subset are selected as the testing set, while the remaining subsets are selected as the training set. After five times, all subsets are used up to be testing sets, and the predicting results are produced.

In order to avoid using the testing information in the training procedure and make the testing procedure more accurate, the DDIs in the testing set are deleted from G_I and G_D at each training.

In experiments, the area under receiver operating characteristic curve (AUC-ROC) and area under precision-recall curve (AUC-PR) are used to measure the performance of results. The higher the values are, the more reliable the model is.

We adjust the parameters in order to achieve optimal performances. For the learning rate lr, L_2-regularization coefficient λ, and embedding size d, we search for the optimal values with the nominal values lr=0.0005, λ=0.0005, d=128. When optimizing the influence of a specific parameter, the other two parameters are set to be the nominal values. After optimization, its optimal value is used to update its nominal value. In those experiments, the target information is used to construct the drug feature matrix H^0.

The learning rate $lr \in (0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001)$. After achieving that the optimal value is around 0.001, we set the learning rate to be in a refined range (0.0001,0.0002,...,0.0009,0.001,0.002,...,0.009). In order to show them clearly, we use two histograms to depict the AUC-ROC and AUC-PR values under different lr values, as shown in Figure 6.2. When lr increases from 0.000001 to 0.002, the general trend of AUC-ROC and AUC-PR is ascending. When lr is larger than 0.002, the AUC-ROC and AUC-PR are reduced. Therefore, we set the learning rate lr to be 0.002 in our proposed GCNMK model.

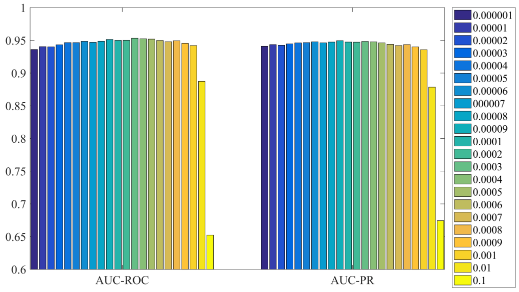


Figure 6.3: The influence of L_2-regularization coefficient λ.

The L_2-regularization coefficient $\lambda \in$ (0.1,0.01,0.001,0.0001,0.00001,0.000001). The optimal value is around 0.0001. Then λ is set to be in a refined range (0.00001,0.00002,...,0.00009,0.0001,0.0002,...,0.0009).

All the AUC-ROC and AUC-PR values are shown in Figure 6.3. When λ increases from 0.000001 to 0.0003, the AUC-ROC and AUC-PR increase slightly. When λ is larger than 0.0003, the AUC-ROC and AUC-PR are decreasing. Therefore, we set λ to be 0.0003 in our proposed GCNMK model.

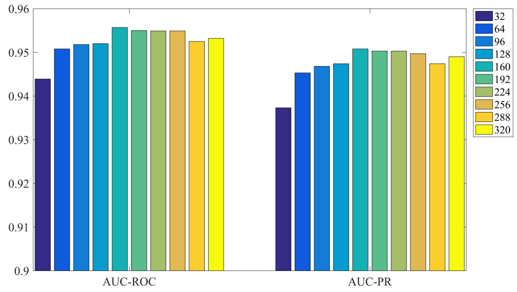


Figure 6.4: The influence of embedding size d.

The embedding size $d \in$ (32,64,96,128,160,192,224,256,288,320). The prediction performance changes a little when the embedding size varies, as depicted in Figure 6.4. When d is increasing from 32 to 160, the AUC-ROC and AUC-PR are increased When d is larger than 160, the AUC-ROC and AUC-PR are becoming smaller. We set the optimal embedding size d to be 160 in our GCNMK model.

Various types of features are used in our GCNMK model. The histograms of their prediction performance are shown in Figure 6.5. Although the node2vec feature has a problem of information leak, its prediction performance is the worst among the eight types of features. The PRL feature produces the second-worst prediction results. The differences of the AUC-ROC and AUC-PR of the other six types of features are not large, and the target feature of drugs achieves the best prediction performance among them. Therefore, in the following comparison, we use the target feature of drugs in our GCNMK model.

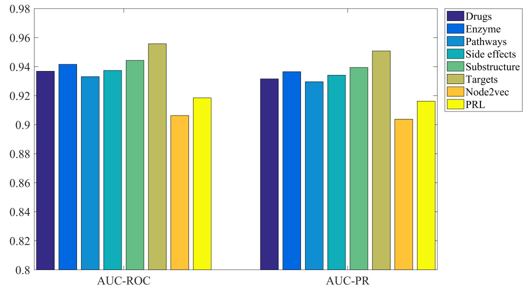


Figure 6.5: The influence of feature type.

We compare our methods with three DDI prediction methods, which are DPDDI [76], SkipGNN [385], and MDAE [10]. The parameters are set to be the optimal values as described in their methods. The type of feature used in DPDDI is the associated drugs. In SkipGNN, it is node2vec. Five types of features are used in MDAE, including associated drugs, enzymes, pathways, targets, and substructures. Additionally, the same five types of features are used in our GCNMK model, which is represented as GCNMK-5 in Table 6.2.

6.3.2 Visualization analysis of embedding features

In order to study the embedding performance of our proposed model, we employ t-distributed stochastic neighbor embedding (t-SNE) [391] to visualize DDIs based on the embedding features learned from our model. t-SNE is applied to reduce the dimensionality of embedding features to 2 and plot a 2-D figure, as shown in Figure 6.6. The green dots are known DDIs, while the red dots are unknown DDIs. Based on Figure 6.6, we can see that most of the dots are gathered in two areas. Especially, the known DDIs are located at the lower half of the figure, while the unknown DDIs are located on the upper right quarter of the figure, which can explain the performance of our model.

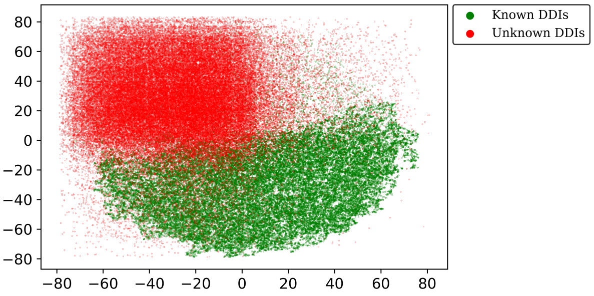


Figure 6.6: The visualization analysis of embedding features.

6.3.3 Results

The prediction performances of all competing methods are listed in Table 6.2. Each method is repeated ten times to generate an average value and a standard deviation of the AUC-ROC and AUC-PR metrics. The GCNMK and GCNMK-5, whose performance ranks are 1 and 2 in terms of AUC-ROC and AUC-PR, respectively, are our proposed methods. The ranks of the other three competing methods are from 3 to 5.

We compare our GCNMK model with others in different aspects. There is only one graph kernel in DPDDI method [76], which is the graph of all known DDIs $G_{all} = G_I \cup G_D$. The AUC-ROC and AUC-RP values produced by GCNMK model are about 4% larger than those of DPDDI. Referring to the results in Figure 6.5, our GCNMK model still achieves better performance than DPDDI when using the same type of feature. The results indicate that using the increased-decreased graphs G_I and G_D can improve the prediction performance.

There are two graph kernels in SkipGNN [385], that one kernel is G_{all} and another kernel G_{skip} is based on G_{all}. The GCNMK generates 10% larger AUC-ROC and AUC-RP values than SkipGNN. In this way, the graphs G_I and G_D work better in predicting potential DDIs. One possible reason is that the ratio of edges in G_{all} is about 43% in our datasets, and it is nearly 95% in G_{skip}. Adding such an almost fully connected graph can not improve the prediction performance.

Five types of features are used to identify the drug representation feature vectors in GCNMK-5 and MDAE [10]. In the results, the GCNMK-5 outperforms MDAE. Furthermore, the GCNMK achieves better

Table 6.2: The prediction performances of the competing methods.

Methods	AUC-ROC			AUC-PR		
	Ave.	Std.	Rank	Ave.	Std.	Rank
GCNMK	**0.9557**	0.0017	1	**0.9508**	0.0012	1
GCNMK-5	0.9337	0.0042	2	0.9292	0.0048	2
DPDDI	0.9126	0.0003	3	0.9131	0.0003	4
SkipGNN	0.8589	0.0005	5	0.8604	0.0005	5
MDAE	0.8981	0.0015	4	0.9232	0.0013	3

Ave.: The average value across ten repeats.

Std.: The standard deviation across ten repeats.

Rank: The ranks are based on the average values.

prediction performance than GCNMK-5, which indicates that multiple types of features do not achieve better results than a single type of feature.

In summary, our proposed GCNMK model achieves the best prediction performance among all competing methods in terms of AUC-ROC and AUC-PR.

Table 6.3: The top 20 predicted DDIs.

Rank	Drug A	Drug B	Evidence Source	Description
1	Imipramine	Olanzapine	Drugs.com	Using imipramine together with olanzapine may increase side effects such as drowsiness.
2	Olanzapine	Theophylline	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
3	Desipramine	Olanzapine	Drugs.com	Using desipramine together with olanzapine may increase side effects such as drowsiness.
4	Sulfadiazine	Trimethoprim	TWOSIDE	Using the drug combination may increase the side effect of anaemia.

5	Cimetidine	Tramadol	Drugs.com	Cimetidine may increase the blood levels and effects of tramadol.
6	Sulfamethoxazole	Trimethoprim	TWOSIDE	Using the drug combination may increase the side effect of anaemia folate deficiency.
7	Hydrochlorothiazide	Metoprolol	Drugs.com	Using metoprolol and hydrochlorothiazide together may lower your blood pressure and slow your heart rate.
8	Ofloxacin	Ticlopidine	N.A.	N.A.
9	Dextromethorphan	Quinidine	Drugs.com	Using dextromethorphan together with quinidine may increase the effects of dextromethorphan.
10	Tolbutamide	Vincristine	N.A.	N.A.
11	Estradiol	Progesterone	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
12	Fosinopril	Hydrochlorothiazide	Drugs.com	Their effects may be additive on lowering your blood pressure.
13	Nicotine	Vincristine	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
14	Hydrochlorothiazide	Pindolol	Drugs.com	Using pindolol and hydrochlorothiazide together may lower your blood pressure and slow your heart rate.
15	Lorazepam	Ranitidine	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
16	Promethazine	Pseudoephedrine	TWOSIDE	Using the drug combination may increase the side effect of anaemia.

17	Theophylline	Vincristine	TWOSIDE	Using the drug combination may increase the side effect of neutropenia.
18	Panobinostat	Rosiglitazone	N.A.	N.A.
19	Hydralazine	Reserpine	N.A.	N.A.
20	Ranitidine	Teniposide	N.A.	N.A.

N.A.: The evidence of the given DDI is not available till now.

Table 6.4: The top ten predicted DDIs of breast neoplasms-related drugs.

Rank	Drug A	Drug B	Evidence Source	Description
1	**Verapamil**	Mefloquine	Drugs.com	Using mefloquine together with verapamil can increase the risk of irregular heart rhythm that may be serious and potentially life-threatening.
2	**Sulindac**	Methazolamide	N.A.	N.A.
3	**Ranitidine**	**Vinblastine**	TWOSIDE	Using the drug combination may increase the side effect of neutropenia.
4	**Rosiglitazone**	Metformin	TWOSIDE	Using the drug combination may increase the side effect of anaemia vitamin b12 deficiency.
5	**Quinine**	Nizatidine	TWOSIDE	Using the drug combination may increase the side effect of chest pain.
6	**Sulindac**	Theobromine	N.A.	N.A.
7	**Ranitidine**	Sunitinib	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
8	**Ranitidine**	Teniposide	N.A.	N.A.
9	**Ranitidine**	**Vinorelbine**	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
10	**Sulfasalazine**	Isosorbide	TWOSIDE	Using the drug combination may increase the side effect of anaemia.

The breast neoplasms-related drugs are in bold.

6.3.4 Case studies

In case studies, all 106,876 unknown DDIs are fed into our GCNMK model. A larger prediction score of two drugs suggests that they have a higher probability of having an interaction. We generate a ranked list of DDIs in descending order according to their prediction scores.

The top 20 predicted DDIs are listed in Table 6.3. We verify them with TWOSIDE database [392] and Drug Interactions Checker of Drugs.com [393], and collect the descriptions about their interactions. For instance, the description of "Imipramine-Olanzapine" is "Using imipramine together with olanzapine may increase side effects such as drowsiness". We can see that 15 DDIs are confirmed in either Drugs.com or TWOSIDE. The results indicate that our proposed GCNMK model is effective in predicting novel DDIs. Other five DDIs, "Ofloxacin-Ticlopidine", "Tolbutamide-Vincristine", "Panobinostat-Rosiglitazone", "Hydralazine-Reserpine", and "Ranitidine-Teniposide", deserve to be confirmed by further experiments. Additionally, the drug "Vincristine" appears in three predicted DDIs, two of which have been confirmed. More attention should be paid to "Tolbutamide-Vincristine".

Especially, in order to study the potential DDIs which are related to a given disease, we generate the disease-related drugs from CTD database. Those drugs have been used to treat the given disease. In our datasets, the numbers of breast, colorectal, and lung neoplasms-related drugs are 64, 31, and 36, respectively. The unknown DDIs which are connected with those drugs are predicted. The predicted results are listed in Tables 6.4, 6.5, and 6.6.

In the predicted results of breast neoplasms-related DDIs, seven out of ten DDIs have been confirmed to have interactions in either TWOSIDE or Drugs.com. Especially, there are two confirmed DDIs, each of which consists of two breast neoplasms-related drugs. The other three DDIs, "Sulindac-Methazolamide", "Sulindac-Theobromine", and "Ranitidine-Teniposide", deserve to be confirmed by further experiments. Especially, among the ten predicted DDIs, the drug "Ranitidine" appears in four DDIs, while three DDIs have been confirmed. The DDI "Ranitidine-Teniposide" should attract more attention.

In the predicted results of colorectal neoplasms-related DDIs, seven out of ten DDIs have been confirmed to have interactions in TWOSIDE. The other three interactions, "Dacarbazine-Phenytoin", "Fluorouracil-Oxymetholone", and "Doxorubicin-Lynestrenol", could be potential DDIs.

In the predicted results of lung neoplasms-related DDIs, eight out of ten DDIs have been confirmed to have interactions in either TWOSIDE or Drugs.com. The other two DDIs, "Sulindac-Methazolamide" and "Sulindac-Theobromine", are also on the predicted list of breast neoplasms.

These neoplasms-related case studies demonstrate the usefulness of our GCNMK model in identifying potential DDIs for specific disease-related drugs.

6.4 Conclusion

In this study, we have proposed a GCNMK model for predicting DDIs. The “increase”-related DDIs and “decrease”-related DDIs are used to construct two DDI graphs, which are the graph kernels in our model. Then novel embeddings of drugs are produced by three GCN blocks. A DDI feature vector is the concatenation of two drug feature vectors. A block of three fully connected layers is used as a predictor. Comprehensive experiments have been conducted to evaluate the performance of GCNMK and other methods. In the experiments, our GCNMK model outperforms all other methods. In the case studies, most of the predicted DDIs have evidence to support the existence of their interactions. Therefore, benefiting from the two graph kernels, our GCNMK model can be used to predict DDIs effectively.

Even so, there is a limitation in our proposed model. When constructing the DDI graphs and generating the set of drugs, the drugs in the experiment have at least one DDI. We remove the drugs which do not have any known DDIs. As a result, our model can not identify DDIs among isolated drugs.

There are several directions of future work along with this study. In the DDI graphs of GCNMK, the edges belong to the same type. We could adapt this to any heterogeneous network, such as the drug-disease network. The descriptions of drug-diseases associations consist of two types: therapeutic and marker/mechanism, which may be useful for employing a GCN model. Another future direction is to distinguish more types of predicted DDIs. According to their functions, each type of DDI may be used to construct a graph kernel, and the novel model has the potential to identify the specific type of a predicted DDI.

Table 6.5: The top ten predicted DDIs of colorectal neoplasms-related drugs.

Rank	Drug A	Drug B	Evidence Source	Description
1	**Simvastatin**	Niacin	TWOSIDE	Using the drug combination may increase the side effect of iron deficiency anaemia.
2	**Fluorouracil**	Lorazepam	TWOSIDE	Using the drug combination may increase the side effect of iron deficiency anaemia.
3	**Meloxicam**	**Methotrexate**	TWOSIDE	Using the drug combination may increase the side effect of iron deficiency anaemia.
4	**Fluorouracil**	Tramadol	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
5	**Famotidine**	Primidone	TWOSIDE	Using the drug combination may increase the side effect of haemorrhagic anaemia.
6	**Dacarbazine**	Phenytoin	N.A.	N.A.
7	**Famotidine**	Progesterone	TWOSIDE	Using the drug combination may increase the side effect of atrial fibrillation.
8	**Fluorouracil**	Oxymetholone	N.A.	N.A.
9	**Doxorubicin**	Lynestrenol	N.A.	N.A.
10	**Simvastatin**	Trifluoperazine	TWOSIDE	Using the drug combination may increase the side effect of pancytopenia.

The colorectal neoplasms-related drugs are in bold.

Table 6.6: The top ten predicted DDIs of lung neoplasms-related drugs.

Rank	Drug A	Drug B	Evidence Source	Description
1	**Sulindac**	Methazolamide	N.A.	N.A.
2	**Rosiglitazone**	Metformin	TWOSIDE	Using the drug combination may increase the side effect of anaemia vitamin b12 deficiency.
3	**Theophylline**	**Vincristine**	TWOSIDE	Using the drug combination may increase the side effect of neutropenia.
4	**Sulindac**	Theobromine	N.A.	N.A.
5	**Methotrexate**	Meloxicam	TWOSIDE	Using the drug combination may increase the side effect of iron deficiency anaemia.
6	**Theophylline**	Thalidomide	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
7	**Ifosfamide**	Ofloxacin	Drugs.com	Chemotherapy with ifosfamide may reduce the plasma concentrations of oral ofloxacin.
8	**Theophylline**	Olanzapine	TWOSIDE	Using the drug combination may increase the side effect of anaemia.
9	**Sulindac**	Isosorbide	TWOSIDE	Using the drug combination may increase the side effect of pancytopenia.
10	**Melatonin**	Tacrolimus	TWOSIDE	Using the drug combination may increase the side effect of pancytopenia.

The lung neoplasms-related drugs are in bold.

7 Summary, limitations, and future work

7.1 Summary

Computational drug repositioning is a critical yet challenging issue. The datasets are vast, and the computational methods are numerous. Besides generating a list of potential drugs for a given disease, more descriptions about the potential treatments, such as DDIs, are useful. This dissertation aims to identify a list of potential drugs for several types of cancers and predict potential DDIs. In total, five objectives are proposed in Chapter 1, and Chapter 2 to 6 have achieved these objectives.

Chapter 2 comprehensively reviews some latest studies in predicting novel treatments for existing drugs. The widely used databases and pre-processing steps are firstly introduced. Some types of algorithms, such as signature-based, network-based, basic ML, and DL methods are discussed. Moreover, three scenarios about DDAs, DDIs, and DTIs are presented.

Chapter 3 designs a weighting strategy to identify gene signatures from heterogeneous datasets of multiple types of cancers. A sample clustering procedure is applied on the datasets, while the existing DEG approach are proposed to identify a list of DEGs from each cluster. Then an integrated gene signature is constructed from all lists through a weighting strategy.

Chapter 4 proposes a type of human protein complex signature instead of a gene signature for identifying potential drugs. The gene expression profiles of both diseases and drugs are transformed into the form of human protein complexes. The novel profiles are applied to identify signature and predict potential drugs for several types of cancers.

Chapter 5 proposes a drug signature strategy for personalized cancer medicine. This strategy identifies a signature for each drugs. Depending on the drug signatures, a single patient is given a list of drug candidates. For the specific type of cancer, a frequency analysis is proposed to identify potential drugs from all list of drugs of patients.

Chapter 6 proposes a graph convolutional network with multi-kernel (GCNMK) to identify potential DDIs. In the GCNMK model, the known DDIs are divided into two graphs based on their clinical responses. The proposed model concatenates those two graph kernels together and achieves improved performance.

With my proposed algorithms, the accuracy of both signature-based and DDI prediction methods is improved. In the experiments, the signature-based methods proposed lists of potential drugs for several types of cancers, and the GCNMK model predicted potential DDIs for cancer-related drugs. Those studies enhance my understandings of drug repositioning.

7.2 Limitations

In previous section, I have discussed the performance of my proposed methods. However, they are not perfect. Each of them has some limitations. In Chapter 3, the clustering strategy is proposed to identify two subgroups from heterogeneous datasets. However, two clusters may not be optimal for other datasets.

Additionally, in Chapter 3, genes are treated as independent elements. However, they cooperate in disease conditions. In order to reflect their dependencies, I utilize HPC information in Chapter 4. The gene signatures of diseases are transformed to HPC signatures.

In Chapters 3 and 4, both gene signatures and HPC signatures are about diseases. A number of patient profiles are essential to identify disease signatures. Therefore, they can not work well when dealing with a single patient in practice. In order to address this limitation, I construct drug signatures and predict potential drugs for personalized treatment in Chapter 5. However, a drug signature would be not reliable if the number of drug induced expression profiles is small.

In Chapter 6, DDIs are divided into two groups according to their “increase” and “decrease” responses. However, DDIs are more heterogeneous, and more types of DDIs should be distinguished. Additionally, the networks are based on known DDIs. Therefore, the proposed method can not identify interactions between isolated drugs.

7.3 Future work

Based on the studies proposed in this dissertation, several future directions for drug repositioning are proposed as follows:

1. Using multiple types of data to identify a signature of either a drug or a disease.
   Multiple types of data characterize different aspects of drugs and diseases. Analyzing more data may help describe a disease condition or drug perturbation more accurately. However, in signature-based methods, only the transcriptomic data are commonly employed to identify a signature. Therefore, new methods should use other types of data, such as disease-gene associations, to study their applications in identifying signatures.

2. Using multiple types of biomedical entities and associations to enrich the drug-related network.
   More types of biomedical entities and associations can be used to construct a heterogeneous network, such as a drug-target-disease network. Predicting various types of missing links, such as DTIs, enhances the understanding of drug repositioning.

3. Using sparse networks to describe multiple functions of a specific type of interaction.
   Instead of constructing a large comprehensive heterogeneous network, some sparse networks can be used

to predict different functions. For instance, each function of DDI can have a corresponding network, while they can be concatenated to predict potential functions of unknown DDIs.

References

[1] Tonse NK Raju. The Nobel chronicles. *The Lancet*, 354(9175):1–2, 1999.

[2] Beste Turanli, Ozlem Altay, Jan Borén, Hasan Turkez, Jens Nielsen, Mathias Uhlen, Kazim Yalcin Arga, and Adil Mardinoglu. Systems biology based drug repositioning for development of cancer therapy. In *Seminars in Cancer Biology*, volume 68, pages 47–58. Elsevier, 2021.

[3] Fei Wang, Xiujuan Lei, Bo Liao, and Fang-Xiang Wu. Human protein complex signatures for drug repositioning. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 42–50, 2019.

[4] Chien-Hung Huang, Peter Mu-Hsin Chang, Chia-Wei Hsu, Chi-Ying F Huang, and Ka-Lok Ng. Drug repositioning for non-small cell lung cancer by using machine learning algorithms and topological graph theory. *BMC Bioinformatics*, 17(1):13–26, 2016.

[5] Swarnaseetha Adusumalli, Zhen-Kai Ngian, Wei-Qi Lin, Touati Benoukraf, and Chin-Tong Ong. Increased intron retention is a post-transcriptional signature associated with progressive aging and Alzheimer's disease. *Aging Cell*, 18(3):e12928, 2019.

[6] Nelson RC Monteiro, Bernardete Ribeiro, and Joel Arrais. Drug-target interaction prediction: end-to-end deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[7] Min Oh, Jaegyoon Ahn, and Youngmi Yoon. A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. *PLoS One*, 9(10):1–12, 2014.

[8] Andrej Kastrin, Polonca Ferk, and Brane Leskošek. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. *PLoS One*, 13(5):1–23, 2018.

[9] Narjes Rohani and Changiz Eslahchi. Drug-drug interaction predicting by neural network using integrated similarity. *Scientific Reports*, 9(1):1–11, 2019.

[10] Yang Zhang, Yang Qiu, Yuxin Cui, Shichao Liu, and Wen Zhang. Predicting drug-drug interactions using multi-modal deep auto-encoders based network embedding and positive-unlabeled learning. *Methods*, 179:37–46, 2020.

[11] Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Xiaoqing Peng, Fang-Xiang Wu, and Yi Pan. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, 32(17):2664–2671, 2016.

[12] Zui Xuan Xiao, Ruo Qiao Chen, Dian Xing Hu, Xiao Qiang Xie, Shang Bin Yu, and Xiao Qian Chen. Identification of repaglinide as a therapeutic drug for glioblastoma multiforme. *Biochemical and Biophysical Research Communications*, 488(1):33–39, 2017.

[13] Vijayendran Chandran, Giovanni Coppola, Homaira Nawabi, Takao Omura, Revital Versano, Eric A Huebner, Alice Zhang, Michael Costigan, Ajay Yekkirala, Lee Barrett, et al. A systems-level analysis of the peripheral nerve intrinsic axonal growth program. *Neuron*, 89(5):956–970, 2016.

[14] Qing Wen, Paul O'reilly, Philip D Dunne, Mark Lawler, Sandra Van Schaeybroeck, Manuel Salto-Tellez, Peter Hamilton, and Shu-Dong Zhang. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC Systems Biology*, 9(5):1–11, 2015.

[15] Eunyoung Kim, A-sol Choi, and Hojung Nam. Drug repositioning of herbal compounds via a machine-learning approach. *BMC Bioinformatics*, 20(10):33–43, 2019.

[16] Taekeon Lee and Youngmi Yoon. Drug repositioning using drug-disease vectors based on an integrated network. *BMC Bioinformatics*, 19(1):1–12, 2018.

[17] Reza Ferdousi, Reza Safdari, and Yadollah Omidi. Computational prediction of drug-drug interactions based on drugs functional similarities. *Journal of Biomedical Informatics*, 70:54–64, 2017.

[18] Cheng Yan, Guihua Duan, Yayan Zhang, Fang-Xiang Wu, Yi Pan, and Jianxin Wang. Predicting drug-drug interactions based on integrated similarity and semi-supervised learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[19] Cheng Yan, Guihua Duan, Yi Pan, Fang-Xiang Wu, and Jianxin Wang. DDIGIP: predicting drug-drug interactions based on gaussian interaction profile kernels. *BMC Bioinformatics*, 20(15):1–10, 2019.

[20] Yi Zheng, Hui Peng, Xiaocai Zhang, Zhixun Zhao, Xiaoying Gao, and Jinyan Li. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions. *BMC Bioinformatics*, 20(19):1–12, 2019.

[21] Yu Wang, Yanzhi Guo, Qifan Kuang, Xuemei Pu, Yue Ji, Zhihang Zhang, and Menglong Li. A comparative study of family-specific protein–ligand complex affinity prediction based on random forest approach. *Journal of Computer-Aided Molecular Design*, 29(4):349–360, 2015.

[22] Xiangxiang Zeng, Siyi Zhu, Yuan Hou, Pengyue Zhang, Lang Li, Jing Li, L Frank Huang, Stephen J Lewis, Ruth Nussinov, and Feixiong Cheng. Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics*, 36(9):2805–2812, 2020.

[23] Yanyi Chu, Aman Chandra Kaushik, Xiangeng Wang, Wei Wang, Yufang Zhang, Xiaoqi Shan, Dennis Russell Salahub, Yi Xiong, and Dong-Qing Wei. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. *Briefings in Bioinformatics*, 22(1):451–462, 2021.

[24] Yu-Ting Lin, Sheh-Yi Sheu, and Chen-Ching Lin. Prediction of drug-protein interaction and drug repositioning using machine learning model. *bioRxiv*, 2020.

[25] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research*, 16(4):1401–1409, 2017.

[26] Frank Emmert-Streib, Shailesh Tripathi, Ricardo de Matos Simoes, Ahmed F Hawwa, and Matthias Dehmer. The human disease network: Opportunities for classification, diagnosis, and prediction of disorders and disease genes. *Systems Biomedicine*, 1(1):20–28, 2013.

[27] Holly Matthews, James Hanison, and Niroshini Nirmalan. “omics”-informed drug and biomarker discovery: opportunities, challenges and future perspectives. *Proteomes*, 4(3):1–12, 2016.

[28] Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery*, 18(1):41–58, 2019.

[29] Guangxu Jin and Stephen TC Wong. Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug Discovery Today*, 19(5):637–644, 2014.

[30] David C Swinney and Jason Anthony. How were new medicines discovered? *Nature Reviews Drug Discovery*, 10(7):507–519, 2011.

[31] Mark R Hurle, Lun Yang, Qing Xie, Deepak K Rajpal, Philippe Sanseau, and Pankaj Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335–341, 2013.

[32] Mihaly Szabo, Sara Svensson Akusjärvi, Ankur Saxena, Jianping Liu, Gayathri Chandrasekar, and Satish S Kitambi. Cell and small animal models for phenotypic drug discovery. *Drug Design, Development and Therapy*, 11:1957, 2017.

[33] Rita Santos, Oleg Ursu, Anna Gaulton, A Patrícia Bento, Ramesh S Donadi, Cristian G Bologa, Anneli Karlsson, Bissan Al-Lazikani, Anne Hersey, Tudor I Oprea, et al. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017.

[34] Gordon K Smyth, Joëlle Michaud, and Hamish S Scott. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics*, 21(9):2067–2075, 2005.

[35] Olga G Troyanskaya, Mitchell E Garber, Patrick O Brown, David Botstein, and Russ B Altman. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461, 2002.

[36] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[37] Jiao Li and Zhiyong Lu. A new method for computational drug repositioning using drug pairwise similarity. In *2012 IEEE International Conference on Bioinformatics and Biomedicine*, pages 1–4. IEEE, 2012.

[38] Kalum Clayton, Marta E Polak, Christopher H Woelk, and Paul Elkington. Gene expression signatures in tuberculosis have greater overlap with autoimmune diseases than with infectious diseases. *American Journal of Respiratory and Critical Care Medicine*, 196(5):655–656, 2017.

[39] Adi L Tarca, Xiaofeng Gong, Roberto Romero, Wenxin Yang, Zhongqu Duan, Hao Yang, Chengfang Zhang, and Peixuan Wang. Human blood gene signature as a marker for smoking exposure: computational approaches of the top ranked teams in the sbv IMPROVER systems toxicology challenge. *Computational Toxicology*, 5:31–37, 2018.

[40] Gang Chen, Jianxin Wang, Yi Pan, and Jianer Chen. Identification of breast cancer gene signature in protein interaction network using graph centrality. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 402–405. IEEE, 2011.

[41] Alan Mathison Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[42] Zhanchao Li, Qixing Huang, Xingyu Chen, Yang Wang, Jinlong Li, Yun Xie, Zong Dai, and Xiaoyong Zou. Identification of drug-disease associations using information of molecular structures and clinical symptoms via deep convolutional neural network. *Frontiers in Chemistry*, 7:924, 2020.

[43] Hafez Eslami Manoochehri, Susmitha Sri Kadiyala, and Mehrdad Nourani. Predicting drug-target interactions using Weisfeiler-Lehman neural network. In *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 1–4. IEEE, 2019.

[44] Hafez Eslami Manoochehri and Mehrdad Nourani. Drug-target interaction prediction using semi-bipartite graph model and deep learning. *BMC Bioinformatics*, 21(4):1–16, 2020.

[45] ShanShan Hu, Chenglin Zhang, Peng Chen, Pengying Gu, Jun Zhang, and Bing Wang. Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC Bioinformatics*, 20(25):1–12, 2019.

[46] Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23):5545–5547, 2020.

[47] Ingoo Lee, Jongsoo Keum, and Hojung Nam. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Computational Biology*, 15(6):1–21, 2019.

[48] Mingjian Jiang, Zhen Li, Shugang Zhang, Shuang Wang, Xiaofeng Wang, Qing Yuan, and Zhiqiang Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC Advances*, 10(35):20701–20712, 2020.

[49] Wen Torng and Russ B Altman. Graph convolutional neural networks for predicting drug-target interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019.

[50] Yingdong Wang, Gaoshan Deng, Nianyin Zeng, Xiao Song, and Yuanying Zhuang. Drug-disease association prediction based on neighborhood information aggregation in neural networks. *IEEE Access*, 7:50581–50587, 2019.

[51] Renyi Zhou, Zhangli Lu, Huimin Luo, Ju Xiang, Min Zeng, and Min Li. NEDD: a network embedding based method for predicting drug-disease associations. *BMC Bioinformatics*, 21(13):1–12, 2020.

[52] Hui Liu, Wenhao Zhang, Yinglong Song, Lei Deng, and Shuigeng Zhou. HNet-DNN: Inferring new drug–disease associations with deep neural network based on heterogeneous network features. *Journal of Chemical Information and Modeling*, 60(4):2367–2376, 2020.

[53] Han-Jing Jiang, Yu-An Huang, and Zhu-Hong You. Predicting drug-disease associations via using gaussian interaction profile and kernel-based autoencoder. *BioMed Research International*, 2019, 2019.

[54] Han-Jing Jiang, Zhu-Hong You, and Yu-An Huang. Predicting drug- disease associations via sigmoid kernel-based convolutional neural networks. *Journal of Translational Medicine*, 17(1):1–11, 2019.

[55] Lei Wang, Zhu-Hong You, Xing Chen, Shi-Xiong Xia, Feng Liu, Xin Yan, Yong Zhou, and Ke-Jian Song. A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology*, 25(3):361–373, 2018.

[56] Bowen Kuo, Yihuang Kang, Pinghsung Wu, Sheng-Tai Huang, and Yajie Huang. Discovering drug-drug and drug-disease interactions inducing acute kidney injury using deep rule forests. In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 385–390. IEEE, 2020.

[57] Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Peng Ni, Kaijie Zhao, Fang-Xiang Wu, and Yi Pan. Computational drug repositioning with random walk on a heterogeneous network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(6):1890–1900, 2018.

[58] Wen Zhang, Yanlin Chen, Feng Liu, Fei Luo, Gang Tian, and Xiaohong Li. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinformatics*, 18(1):1–12, 2017.

[59] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710, 2014.

[60] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8(1):1–13, 2017.

[61] Ingoo Lee and Hojung Nam. Identification of drug-target interaction by a random walk with restart method on an interactome network. *BMC Bioinformatics*, 19(8):9–18, 2018.

[62] Ping Xuan, Chang Sun, Tiangang Zhang, Yilin Ye, Tonghui Shen, and Yihua Dong. Gradient boosting decision tree-based method for predicting interactions between target genes and drugs. *Frontiers in Genetics*, 10:459, 2019.

[63] Poorya Parvizi, Francisco Azuaje, Evropi Theodoratou, and Saturnino Luz. A network-based embedding method for drug-target interaction prediction. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 5304–5307. IEEE, 2020.

[64] Chang Sun, Ping Xuan, Tiangang Zhang, and Yilin Ye. Graph convolutional autoencoder and generative adversarial network-based method for predicting drug-target interactions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[65] Huiqing Wang, Jingjing Wang, Chunlin Dong, Yuanyuan Lian, Dan Liu, and Zhiliang Yan. A novel approach for drug-target interactions prediction based on multimodal deep autoencoder. *Frontiers in Pharmacology*, 10:1–19, 2020.

[66] Dalong Song, Yao Chen, Qian Min, Qingrong Sun, Kai Ye, Changjiang Zhou, Shengyue Yuan, Zhaolin Sun, and Jun Liao. Similarity-based machine learning support vector machine predictor of drug-drug interactions with improved accuracies. *Journal of Clinical Pharmacy and Therapeutics*, 44(2):268–275, 2019.

[67] Sathien Hunta, Nattapol Aunsri, and Thongchai Yooyativong. Drug-drug interactions prediction from enzyme action crossing through machine learning approaches. In *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1–4. IEEE, 2015.

[68] Wen Zhang, Xiang Yue, Feng Huang, Ruoqi Liu, Yanlin Chen, and Chunyang Ruan. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods*, 145:51–59, 2018.

[69] Tamer N Jarada, Jon G Rokne, and Reda Alhajj. SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder. *Knowledge-Based Systems*, 22:1–20, 2021.

[70] Geonhee Lee, Chihyun Park, and Jaegyoon Ahn. Novel deep learning model for more accurate prediction of drug-drug interaction effects. *BMC Bioinformatics*, 20(1):1–8, 2019.

[71] Xiangxiang Zeng, Siyi Zhu, Weiqiang Lu, Zehui Liu, Jin Huang, Yadi Zhou, Jiansong Fang, Yin Huang, Huimin Guo, Lang Li, et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science*, 11(7):1775–1797, 2020.

[72] Tamer N Jarada, Jon G Rokne, and Reda Alhajj. SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks. *BMC Bioinformatics*, 22(1):1–20, 2021.

[73] Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning improves prediction of drug–drug and drug–food interactions. *Proceedings of the National Academy of Sciences*, 115(18):E4304–E4311, 2018.

[74] Prashant Kumar Shukla, Piyush Kumar Shukla, Poonam Sharma, Paresh Rawat, Jashwant Samar, Rahul Moriwal, and Manjit Kaur. Efficient prediction of drug–drug interaction using deep learning models. *IET Systems Biology*, 14(4):211–216, 2020.

[75] Yifan Deng, Xinran Xu, Yang Qiu, Jingbo Xia, Wen Zhang, and Shichao Liu. A multimodal deep learning framework for predicting drug–drug interaction events. *Bioinformatics*, 36(15):4316–4322, 2020.

[76] Yue-Hua Feng, Shao-Wu Zhang, and Jian-Yu Shi. DPDDI: a deep predictor for drug-drug interactions. *BMC Bioinformatics*, 21(1):1–15, 2020.

[77] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[78] Tianyi Zhao, Yang Hu, Linda R Valsdottir, Tianyi Zang, and Jiajie Peng. Identifying drug–target interactions based on graph convolutional network and deep neural network. *Briefings in Bioinformatics*, 22(2):2141–2150, 2021.

[79] Hanbi Lee and Wankyu Kim. Comparison of target features for predicting drug-target interactions by deep neural network based on large-scale drug-induced transcriptome data. *Pharmaceutics*, 11(8):1–11, 2019.

[80] Jiajie Peng, Jingyi Li, and Xuequn Shang. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. *BMC Bioinformatics*, 21(13):1–13, 2020.

[81] Ping Xuan, Yilin Ye, Tiangang Zhang, Lianfeng Zhao, and Chang Sun. Convolutional neural network and bidirectional long short-term memory-based method for predicting drug–disease associations. *Cells*, 8(7):1–15, 2019.

[82] Ping Xuan, Hui Cui, Tonghui Shen, Nan Sheng, and Tiangang Zhang. Heterodualnet: A dual convolutional neural network with heterogeneous layers for drug-disease association prediction via chou's five-step rule. *Frontiers in Pharmacology*, 10:1–12, 2019.

[83] Ping Xuan, Lianfeng Zhao, Tiangang Zhang, Yilin Ye, and Yan Zhang. Inferring drug-related diseases based on convolutional neural network and gated recurrent unit. *Molecules*, 24(15):1–15, 2019.

[84] Ping Xuan, Ling Gao, Nan Sheng, Tiangang Zhang, and Toshiya Nakaguchi. Graph convolutional autoencoder and fully-connected autoencoder with attention mechanism based method for predicting drug-disease associations. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1793–1804, 2020.

[85] Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. DeepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.

[86] Caterina Palleria, Antonello Di Paolo, Chiara Giofrè, Chiara Caglioti, Giacomo Leuzzi, Antonio Siniscalchi, Giovambattista De Sarro, and Luca Gallelli. Pharmacokinetic drug-drug interaction and their implication in clinical management. *Journal of Research in Medical Sciences: the Official Journal of Isfahan University of Medical Sciences*, 18(7):1–10, 2013.

[87] David Altshuler, Mark Daly, and Leonid Kruglyak. Guilt by association. *Nature Genetics*, 26(2):135–137, 2000.

[88] Stephen Oliver. Guilt-by-association goes global. *Nature*, 403(6770):601–602, 2000.

[89] Shan-Shan Hu, Peng Chen, Bing Wang, and Jinyan Li. Protein binding hot spots prediction from sequence only by a new ensemble learning method. *Amino Acids*, 49(10):1773–1785, 2017.

[90] Isabella Gashaw, Peter Ellinghaus, Anette Sommer, and Khusru Asadullah. What makes a good drug target? *Drug Discovery Today*, 16(23-24):1037–1043, 2011.

[91] Huimin Luo, Min Li, Mengyun Yang, Fang-Xiang Wu, Yaohang Li, and Jianxin Wang. Biomedical data and computational models for drug repositioning: a comprehensive review. *Briefings in Bioinformatics*, 22(2):1604–1619, 2021.

[92] Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The Chemistry Development Kit (CDK): An open-source java library for chemo-and bioinformatics. *Journal of Chemical Information and Computer Sciences*, 43(2):493–500, 2003.

[93] Masahiro Hattori, Yasushi Okuno, Susumu Goto, and Minoru Kanehisa. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society*, 125(39):11853–11865, 2003.

[94] Hisashi Kashima, Koji Tsuda, and Akihiro Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 321–328, 2003.

[95] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):1–13, 2015.

[96] Günter Klambauer, Martin Wischenbart, Michael Mahr, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map. *Bioinformatics*, 31(20):3392–3394, 2015.

[97] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*, 1995.

[98] Monica Campillos, Michael Kuhn, Anne-Claude Gavin, Lars Juhl Jensen, and Peer Bork. Drug target identification using side-effect similarity. *Science*, 321(5886):263–266, 2008.

[99] Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto, and Yoshihiro Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, 2012.

[100] Liang Cheng, Jie Li, Peng Ju, Jiajie Peng, and Yadong Wang. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PloS One*, 9(6):1–11, 2014.

[101] Jörg Menche, Amitabh Sharma, Maksim Kitsak, Susan Dina Ghiassian, Marc Vidal, Joseph Loscalzo, and Albert-László Barabási. Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1–15, 2015.

[102] Guangchuang Yu, Li-Gen Wang, Guang-Rong Yan, and Qing-Yu He. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, 31(4):608–609, 2015.

[103] Sachin Mathur and Deendayal Dinakarpandian. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, 45(2):363–371, 2012.

[104] Hyojung Paik, Hyoung-Sam Heo, Hyo-jeong Ban, and Seong Beom Cho. Unraveling human protein interaction networks underlying co-occurrences of diseases and pathological conditions. *Journal of Translational Medicine*, 12(1):1–8, 2014.

[105] Steven B Smith, William Dampier, Aydin Tozeren, James R Brown, and Michal Magid-Slav. Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. *PLoS One*, 7(3):1–15, 2012.

[106] Johannes Palme, Sepp Hochreiter, and Ulrich Bodenhofer. KeBABS: an R package for kernel-based analysis of biological sequences. *Bioinformatics*, 31(15):2574–2576, 2015.

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

[108] Hawoong Jeong, Sean P Mason, A-L Barabási, and Zoltan N Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, 2001.

[109] Maliackal Poulo Joy, Amy Brock, Donald E Ingber, and Sui Huang. High-betweenness proteins in the yeast protein interaction network. *Journal of Biomedicine and Biotechnology*, 2005(2):96, 2005.

[110] Mahdi Jalili, Ali Salehzadeh-Yazdi, Shailendra Gupta, Olaf Wolkenhauer, Marjan Yaghmaie, Osbaldo Resendis-Antonio, and Kamran Alimoghaddam. Evolution of centrality measurements for the detection of essential proteins in biological networks. *Frontiers in Physiology*, 7:375, 2016.

[111] Caroline C Friedel and Ralf Zimmer. Inferring topology from clustering coefficients in protein-protein interaction networks. *BMC Bioinformatics*, 7(1):519, 2006.

[112] Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer Science & Business Media, 2013.

[113] Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[114] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[115] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[116] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[117] Yanjun Qi. Random forest for bioinformatics. In *Ensemble Machine Learning*, pages 307–323. Springer, 2012.

[118] Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41(3):552–568, 2010.

[119] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[120] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378, 2002.

[121] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[122] Fei Wang, Yulian Ding, Xiujuan Lei, Bo Liao, and Fangxiang Wu. Identifying gene signatures for cancer drug repositioning based on sample clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.

[123] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[124] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[125] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318. PMLR, 2013.

[126] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(12):3371–3408, 2010.

[127] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[128] Kunihiko Fukushima and Sei Miyake. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and Cooperation in Neural Nets*, pages 267–285. Springer, 1982.

[129] Yan-Bin Wang, Zhu-Hong You, Shan Yang, Hai-Cheng Yi, Zhan-Heng Chen, and Kai Zheng. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Medical Informatics and Decision Making*, 20(2):1–9, 2020.

[130] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

[131] Brighter Agyemang, Wei-Ping Wu, Michael Y Kpiebaareh, and Ebenezer Nanor. Drug-target indication prediction by integrating end-to-end learning and fingerprints. In *2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pages 266–272. IEEE, 2019.

[132] Kelli L Goss and David J Gordon. Gene expression signature based screening identifies ribonucleotide reductase as a candidate therapeutic target in ewing sarcoma. *Oncotarget*, 7(39):63003, 2016.

[133] Ziyan Pessetto, Bin Chen, Hani Alturkmani, Stephen Hyter, Colleen A Flynn, Michael Baltezor, Yan Ma, Howard G Rosenthal, Kathleen A Neville, Scott J Weir, et al. In silico and in vitro drug screening identifies new therapeutic approaches for Ewing sarcoma. *Oncotarget*, 8(3):4079, 2017.

[134] Lan Huang, HuiMin Luo, Mengyun Yang, Fang-Xiang Wu, and Jianxin Wang. Drug and disease similarity calculation platform for drug repositioning. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 124–129. IEEE, 2019.

[135] Lan Huang, Huimin Luo, Suning Li, Fang-Xiang Wu, and Jianxin Wang. Drug–drug similarity measure and its applications. *Briefings in Bioinformatics*, 22(4):bbaa265, 2021.

[136] Carolyn E Lipscomb. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[137] Marc A Van Driel, Jorn Bruggeman, Gert Vriend, Han G Brunner, and Jack AM Leunissen. A text-mining analysis of the human phenome. *European Journal of Human Genetics*, 14(5):535–542, 2006.

[138] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526, 2007.

[139] Xing Chen, Chenggang Clarence Yan, Xu Zhang, Zhu-Hong You, Lixi Deng, Ying Liu, Yongdong Zhang, and Qionghai Dai. WBSMDA: within and between score for MiRNA-disease association prediction. *Scientific Reports*, 6(1):1–9, 2016.

[140] Assaf Gottlieb, Gideon Y Stein, Eytan Ruppin, and Roded Sharan. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Molecular Systems Biology*, 7(1):496, 2011.

[141] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988.

[142] XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. Human symptoms–disease network. *Nature Communications*, 5(1):1–10, 2014.

[143] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[144] Bei Wang, Xiaoqing Lyu, Jingwei Qu, Haowen Sun, Zehua Pan, and Zhi Tang. GNDD: A graph neural network-based method for drug-disease association prediction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1253–1255. IEEE, 2019.

[145] Zhouxin Yu, Feng Huang, Xiaohan Zhao, Wenjie Xiao, and Wen Zhang. Predicting drug–disease associations through layer attention graph convolutional network. *Briefings in Bioinformatics*, 22(4):bbaa243, 2021.

[146] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795):1929–1935, 2006.

[147] Justin Lamb. The Connectivity Map: a new tool for biomedical research. *Nature Reviews Cancer*, 7(1):54, 2007.

[148] Wei-Qi Wei, Robert M Cronin, Hua Xu, Thomas A Lasko, Lisa Bastarache, and Joshua C Denny. Development and evaluation of an ensemble resource linking medications to their indications. *Journal of the American Medical Informatics Association*, 20(5):954–961, 2013.

[149] Xin Bi, He Ma, Jianhua Li, Yuliang Ma, and Deyang Chen. A positive and unlabeled learning framework based on extreme learning machine for drug-drug interactions discovery. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–12, 2018.

[150] Gao Huang, Shiji Song, Jatinder ND Gupta, and Cheng Wu. Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics*, 44(12):2405–2417, 2014.

[151] Rawan S Olayan, Haitham Ashoor, and Vladimir B Bajic. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics*, 34(7):1164–1173, 2018.

[152] Bo Wang, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[153] Bin Zhou, Rong Wang, Ping Wu, and De-Xin Kong. Drug repurposing based on drug–drug interaction. *Chemical Biology & Drug Design*, 85(2):137–144, 2015.

[154] Anum Munir, Sana Elahi, and Nayyer Masood. Clustering based drug-drug interaction networks for possible repositioning of drugs against egfr mutations: clustering based ddi networks for egfr mutations. *Computational Biology and Chemistry*, 75:24–31, 2018.

[155] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, 2006.

[156] Bo Peng and Xia Ning. Deep learning for high-order drug-drug interaction prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 197–206, 2019.

[157] Xuan Lin, Zhe Quan, Zhi-Jie Wang, Tengfei Ma, and Xiangxiang Zeng. KGNN: Knowledge graph neural network for drug-drug interaction prediction. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence*, volume 380, pages 2739–2745, 2020.

[158] Tom Lynch and Amy L Price. The effect of cytochrome P450 metabolism on drug response, interactions, and adverse effects. *American Family Physician*, 76(3):391–396, 2007.

[159] Zhi-Hua Zhou and Ji Feng. Deep forest. *National Science Review*, 6(1):74–86, 2019.

[160] Nansu Zong, Hyeoneui Kim, Victoria Ngo, and Olivier Harismendy. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics*, 33(15):2337–2344, 2017.

[161] Feixiong Cheng, Chuang Liu, Jing Jiang, Weiqiang Lu, Weihua Li, Guixia Liu, Weixing Zhou, Jin Huang, and Yun Tang. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology*, 8(5):e1002503, 2012.

[162] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229, 2015.

[163] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1, 2010.

[164] Jiaying You, Robert D McLeod, and Pingzhao Hu. Predicting drug-target interaction network using deep learning model. *Computational Biology and Chemistry*, 80:90–101, 2019.

[165] Chun Wei Yap. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.

[166] Shuichi Kawashima and Minoru Kanehisa. AAindex: amino acid index database. *Nucleic Acids Research*, 28(1):374–374, 2000.

[167] Mirco Michel, David Menéndez Hurtado, and Arne Elofsson. PconsC4: fast, accurate and hassle-free contact predictions. *Bioinformatics*, 35(15):2677–2679, 2019.

[168] Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988, 2019.

[169] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

[170] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, Nadezhda T Doncheva, John H Morris, Peer Bork, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[171] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[172] Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal*, 18:784–790, 2020.

[173] Manli Wang, Ruiyuan Cao, Leike Zhang, Xinglou Yang, Jia Liu, Mingyue Xu, Zhengli Shi, Zhihong Hu, Wu Zhong, and Gengfu Xiao. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research*, 30(3):269–271, 2020.

[174] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S Yu. Building text classifiers using positive and unlabeled examples. In *Third IEEE International Conference on Data Mining*, pages 179–186. IEEE, 2003.

[175] Wei Lan, Jianxin Wang, Min Li, Jin Liu, Yaohang Li, Fang-Xiang Wu, and Yi Pan. Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing*, 206:50–57, 2016.

[176] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[177] Wenhui Wang, Sen Yang, Xiang Zhang, and Jing Li. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, 30(20):2923–2930, 2014.

[178] Steven M Paul, Daniel S Mytelka, Christopher T Dunwiddie, Charles C Persinger, Bernard H Munos, Stacy R Lindborg, and Aaron L Schacht. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010.

[179] Maryam Lotfi Shahreza, Nasser Ghadiri, Sayed Rasoul Mousavi, Jaleh Varshosaz, and James R Green. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics*, 19(5):878–892, 2018.

[180] Mitradev Boolell, Michael J Allen, Stephen A Ballard, Sam Gepi-Attee, Gary J Muirhead, Alasdair M Naylor, Ian H Osterloh, and Clive Gingell. Sildenafil: an orally active type 5 cyclic GMP-specific phosphodiesterase inhibitor for the treatment of penile erectile dysfunction. *International Journal of Impotence Research*, 8(2):47–52, 1996.

[181] Jeffrey K Aronson. Old drugs–new uses. *British Journal of Clinical Pharmacology*, 64(5):563–565, 2007.

[182] Zikai Wu, Yong Wang, and Luonan Chen. Network-based drug repositioning. *Molecular BioSystems*, 9(6):1268–1281, 2013.

[183] Fei Wang, Xiujuan Lei, and Fang-Xiang Wu. A review of drug repositioning based chemical-induced cell line expression data. *Current Medicinal Chemistry*, 27(32):5340–5350, 2020.

[184] Alexandra B Keenan, Sherry L Jenkins, Kathleen M Jagodnik, Simon Koplev, Edward He, Denis Torre, Zichen Wang, Anders B Dohlman, Moshe C Silverstein, Alexander Lachmann, et al. The library of integrated network-based cellular signatures NIH program: system-level cataloging of human cells response to perturbations. *Cell Systems*, 6(1):13–24, 2018.

[185] Xianxiao Zhou, Minghui Wang, Igor Katsyv, Hanna Irie, and Bin Zhang. EMUDRA: Ensemble of multiple drug repositioning approaches to improve prediction accuracy. *Bioinformatics*, 34(18):3151–3159, 2018.

[186] Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici. A novel computational approach for drug repurposing using systems biology. *Bioinformatics*, 34(16):2817–2825, 2018.

[187] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, 2012.

[188] Mark Schena, Dari Shalon, Renu Heller, Andrew Chai, Patrick O Brown, and Ronald W Davis. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences*, 93(20):10614–10619, 1996.

[189] Joseph DeRisi, Lolita Penland, Patrick O Brown, Michael L Bittner, Paul S Meltzer, Michael Ray, Yidong Chen, Yan A Su, and Jeffery M Trent. Use of a cDNA microarray to analyse gene expression. *Nature Genetics*, 14:457–460, 1996.

[190] Davis J McCarthy and Gprdon K Smyth. Testing significance relative to a fold-change threshold is a TREAT. *Bioinformatics*, 25(6):765–771, 2009.

[191] Paul O'Reilly, Qing Wen, Peter Bankhead, Philip D Dunne, Darragh G McArt, Suzanne McPherson, Peter W Hamilton, Ken I Mills, and Shu-Dong Zhang. QUADrATiC: scalable gene expression connectivity mapping for repurposing FDA-approved therapeutics. *BMC Bioinformatics*, 17(1):1–15, 2016.

[192] Jianzhen Xu and Yongjin Li. Discovering disease-genes by topological features in human protein–protein interaction network. *Bioinformatics*, 22(22):2800–2805, 2006.

[193] Eric March-Vila, Luca Pinzi, Noé Sturm, Annachiara Tinivella, Ola Engkvist, Hongming Chen, and Giulio Rastelli. On the integration of in silico drug design methods for drug repurposing. *Frontiers in Pharmacology*, 8:298, 2017.

[194] Xiaoping Liu, Xiao Chang, Rui Liu, Xiangtian Yu, Luonan Chen, and Kazuyuki Aihara. Quantifying critical states of complex diseases using single-sample dynamic network biomarkers. *PLoS Computational Biology*, 13(7):e1005633, 2017.

[195] Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825–2827, 2006.

[196] Kentaro Inamura. Lung cancer: understanding its molecular pathology and the 2015 WHO classification. *Frontiers in Oncology*, 7:193, 2017.

[197] Lisandra West, Smruti J Vidwans, Nicholas P Campbell, Jeff Shrager, George R Simon, Raphael Bueno, Phillip A Dennis, Gregory A Otterson, and Ravi Salgia. A novel classification of lung cancer into molecular subtypes. *PLoS One*, 7(2):e31906, 2012.

[198] Shai Ben-David and Margareta Ackerman. Measures of clustering quality: A working set of axioms for clustering. In *Advances in Neural Information Processing Systems*, volume 21, pages 121–128, 2008.

[199] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[200] Andrew Chatr-Aryamontri, Bobby-Joe Breitkreutz, Rose Oughtred, Lorrie Boucher, Sven Heinicke, Daici Chen, Chris Stark, Ashton Breitkreutz, Nadine Kolas, Lara O'Donnell, et al. The BioGRID interaction database: 2015 update. *Nucleic Acids Research*, 43(D1):D470–D478, 2015.

[201] Wei Lan, Jianxin Wang, Min Li, Wei Peng, and Fangxiang Wu. Computational approaches for prioritizing candidate disease genes based on PPI networks. *Tsinghua Science and Technology*, 20(5):500–512, 2015.

[202] Xiujuan Lei, Jianfang Tian, Liang Ge, and Aidong Zhang. The clustering model and algorithm of PPI network based on propagating mechanism of artificial bee colony. *Information Sciences*, 247:21–39, 2013.

[203] Woochang Hwang, Young-rae Cho, Aidong Zhang, and Murali Ramanathan. Bridging centrality: identifying bridging nodes in scale-free networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 20–23, 2006.

[204] James X Hu, Hongyu Zhao, and Harrison H Zhou. False discovery rate control with groups. *Journal of the American Statistical Association*, 105(491):1215–1227, 2010.

[205] Shu-Dong Zhang and Timothy W Gant. sscMap: an extensible java application for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics*, 10(1):1–4, 2009.

[206] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

[207] Robert L Grossman, Allison P Heath, Vincent Ferretti, Harold E Varmus, Douglas R Lowy, Warren A Kibbe, and Louis M Staudt. Toward a shared vision for cancer genomic data. *New England Journal of Medicine*, 375(12):1109–1112, 2016.

[208] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[209] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[210] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[211] Mahvash Zakikhani, Ryan Dowling, I George Fantus, Nahum Sonenberg, and Michael Pollak. Metformin is an AMP kinase–dependent growth inhibitor for breast cancer cells. *Cancer Research*, 66(21):10269–10273, 2006.

[212] Edna Ayerim Mandujano-Tinoco, Juan Carlos Gallardo-Pérez, Alvaro Marín-Hernández, Rafael Moreno-Sánchez, and Sara Rodríguez-Enríquez. Anti-mitochondrial therapy in human breast cancer multi-cellular spheroids. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, 1833(3):541–551, 2013.

[213] Raoul Charles Coombes, David Dearnaley, Jonathan Humphreys, J. C. Gazet, H. T. Ford, A. G. Nash, K. Mashiter, and T. J. Powles. Danazol treatment of advanced breast cancer. *Cancer Treatment Reports*, 64(10-11):1073–1076, 1980.

[214] Manali Mody, Nachiket Dharker, Mark Bloomston, Pei-Shan Wang, Fu-Sheng Chou, Theodore S Glickman, Timothy McCaffrey, Zhaoqing Yang, Anne Pumfery, Daniel Lee, et al. Rosiglitazone sensitizes MDA-MB-231 breast cancer cells to anti-tumour effects of tumour necrosis factor-α, CH11 and CYC202. *Endocrine-Related Cancer*, 14(2):305–315, 2007.

[215] Rakesh K Srivastava, Razelle Kurzrock, and Sharmila Shankar. MS-275 sensitizes TRAIL-resistant breast cancer cells, inhibits angiogenesis and metastasis, and reverses epithelial-mesenchymal transition in vivo. *Molecular Cancer Therapeutics*, 9(12):3254–3266, 2010.

[216] Palimecio G Guerrero Jr, Paulo R de Oliveira, Adriano CM Baroni, Francisco A Marques, Ricardo Labes, Gabriela R Hurtado, and Miguel J Dabdoub. Synthesis of arotinoid acid and temarotene using mixed (Z)-1, 2-bis (organylchalcogene)-1-alkene as precursor. *Tetrahedron Letters*, 53(39):5302–5305, 2012.

[217] Rita S Mehta, William E Barlow, Kathy S Albain, Ted A Vandenberg, Shaker R Dakhil, Nagendra R Tirumali, Danika L Lew, Daniel F Hayes, Julie R Gralow, Robert B Livingston, et al. Combination anastrozole and fulvestrant in metastatic breast cancer. *New England Journal of Medicine*, 367(5):435–444, 2012.

[218] Leri Septiani Faried, Aaried Faried, Tatuya Kanuma, Takayuki Nakazato, Tomohide Tamura, Hiroyuki Kuwano, and Takashi Minegishi. Inhibition of the mammalian target of rapamycin (mTOR) by rapamycin increases chemosensitivity of caski cells to paclitaxel. *European Journal of Cancer*, 42(7):934–947, 2006.

[219] Christopher M Lee, Christa B Fuhrman, Vicente Planelles, Morgan R Peltier, David K Gaffney, Andrew P Soisson, Mark K Dodson, H Dennis Tolley, Christopher L Green, and Karen A Zempolich. Phosphatidylinositol 3-kinase inhibition by LY294002 radiosensitizes human cervical cancer cell lines. *Clinical Cancer Research*, 12(1):250–256, 2006.

[220] Shuyu Feng, Yue Yang, Jingyi Lv, Lichun Sun, and Mingqiu Liu. Valproic acid exhibits different cell growth arrest effect in three HPV-positive/negative cervical cancer cells and possibly via inducing Notch1 cleavage and E6 downregulation. *International Journal of Oncology*, 49(1):422–430, 2016.

[221] Bai-Cheng He, Jian-Li Gao, Bing-Qiang Zhang, Qing Luo, Qiong Shi, Stephanie H Kim, Enyi Huang, Yanhong Gao, Ke Yang, Eric R Wagner, et al. Tetrandrine inhibits Wnt/β-catenin signaling and suppresses tumor growth of human colorectal cancer. *Molecular Pharmacology*, 79(2):211–219, 2011.

[222] Hong-Mei Wang and Gui-Ying Zhang. Indomethacin suppresses growth of colon cancer via inhibition of angiogenesis in vivo. *World Journal of Gastroenterology: WJG*, 11(3):340, 2005.

[223] Xufeng Chen, Patty Wong, Eric Radany, and Jeffrey YC Wong. HDAC inhibitor, valproic acid, induces p53-dependent radiosensitization of colon cancer cells. *Cancer Biotherapy and Radiopharmaceuticals*, 24(6):689–699, 2009.

[224] Haizhong Huo, Zhiyuan Zhou, Jian Qin, Wenyong Liu, Bing Wang, and Yan Gu. Erastin disrupts mitochondrial permeability transition pore (mPTP) and induces apoptotic death of colorectal cancer cells. *PLoS One*, 11(5):e0154605, 2016.

[225] Shuho Semba, Nanami Itoh, Masafumi Ito, Masaru Harada, and Mitsunori Yamakawa. The in vitro and in vivo effects of 2-(4-morpholinyl)-8-phenyl-chromone (LY294002), a specific inhibitor of phosphatidylinositol 3-kinase, in human colon cancer cells. *Clinical Cancer Research*, 8(6):1957–1963, 2002.

[226] Chen Zhang, Ping Gong, Pengfei Liu, Ning Zhou, Yulai Zhou, and Yi Wang. Thioridazine elicits potent antitumor effects in colorectal cancer stem cells. *Oncology Reports*, 37(2):1168–1174, 2017.

[227] Xiaoxiong Wang, Jun Xu, Hao Wang, Long Wu, Weiqi Yuan, Jun Du, and Shaohui Cai. Trichostatin A, a histone deacetylase inhibitor, reverses epithelial–mesenchymal transition in colorectal cancer SW480 and prostate cancer PC3 cells. *Biochemical and Biophysical Research Communications*, 456(1):320–326, 2015.

[228] Luciana Tessitore, Annalisa Davit, Ivana Sarotto, and Giovanna Caderni. Resveratrol depresses the growth of colorectal aberrant crypt foci by affecting bax and p21 CIP expression. *Carcinogenesis*, 21(8):1619–1622, 2000.

[229] John C Marsh, JR Bertino, KH Katz, Carol A Davis, Henry J Durivage, Lisa S Rome, F Richards 2nd, Robert L Capizzi, Leonard R Farber, and Dominick N Pasquale. The influence of drug interval on the effect of methotrexate and fluorouracil in the treatment of advanced colorectal cancer. *Journal of Clinical Oncology*, 9(3):371–380, 1991.

[230] Celeste B Burness and Sean T Duggan. Trifluridine/tipiracil: a review in metastatic colorectal cancer. *Drugs*, 76(14):1393–1402, 2016.

[231] Fatemehsadat Amiri, Amir-Hassan Zarnani, Hamid Zand, Fariba Koohdani, Mahmood Jeddi-Tehrani, and Mohammadreza Vafa. Synergistic anti-proliferative effect of resveratrol and etoposide on human hepatocellular and colon cancer cell lines. *European Journal of Pharmacology*, 718(1-3):34–40, 2013.

[232] Leonard B Saltz, John V Cox, Charles Blanke, Lee S Rosen, Louis Fehrenbacher, Malcolm J Moore, Jean A Maroun, Stephen P Ackland, Paula K Locker, Nicoletta Pirotta, et al. Irinotecan plus fluorouracil and leucovorin for metastatic colorectal cancer. *New England Journal of Medicine*, 343(13):905–914, 2000.

[233] Kenji Kuroda, Akio Horiguchi, Makoto Sumitomo, Takako Asano, Keiichi Ito, Masamichi Hayakawa, and Tomohiko Asano. Activated akt prevents antitumor activity of gefitinib in renal cancer cells. *Urology*, 74(1):209–215, 2009.

[234] Bo Ram Seo, Kyoung-jin Min, Shin Kim, Jong-Wook Park, Won-Kyun Park, Tae-Jin Lee, and Taeg Kyu Kwon. Anisomycin treatment enhances TRAIL-mediated apoptosis in renal carcinoma cells through the down-regulation of Bcl-2, c-FLIP (L) and Mcl-1. *Biochimie*, 95(4):858–865, 2013.

[235] Paolo Bidoli, Diego L Cortinovis, Ilaria Colombo, Alessandra Crippa, Federica Cicchiello, Federica Villa, Marina E Cazzaniga, and Gianfranco Altomare. Isotretinoin plus clindamycin seem highly effective against severe erlotinib-induced skin rash in advanced non-small cell lung cancer. *Journal of Thoracic Oncology*, 5(10):1662–1663, 2010.

[236] Léa Payen, Laurence Delugin, Arnaud Courtois, Yolande Trinquart, André Guillouzo, and Olivier Fardel. The sulphonylurea glibenclamide inhibits multidrug resistance protein (MRP1) activity in human lung cancer cells. *British Journal of Pharmacology*, 132(3):778–784, 2001.

[237] Weiguo Zhao, Pengtao Bao, Haowen Qi, and Houcheng You. Resveratrol down-regulates survivin and induces apoptosis in human multidrug-resistant SPC-A-1/CDDP cells. *Oncology Reports*, 23(1):279–286, 2010.

[238] Peipei Nie, Weicheng Hu, Tao Zhang, Yijiu Yang, Benxin Hou, and Zhengzhi Zou. Synergistic induction of erlotinib-mediated apoptosis by resveratrol in human non-small-cell lung cancer cells by down-regulating survivin and up-regulating PUMA. *Cellular Physiology and Biochemistry*, 35(6):2255–2271, 2015.

[239] Derk Jan A De Groot, Margaretha Van Der Deen, Trong Khoan P Le, Anouk Regeling, Steven De Jong, and Elisabeth G E De Vries. Indomethacin induces apoptosis via a MRP1-dependent mechanism in doxorubicin-resistant small-cell lung cancer cells overexpressing MRP1. *British Journal of Cancer*, 97(8):1077–1083, 2007.

[240] Minyoung Lim, Maya Otto-Duessel, Miaoling He, Leila Su, Dan Nguyen, Emily Chin, Tamara Alliston, and Jeremy O Jones. Ligand-independent and tissue-selective androgen receptor inhibition by pyrvinium. *ACS Chemical Biology*, 9(3):692–702, 2014.

[241] Omar Alqawi, Mohammad Javad Moghaddas, and Gurmit Singh. Effects of geldanamycin on HIF-1 α mediated angiogenesis and invasion in prostate cancer cells. *Prostate Cancer and Prostatic Diseases*, 9(2):126–135, 2006.

[242] Hong-Chiang Chang, Chorng-Chih Huang, Chun-Jen Huang, Jin-Shiung Cheng, Shiuh-In Liu, Jeng-Yu Tsai, Hong-Tai Chang, Jong-Khing Huang, Chiang-Ting Chou, and Chung-Ren Jan. Desipramine-induced apoptosis in human PC3 prostate cancer cells: activation of JNK kinase and caspase-3 pathways and a protective role of [Ca2+] i elevation. *Toxicology*, 250(1):9–14, 2008.

[243] Ahmet Imrali, Xueying Mao, Marc Yeste-Velasco, Jonathan Shamash, and Yongjie Lu. Rapamycin inhibits prostate cancer cell growth through cyclin D1 and enhances the cytotoxic efficacy of cisplatin. *American Journal of Cancer Research*, 6(8):1772, 2016.

[244] Ram V Roy, Suman Suman, Trinath P Das, Joe E Luevano, and Chendil Damodaran. Withaferin A, a steroidal lactone from withania somnifera, induces mitotic catastrophe and growth arrest in prostate cancer cells. *Journal of Natural Products*, 76(10):1909–1915, 2013.

[245] Jacques Gilloteaux, James M Jamison, Deborah Neal, and Jack L Summers. Synergistic antitumor cytotoxic actions of ascorbate and menadione on human prostate (DU145) cancer cells in vitro: nucleus and other injuries preceding cell death by autoschizis. *Ultrastructural Pathology*, 38(2):116–140, 2014.

[246] Vibha Singh, Praveen Kumar Jaiswal, Ishita Ghosh, Hari K Koul, Xiuping Yu, and Arrigo De Benedetti. Targeting the TLK1/NEK1 DDR axis with Thioridazine suppresses outgrowth of androgen independent prostate tumors. *International Journal of Cancer*, 145(4):1055–1067, 2019.

[247] Yang Meng, Wenhua Tang, Yao Dai, Xiaoqing Wu, Meilan Liu, Qing Ji, Min Ji, Kenneth Pienta, Theodore Lawrence, and Liang Xu. Natural BH3 mimetic (-)-gossypol chemosensitizes human prostate cancer via Bcl-xl inhibition accompanied by increase of Puma and Noxa. *Molecular Cancer Therapeutics*, 7(7):2192–2202, 2008.

[248] Dimitrios Zgouras, Ute Becker, Stefan Loitsch, and Jürgen Stein. Modulation of angiogenesis-related protein synthesis by valproic acid. *Biochemical and Biophysical Research Communications*, 316(3):693–697, 2004.

[249] Caroline Habold, Angela Poehlmann, Khuloud Bajbouj, Roland Hartig, Kemal Sami Korkmaz, Albert Roessner, and Regine Schneider-Stock. Trichostatin A causes p53 to switch oxidative-damaged colorectal cancer cells from cell cycle arrest into apoptosis. *Journal of Cellular and Molecular Medicine*, 12(2):607–621, 2008.

[250] Qing Ji, Xuan Liu, Xiaoling Fu, Long Zhang, Hua Sui, Lihong Zhou, Jian Sun, Jianfeng Cai, Jianmin Qin, Jianlin Ren, et al. Resveratrol inhibits invasion and metastasis of colorectal cancer cells via MALAT1 mediated Wnt/β-catenin signal pathway. *PLoS One*, 8(11):e78700, 2013.

[251] Jin Young Park, Pei Yin Lin, and Robert H Weiss. Targeting the PI3K–Akt pathway in kidney cancer. *Expert Review of Anticancer Therapy*, 7(6):863–870, 2007.

[252] Amir A Momtazi-borojeni, Elham Abdollahi, Faezeh Ghasemi, Michele Caraglia, and Amirhossein Sahebkar. The novel role of pyrvinium in cancer therapy. *Journal of Cellular Physiology*, 233(4):2871–2881, 2018.

[253] Jon Jones, Eva Juengel, Ausra Mickuckyte, Lukasz Hudak, Steffen Wedel, Dietger Jonas, and Roman A Blaheta. The histone deacetylase inhibitor valproic acid alters growth properties of renal cell carcinoma in vitro and in vivo. *Journal of Cellular and Molecular Medicine*, 13(8b):2376–2385, 2009.

[254] Hong Zhao, Guangxu Jin, Kemi Cui, Ding Ren, Timothy Liu, Peikai Chen, Solomon Wong, Fuhai Li, Yubo Fan, Angel Rodriguez, et al. Novel modeling of cancer cell signaling pathways enables systematic drug repositioning for distinct breast cancer metastases. *Cancer research*, 73(20):6149–6163, 2013.

[255] Sze Kiat Tan, Anna Jermakowicz, Adnan K Mookhtiar, Charles B Nemeroff, Stephan C Schürer, and Nagi G Ayad. Drug repositioning in glioblastoma: A pathway perspective. *Frontiers in Pharmacology*, 9:218, 2018.

[256] Ping Xuan, Yangkun Cao, Tiangang Zhang, Xiao Wang, Shuxiang Pan, and Tonghui Shen. Drug repositioning through integration of prior knowledge and projections of drugs and diseases. *Bioinformatics*, 35(20):4108–4119, 2019.

[257] Chao-Kun Yan, Wen-Xiu Wang, Ge Zhang, Jian-Lin Wang, and Ashutosh Patel. BiRWDDA: a novel drug repositioning method based on multisimilarity fusion. *Journal of Computational Biology*, 26(11):1230–1242, 2019.

[258] Francesco Iorio, Julio Saez-Rodriguez, and Diego Di Bernardo. Network based elucidation of drug response: from modulators to targets. *BMC Systems Biology*, 7(1):1–9, 2013.

[259] Xu Zhou, Enyu Dai, Qian Song, Xueyan Ma, Qianqian Meng, Yongshuai Jiang, and Wei Jiang. In silico drug repositioning based on drug-miRNA associations. *Briefings in Bioinformatics*, 21(2):498–510, 2020.

[260] Christopher C Yang and Mengnan Zhao. Mining heterogeneous network for drug repositioning using phenotypic information extracted from social media and pharmaceutical databases. *Artificial Intelligence in Medicine*, 96:80–92, 2019.

[261] Henry Haeberle, Joel T Dudley, Jonathan T Liu, Atul J Butte, and Christopher H Contag. Identification of cell surface targets through meta-analysis of microarray data. *Neoplasia*, 14(7):666–669, 2012.

[262] Philippe Sanseau, Pankaj Agarwal, Michael R Barnes, Tomi Pastinen, J Brent Richards, Lon R Cardon, and Vincent Mooser. Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, 30(4):317, 2012.

[263] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *The Two-Sample Dispersion Problem and Other Two-Sample Problems.* Wiley Online Library, 2015.

[264] Shu-Dong Zhang and Timothy W Gant. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics*, 9(1):258, 2008.

[265] Junjiang Fu, Li Qin, Tao He, Jun Qin, Jun Hong, Jiemin Wong, Lan Liao, and Jianming Xu. The TWIST/Mi2/NuRD protein complex and its essential role in cancer metastasis. *Cell Research*, 21(2):275, 2011.

[266] David M Sabatini. mTOR and cancer: insights into a complex relationship. *Nature Reviews Cancer*, 6(9):729–734, 2006.

[267] Mark DM Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2):106–114, 2015.

[268] Xiaokai Li, Teresa Colvin, Jennifer N Rauch, Diego Acosta-Alvear, Martin Kampmann, Bryan Dunyak, Byron Hann, Blake T Aftab, Megan Murnane, Min Cho, et al. Validation of the Hsp70–Bag3 protein–protein interaction as a potential therapeutic target in cancer. *Molecular Cancer Therapeutics*, 14(3):642–648, 2015.

[269] Bolin Chen, Jinhong Shi, Shenggui Zhang, and Fang-Xiang Wu. Identifying protein complexes in protein–protein interaction networks by using clique seeds and graph entropy. *Proteomics*, 13(2):269–277, 2013.

[270] Xiujuan Lei, Fei Wang, Fang-Xiang Wu, Aidong Zhang, and Witold Pedrycz. Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. *Information Sciences*, 329:303–316, 2016.

[271] Chao Wu, Jun Zhu, and Xuegong Zhang. Integrating gene expression and protein-protein interaction network to prioritize cancer-associated genes. *BMC Bioinformatics*, 13(1):1–10, 2012.

[272] Andrei A Ivanov, Fadlo R Khuri, and Haian Fu. Targeting protein–protein interactions as an anticancer strategy. *Trends in Pharmacological Sciences*, 34(7):393–400, 2013.

[273] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):1–27, 2003.

[274] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2017.

[275] Palwinder Singh and Atul Bhardwaj. Mechanism of action of key enzymes associated with cancer propagation and their inhibition by various chemotherapeutic agents. *Mini Reviews in Medicinal Chemistry*, 8(4):388–398, 2008.

[276] P. E. Lønning and S. Kvinnsland. Mechanisms of action of aminoglutethimide as endocrine therapy of breast cancer. *Drugs*, 35(6):685–710, 1988.

[277] Hui Wang, Ying Wang, Elizabeth R Rayburn, Donald L Hill, John J Rinehart, and Ruiwen Zhang. Dexamethasone as a chemosensitizer for breast cancer chemotherapy: potentiation of the antitumor activity of adriamycin, modulation of cytokine expression, and pharmacokinetics. *International Journal of Oncology*, 30(4):947–953, 2007.

[278] Mylène Honorat, Aurélia Mesnier, Attilio Di Pietro, Valérie Lin, Pascale Cohen, Charles Dumontet, and Léa Payen. Dexamethasone down-regulates ABCG2 expression levels in breast cancer cells. *Biochemical and Biophysical Research Communications*, 375(3):308–314, 2008.

[279] Stina Garvin, Karin Öllinger, and Charlotta Dabrosin. Resveratrol induces apoptosis and inhibits angiogenesis in human breast cancer xenografts in vivo. *Cancer Letters*, 231(1):113–122, 2006.

[280] Pamela N Munster, Kenneth Ted Thurn, Scott Thomas, Paromita Raha, Mensura Lacevic, Aidan Miller, Michele Melisko, Roohi Ismail-Khan, Hope Rugo, Mark Moasser, et al. A phase ii study of the histone deacetylase inhibitor vorinostat combined with tamoxifen for the treatment of patients with hormone therapy-resistant breast cancer. *British Journal of Cancer*, 104(12):1828, 2011.

[281] Renae D Van Wyhe, Omar M Rahal, and Wendy A Woodward. Effect of statins on breast cancer recurrence and mortality: a review. *Breast Cancer: Targets and Therapy*, 9:559, 2017.

[282] Di Chen, Qiuzhi Cui, Huanjie Yang, and Q Ping Dou. Disulfiram, a clinically used anti-alcoholism drug and copper-binding agent, induces apoptotic cell death in breast cancer cultures and xenografts via inhibition of the proteasome activity. *Cancer Research*, 66(21):10425–10433, 2006.

[283] Xiaoya Wang, Sanhua Wei, Yong Zhao, Changhong Shi, Peijuan Liu, Caiqin Zhang, Yingfeng Lei, Bo Zhang, Bing Bai, Yong Huang, et al. Anti-proliferation of breast cancer cells with itraconazole: Hedgehog pathway inhibition induces apoptosis and autophagic cell death. *Cancer Letters*, 385:128–136, 2017.

[284] Ashish Juvekar, Laura N Burga, Hai Hu, Elaine P Lunsford, Yasir H Ibrahim, Judith Balmañà, Anbazhagan Rajendran, Antonella Papa, Katherine Spencer, Costas A Lyssiotis, et al. Combining a PI3K inhibitor with a PARP inhibitor provides an effective therapy for BRCA1-related breast cancer. *Cancer Discovery*, 2(11):1048–1063, 2012.

[285] Katarzyna Winnicka, Krzysztof Bielawski, Anna Bielawska, and Wojciech Miltyk. Apoptosis-mediated cytotoxicity of ouabain, digoxin and proscillaridin A in the estrogen independent MDA-MB-231 breast cancer cells. *Archives of Pharmacal Research*, 30(10):1216–1224, 2007.

[286] Hua Fan-Minogue, Sandhya Bodapati, David Solow-Cordero, Alice Fan, Ramasamy Paulmurugan, Tarik F Massoud, Dean W Felsher, and Sanjiv S Gambhir. A c-Myc activation sensor-based high-throughput drug screening identifies an antineoplastic effect of nitazoxanide. *Molecular Cancer Therapeutics*, 12(9):1896–1905, 2013.

[287] Bal L Lokeshwar, Marie G Selzer, Bao-Qian Zhu, Norman L Block, and Lorne M Golub. Inhibition of cell proliferation, invasion, tumor growth and metastasis by an oral non-antimicrobial tetracycline analog (COL-3) in a metastatic prostate cancer model. *International Journal of Cancer*, 98(2):297–309, 2002.

[288] Toshinao Onoda, Takashi Ono, Dipok Kumar Dhar, Akira Yamanoi, and Naofumi Nagasue. Tetracycline analogues (doxycycline and COL-3) induce caspase-dependent and-independent apoptosis in human colon cancer cells. *International Journal of Cancer*, 118(5):1309–1315, 2006.

[289] Yoh Watanabe, Hiroshi Hoshiai, Toru Nakanishi, Naoki Kawamura, Naotake Tanaka, Keiichi Isaka, Shoji Kamiura, Masahide Ohmichi, Masayuki Hatae, and Kazunori Ochiai. Evaluation of oral etoposide in combination with cisplatin for patients with recurrent cervical cancer: long-term follow-up results of a japanese multicenter study. *Anticancer Research*, 31(9):3063–3067, 2011.

[290] Su-Hyeon Kim, Su-Hyeong Kim, Yong-Beom Kim, Yong-Tark Jeon, Sang-Chul Lee, and Yong-Sang Song. Genistein inhibits cell growth by modulating various mitogen-activated protein kinases and AKT in cervical cancer cells. *Annals of the New York Academy of Sciences*, 1171(1):495–500, 2009.

[291] Liping Chen, Li Wang, Haibin Shen, Hui Lin, and Dan Li. Anthelminthic drug niclosamide sensitizes the responsiveness of cervical cancer cells to paclitaxel via oxidative stress-mediated mTOR inhibition. *Biochemical and Biophysical Research Communications*, 484(2):416–421, 2017.

[292] Sokbom Kang, Seung Myung Dong, Boh-Ram Kim, Mi Sun Park, Barry Trink, Hyun-Jung Byun, and Seung Bae Rho. Thioridazine induces apoptosis by targeting the PI3K/Akt/mTOR pathway in cervical and endometrial cancer cells. *Apoptosis*, 17(9):989–997, 2012.

[293] Commonly used drugs in treatment of prostate cancer, 2017. https://www.prostate.org.au/awareness/further-detailed-information/commonly-used-drugs-in-treatment-of-prostate-cancer/mitoxantrone/.

[294] Ho Lin, Jyn-Lyh Juang, and Paulus S Wang. Involvement of Cdk5/p25 in digoxin-triggered prostate cancer cell apoptosis. *Journal of Biological Chemistry*, 279(28):29302–29307, 2004.

[295] Yi-Long Wu, Caicun Zhou, Cheng-Ping Hu, Jifeng Feng, Shun Lu, Yunchao Huang, Wei Li, Mei Hou, Jian Hua Shi, Kye Young Lee, et al. Afatinib versus cisplatin plus gemcitabine for first-line treatment of asian patients with advanced non-small-cell lung cancer harbouring EGFR mutations (LUX-Lung 6): an open-label, randomised phase 3 trial. *The Lancet Oncology*, 15(2):213–222, 2014.

[296] Eduarda Schultze, Aline Ourique, Virginia Campello Yurgel, Karine Rech Begnini, Helena Thurow, Priscila Marques Moura de Leon, Vinicius Farias Campos, Odir Antônio Dellagostin, Silvia R Guterres, Adriana R Pohlmann, et al. Encapsulation in lipid-core nanocapsules overcomes lung cancer cell resistance to tretinoin. *European Journal of Pharmaceutics and Biopharmaceutics*, 87(1):55–63, 2014.

[297] Bandaru S Reddy, Chung Xiou Wang, Ah-Ng Kong, Tin Oo Khor, Xi Zheng, Vernon E Steele, Levy Kopelovich, and Chinthalapally V Rao. Prevention of azoxymethane-induced colon cancer by combination of low doses of atorvastatin, aspirin, and celecoxib in f 344 rats. *Cancer Research*, 66(8):4542–4546, 2006.

[298] Todd M Pitts, Mark Morrow, Sara A Kaufman, John J Tentler, and S Gail Eckhardt. Vorinostat and bortezomib exert synergistic antiproliferative and proapoptotic effects in colon cancer cell models. *Molecular Cancer Therapeutics*, 8(2):342–349, 2009.

[299] Najme Faham and Alana L Welm. RON signaling is a key mediator of tumor progression in many human cancers. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume 81, pages 177–188. Cold Spring Harbor Laboratory Press, 2016.

[300] Sandra Van Schaeybroeck, Anthi Karaiskou-McCaul, Donal Kelly, Daniel Longley, Leeona Galligan, Eric Van Cutsem, and Patrick Johnston. Epidermal growth factor receptor activity determines response of colorectal cancer cells to gefitinib alone and in combination with chemotherapy. *Clinical Cancer Research*, 11(20):7480–7489, 2005.

[301] Giuseppe Cornelia, Rossana Casaretti, Pasquale Cornelia, Antonio Daponte, Alberto Parziale, Vincenzo Iervolino, Gaetano Santillo, and Donato Zarrilli. Treatment of advanced colorectal cancer with mitoxantrone, high dose folinic acid and fluorouracil. *Tumori Journal*, 77(5):445–446, 1991.

[302] Chen Wang, Nicholas Jette, Daniel Moussienko, D Gwyn Bebb, and Susan P Lees-Miller. ATM-deficient colorectal cancer cells are sensitive to the PARP inhibitor olaparib. *Translational Oncology*, 10(2):190–196, 2017.

[303] John J Arcaroli, Basel M Touban, Aik Choon Tan, Marileila Varella-Garcia, Rebecca W Powell, S Gail Eckhardt, Paul Elvin, Dexiang Gao, and Wells A Messersmith. Gene array and fluorescence in situ hybridization biomarkers of activity of saracatinib (AZD0530), a Src inhibitor, in a preclinical model of colorectal cancer. *Clinical Cancer Research*, 16(16):4165–4177, 2010.

[304] Pei-Ming Yang, Yi-Ting Lin, Chia-Tung Shun, Shan-Hu Lin, Tzu-Tang Wei, Shu-Hui Chuang, Ming-Shiang Wu, and Ching-Chow Chen. Zebularine inhibits tumorigenesis and stemness of colorectal cancer via p53-dependent endoplasmic reticulum stress. *Scientific Reports*, 3:3219, 2013.

[305] Sharad Sharma, Hang-Ping Yao, Yong-Qing Zhou, Jianwei Zhou, Ruiwen Zhang, and Ming Hai Wang. Prevention of BMS-777607-induced polyploidy/senescence by mTOR inhibitor AZD8055 sensitizes breast cancer cells to cytotoxic chemotherapeutics. *Molecular Oncology*, 8(3):469–482, 2014.

[306] Srikala S Sridhar, Mary J Mackenzie, Sebastien J Hotte, Som D Mukherjee, Ian F Tannock, Nevin Murray, Christian Kollmannsberger, Masoom A Haider, Eric X Chen, Robert Halford, et al. A phase II study of cediranib (AZD 2171) in treatment naive patients with progressive unresectable recurrent or metastatic renal cell carcinoma. A trial of the PMH phase 2 consortium. *Investigational New Drugs*, 31(4):1008–1015, 2013.

[307] Kazuki Okubo, Makoto Isono, Takako Asano, and Akinori Sato. Panobinostat and nelfinavir inhibit renal cancer growth by inducing endoplasmic reticulum stress. *Anticancer Research*, 38(10):5615–5626, 2018.

[308] Bernard Escudier, Camillo Porta, Tim Eisen, Jonathan Belsey, Damilola Gibson, Jonathan Morgan, and Robert Motzer. The role of tivozanib in advanced renal cell carcinoma therapy. *Expert review of anticancer therapy*, 18(11):1113–1124, 2018.

[309] H. J. Hammers, H. Verheul, B. Wilky, B. Salumbides, J. Holleran, M. J. Egorin, M. Lodge, R. L. Wahl, J. A. Zwiebel, M. A. Carducci, et al. Phase I safety and pharmacokinetic/pharmacodynamic results of the histone deacetylase inhibitor vorinostat in combination with bevacizumab in patients with kidney cancer. *Journal of Clinical Oncology*, 26(15_suppl):16094–16094, 2008.

[310] Christine Y Shiang, Yuan Qi, Bailiang Wang, Vladimir Lazar, Jing Wang, Fraser Symmans, Gabriel N Hortobagyi, Fabrice Andre, and Lajos Pusztai. Amplification of fibroblast growth factor receptor-1 in breast cancer and the effects of brivanib alaninate. *Breast Cancer Research and Treatment*, 123(3):747–755, 2010.

[311] Paul E Goss, James N Ingle, José E Alés-Martínez, Angela M Cheung, Rowan T Chlebowski, Jean Wactawski-Wende, Anne McTiernan, John Robbins, Karen C Johnson, Lisa W Martin, et al. Exemestane for breast-cancer prevention in postmenopausal women. *New England Journal of Medicine*, 364(25):2381–2391, 2011.

[312] Hongyang Liang and Antoinette R Tan. Iniparib, a PARP1 inhibitor for the potential treatment of cancer, including triple-negative breast cancer. *IDrugs: the Investigational Drugs Journal*, 13(9):646–656, 2010.

[313] Karthic Chandran, Sudeshna Goswami, and Neelam Sharma-Walia. Implications of a peroxisome proliferator-activated receptor alpha (PPARα) ligand clofibrate in breast cancer. *Oncotarget*, 7(13):15577–15599, 2016.

[314] Chlorambucil, 2021. https://www.drugbank.ca/drugs/DB00291.

[315] Stewart H Jones. Nitrogen mustard with corticosteroid and chlortetracycline for far-advanced metastatic cancer. *Postgraduate Medicine*, 21(5):520–525, 1957.

[316] Carmen S Tellez, Marcie J Grimes, Maria A Picchi, Yushi Liu, Thomas H March, Matthew D Reed, Aram Oganesian, Pietro Taverna, and Steven A Belinsky. SGI-110 and entinostat therapy reduces lung tumor burden and reprograms the epigenome. *International Journal of Cancer*, 135(9):2223–2231, 2014.

[317] Hiroko Endo, Momoko Yano, Yuushi Okumura, and Hiroshi Kido. Ibuprofen enhances the anticancer activity of cisplatin in lung cancer cells by inhibiting the heat shock protein 70. *Cell Death & Disease*, 5(1):e1027, 2014.

[318] Nithya Ramnath, Stephanie Daignault-Newton, Grace K Dy, Josephia Muindi, Araba Adjei, Gregory Peter Kalemkerian, Kemp Bailey Cease, Philip J Stella, Dean E Brenner, Candace S Johnson, et al. A phase I/II clinical trial of intravenous (IV) calcitriol with fixed dose of cisplatin and docetaxel in advanced non-small cell lung cancer. *Ann Arbor*, 1001:48109–5848, 2012.

[319] Daisuke Minami, Nagio Takigawa, Hiromasa Takeda, Minoru Takata, Nobuaki Ochi, Eiki Ichihara, Akiko Hisamoto, Katsuyuki Hotta, Mitsune Tanimoto, and Katsuyuki Kiura. Synergistic effect of olaparib with combination of cisplatin on PTEN-deficient lung cancer cells. *Molecular Cancer Research*, 11(2):140–148, 2013.

[320] Danièle Taras, Jean-Frédéric Blanc, Anne Rullier, Nathalie Dugot-Senant, Ingrid Laurendeau, Michel Vidaud, and Jean Rosenbaum. Pravastatin reduces lung metastasis of rat hepatocellular carcinoma via a coordinated decrease of MMP expression and activity. *Journal of Hepatology*, 46(1):69–76, 2007.

[321] Guo Chen, B. Zhang, Haidong Xu, Youwei Sun, Y. Shi, Y. Luo, H. Jia, and Fu Wang. Suppression of Sirt1 sensitizes lung cancer cells to WEE1 inhibitor MK-1775-induced DNA damage and apoptosis. *Oncogene*, 36(50):6863, 2017.

[322] Tacedinaline, 2021. https://www.drugbank.ca/drugs/DB12291.

[323] Natasha B Leighl, Ming-Sound Tsao, Geoffrey Liu, Dongsheng Tu, Cheryl Ho, Frances A Shepherd, Nevin Murray, John R Goffin, Garth Nicholas, Shingo Sakashita, et al. A phase I study of foretinib plus erlotinib in patients with previously treated advanced non-small cell lung cancer: Canadian cancer trials group IND. 196. *Oncotarget*, 8(41):69651, 2017.

[324] Meredith A Tennis, Michelle Van Scoyk, Lynn E Heasley, Katherine Vandervest, Mary Weiser-Evans, Scott Freeman, Robert L Keith, Pete Simpson, Raphael A Nemenoff, and Robert A Winn. Prostacyclin inhibits non-small cell lung cancer growth by a frizzled 9-dependent pathway that is blocked by secreted frizzled-related protein 1. *Neoplasia*, 12(3):244–IN6, 2010.

[325] Mingyue Li, Tak W Lee, Anthony P Yim, Tony S Mok, and George G Chen. Apoptosis induced by troglitazone is both peroxisome proliferator-activated receptor-γ-and ERK-dependent in human non-small lung cancer cells. *Journal of Cellular Physiology*, 209(2):428–438, 2006.

[326] Leo R Zacharski, Paolo Prandoni, and Manuel Monreal. Warfarin versus low-molecular-weight heparin therapy in cancer patients. *The Oncologist*, 10(1):72–79, 2005.

[327] Joong Sup Shim and Jun O Liu. Recent advances in drug repositioning for the discovery of new anticancer drugs. *International Journal of Biological Sciences*, 10(7):654–663, 2014.

[328] Ted T Ashburn and Karl B Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004.

[329] Mathias Rask-Andersen, Markus Sällman Almén, and Helgi B Schiöth. Trends in the exploitation of novel drug targets. *Nature Reviews Drug Discovery*, 10(8):579–590, 2011.

[330] Rammohan Shukla, Nicholas D Henkel, Khaled Alganem, Abdul-rizaq Hamoud, James Reigle, Rawan S Alnafisah, Hunter M Eby, Ali S Imami, Justin F Creeden, Scott A Miruzzi, et al. Signature-based approaches for informed drug repurposing: Targeting CNS disorders. *Neuropsychopharmacology*, 46(1):116–130, 2021.

[331] Sipko Van Dam, Urmo Vosa, Adriaan van der Graaf, Lude Franke, and Joao Pedro de Magalhaes. Gene co-expression analysis for functional classification and gene–disease predictions. *Briefings in Bioinformatics*, 19(4):575–592, 2018.

[332] Victor Spirin and Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proceedings of the National Academy of Sciences*, 100(21):12123–12128, 2003.

[333] Min Li, Weijie Chen, Jianxin Wang, Fang-Xiang Wu, and Yi Pan. Identifying dynamic protein complexes based on gene expression profiles and PPI networks. *BioMed Research International*, 2014:1–10, 2014.

[334] Xiujuan Lei, Huan Li, Aidong Zhang, and Fang-Xiang Wu. iOPTICS-GSO for identifying protein complexes from dynamic PPI networks. *BMC Medical Genomics*, 10(5):55–66, 2017.

[335] Qianghua Xiao, Ping Luo, Min Li, Jianxin Wang, and Fang-Xiang Wu. A novel core-attachment–based method to identify dynamic protein complexes based on gene expression profiles and PPI networks. *Proteomics*, 19(5):1–7, 2019.

[336] Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes—2019. *Nucleic Acids Research*, 47(D1):D559–D563, 2019.

[337] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34(suppl_1):D668–D672, 2006.

[338] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 68(6):394–424, 2018.

[339] Francesco Iorio, Roberta Bosotti, Emanuela Scacheri, Vincenzo Belcastro, Pratibha Mithbaokar, Rosa Ferriero, Loredana Murino, Roberto Tagliaferri, Nicola Brunetti-Pierri, Antonella Isacchi, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences*, 107(33):14621–14626, 2010.

[340] Francesco Iorio, Roshan L Shrestha, Nicolas Levin, Viviane Boilot, Mathew J Garnett, Julio Saez-Rodriguez, and Viji M Draviam. A semi-supervised approach for refining transcriptional signatures of drug response and repositioning predictions. *PLoS One*, 10(10):1–21, 2015.

[341] Hongyu Wu, Jinjiang Huang, Yang Zhong, and Qingshan Huang. DrugSig: A resource for computational drug repositioning utilizing gene expression signatures. *PLoS One*, 12(5):1–11, 2017.

[342] In-Wha Kim, Hayoung Jang, Jae Hyun Kim, Myeong Gyu Kim, Sangsoo Kim, and Jung Mi Oh. Computational drug repositioning for gastric cancer using reversal gene expression profiles. *Scientific Reports*, 9(1):1–10, 2019.

[343] Theresa A Reno, Jae Y Kim, and Dan J Raz. Triptolide inhibits lung cancer cell migration, invasion, and metastasis. *The Annals of Thoracic Surgery*, 100(5):1817–1825, 2015.

[344] Elizabeth C Halvorsen, Melisa J Hamilton, Ada Young, Brennan J Wadsworth, Nancy E LePard, Ha-Na Lee, Natalie Firmino, Jenna L Collier, and Kevin L Bennewith. Maraviroc decreases CCL8-mediated migration of CCR5+ regulatory t cells and reduces metastatic tumor growth in the lungs. *Oncoimmunology*, 5(6):1–15, 2016.

[345] Jianya Zhou, Shumeng Zhang, Xi Chen, Xianan Zheng, Yinan Yao, Guohua Lu, and Jianying Zhou. Palbociclib, a selective CDK4/6 inhibitor, enhances the effect of selumetinib in RAS-driven non-small cell lung cancer. *Cancer Letters*, 408:130–137, 2017.

[346] Alice T Shaw, Dong-Wan Kim, Kazuhiko Nakagawa, Takashi Seto, Lucio Crinó, Myung-Ju Ahn, Tommaso De Pas, Benjamin Besse, Benjamin J Solomon, Fiona Blackhall, et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *New England Journal of Medicine*, 368(25):2385–2394, 2013.

[347] Prithviraj Bose and Howard Ozer. Neratinib: an oral, irreversible dual EGFR/HER2 inhibitor for breast and non-small cell lung cancer. *Expert Opinion on Investigational Drugs*, 18(11):1735–1751, 2009.

[348] Jinhui Shao and Guihua Feng. Selective killing effect of oxytetracycline, propafenone and metamizole on A549 or hela cells. *Chinese Journal of Cancer Research*, 25(6):662–670, 2013.

[349] Astrid Nehlig, Jean-Luc Daval, and Gérard Debry. Caffeine and the central nervous system: mechanisms of action, biochemical, metabolic and psychostimulant effects. *Brain Research Reviews*, 17(2):139–170, 1992.

[350] Gan Wang, Vanitha Bhoopalan, David Wang, Le Wang, and Xiaoxin Xu. The effect of caffeine on cisplatin-induced apoptosis of lung cancer cells. *Experimental Hematology & Oncology*, 4(1):1–9, 2015.

[351] Swei Sunny Hann, Qing Tang, Fang Zheng, Shunyu Zhao, Jianping Chen, and ZhiYu Wang. Repression of phosphoinositide-dependent protein kinase 1 expression by ciglitazone via Egr-1 represents a new approach for inhibition of lung cancer cell growth. *Molecular Cancer*, 13(1):1–13, 2014.

[352] Gregory P Kalemkerian, Rodney Slusher, Sakkaraiappan Ramalingam, Shirish Gadgeel, and Mack Mabry. Growth inhibition and induction of apoptosis by fenretinide in small-cell lung cancer cell lines. *JNCI: Journal of the National Cancer Institute*, 87(22):1674–1680, 1995.

[353] Apar Pataer, Dora Bocangel, Sunil Chada, Jack A Roth, Kelly K Hunt, and Stephen G Swisher. Enhancement of adenoviral MDA-7-mediated cell killing in human lung cancer cells by geldanamycin and its 17-allyl-amino-17-demethoxy analogue. *Cancer Gene Therapy*, 14(1):12–18, 2007.

[354] Nobuaki Shiraki, Akinobu Hamada, Takafumi Ohmura, Jin Tokunaga, Naoki Oyama, and Masahiro Nakano. Increase in doxorubicin cytotoxicity by inhibition of P-glycoprotein activity with lomerizine. *Biological and Pharmaceutical Bulletin*, 24(5):555–557, 2001.

[355] GSK-1059615, 2021. https://go.drugbank.com/drugs/DB11962.

[356] Francesco Serra, Pietro Lapidari, Erica Quaquarini, Barbara Tagliaferri, Federico Sottotetti, and Raffaella Palumbo. Palbociclib in metastatic breast cancer: current evidence and real-life data. *Drugs in Context*, 8, 2019.

[357] Ben O'Leary, Sarah Hrebien, James P Morden, Matthew Beaney, Charlotte Fribbens, Xin Huang, Yuan Liu, Cynthia Huang Bartlett, Maria Koehler, Massimo Cristofanilli, et al. Early circulating tumor DNA dynamics and clonal selection with palbociclib and fulvestrant for breast cancer. *Nature Communications*, 9(1):1–10, 2018.

[358] Xiao-He Yang, Todd L Sladek, Xuesong Liu, Bryn R Butler, Christopher J Froelich, and Ann D Thor. Reconstitution of caspase 3 sensitizes MCF-7 breast cancer cells to doxorubicin-and etoposide-induced apoptosis. *Cancer Research*, 61(1):348–354, 2001.

[359] Eduarda Schultze, Julieti Buss, Karine Coradini, Karine Rech Begnini, Silvia S Guterres, Tiago Collares, Ruy Carlos Ruver Beck, Adriana R Pohlmann, and Fabiana Kömmling Seixas. Tretinoin-loaded lipid-core nanocapsules overcome the triple-negative breast cancer cell resistance to tretinoin and show synergistic effect on cytotoxicity induced by doxorubicin and 5-fluororacil. *Biomedicine & Pharmacotherapy*, 96:404–409, 2017.

[360] Bingyang Chu, Shuai Shi, Xingyi Li, Lufeng Hu, Lu Shi, Haina Zhang, Qiaoqiao Xu, Lei Ye, Guanyang Lin, Nansheng Zhang, et al. Preparation and evaluation of teniposide-loaded polymeric micelles for breast cancer therapy. *International Journal of Pharmaceutics*, 513(1-2):118–129, 2016.

[361] Xiqian Han, Xiaobing Zhang, Hui Li, Shengshi Huang, Shu Zhang, Fengshan Wang, and Yikang Shi. Tunicamycin enhances the antitumor activity of trastuzumab on breast cancer in vitro and in vivo. *Oncotarget*, 6(36):38912–38925, 2015.

[362] Han Li, Guo-feng Pan, Zhen-zhou Jiang, Jing Yang, Li-xin Sun, and Lu-yong Zhang. Triptolide inhibits human breast cancer MCF-7 cell growth via downregulation of the ERα-mediated signaling pathway. *Acta Pharmacologica Sinica*, 36(5):606–613, 2015.

[363] Idarubicin, 2021. https://go.drugbank.com/drugs/DB01177.

[364] Ufuk Gunduz, Tugba Keskin, Gulistan Tansık, Pelin Mutlu, Serap Yalcın, Gozde Unsoy, Arzu Yakar, Rouhollah Khodadust, and Gungor Gunduz. Idarubicin-loaded folic acid conjugated magnetic nanoparticles as a targetable drug delivery system for breast cancer. *Biomedicine & Pharmacotherapy*, 68(6):729–736, 2014.

[365] Elena Laakmann, Isabell Witzel, and Volkmar Müller. Efficacy of liposomal cytarabine in the treatment of leptomeningeal metastasis of breast cancer. *Breast Care*, 12(3):165–167, 2017.

[366] Sadhna R Vora, Dejan Juric, Nayoon Kim, Mari Mino-Kenudson, Tiffany Huynh, Carlotta Costa, Elizabeth L Lockerman, Sarah F Pollack, Manway Liu, Xiaoyan Li, et al. CDK 4/6 inhibitors sensitize PIK3CA mutant breast cancer to PI3K inhibitors. *Cancer Cell*, 26(1):136–149, 2014.

[367] Renate Schmidt, Frank Baumann, Heike Knüpfer, Michael Brauckhoff, LC Horn, M Schönfelder, Uwe Köhler, and Rainer Preiss. CYP3A4, CYP2C9 and CYP2B6 expression and ifosfamide turnover in breast cancer tissue microsomes. *British Journal of Cancer*, 90(4):911–916, 2004.

[368] Channa Keshava, Erin C McCanlies, and Ainsley Weston. CYP3A4 polymorphisms—potential risk factors for breast and prostate cancer: a HuGE review. *American Journal of Epidemiology*, 160(9):825–841, 2004.

[369] Xiaodong Wang, Yuwen Diao, Yu Liu, Ningning Gao, Dong Gao, Yanyan Wan, Jingjing Zhong, and Guangyi Jin. Synergistic apoptosis-inducing effect of aspirin and isosorbide mononitrate on human colon cancer cells. *Molecular Medicine Reports*, 12(3):4750–4758, 2015.

[370] Lin Zhao, Peng Wu, Pinggui Zhang, Daze Xie, Dian Gao, and Nanjin Zhou. Effect of triptolide on human colorectal cancer HCT116 cell proliferation, autophagy and apoptosis. *Chinese Pharmacological Bulletin*, 32(10):1399–1403, 2016.

[371] Asim Pervaiz, Shariq Ansari, Martin R Berger, and Hassan Adwan. CCR5 blockage by maraviroc induces cytotoxic and apoptotic effects in colorectal cancer cells. *Medical Oncology*, 32(5):1–10, 2015.

[372] Jun Zhang, Lanlan Zhou, Shuai Zhao, David T Dicker, and Wafik S El-Deiry. The CDK4/6 inhibitor palbociclib synergizes with irinotecan to promote colorectal cancer cell death under hypoxia. *Cell Cycle*, 16(12):1193–1200, 2017.

[373] Brian M Wolpin, Kimmie Ng, Andrew X Zhu, Thomas Abrams, Peter C Enzinger, Nadine J McCleary, Deborah Schrag, Eunice L Kwak, Jill N Allen, Pankaj Bhargava, et al. Multicenter phase II study of tivozanib (AV-951) and everolimus (RAD001) for patients with refractory, metastatic colorectal cancer. *The Oncologist*, 18(4):377–378, 2013.

[374] Ryan B Corcoran, Chloe E Atreya, Gerald S Falchook, Eunice L Kwak, David P Ryan, Johanna C Bendell, Omid Hamid, Wells A Messersmith, Adil Daud, Razelle Kurzrock, et al. Combined BRAF and MEK inhibition with dabrafenib and trametinib in BRAF V600–mutant colorectal cancer. *Journal of Clinical Oncology*, 33(34):4023–4031, 2015.

[375] Erdem Bangi, Celina Ang, Peter Smibert, Andrew V Uzilov, Alexander G Teague, Yevgeniy Antipin, Rong Chen, Chana Hecht, Nelson Gruszczynski, Wesley J Yon, et al. A personalized platform identifies trametinib plus zoledronate for a patient with KRAS-mutant metastatic colorectal cancer. *Science Advances*, 5(5):eaav6528, 2019.

[376] Donghong Ju, Xiaogang Wang, and Youming Xie. Dyclonine and alverine citrate enhance the cytotoxic effects of proteasome inhibitor MG132 on breast cancer cells. *International Journal of Molecular Medicine*, 23(2):205–209, 2009.

[377] Palbociclib in patients with metastatic castration-resistant prostate cancer, 2021. https://clinicaltrials.gov/ct2/show/NCT02905318.

[378] Weiwei Huang, Tiantian He, Chengsen Chai, Yuan Yang, Yahong Zheng, Pei Zhou, Xiaoxia Qiao, Bin Zhang, Zengzhen Liu, Junru Wang, et al. Triptolide inhibits the proliferation of prostate cancer cells and down-regulates SUMO-specific protease 1 expression. *PLoS One*, 7(5):1–17, 2012.

[379] Daniela Sicoli, Xuanmao Jiao, Xiaoming Ju, Marco Velasco-Velazquez, Adam Ertel, Sankar Addya, Zhiping Li, Sebastiano Andò, Alessandro Fatatis, Bishnuhari Paudyal, et al. CCR5 receptor antagonists block metastasis to bone of v-Src oncogene–transformed metastatic prostate cancer cell lines. *Cancer Research*, 74(23):7103–7114, 2014.

[380] Takeo Nomura, Mutsushi Yamasaki, Yoshio Nomura, and Hiromitsu Mimata. Expression of the inhibitors of apoptosis proteins in cisplatin-resistant prostate cancer cells. *Oncology Reports*, 14(4):993–997, 2005.

[381] Shanta Dhar, Frank X Gu, Robert Langer, Omid C Farokhzad, and Stephen J Lippard. Targeted delivery of cisplatin to prostate cancer cells by aptamer functionalized Pt (IV) prodrug-PLGA–PEG nanoparticles. *Proceedings of the National Academy of Sciences*, 105(45):17356–17361, 2008.

[382] Wassim Abida, Akash Patnaik, David Campbell, Jeremy Shapiro, Alan H Bryce, Ray McDermott, Brieuc Sautois, Nicholas J Vogelzang, Richard M Bambury, Eric Voog, et al. Rucaparib in men with metastatic castration-resistant prostate cancer harboring a BRCA1 or BRCA2 gene alteration. *Journal of Clinical Oncology*, 38(32):3763–3772, 2020.

[383] Bruno Blaya, Francesca Nicolau-Galmés, Shawkat M Jangi, Idoia Ortega-Martínez, Erika Alonso-Tejerina, Juan Burgos-Bretones, Gorka Pérez-Yarza, Aintzane Asumendi, and María D Boyano. Histamine and histamine receptor antagonists in cancer biology. *Inflammation & Allergy-Drug Targets (Formerly Current Drug Targets-Inflammation & Allergy)(Discontinued)*, 9(3):146–157, 2010.

[384] Mary E Law, Patrick E Corsino, Satya Narayan, and Brian K Law. Cyclin-dependent kinase inhibitors as anticancer therapeutics. *Molecular Pharmacology*, 88(5):846–852, 2015.

[385] Kexin Huang, Cao Xiao, Lucas M Glass, Marinka Zitnik, and Jimeng Sun. SkipGNN: predicting molecular interactions with skip-graph networks. *Scientific Reports*, 10(1):1–16, 2020.

[386] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018.

[387] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Jolene Wiegers, Thomas C Wiegers, and Carolyn J Mattingly. Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143, 2021.

[388] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1):D545–D551, 2021.

[389] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The SIDER database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079, 2016.

[390] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[391] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11):2579–2605, 2008.

[392] Nicholas P Tatonetti, P Ye Patrick, Roxana Daneshjou, and Russ B Altman. Data-driven prediction of drug effects and interactions. *Science Translational Medicine*, 4(125):1–26, 2012.

[393] Drug interactions checker, 2021. https://www.drugs.com/drug_interactions.html.

Appendix A

List of Publications

Referred journal publications:

1 **Fei Wang**, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Predicting drug-drug interactions by graph convolutional network with multi-kernel, *Briefings in Bioinformatics*, 2021. DOI: 10.1093/bib/bbab511.

2 **Fei Wang**, Yulian Ding, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Machine learning and deep learning strategies in drug repositioning, *Current bioinformatics*, 2021, Accepted.

3 **Fei Wang**, Yulian Ding, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Human protein complex-based drug signatures for personalized cancer medicine. *IEEE Journal of Biomedical and Health Informatics*, 25(11): 4079-4088, 2021. DOI: 10.1109/JBHI.2021.3120933.

4 **Fei Wang**, Yulian Ding, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Identifying gene signatures for cancer drug repositioning based on sample clustering. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020, DOI: 10.1109/TCBB.2020.3019781.

5 **Fei Wang**, Xiujuan Lei, and Fang-Xiang Wu. A review of drug repositioning based chemical-induced cell line expression data. *Current medicinal chemistry*, 27(32): 5340-5350, 2020. DOI: 10.2174/09298673 25666181101115801.

6 Yulian Ding, **Fei Wang**, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Deep belief network-based matrix factorization model for microRNA-disease associations prediction. *Evolutionary Bioinformatics*, 2020. DOI: 10.1177/1176934320919707.

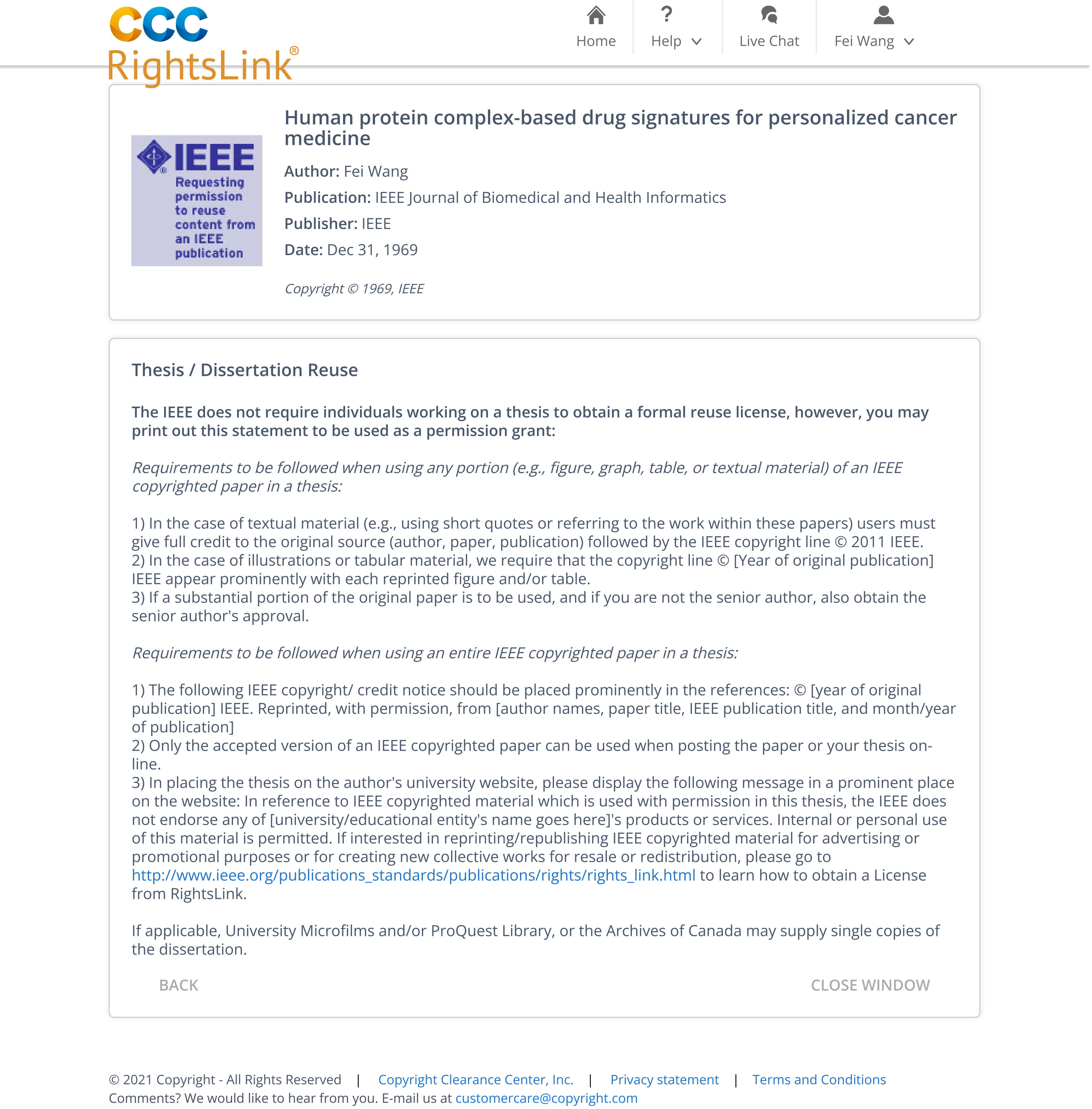
Referred conference publication:

1 **Fei Wang**, Xiujuan Lei, Bo Liao, Fang-Xiang Wu. Human protein complex signatures for drug repositioning. *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 42–50, 2019. DOI: 10.1145/3307339.3342132.

Appendix B

Copyright Permissions

Copyright forms of thesis-related publications are attached in the following pages.


CCC RightsLink®
Home
Help
Live Chat
Fei Wang
IEEE
Requesting permission to reuse content from an IEEE publication

Human protein complex-based drug signatures for personalized cancer medicine

Author: Fei Wang

Publication: IEEE Journal of Biomedical and Health Informatics

Publisher: IEEE

Date: Dec 31, 1969

Copyright © 1969, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

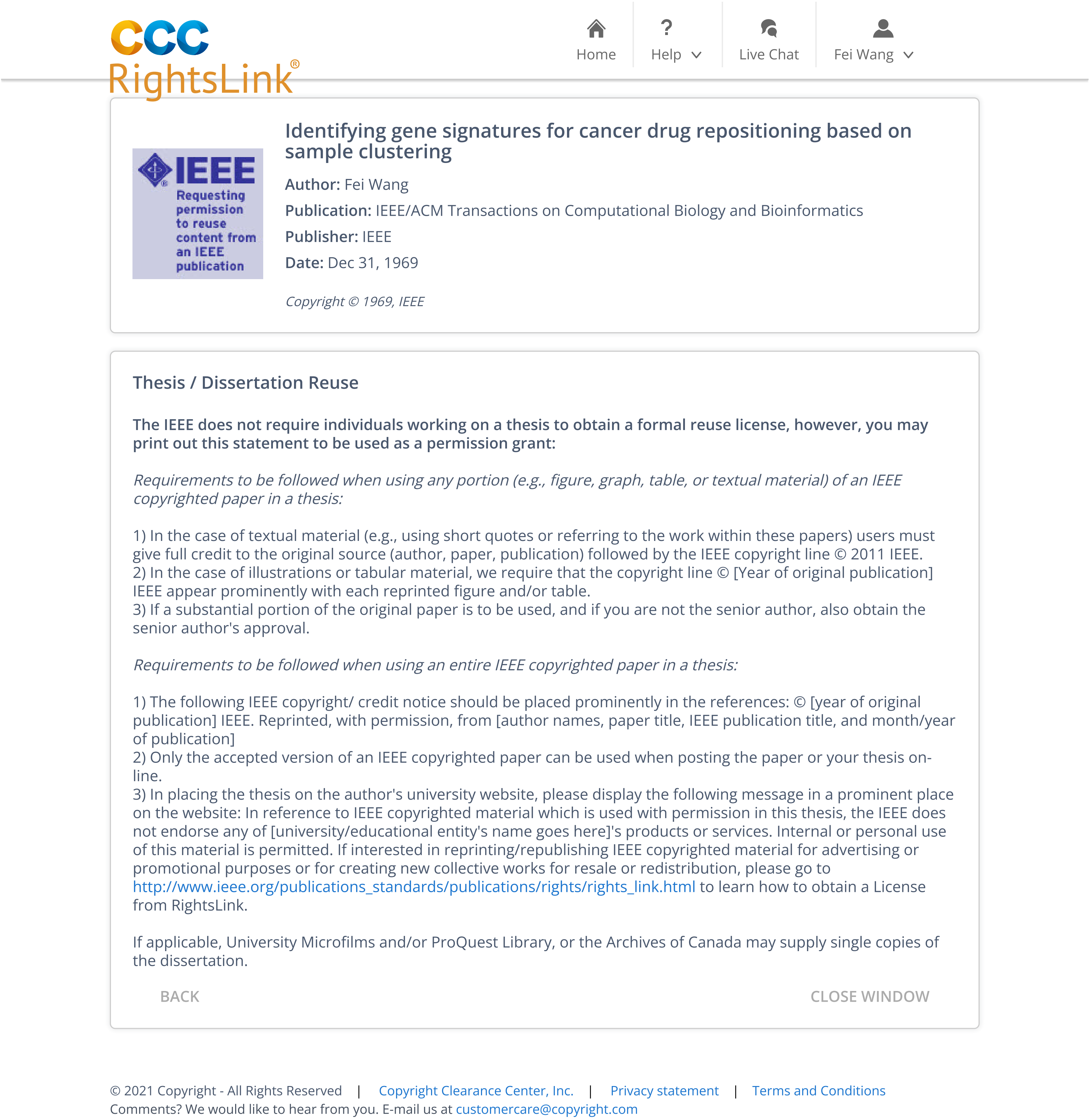
Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.



CCC
RightsLink®
Home
Help
Live Chat
Fei Wang

IEEE
Requesting permission to reuse content from an IEEE publication

Identifying gene signatures for cancer drug repositioning based on sample clustering

Author: Fei Wang

Publication: IEEE/ACM Transactions on Computational Biology and Bioinformatics

Publisher: IEEE

Date: Dec 31, 1969

Copyright © 1969, IEEE

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

CCC | Marketplace™

Order Number: 1157486
Order Date: 27 Oct 2021

Payment Information

Fei Wang fei.wang@usask.ca **Payment method:** Invoice	**Billing Address:** Fei Wang University of Saskatchewa n 57 Campus Dr Saskatoon, SK S7N 5A9 Canada +1 (306) 270-3923 fei.wang@usask.ca	**Customer Location:** Fei Wang University of Saskatchewa n 57 Campus Dr Saskatoon, SK S7N 5A9 Canada

Order Details

1. CURRENT MEDICINAL CHEMISTRY

Billing Status:
Open

Order License ID	1157486-1	**Type of use**	Republish in a thesis/dissertation
Order detail status	Completed	**Publisher**	BENTHAM SCIENCE PUBLISHERS LTD.
ISSN	0929-8673	**Portion**	Chapter/article

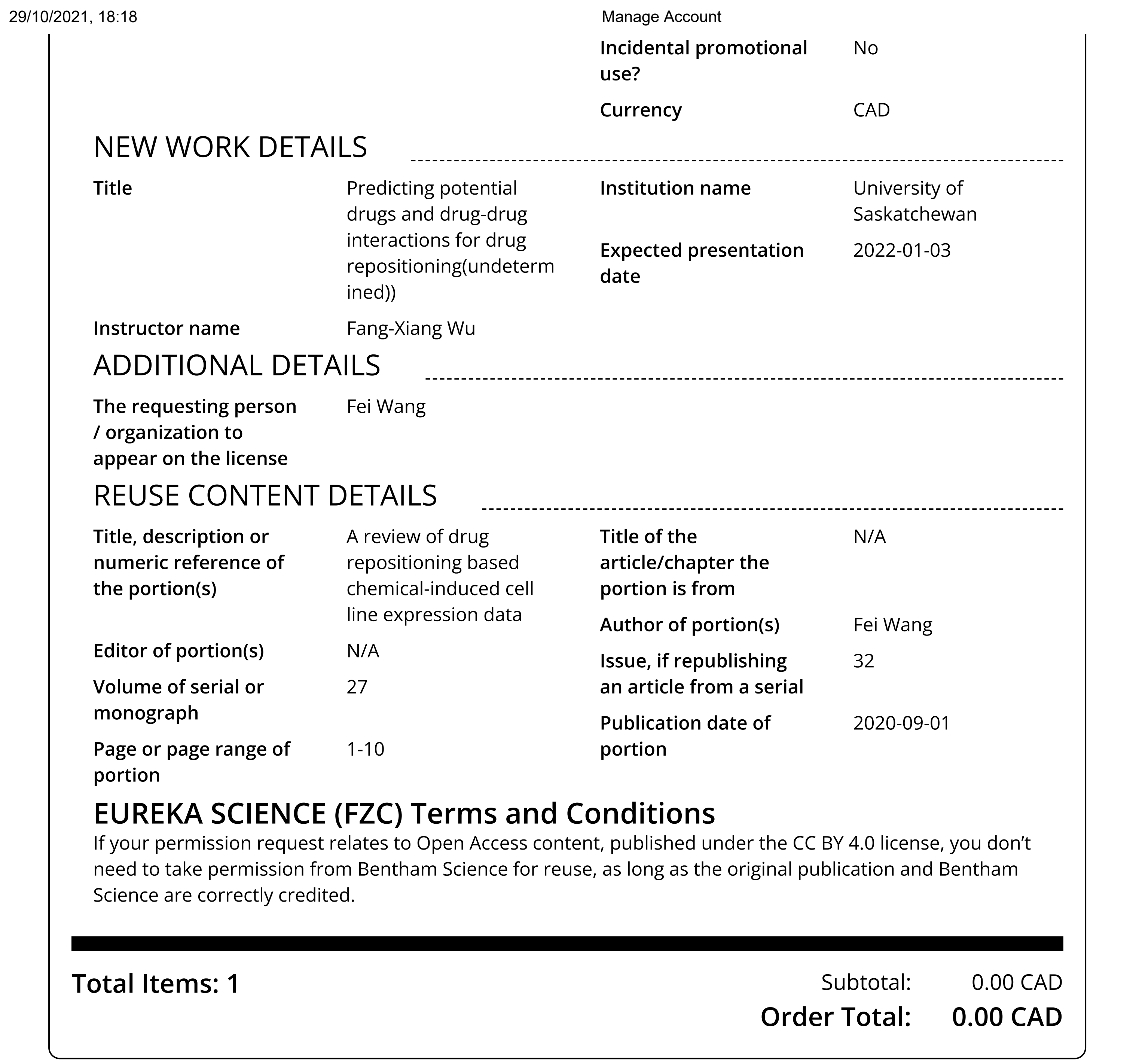
0.00 CAD
Republication Permission

LICENSED CONTENT

Publication Title	CURRENT MEDICINAL CHEMISTRY	**Country**	Netherlands
		Rightsholder	EUREKA SCIENCE (FZC)
Date	12/31/1993	**Publication Type**	Journal
Language	English		

REQUEST DETAILS

Portion Type	Chapter/article	**Rights Requested**	Main product
Page range(s)	1-10	**Distribution**	Canada
Total number of pages	10	**Translation**	Original language of publication
Format (select all that apply)	Electronic	**Copies for the disabled?**	No
Who will republish the content?	Academic institution	**Minor editing privileges?**	No
Duration of Use	Life of current edition		
Lifetime Unit Quantity	Up to 499		



		Incidental promotional use?	No
		Currency	CAD

NEW WORK DETAILS

Title	Predicting potential drugs and drug-drug interactions for drug repositioning(undetermined))	**Institution name**	University of Saskatchewan
		Expected presentation date	2022-01-03
Instructor name	Fang-Xiang Wu		

ADDITIONAL DETAILS

The requesting person / organization to appear on the license	Fei Wang

REUSE CONTENT DETAILS

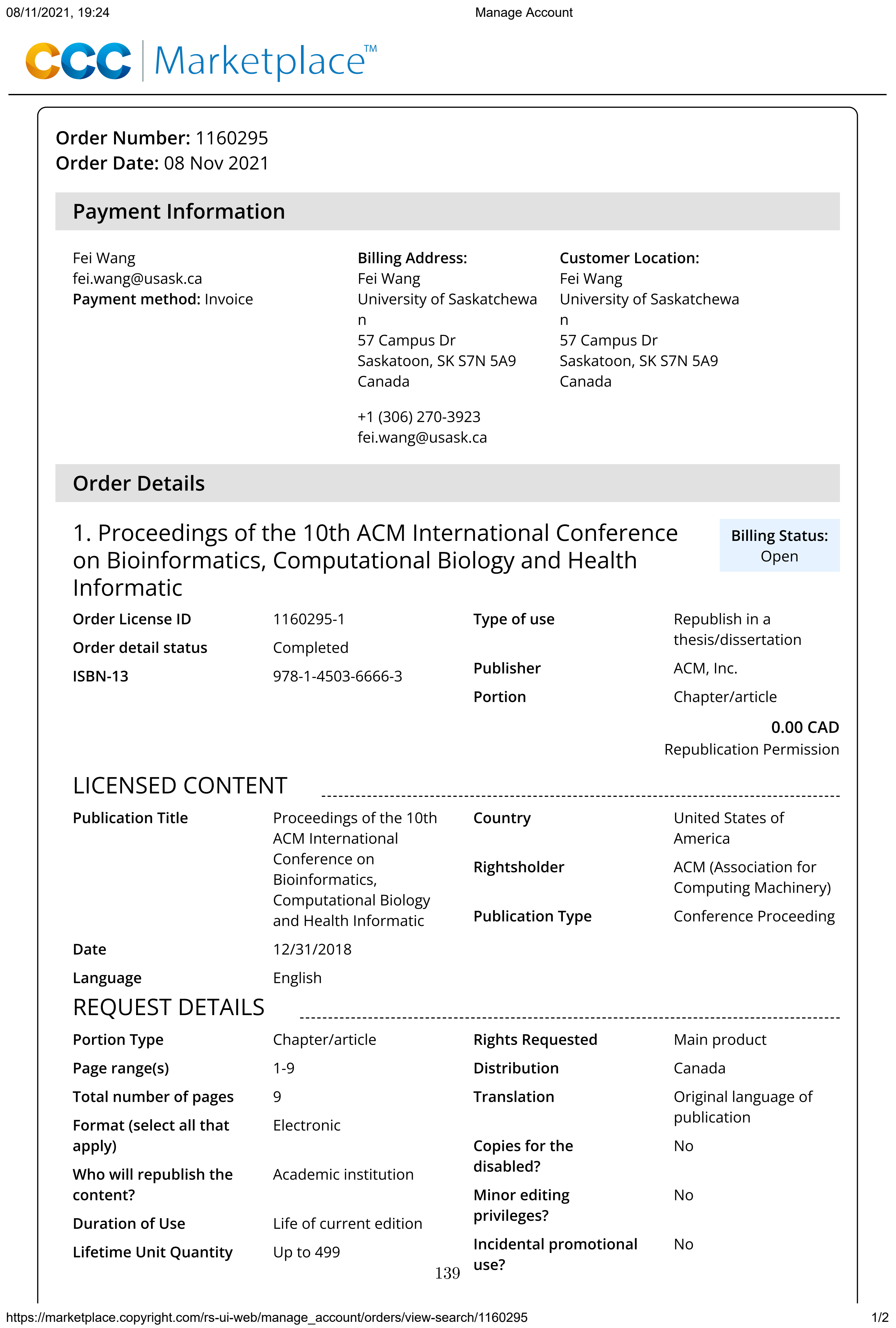
Title, description or numeric reference of the portion(s)	A review of drug repositioning based chemical-induced cell line expression data	**Title of the article/chapter the portion is from**	N/A
		Author of portion(s)	Fei Wang
Editor of portion(s)	N/A	**Issue, if republishing an article from a serial**	32
Volume of serial or monograph	27	**Publication date of portion**	2020-09-01
Page or page range of portion	1-10		

EUREKA SCIENCE (FZC) Terms and Conditions

If your permission request relates to Open Access content, published under the CC BY 4.0 license, you don’t need to take permission from Bentham Science for reuse, as long as the original publication and Bentham Science are correctly credited.

Total Items: 1

Subtotal:	0.00 CAD
Order Total:	**0.00 CAD**



08/11/2021, 19:24 Manage Account

CCC Marketplace™

Order Number: 1160295
Order Date: 08 Nov 2021

Payment Information

Fei Wang
fei.wang@usask.ca
Payment method: Invoice

Billing Address:
Fei Wang
University of Saskatchewa
n
57 Campus Dr
Saskatoon, SK S7N 5A9
Canada

+1 (306) 270-3923
fei.wang@usask.ca

Customer Location:
Fei Wang
University of Saskatchewa
n
57 Campus Dr
Saskatoon, SK S7N 5A9
Canada

Order Details

1. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatic

Billing Status:
Open

Order License ID	1160295-1	**Type of use**	Republish in a thesis/dissertation
Order detail status	Completed	**Publisher**	ACM, Inc.
ISBN-13	978-1-4503-6666-3	**Portion**	Chapter/article

0.00 CAD
Republication Permission

LICENSED CONTENT

Publication Title	Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatic	**Country**	United States of America
		Rightsholder	ACM (Association for Computing Machinery)
		Publication Type	Conference Proceeding
Date	12/31/2018		
Language	English		

REQUEST DETAILS

Portion Type	Chapter/article	**Rights Requested**	Main product
Page range(s)	1-9	**Distribution**	Canada
Total number of pages	9	**Translation**	Original language of publication
Format (select all that apply)	Electronic	**Copies for the disabled?**	No
Who will republish the content?	Academic institution	**Minor editing privileges?**	No
Duration of Use	Life of current edition	**Incidental promotional use?**	No
Lifetime Unit Quantity	Up to 499		

https://marketplace.copyright.com/rs-ui-web/manage_account/orders/view-search/1160295 1/2

		Currency	CAD

NEW WORK DETAILS

Title	Predicting potential drugs and drug-drug interactions for drug repositioning(undetermined)	**Institution name**	University of Saskatchewan
		Expected presentation date	2022-01-03
Instructor name	Fang-Xiang Wu		

ADDITIONAL DETAILS

The requesting person / organization to appear on the license	Fei Wang

REUSE CONTENT DETAILS

Title, description or numeric reference of the portion(s)	Human Protein Complex Signatures for Drug Repositioning	**Title of the article/chapter the portion is from**	N/A
Editor of portion(s)	N/A	**Author of portion(s)**	Fei Wang
Volume of serial or monograph	N/A	**Publication date of portion**	2019-09-04
Page or page range of portion	42-50		

Total Items: 1

Subtotal: 0.00 CAD

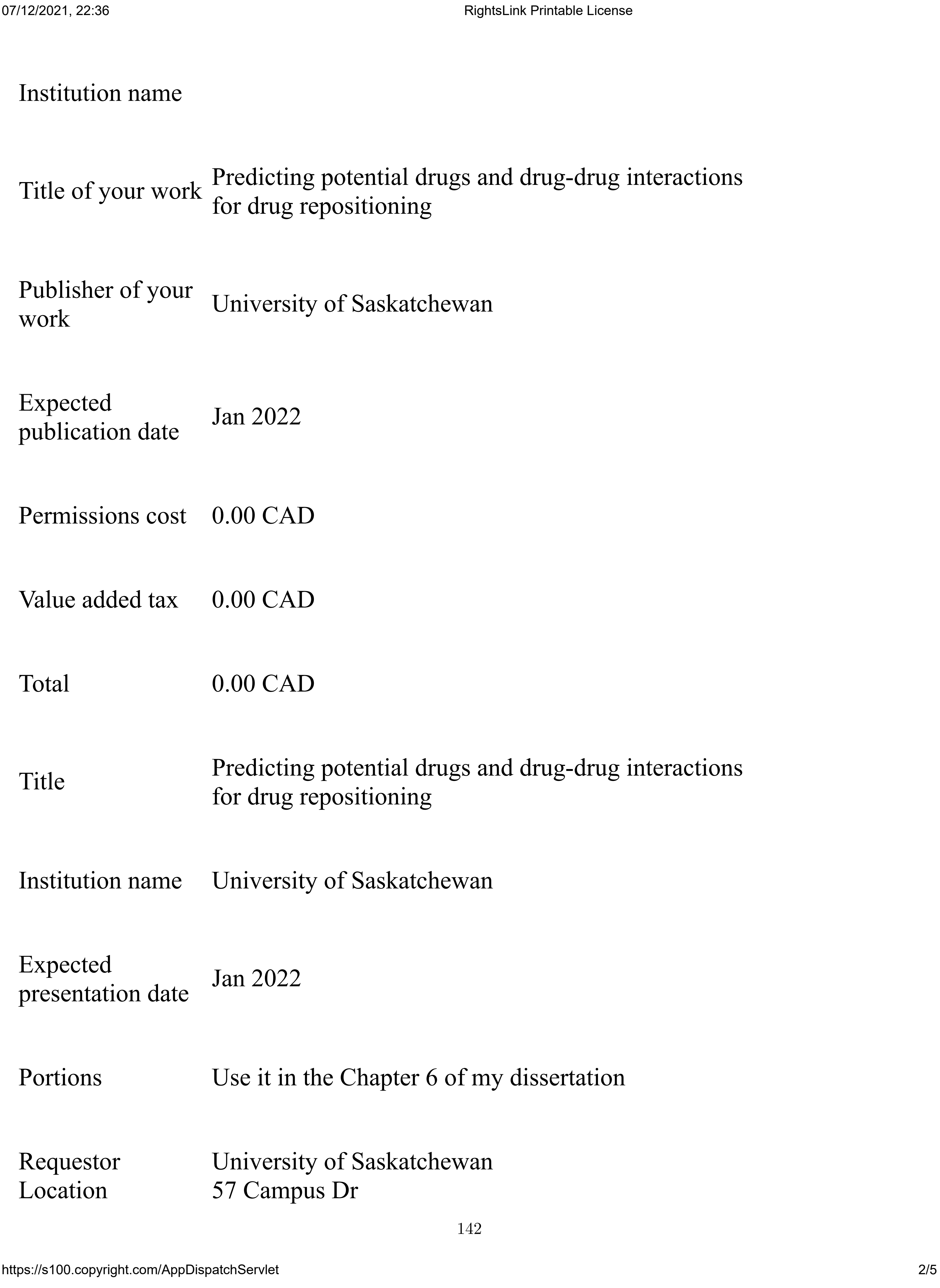
Order Total: 0.00 CAD

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Dec 07, 2021

This Agreement between University of Saskatchewan -- Fei Wang ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	5203990060093
License date	Dec 07, 2021
Licensed content publisher	Oxford University Press
Licensed content publication	Briefings in Bioinformatics
Licensed content title	Predicting drug–drug interactions by graph convolutional network with multi-kernel
Licensed content author	Wang, Fei; Lei, Xiujuan
Licensed content date	Dec 2, 2021
Type of Use	Thesis/Dissertation



Institution name	
Title of your work	Predicting potential drugs and drug-drug interactions for drug repositioning
Publisher of your work	University of Saskatchewan
Expected publication date	Jan 2022
Permissions cost	0.00 CAD
Value added tax	0.00 CAD
Total	0.00 CAD
Title	Predicting potential drugs and drug-drug interactions for drug repositioning
Institution name	University of Saskatchewan
Expected presentation date	Jan 2022
Portions	Use it in the Chapter 6 of my dissertation
Requestor Location	University of Saskatchewan 57 Campus Dr

Saskatoon, SK S7N 5A9
Canada
Attn: University of Saskatchewan

Publisher Tax ID GB125506730

Total 0.00 CAD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a

society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press’s written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC’s Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

OXFORD UNIVERSITY PRESS LICENSE TERMS AND CONDITIONS

Dec 07, 2021

This Agreement between University of Saskatchewan -- Fei Wang ("You") and Oxford University Press ("Oxford University Press") consists of your license details and the terms and conditions provided by Oxford University Press and Copyright Clearance Center.

License Number	5203990366405
License date	Dec 07, 2021
Licensed content publisher	Oxford University Press
Licensed content publication	Briefings in Bioinformatics
Licensed content title	Predicting drug–drug interactions by graph convolutional network with multi-kernel
Licensed content author	Wang, Fei; Lei, Xiujuan
Licensed content date	Dec 2, 2021
Type of Use	Thesis/Dissertation

Institution name	
Title of your work	Predicting potential drugs and drug-drug interactions for drug repositioning
Publisher of your work	University of Saskatchewan
Expected publication date	Jan 2022
Permissions cost	0.00 CAD
Value added tax	0.00 CAD
Total	0.00 CAD
Title	Predicting potential drugs and drug-drug interactions for drug repositioning
Institution name	University of Saskatchewan
Expected presentation date	Jan 2022
Portions	Use the figures and tables in Chapter 6 of my dissertation.
Requestor Location	University of Saskatchewan 57 Campus Dr

	Saskatoon, SK S7N 5A9 Canada Attn: University of Saskatchewan
Publisher Tax ID	GB125506730
Total	0.00 CAD

Terms and Conditions

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a

society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press’s written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC’s Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.