Predicting drug - target interaction network using deep learning models for drug repurposing through genetic information

by

Jiaying You
A Thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada

Abstract

Traditional methods for drug discovery are time-consuming and expensive, so new efforts have been made to perform drug repurposing, which can develop new uses from already approved drugs. In order to explore new ways for this innovation, many computational approaches have been proposed to predict drug-target interactions (DTIs). However, due to the high-dimensional nature of the data sets extracted from drugs and targets, traditional machine learning approaches such as logistic regression analysis could not analyze these data efficiently. In this thesis, we aim to propose LASSO-based regularized linear classification models and a deep neural network (DNN) model to predict DTIs and apply clustering analysis to the new predictions for exploring potential drugs in enriched clusters for inflammatory bowel disease (IBD) and breast cancer.

We collected drug descriptors, protein sequence data from Drugbank and protein domain information from the National Center for Biotechnology Information (NCBI). Validated DTIs were downloaded from Drugbank. A new similarity-based approach was developed to build the negative DTIs. We proposed multiple LASSO models to integrate different combinations of feature sets to explore the prediction power and predict DTIs. Furthermore, building on the features extracted from the LASSO models with the best performance, we also introduced a DNN model to predict DTIs. The performance of the DNN model was compared with the LASSO models and a standard logistic regression (SLR) model. Furthermore, we proposed the clustering analysis based on the DNN predictions and observed DTIs that in enriched clusters for repurposing existing drugs for breast cancer and IBD based on their risk gene information.

Experimental results showed that the DNN model outperformed the SLR and LASSO models. Particularly, LASSO models with protein tripeptide composition (TC) features and domain features were superior to those that contained other protein information in LASSO models, which may imply that TC and domain information could be better representations of proteins. Furthermore, we showed that using DTIs predicted by the DNN model, the potential drugs we collected for IBD and breast cancer from the enriched clusters can potentially be repurposed for IBD and breast cancer treatment.

.

Acknowledgement

This thesis was completed with the encouragement and motivation of many people and my lovely pet, and I would like to express my sincere gratitude to all of them.

I would first like to thank my supervisors, Dr. Pingzhao Hu and Dr. Robert McLeod for their continuous support in my M.Sc. study and thesis completion. Their professional insights and expertise in academic steered me in the right direction to acquire wisdom and knowledge. They were always willing to help when I met obstacles during my research and thesis writing. They are both dedicated professors that I could learn a lot from. They have made it possible for me to further pursue my dreams down the road.

I am grateful to the lab members at The Hu Lab: Md. Mohaiminul Islam, Qian Liu, Rayhan Shikdar, Chen Chi, Ye Tian, Yan Cheng, Svetlana Frenkel, Xinyu Hou and Nikta Feizi, for their friendship and research discussions. I would also like to genuinely thank all the other faculty members and staff at the Department of Electrical and Computer Engineering and the Department of Biochemistry and Medical Genetics for their fully support during my M.Sc. degree completion in various ways.

I also thank my thesis committee members, Dr. Yang Wang and Dr. Dr. Ahmed Ashraf, for their insightful and stimulating comments on my thesis.

Last but not the least, I would like to thank my family: my parents Xianghong Zhu and Li You, for supporting me emotionally and financially throughout my life. I would also like to express my special thanks to my pet hamster for always being there for me through ups and downs, I love you like no other.

Publications and Contributions

1. **J You**, MM Islam, L Grenier, Q Kuang, R McLeod, P Hu. Drug-target Interaction Network Predictions for Drug Repurposing Using LASSO-based Regularized Linear Classification Model. In: Bagheri E., Cheung JCK. (eds) Advances in Artificial Intelligence. AI 2018. Lecture Notes in Computer Science, Vol 10832. Springer, Cham. This paper is presented in **Chapter 3**.
2. **J You**, R McLeod, P Hu. Predicting Drug-Target Interaction Network Using Deep Learning Models. Submitted to Asia Pacific Bioinformatics Conference 2019. This paper is presented in **Chapter 3** and part of **Chapter 4**.

I, Jiaying You generated the interaction dataset, designed, and implemented the experiments and drafted the manuscripts. Dr. Pingzhao Hu generated IBD and breast cancer risk gene data from Genome-wide association study (GWAS). Dr. Pingzhao Hu, Dr. Yang Wang and Dr. Robert McLeod provided their insightful suggestions during the course of designing and implementing the experiments. Dr. Pingzhao Hu supervised and monitored the whole project.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACC	Amino Acid Composition
AUC	Area Under the Curve
CD	Crohn's Disease
CNN	Convolutional Neural Networks
DC	Dipeptide Composition
DL	Deep Learning
DNN	Deep Neural Network
DTI	Drug-Target Interaction
FNR	False Negative Rate
FPR	False Negative Rate
GBM	Gradient Boosted Machine with Trees
GWAS	Genome-Wide Association Study
IBD	Inflammatory bowel disease
RBM	Restricted Boltzmann machine
ReLU	Rectified Linear Units
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristics
SE	Disease-Side Effect
SLR	Standard Logistic Regression
SNP	Single Nucleotide Polymorphism

SVM	Support Vector Machine
TC	Tripeptide Composition
TNR	True Negative Rate
TPR	True Positive Rate
UC	Ulcerative Colitis

Chapter 1 Introduction

1.1 Drug discovery and drug repurposing

Drug development is usually a high cost and time-consuming procedure. According to an estimate raised by a workshop in 2014, bringing a new drug into market successfully could take 10 years and cost at least $1 billion [1]. For years in the biopharmaceutical industry, companies have been making efforts to increase the productivity by new drug discovery methods. One example is to discover new uses of existing drugs, which is known as drug repositioning or drug repurposing. These terminologies are used for finding potential drugs that could be applied for new indications or applications.

From an industry perspective, most of the concern comes from the risk versus return trade-off in research and development. Traditional drug discovery has high risks with low probability albeit of very high reward. Industries are trying to increase the productivity and have made enormous investments to minimize the trade-off risk by novel discovery technologies, such as structure-based drug design, combinatorial chemistry, high-throughput screening (HTS) and genomics[2]. However, to date the results are not so satisfying with only few products developed [3]. Drug repositioning overcomes the drawbacks of the traditional drug discovery process. Traditional de novo drug discovery is well-known typically requiring 10 years of process yet the success is lower than 10%, drug repurposing offers a much cheaper and faster way for drug discovery (**Fig. 1**)[4]. Advantages in drug repurposing compared to traditional methods showed significance savings in time and cost differences since the repurposed candidates have already been through the clinical usage and thus could substantially be bypassed in terms of safety

concerns. These appealing characteristics led to the increasing interests for biopharmaceutical companies to scan the existing pharmaceuticals for repurposing candidates and the technologies of drug repositioning are developing rapidly. A great number of successes have been achieved by biopharmaceutical companies, such as thalidomide for severe erythema nodosum leprosum offered by Celgene company [1].

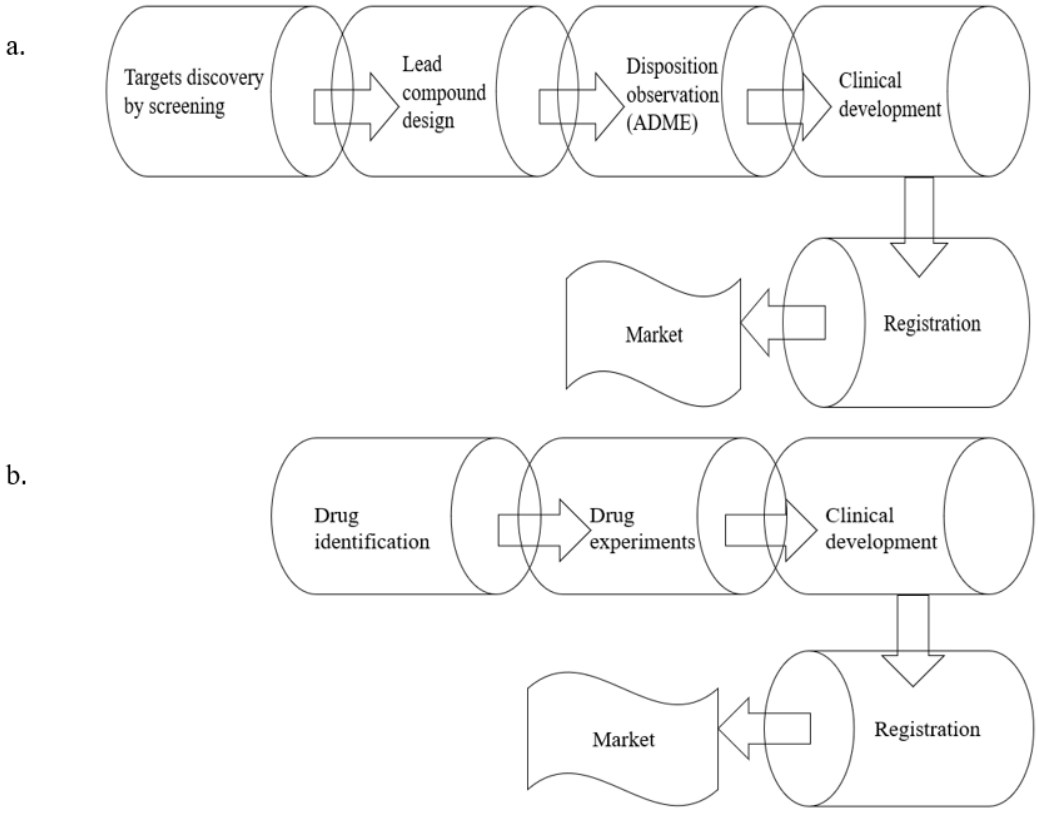


Figure 1- General schematic flowcharts of traditional de novo drug discovery (a) and drug repurposing (b). Traditional drug discovery involves a longer process as well as risks in safety. Drug repurposing reduces time and risk concerns due to the reuse of existed drugs.

The great benefits of drug repositioning have attracted attentions from many groups including academic researches, government institutions, such as the National Center for Advancing Translational Science in the United States and the Medical Research Council in UK.

As summarized in 2012, 30% of the FDA-approved new drug products were discovered by drug repurposing[5]. This leads us to ask, will it ever end? With the booming development of the current drug repositioning market, will the potential chemical candidates ever be exhausted and have nothing to take advantage of? Luckily, Researchers keep optimistic to this question and it is believed that the number of potential drugs for indications exceeds the majority's screening capacity[2].

1.2 Drug repositioning using traditional machine learning strategy

Machine learning (ML) is an application of artificial intelligence that expects machines to improve their performance as a result of a learning procedure. Like black-box operations, machine learning models obtain optimizations of given problems ruled by algorithms and parameters that are tune automatically, trying to find the potential patterns that are usually hidden and hard to observe from voluminous data. The advantages of ML models are that they can exceed human performance and change the ways of improving personalized products in many fields including, but not limited to, house price prediction, travel salesman problem (TSP), advertisement recommendation, language translation and image classification.

Widely acknowledged machine learning methods, such as support vector machine (SVM), random forest (RF), simulated annealing (SA), shallow artificial neural network (ANN) etc., have been shown to be efficient in different applications. For example, SVM models are proven to have better performances in traditional text categorization than Naïve Bayes models [6]. However, the limitations of these traditional machine learning techniques are that they usually have shallow architectures and could only exploit limited nonlinear functions. As such, support

vector machines depend on the use of kernel tricks to generate the feature transformation layer [7].

As the biological and experimental methods of drug repurposing are time-consuming, expensive and difficult [8], shortcomings of the traditional methods have led to increasing interests in applying sophisticated analysis methods, such as machine learning, to analyze the vast amount of complex data available to researchers. Machine learning has been used in discovering patterns from diverse data sources and providing excellent versatility in its applications for many years and its technologies are proven to be highly useful in identifying drug-target interactions [9]. Many researches have applied machine learning technologies in drug repositioning. A study by Kai Zhao et.al applied several machine learning approaches, such as SVM, RF, gradient boosted machine with trees (GBM) and logistic regression approaches as baseline models compared to deep neural network models for drug repurposing in psychiatry using drug expression profiles, which express the transcriptomic changes when the specific cells lines were affected with drugs or chemical [10]. Given the known interactions between drugs and genes, supervised learning has been applied to discover unknown genes that could be affected by compounds. Drugs that have high prediction probability values in the models were believed to be good candidates for repositioning. As a result, they focused on the top repurposing candidates that were enriched for psychiatric disease for clinical experiments, such as Cyproheptadine, Chlorcyclizine, Pizotifen and TrichostatinA. These selected drugs were supported by literature on psychiatric disease treatment. For example, Metformin, a repurposed drug selected by logistic regression, has shown clinical evidence of reducing depression risks. They also found support for of some medication as evidenced in human and animal clinical trials.

Yamanishi et al. used machine learning tools to develop a kernel regression-based approach for predicting drug-target interactions that can identify previously unknown drug-target interactions from various types of biological data [11], such as chemical structures, drug side effects, amino acid sequences and protein domains. They used these features to generate a kernel similarity matrix, where each descriptor corresponds to a drug-drug similarity or protein-protein similarity. The model assumes that similar compounds are likely to interact with similar proteins to generate their prediction model. Generally speaking, drugs with high similarities are considered to have shared target proteins and new interactions observed are used for drug repurposing. They applied analysis in five diseases: Huntington disease, two subtypes of lung cancer, alcohol dependence and polysubstance abuse and reported the most significant candidates for each disease. For example, Olanzapine showed very likely prediction in their newly designed three-layer model with the known drug Baclofen, which indicates a potential possibility for treatment of Huntington disease.

Additionally, Chiang and Butte suggested new indications for existing drugs based on theirtherapeutic effect [12]. A study in 2011 by Lun Yang [13] identified a disease-side effect (SE) weighted network based on given disease-drug and drug-SE associations where Matthews correlation coefficient (MCC) was tested as the evaluation of a node's association strength. From that study, 3,175 predicted relationships between diseases and SEs were extracted and some of the observations were found in compelling published clinical trials.

Other drug-target interaction (DTI) methods, such as network-based and phenotype-based algorithms are developing rapidly. A network-based inference method developed by Cheng et al.

[14] designed three inference algorithms for DTI prediction. Two of them were associated with drug similarity and protein similarity network respectively, and the last one is simply based on the known drug target bipartite interaction. Generally speaking, for the first similarity check of drugs and proteins, the author took the chemical structure and genomic sequence to quantify the similarity between them and assumed similar drugs may share the same targets and vise-versa. Although the result from drug similarity network and protein similarity network seem poor in (Area Under Curve) AUC values, the simplest network-based technique shows the best results among three. One suspicion pointed to the redundancy issue, in which chemical structure does not accurately present drug similarity. Through the network-based inference method, the authors calculated the top potential interaction pairs for the whole dataset and published the full prediction online: http:// www.lmmd.org/database/dti/. By vitro assay experiment of the prediction interactions, four drugs show positive features on estrogen receptors or dipeptidyl peptidase-IV and two drugs show activities on a breast cancer cell line.

Considerable number of state-of-the-art machine learning approaches have been widely applied in drug repositioning using biological data, such as gene expression measurements after treatment of certain cell lines or integrated drug information. However, with data that are potentially high in background noise, researchers have concerns if traditional machine learning algorithms could detect and reduce the effects of noise, and thus explore the full information of input data.

1.3 Drug repositioning using deep learning

1.3.1 Deep learning introduction

With the development of data collection, the amount of data we can access has been massively increased in the past few decades. Therefore, traditional machine learning approaches may not be capable for such high-dimensional and noisy data, which indicates that the tremendous information stored in big data may not be completely discovered and exploited.

Deep learning is an extension of artificial neural network-based machine learning that simulates the decision making of a human brain. Deep learning, as a subfield of machine learning, has more capabilities than traditional machine learning algorithms due to its high complexity and non-linearity. A DL model usually has a series of hidden layers connecting to each other between input data and predicted outcomes. These hidden layers transform information forwardly with customized architectures, such as losing connections or normalizations. Such a complex network generally provides more flexibility with respect to raw data than a shallow neural network since it automatically abstracts necessary representations from layer to layer, which also gives us the opportunity to obtain a good performance with some noisy data. The deep neural network models automatically and ideally learn the abstracted features from the information and pass these abstractions on to next layers themselves without of requiring human intervention, which makes problems that need massive computations accessible for engineers.

The efficiency and applicability to use deep learning models in multiple fields, such as language translation, image prediction and objects detection has clearly been demonstrated. As the nature of biomedical data is usually high dimensional and noisy [15], it is recommended to apply deep learning methods in bioinformatics.

1.3.2 Deep learning architectures

Traditional and deep neural networks can be trained with multiple layers and optimized by techniques such as stochastic gradient descent. Weights can then be computed or adjusted using a backpropagation process. Traditional backpropagation computes the derivative of each weigh backwardly through the network based on an objective or cost function. Typically, the cost function expresses the difference between the predicted and expected values. To minimize the difference for better fitting, the backpropagation is applied gradually from the output to the input feature vector. In practice, issues like over-fitting occur often especially when the training dataset size is small. Hence, the selection of hyper-parameters or parameters initialization should be carefully considered, but this tuning heavily relies on the researchers' experience.

Parallel processors like graphics processing units (GPUs) offer more efficient implementations in practice, which allow researchers to deep mine high throughput data. It has been shown that speech recognition and image classification have benefited from GPUs using deep neural networks [16]. A fast GPU fundamentally changes the experience in practical deep learning, such as shortening the learning iteration time for choosing, tuning, and training parameters. For example, intensive tasks that have high requirements of computations and memory storage like convolutional neural network (CNN) achieve good performance by GPUs-

based computational platforms. Researchers developed a deep convolutional neural network for image classification using 1.2 million high-resolution images from LSVRC-2010, and classified them into 1000 categories [17]. The efficient GPU makes the complex network with 60 million parameters and 650,000 neurons train much faster. The group reportedconsiderably better results than previous machine learning approaches. The architecture of this CNN model contains typical convolutional layers, max pooling layers, normalization layers and fully connected layers with some drop-outs to avoid overfitting.

Another classical deep learning approach is a Deep Belief Network (DBN), which could be considered as a preprocess procedure for feature selection used in classification. With one input layer, one output layer and several hidden layers, DBN structure stacks Restricted Boltzmann Machines (RBMs) [18] and learns to extract deep features of input data. Each RBM contains two adjunct layers as one visible layer and one hidden layer, parameters of weights between two layers and biases are learnt by stochastic gradient descent and then the current hidden layer is treated as visible layer for next RMB to start a new training. It is reported that it worked well on MNIST handwritten digits using DBN and good results are also observed on phoneme recognition[19].

Applications of deep neural networks have been benefitted lots of domains, such as image recognition[17], language processing[20], business advertising[21] and bioinformatics[22]. It also plays an important role in computational drug discovery.

1.3.3 Deep learning in bioinformatics and drug repurposing

Since deep learning transforms massive data to useful information efficiently in various aspects, many researches start to adopt deep learning methods for biological and healthcare studies [23]. For example, the company Google DeepMind has devoted resources into a healthcare improvement program after a huge success on AlphaGo, a gaming AI [24]. According to the statistics of published papers on deep learning in general bioinformatics applications [23], the number of publications in deep learning has increased dramatically since 2013 and a huge rise was observed in bioinformatics since then. Additionally, architectures for bioinformatics mostly lie in deep neural networks (DNN), convolutional neural networks (CNN), and recurrent neural networks (RNN). Reviews on these applications were summarised in genomic data [25],image data [26], omics data and signal processing data in recent years.

The rise of deep learning in bioinformatics enriches the repositories and websites. For example, Kipoi provides ready-to-use trained models for regulatory genomics. Taking Kopoi as an example, it was created by three teams from Technical University in Munich, European Bioinformatics Institute and Stanford University in March 2018. It collects 2031 models from 17 groups mostly developed by Keras and TensorFlow, which are open to the public for convenient use. Kopoi is now available in python, Linux command line and R. A popular model, Basenji, in Kopoi contributed by David Kelley[27] applies deep convolutional neural networks (DCNN) to sequential regulatory activity classification. Using DNA sequence data, the model predicts cell type-specific epigenetic and transcriptional profiles. By studying the genotypes, the application predicts phenotypic outcomes and it helps to generate mechanistic hypotheses for disease loci mapping. Another application in predicting variant effect is also accessible in

DeepSEA/varientEffects [28] model contributed by Zhou et al. in 2015. This model was trained on ENCODE and Roadmap Epigenomics data. It predicts the whole 919 categories with a probability score for each category.

Deep learning has also been applied in finding potential indications for existed drugs. A deep learning-based model DeepDTIs was implemented by Wen et al. [22] for drug-target interaction in 2017. The study first generated features by unsupervised pre-training, then applied a deep learning algorithm on known drug-target interactions. The unsupervised training was based on layer-wise RBMs and it went to a standard logistic regression (LG) binary classification. By measuring the accuracy, area under the receiver operator characteristic (AUC), true positive rate (TPR, aka sensitivity) and false positive rate (TNR, aka specificity), the authors reported the highest performance using customised DeepDTIs compared with other baseline models including Bernoulli Naïve Bayesian (BNB), Decision Tree (DT), and Random Forest (RF). The report also suggested potential interacted drug-target pairs based on probability scores the model predicted. They reported top 10 predicted interaction pairs ranked by probability scores. Four of the suggested pairs were already validated by STITCH or other literature. Some drugs were found to have potential for some diseases. For example, 5-hydroxytryptamine receptor 1A,2A,1B,2C,1D was seen in drug Ziprasidone, which were shown to have high sequence homology with 5-hydroxytryptamine receptor 1C.

1.4 Drug-repurposing using genetic information from genome-wide association study (GWAS) findings

1.4.1 GWAS studies and findings

Genome-wide association studies (GWAS) investigates the association of genetic variants diseases and traits . By comparing DNA of control and case group individuals, GWAS identify risk (Single Nucleotide Polymorphism) SNPs or other genomic biomarkers in DNA associated with the diseases and traits. They investigated the entire genome rather than a partial region of the human genome. The first GWA study collected 50 normal individuals and 96 patients with age-related macular degeneration (ARMD) in 2005 [29], and compared the allele frequency between the two groups. Two SNPs were observed significantly different between the two groups. Successful research in GWAS increased rapidly and the relatively large case-control study performed in 2007 involved around 2000 patients for 7 diseases including coronary heart disease, type1/type2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder and hypertension and a shared control group with 2000 individuals [30]. Many significant risk loci have been identified for these diseases in the study. The future of GWAS may be improved by increasing sample size or the use of narrowly defined biomarkers [31].

As of 2017, approximate 1800 diseases or traits have been examined among large populations through GWAS, and thousands of risk SNPs have been found associated with them [32]. However, how to efficiently use these GWAS findings for clinical use, especially for drug discovery, are still underexplored.

1.4.2 Drug repurposing using GWAS findings

It is believed that drugs with associated diseases and traits through genetic studies are twice as likely to be clinically accepted than other drugs [33]. Moreover, the large investments on GWAS also helped the identification of drugs from drug repositioning. Assessment of the GWAS findings for drug repositioning has been investigated. Based on the published GWAS data accessed from US National Human Genome Research Institute, the corresponding catalog contained 796 publications with 4,818 rows of data [34]. The authors selected 991 risk genes associated with diseases and traits from the GWAS. They found 92 individual genes were newly identified as targets by comparing with indications of drugs with known targets. Using the catalog, Wang et.al [35] conducted a pharmacogenomics-based network method to support drug repurposing and next generation of treatment. The known drug Capecitabine, which wass originally used for treating breast cancer, was repurposed by the designed network with other 12 potential indications, such as Bladder Neoplasm. These findings are supported by published results.

A GWAS focused on IBD was conducted to compare the SNP-specific allele frequency between the case and control groups. Two hundred and thirty-three significant risk loci were identified [36]. Based on the identified risk variants, Collij et al. [37] developed a program to link genes including the variants with disease based on Drugbank database, and conducted a protein-protein interaction analysis for drug repositioning. More than 100 promising drug compounds were reported for potential treatment of IBD, such as the newly identified drug INCB3284, which is known as a potent antagonist of protein CCR2. CCR2 has proved to have

expression difference in IBD patients and normal group by previous studies[38].Therefore the authors suggested the candidate INCB3284 could be further studied as a treatment of IBD.

Current studies on using genetic information from GWAS for drug repositioning show great potential. However, it is still lacks principled methods to guide the applications.

Chapter 2 Motivations and Research Objectives

2.1 Motivations

Drug-target based drug repositioning, i.e. finding new indications for existing drugs, relies much on identification of drug-target interactions. Traditional high-cost and time-consuming experiments for drug-target interaction identification were gradually replaced by computational tools. With the fact that the dimension of biological data is quite large, and variables may be highly correlated, e.g. protein sequence data, it may be challenging to apply traditional machine learning models due to over-fitting problems. To overcome the above limitations, other recently developed machine learning approaches, such as deep neural network models, can be developed to integrate multiple data resources for predicting DTIs. Moreover, with recent development in GWAS, many genetic risk variants and genes have been found for different diseases and traits. However, how to efficiently use this genetic information for drug repositioning has not been well-investigated yet. IBD and breast cancer are two complex diseases. They are typical diseases with many risk variants and genes identified from GWAS. Using these diseases as examples for drug repositioning based on the genetic information will provide a model for drug repositioning for other complex diseases. .

2.2 Hypothesis

We hypothesize that drug descriptors encoded information of drugs with amino acid compositions could provide properties of proteins to measure their interactions. LASSO models could accurately learn the drug-protein interaction (DTI) patterns. The improved deep neural network model could have a larger capacity in DTI prediction. Furthermore, we hypothesize that

the bipartite network of drugs and proteins can be clustered efficiently with high modularity and ignorable/negligible loss of interactions. The clusters are enriched for risk genes and drugs, and the drugs in the clusters can be repositioning for the diseases caused by the risk genes.

2.3 Research Objectives

Research objectives of this thesis are to identify DTI and repurpose drugs for IBD disease and breast cancer from the predicted DTI network using risk genes identified from GWAS. Through the bipartite relationships between drugs and proteins we aim to predict interacting pairs or non-interacting pairs. Four specific aims are to:

- Implement standard logistic regression models and improved LASSO regression models to predict DTIs using different features of drug and protein properties.
- Implement DNN models to predict DTIs using integrated data from multiple sources.
- Implement bipartite clustering to the predicted DTI network from the trained DNN model and evaluate the identified clusters through enrichment analysis.
- Repurpose drugs for IBD and breast cancer given risk genes identified from GWAS using the enriched clusters.

Chapter 3 Prediction of drug-target networks using machine learning approaches

3.1 Introduction

Studying drug-target interactions is essential in drug discovery and drug repurposing research. Biological and experimental methods of drug discovery and repurposing are time-consuming, expensive, and difficult. The shortcomings of traditional methods have led to increased interest in applying sophisticated analysis methods, such as machine learning, to analyze the vast amount of complex data available to researchers. Machine learning has been used in discovering patterns from diverse data sources and providing excellent versatility in its applications for many years. Machine learning technologies have proved to be highly useful in identifying drug-target interactions.

To build the machine learning algorithms for predicting drug-target interactions, machine learnable feature vectors should be built. Recent focuses have been on using biological and chemical properties of drugs and targets or proteins. Other necessary information, such as drug-target interaction data, is publicly available. In this study, we use DTI data, drug structure data and protein sequence data from Drugbank to build logistic regression models, LASSO models and a deep neural network model using various integrated features to predict theDTI.

3.2 Material and data collection

3.2.1 Drug-target data collection

The drug and target sets were downloaded from Drugbank [39]. Drugbank is a bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug and target information. Drugbank contains approved small molecule drugs, approved biotech drugs, nutraceutical and experimental drugs with non-redundant protein sequences linked to all the drug entries. Datasets such as drug structures (https://www.drugbank.ca/releases/latest#structures), target sequences (https://www.drugbank.ca/releases/latest#target-sequences), or drug-target interactions (https://www.drugbank.ca/releases/latest#protein-identifiers) can be downloaded from their websites. Previously, Ezzat et al. [9] used DrugBank to collect data consisting of 12,674 drug-target interactions, 5,877 drugs, and 3,348 protein interaction partners. They generated biochemical features for drug-target predictions with a chemogenomic approach through Drugbank. Another resource we can easily access is GVK BIO Online Biomarker Database (GOBIOM). It connects biomarkers associated with different therapeutic areas recorded in international clinical trials or studies[40]. For example, it provides identification profiles of drug-target-disease relationships, which could be used as input data for computational prediction of compound-protein interactions (CPIs). Since the database does not include much drugs- and targets-specific feature information, we will focus only on DrugBank.

Built in 2006, Drugbank contained 841 Food and Drug Administration (FDA) approved drugs as well as 2,133 targets when first released. It grew fast and by the end of 2009, the server

contained 1,467 drugs including both small molecule drugs and biotech drugs. As of now, the version 5.1.0 contains 11,091 drugs and 5,117 linked proteins.

As shown in **Fig. 2**, we used DTIs data under protein identifier (Version5.0.10) and external target drug-uniprot link (Version5.0.10). Biotech drugs were excluded since the Rcpi R package is only applied to small molecular drugs[41]. Thus, 14,792 known drug-target pairs from Drugbank were generated with 5,834 unique drugs and 3,546 unique proteins.

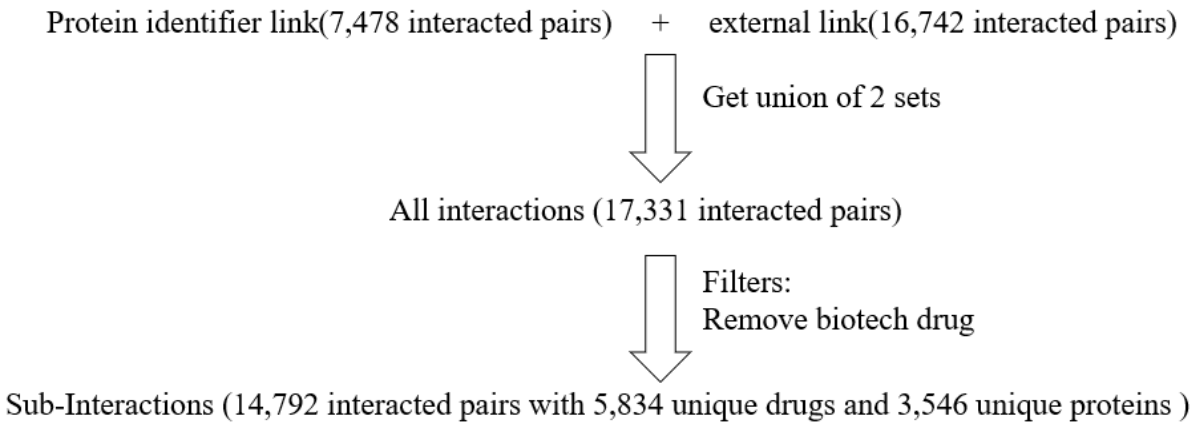


Figure 2– Flowchart of data collection. In this experiment, 14,792 known drug-target pairs from Drugbank were generated with 5,834 unique drugs and 3,546 unique proteins.

3.2.2 Drug and target feature representation

Chemical properties, topological properties and geometrical properties are often referred to as drug descriptors. All these can be considered drug features. We downloaded drug structure information from Drugbank (https://www.drugbank.ca/releases/latest#structures). It should be noted that some drug structures are not available, thus we kept only 5,585 drugs in this procedure. To get drug descriptors from the drug structure information, we used the online sever chemical

database with modeling environment (OCHEM) [42]. Drugs with unavailable descriptors were removed. We retrieved 2,216 drug descriptors for each of the remaining 5,134 drugs. To generate target/protein feature representation, we assumed that the complete information of a target protein is encoded in its sequence [43] and domains, thus we extracted protein features from the commonly recognized features-sequence descriptors and protein domains. The sequence and corresponding domain information of target proteins in our study were downloaded from Drugbank and (National Center for Biotechnology Information Search database [44] (NCBI, https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) respectively. Protein sequence descriptors contain amino acid composition (ACC), dipeptide composition (DC) and tripeptide composition (TC). AAC is the statistic frequency of each amino acid. DC is the statistic frequency of every two amino acid combination while TC is the statistic frequency of every three amino acid combination. Domain information is represented as an adjacency matrix where 1 indicates an interaction and 0 indicates no interaction between a pair of proteins and domains. In total, we extracted 11,943 protein features for each protein target. Therefore, each pair of drug-target/protein (D, P) can be represented by a feature vector as

$$[D_1, \dots, D_{2216}, P_{ACC1}, \dots, P_{ACC20}, P_{DC1,} \dots P_{DC400}, P_{TC1}, \dots P_{TC8000}, P_{domain1,\dots}, P_{domain3523}]$$

which contains 2,216 drug features: $D_1, \dots, D_{2216}$,11,943 target features with 20 ACC expressions $P_{ACC1}, \dots, P_{ACC20}$,400 DC expressions $P_{DC1,} \dots P_{DC400}$,8,000 TC expressions $P_{TC1}, \dots P_{TC8000}$, and 3,523 domain information $P_{domain1,\dots}, P_{domain3523}$.

After all drug and target feature information was generated, the 13,168 known DTIs with 5,132 unique drugs and 3,184 unique proteins were collected in our study as shown in **Fig. 3**. All

these 13,168-known DTIs were treated as positive samples. Since the number of all possible DTIs from the collected drug and target lists is 18,138,422 (5,134*3,533) (**Fig. 3**), 18,125,254 (18,138,422-13,168) of which were considered as unknown DTIs. These can be potential negative or positive DTIs.

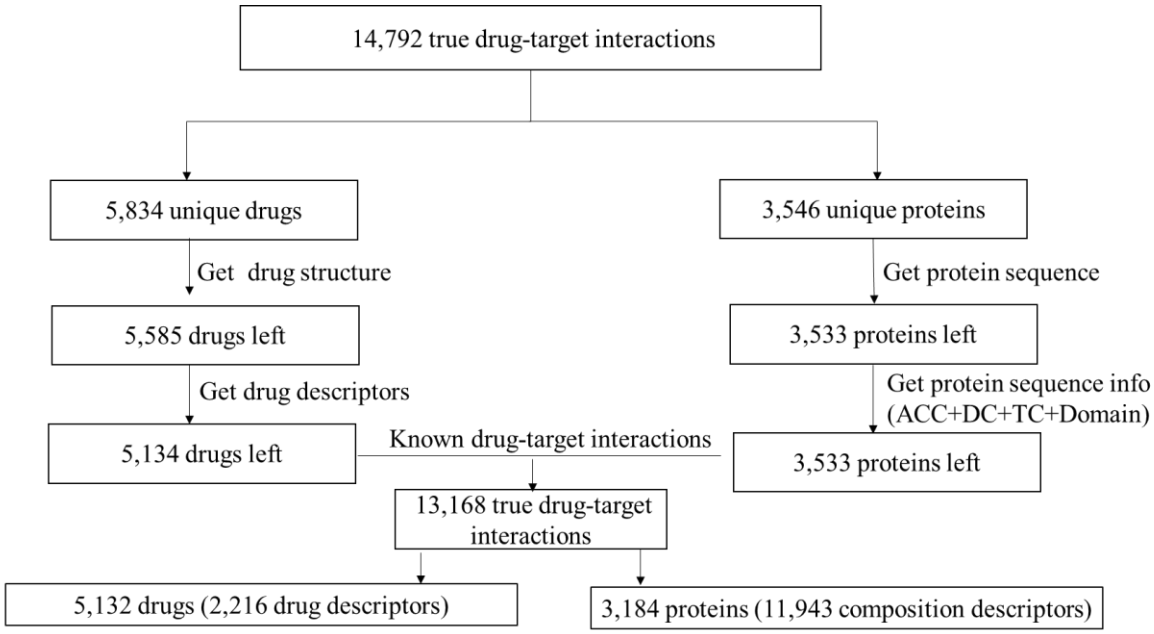


Figure 3 – Flowchart of feature generation. 13,168-known drug-target interaction pairs with 5,132 unique drugs and 3,184 unique proteins are generated as positive samples with drug and protein features. 18,125,254 are considered as unknown interactions. These can be potential negative or positive drug-target pairs.

3.2.3 Construction of negative drug-target interaction pairs

Although the large number of unknown interaction pairs are more likely to be negative since the number of no (negative) interaction pairs is far more than the number of interacted pairs [22], there are probabilities that some of these pairs interact. Hence, instead of randomly selecting negative samples from the remaining 18,125,254 unknown pairs, we proposed a novel method to select the most likely negative DTIs. The assumption of this method is based on

"guilt-by-association", which indicates similar drugs may share similar targets and vice versa [45]. Based on this intuition, we calculated the drug similarity measured by Tanimoto coefficient for each pair of the drugs using R package Rcpi and drugs with Tanimoto coefficients >60% were inferred as similar in this study [45]. We used a well-known method to map molecular structures with bit strings (i.e. molecular fingerprints). The Tanimoto coefficients between drugs for continuous variables as shown in **Equation 1** and dichotomous variables as shown in **Equation 2** are respectively defined by Bajuse et.al. [46]:

$$S_{A,B} = \frac{\left|\sum_{j=1}^{n} x_{jA} x_{jB}\right|}{\left|\sum_{j=1}^{n} (x_{jA})^2 + \sum_{j=1}^{n} (x_{jB})^2 - \sum_{j=1}^{n} x_{jA} x_{jB}\right|} \quad (1)$$

$$S_{A,B} = \frac{c}{a+b-c} \quad (2)$$

Here, S donates similarity between drug A and drug B, x_{jA} donates j-th feature of A and a is the number of on bits in molecule A. x_{jB} donates j-th feature of B and b is the number of on bits in molecule B, while c is the number of bits that are on in both molecules. n is the number of features.

It is believed that similar drugs are more likely to be functionally correlated, thus, targets that interact with one drug are also considered to interact with similar drugs, and these DTIs were referred as potential interacted pairs in our study as shown in **Fig. 4**. Protein similarities were also calculated using Protr R package[47]. They were derived by protein sequence alignment, which is a way to identify regions of similarity of proteins based on their functional, structural, or evolutionary relationships. After obtaining the similarities between proteins, the

same filtering rule was applied to find potential DTIs based on the protein similarity scores. That is, if the similarity score between two proteins (sequence identity score) is above 40% [45], it will be assumed that these proteins share the same interacted drugs. We followed the similar procedure as **Fig. 4** to identify potential DTIs by protein similarity. After we removed all potential positive DTIs, we randomly selected 13,168 negative DTIs (samples) from the remaining unknown DTIs.

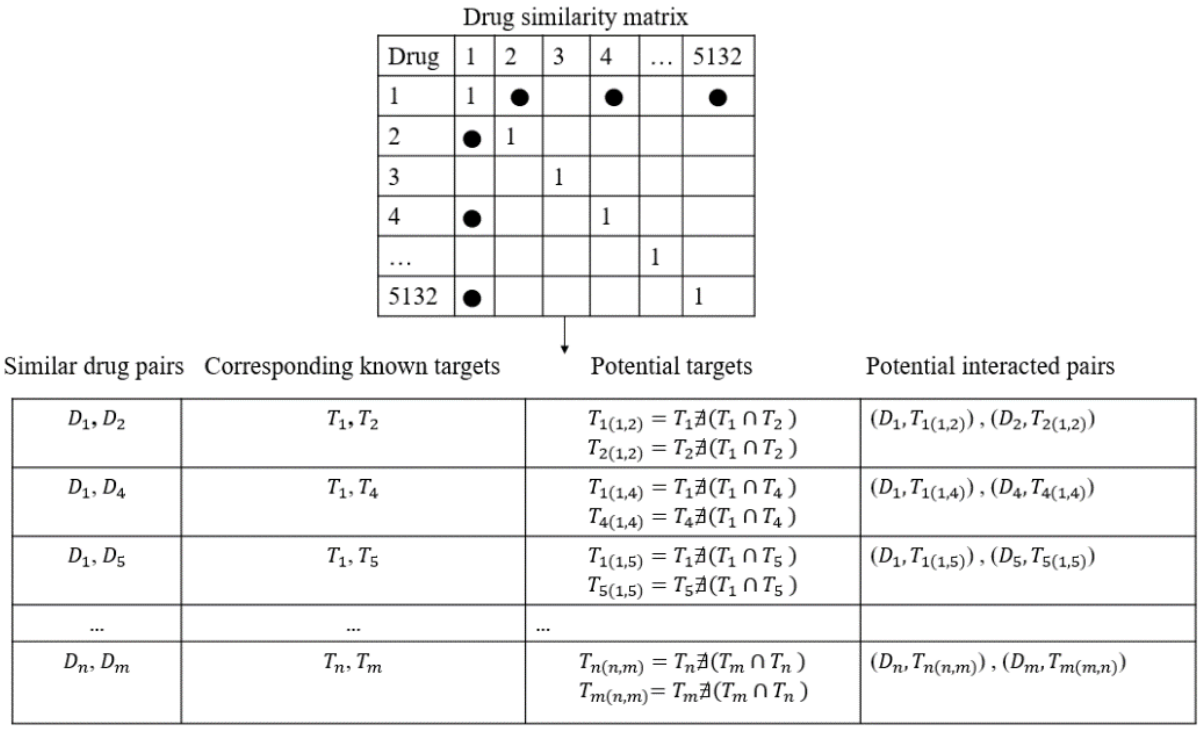


Figure 4 - Flowchart of finding potential DTIs by drug similarities. Bold dots in similarity matrix indicate the Tanimoto coefficients of the drug pairs are larger than 0.6. Thus, the corresponding drug pairs are referred as similar, such as D_1 and D_2, D_1 and D_4. D_n denotes drug n, T_n denotes target list of drug n, and $T_{n(n,m)}$ denotes potential target list of drug n emerged from the high similarity with drug m.

3.3 Logistic regression and LASSO regression

3.3.1 Model selection strategy

We split our samples (positive and negative DTIs) into 2/3 and 1/3 of all samples as our training and test datasets respectively. For each of the training and test sets, we ensured there were the same numbers of positive and negative DTIs. The training set was used for model selection and the test set was used for model performance evaluation.

3.3.2 Logistic regression and LASSO classification models

Standard logistic regression (SLR)-based models are widely used for binary classification problems. SLR takes information from training set and generates a linear pattern to a dependent variable with coded values of 0 or 1. LASSO model is an improved version of SLR. It shrinks the predictors automatically through statistical significance, which reduces over-fitting and tedious training time when high-throughput data are applied [48].

Using the constructed positive and negative DTIs to predict new DTIs, we implemented the SLR and LASSO-based regularized linear classification models. Briefly speaking, the aim of SLR is to find the optimized hyperplane as **Equation 3** that correctly classifies the positive and negative DTIs.

$$\mathrm{z} = \theta^T x \tag{3}$$

Here, x donates feature space and the optimized parameter vector θ fits a curve to separate the training data. To simplify the measurement of its output, we considered the sigmoid function as **Equation 4** to represent the probability of a DTI (y) given one sample x:

$$\mathrm{p}(y = 1|z) = \mathrm{g}(\mathrm{z}) = \frac{1}{1+e^{-z}} \quad (4)$$

Overfitting might be noticeable in the model analysis since we had more than 14,000 features representing drugs and proteins while there were a relatively small number of samples (positive and negative DTIs). It is necessary to select informative features for building the SLR-based classification model to alleviate the overfitting problem. Since the majority of our features were category variables, feature selection based on the training data set was implemented using Wilcoxon rank sum test. The test is a nonparametric statistical test that could be applied when the data population cannot be assumed to follow a normal distribution. It is a simple test ranking all samples (DTIs) from two classes (positive and negative DTIs) and obtaining sum ranks of each class. The significance of each feature can be measured by its P-value, which is the probability of obtaining the estimated parameter θ value under the null hypothesis of θ=0. Thus, to decrease the effect of those less significant features (measured by P-value mentioned above), the improved logistic regression LASSO model was also implemented. It reduces the effect of features who are referred as "not significant" estimated by their P-values. In other words, only features considered "significant" could have contributions on the optimized hyperplane. The decision boundary is now added a penalty term to realize the assumption that the coefficients of those not crucial features are set as zeros as shown in **Equation 5**.

$$z = \theta^T x + \lambda \sum|\theta| \quad (5)$$

Here, λ is selected to minimize the output of sample errors by gird search using cross-validation technology.

In our study, we applied SLR models with the top 100, 200, 500, 1,000, 2,000 of all 14,159 features selected by Wilcoxon rank sum test respectively. We also tried to use more than 2,000 features in the SLR models, but the models could not be fitted due to the large number of features and small number of samples. We implemented 5 LASSO models with all 14,159 features, drug features alone with one single protein feature (ACC/DC/TC/DOMAIN) respectively. Feature vectors for each of the five LASSO models (M1, M2, M3, M4 and M5) were represented as:

M1: $[D_1, \dots, D_{2216}, P_{ACC1}, \dots, P_{ACC20}, P_{DC1,} \dots P_{DC400}, P_{TC1}, \dots P_{TC8000}, P_{domain1,\dots}, P_{domain3523}]$

M2: $[D_1, \dots, D_{2216}, P_{ACC1}, \dots, P_{ACC20}]$

M3: $[D_1, \dots, D_{2216}, P_{DC1,} \dots P_{DC400}]$

M4: $[D_1, \dots, D_{2216}, P_{TC1}, \dots P_{TC8000}]$

M5: $[D_1, \dots, D_{2216}, P_{domain1,\dots}, P_{domain3523}]$

We built these models using R packages e1071[49] for SLR and glmnet[48] for LASSO based on the training set. For LASSO models, we performed 10-fold cross-validation of the training set to select the optimized parameter λ.

3.4 Deep neural network

3.4.1 Introduction

Deep neural network is a kind of artificial neural network (ANN) with more hidden layers and more complex computations between input and output layers, thus are made to recognize more complicated patterns. The basic units of deep neural network are neurons, which are supposed to find patterns just like in human brains do. Deep neural networks have associated many optimized algorithms such as Adaptive Moment Estimation (Adam)[50] and techniques such as dropout[51]. These have been commonly used in computer versions applications like image classification and bioinformatics like compound-protein interaction prediction [52].

A DNN architecture includes an input layer with the same number of neurons as features, several hidden layers, and an output layer (**Fig. 5**). Here, $a_i^{(j)}$ stands for the activation of neuron i in layer j and W_{ik}^{L} means the weigh coming from neuron *k* in layer L to neuron i in layer L+1. In multiclass classification problem, we usually have the same number of neurons in the output layer as the same number of classes assigned for classification. For example, if we have 10 classes to be classified, the initial result $y^{(j)}$ should be defined as {1,0,0,0…0}, {0,1,0,0…0}, {0,0,1,0…0} …, {0,0,0,0…1}, the index of label "1" defines which class the sample belongs to.

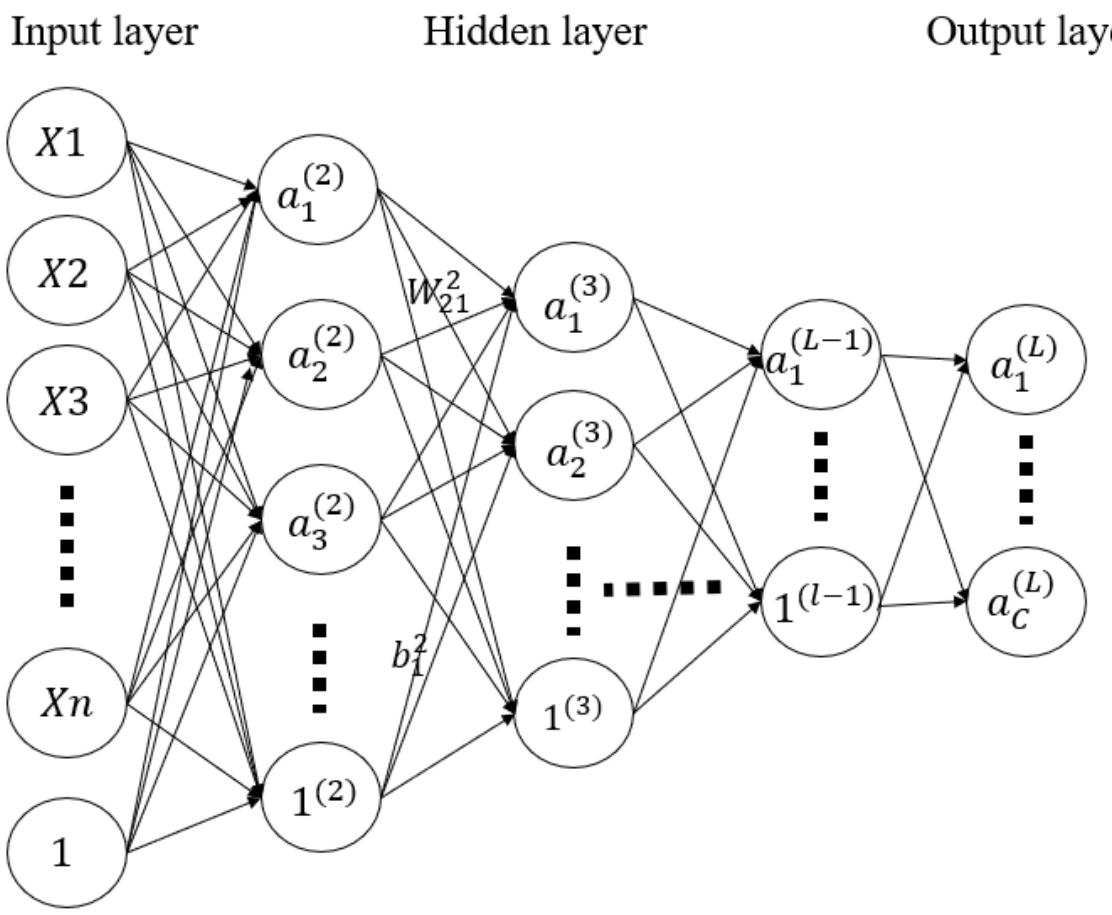


Figure 5 - The architecture of a DNN.

In this study, instead of using traditional gradient descent for fitting the model, we chose to use mortified Adaptive Moment Estimation (Adam) optimizer[50]. The Adam algorithm is basically a combination of momentum and Root mean square prop (RMSprop) optimization methods. The momentum term is widely used in various neural network architectures to improve model performance. The momentum term allows the current direction to be changed somehow based on the previous directions but not totally determined by gradient descent, which may help the model get away from local minima. The percentage of current direction influenced by previous paths and descent will be implemented through a new parameter: the momentum term. Adding the momentum term allows us to change the proportion of decrease coming from a

descent calculation, which can be controlled by the performance on the training set. Meanwhile RMSprop optimization deceases the chance of high oscillations in training performance and helps to learn the patterns faster.

Two moments β1 and β2 are applied in Adam, corresponding to the mentioned momentum optimization and RMSprop optimization respectively. The former is to compute the mean of the derivatives while the later is to compute exponentially weighted average of the squares. On each iteration, the derivatives of weighs dW and bias db are still calculated using gradient descent, and the general difference is that we calculate the momentum exponentially weighted average just as in momentum optimization and a RMSprop update as V follows **Equation 6-7** and S follows **Equation 8-9** respectively, then weights will be upgraded finally based on prepared V and S **Equation 10-11.** Normally, a very small unit ε will be added for mathematical convenience.

$$V_{dw} = \beta 1 * V_{dw} + (1 - \beta 1) * dW \tag{6}$$

$$V_{db} = \beta 1 * V_{db} + (1 - \beta 1) * db \tag{7}$$

$$S_{dw} = \beta 2 * S_{dw} + (1 - \beta 2) * dw^2 \tag{8}$$

$$S_{db} = \beta 2 * S_{db} + (1 - \beta 2) * db^2 \tag{9}$$

$$\mathrm{W} = \mathrm{W} - (\text{learning rate}) * \frac{V_{dw}}{\sqrt{S_{dw}} + \varepsilon} \tag{10}$$

$$\mathrm{b} = \mathrm{b} - (\text{learning rate}) * \frac{V_{db}}{\sqrt{S_{db}} + \varepsilon} \tag{11}$$

3.4.2 Model selection strategy

We split the whole samples (positive and negative DTIs) into 2/3 and 1/3 of all samples as our training and test datasets respectively. For each of the training and test sets, we ensured there were the same numbers of positive and negative DTIs. The training set was used for model selection and the test set was used for model performance evaluation.

3.4.3 Deep neural network-based classification model

3.4.3.1 Feature extraction for deep neural network-based classification model

To build the deep neural network-based classification model, we first extracted drug- and protein-specific features in the training data, which had association with the DTI status using LASSO models. To select significant representations of training samples, we investigated LASSO regression on training data and generated meaningful features. We built two models on the training data using R packages glmnet[48] for LASSO regression based on the drug feature set and protein feature set respectively. The features of these two models are as follows:

$$\text{M6:}\left[P_{ACC1}, \dots, P_{ACC20}, P_{DC1,} \dots P_{DC400}, P_{TC1}, \dots P_{TC8000}, P_{domain1,\dots}, P_{domain3523}\right]$$

$$\text{M7: } [D_1, \dots, D_{2216}]$$

We performed 10-fold cross-validation to explore the optimized fit on the training set, and the features that show statistical significance were recorded for deep neural network implementation as illustrated in **Fig. 6**.

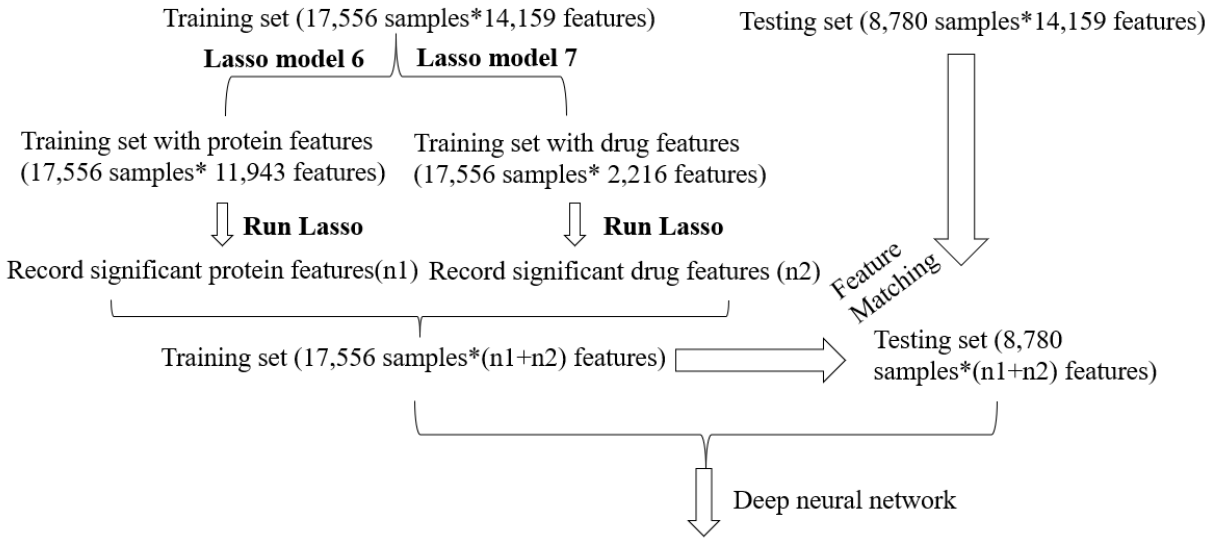


Figure 6 - Feature selection using LASSO regression on the training set. Two models are applied using drug features and protein features respectively. n1 features of drugs and n2 features of proteins are collected through significance filtering. Data with matched (n1+n2) features are generated before feeding into deep neural network model.

3.4.3.2 Model design

The DNN model architecture in our study was presented in **Fig. 7.** We kept features that showed significance in LASSO regression models in the training set and matched these features in test set. We used relu, sigmoid and softmax as activation functions in different layers of the network and the information was dropped out 20% randomly from input layer to the first hidden layer, and from the first hidden layer to its next. Dropout was used to avoid overfitting. It generally drops information randomly from certain layers to avoid obtaining a too close of a fit of the training data, which would otherwise reduce the capacity for generalization.

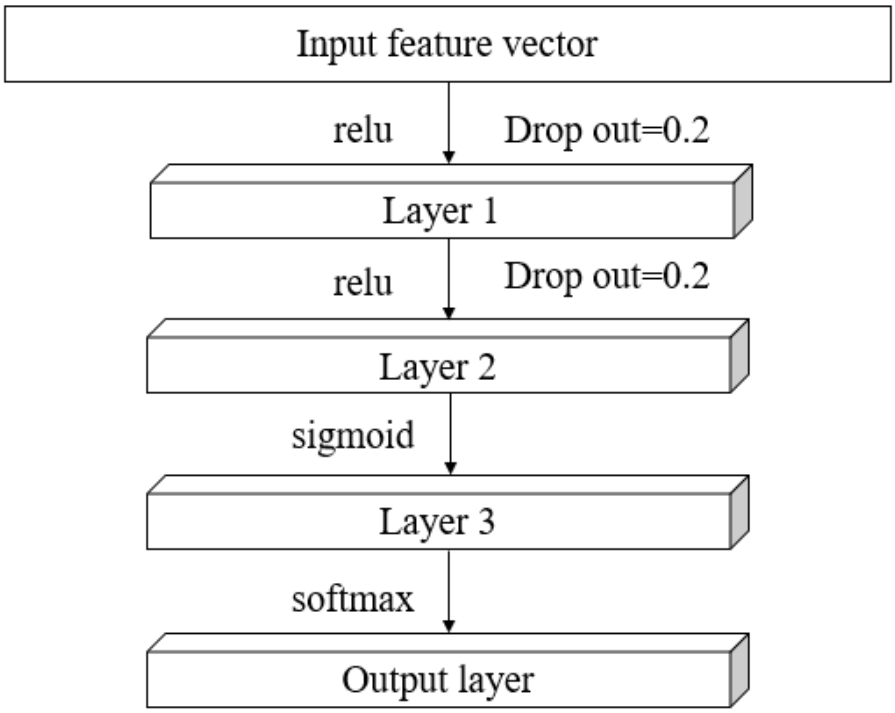


Figure 7 - The architecture of deep neural network in this experiment. Three hidden layers are used in the network with relu, sigmoid and softmax activation functions as well as the dropout technique. The output layer outputs the probability of the classification.

3.4.3.3 Software and model parameters

In deep neural network implementation, we used the TensorFlow [53] library in python. We concatenated significant features extracted from the two LASSO regression models fitted for the drug-specific and protein-specific features respectively. We trained our network using a learning rate=0.00001, a dropout rate=0.8 in first two layers and L2 regularization. We reported the best results after compiling the models with various sets of parameters.

3.5 Model performance evaluation

In our implementations of the SLR, LASSO and DNN models, we used two types of measurements to evaluate the performance of the classifiers.

We computed the overall accuracy, which is the percentage of drug-target pairs with correctly predicted interactions. We also measured Receiver Operating Characteristics (ROC) curve, which illustrates the pattern of sensitivity (1-FNR (False negative rate)) and specificity (1-FPR (False positive rate)) at different cut-offs of the probability to predict a DTI to be an interacted pair. AUC was calculated based on the ROC curve for each model to describe the quality of the classifiers.

3.6 Results

3.6.1 Standard logistic regression and LASSO results

3.6.1.1 Standard logistic regression results

We fitted the logistic regression models using the top 100, 200, 500, 1000 and 2000 features selected by Wilcoxon rank sum test based on the training data. **Table 1** showed the prediction accuracies and AUCs of our test set based on the trained models. Overall, models with more selected features had greater power to predict DTIs. The model had the best prediction performance when 2000 features were used in this study. We also tried models with more than 2000 features, but the models could not converge properly. In general, the model performance decreased when the number of features used in the model decreased.

Table 1. Overall accuracy and AUC of the SLR models of the test set.

Logistic regression model	Numbers of top features selected by Wilcoxon rank sum tests				
	100	200	500	1000	2000
Accuracy	0.64	0.66	0.70	0.72	0.75
AUC	0.70	0.73	0.77	0.80	0.81

3.6.1.2 LASSO regression results

3.6.1.2.1 Parameters selection in training data

Figure 8A-E showed the AUCs of the five models on our training set based on 10-fold cross-validation. The largest AUCs were achieved when X-axis hit points which were called lambda_min. Instead of using lambda_min, we used lambda_1se to fit the models. The Lambda_1se we selected was the largest value of lambda such that AUC was within 1 standard error of the maximum AUC. **Figure 8F** showed the results based on Lambda_1se for the five models. Although the Lambda_1se did not result in the best performance in training set, it considered model uncertainty to make more reliable predictions for the testing set.

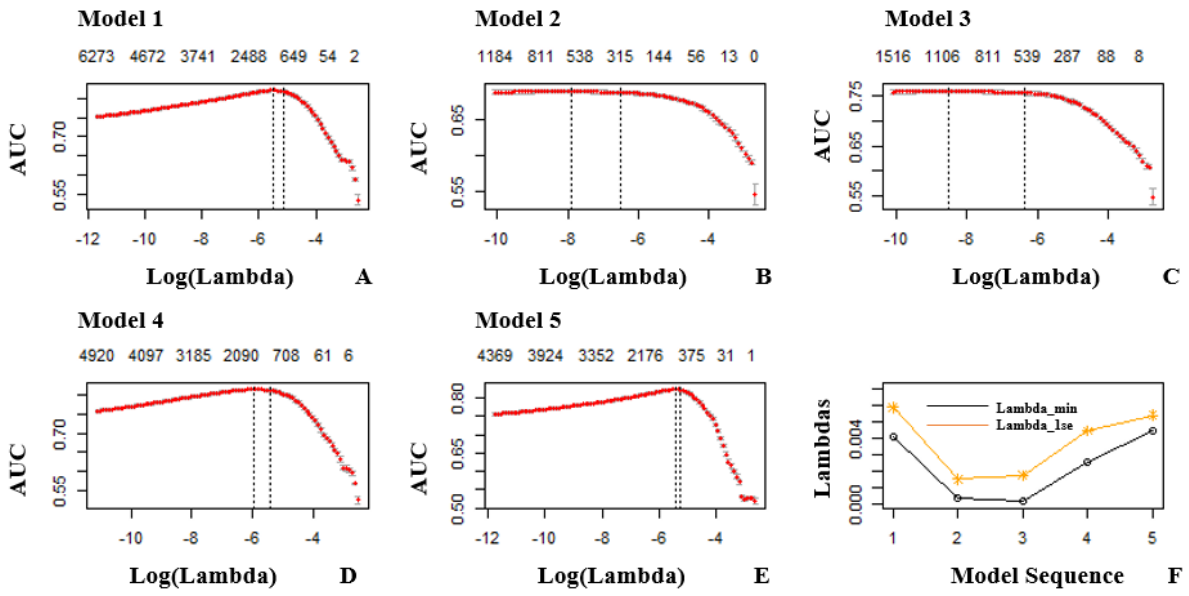


Figure 8 - Lambda selection in 5 models in training. A-E showed AUCs on training data using various lambdas for five models. In this experiment, the largest value of lambda such that AUC was within 1 standard error of the maximum AUC (Lambda_1se) is selected to avoid overfitting issues. F shows the values of Lambda_1se and Lambda_min for five models.

3.6.1.2.2 Performance evaluation of the LASSO models

Given the Lambda_1se we selected for the five models, **Table 2** showed the accuracies and AUCs of the five LASSO models in the training set based on 10-fold cross-validation and the test set. **Figure. 9A** and **Fig. 9B** showed the ROCs of the five models for the training set based on 10-fold cross-validation and the test set respectively. In general, the best model was Model 1 with all five feature sets. Model 2 and Model 3 with fewer numbers of features seemed weaker than other models. However, the model performances using protein TC information (Model 4) and protein domain information (Model 5) were remarkably improved. Better results were obtained from the testing set than training set, which indicated we avoided over-fitting perhaps due to the model lambda selection.

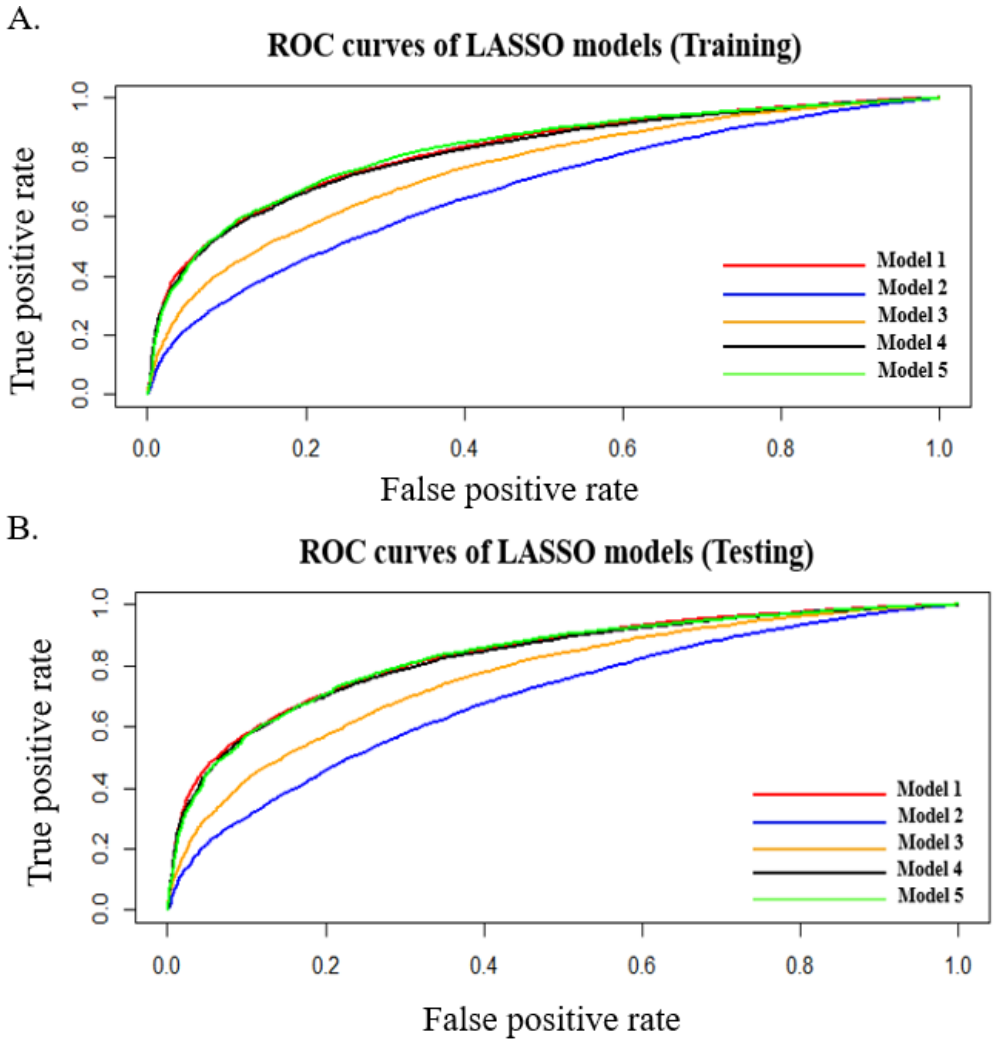


Figure 9 - ROC corves of LASSO models for the training and test set of the five models. Panel A shows ROC corves of the 5 models in the training data Panel. B shows ROC corves of the 5 models in the test data.

Table 2. Accuracies and AUCs of the LASSO models for both training and testing set in five models.

		M1	M2	M3	M4	M5
10-fold cross-validation of training set	Accuracy	0.75	0.63	0.69	0.74	0.74
	AUC	0.82	0.69	0.76	0.81	0.82
Test set	Accuracy	0.75	0.64	0.70	0.75	0.75
	AUC	0.83	0.69	0.77	0.83	0.83

3.6.2 Deep neural network results

3.6.2.1 Feature selection in the training data

To build the deep neural network, we first screened protein- and drug-specific features using the LASSO regression models on the training data, which were named as Model 6 (M6) and Model 7 (M7), respectively. We measured the accuracies, AUC and number of significant features based on different selected lambdas. As shown in **Fig. 10**, The upper half of the graph **(Fig. 10A,10B)** shows the lambdas tested in Models 6 and 7. The best AUC occurs when Lambda_min is applied. However, to ensure the flexibility of the model and avoid overfitting, sometimes lambda_1se was considered. These two mostly used lambdas for both models were checked as shown in **Fig. 10D** and **Fig. 10C,** which illustrate the minor differences in accuracy on training set using these two lambdas. **Table 3** showed the two models' accuracies, AUC and number of significant features based on Lambda_min and Lambda_1se.

We also observed the ROC curves as shown in **Fig. 11** on the training set using different lambdas for both models, lines are extremely close to each other using Lambda_min and lambda_1se for both models. In this experiment, we decided on using Lambda_min in both models for feature selection in view of larger number of significant features obtained while the differences of accuracy were very small for lambda_min and lambad_1se in the two models.

Table 3. Accuracies, AUCs and the number of significant coefficients (features) of the LASSO models in training set for both models

Performance on training using LASSO		Model 6	Model 7
Lambda_min	Accuracy	0.716	0.623
	AUC	0.754	0.659
	Number of significant features	1078	883
Lambda_1se	Accuracy	0.714	0.620
	AUC	0.752	0.657
	Number of significant features	867	596

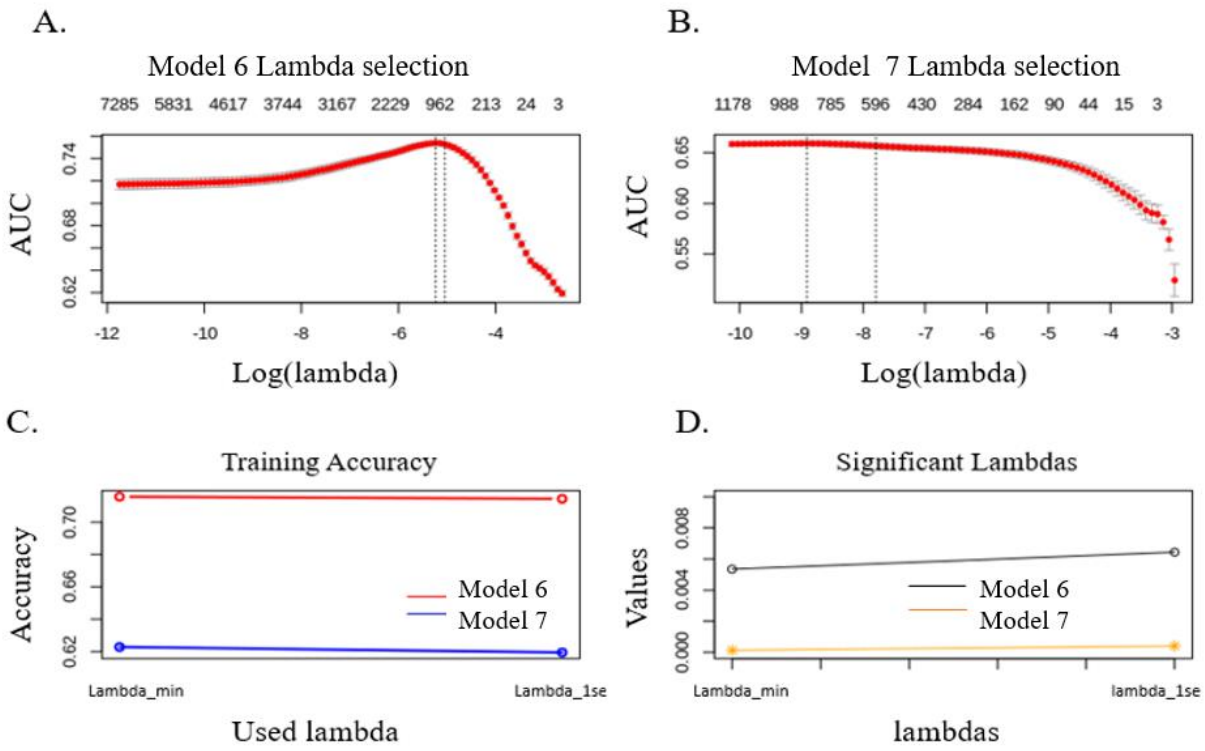


Figure 10 - Lambda selection on the training data. Panel A and B: We tested various lambdas for each model, and the highest AUC were achieved using Lambda_min. Panel C: Tiny differences on accuracies were observed using either Lambda_min or Lambda_1se for both models. Panel D: Values of significant lambdas for both models.

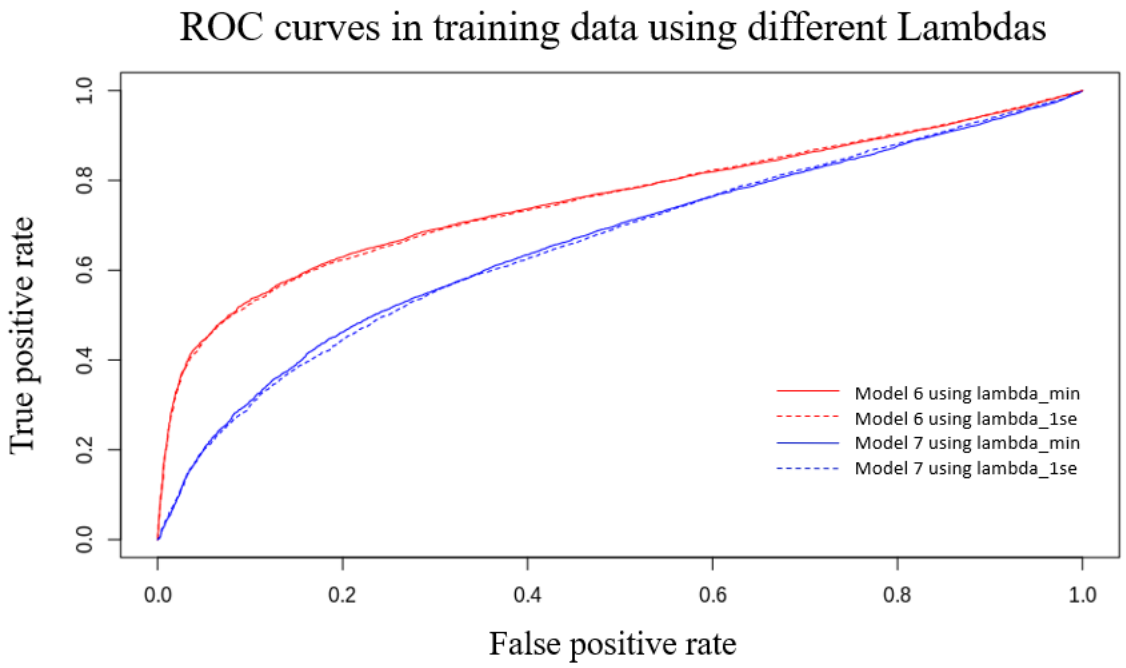


Figure 11 - ROC curves using different lambdas for both models. Minor differences were observed on the ROCs using different Lambdas for both models.

3.6.2.2 Performance evaluation of the deep neural network model

For the 883 drug- and 1078 protein-specific features extracted from the LASSO models, we concatenated them as the input feature set for our DNN model. We trained the network using a learning rate=0.00001, and a dropout rate=0.8 in the first two hidden layers and L2 regularization, and achieved 0.99 AUC value on the training set. The test set was first matched with the extracted features before feeding into the trained DNN model. The accuracy and AUC of the trained DNN model on test data are 81% and 0.89, respectively. The ROC curve on testset is presented in **Fig. 12**.

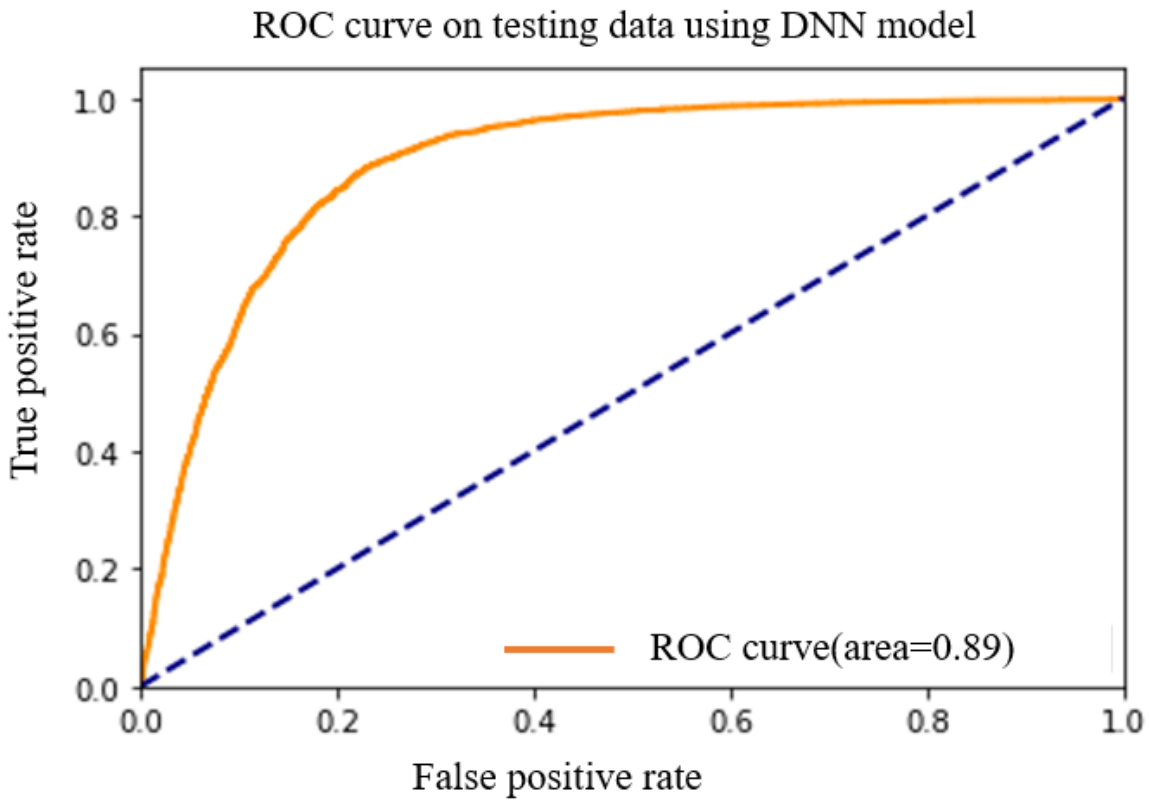


Figure 12 - ROC curves on the test data using DNN model. Area under curve (AUC) is 0.89.

We explored the prediction power of the training DNN model on known positive samples extracted from the training data. We observed the performance on known positive samples with various cut-offs ranging from 0.8 to 0.9999 as shown in **Fig. 13**. Starting from 0.8, we saw that most of the pairs were correctly predicted to be positive (11,943 out of 13,168). However, with the value of the cut-off approaching to 1, the number of correctly categorized training samples decreases notably especially when the cut-off is larger than 0.99.

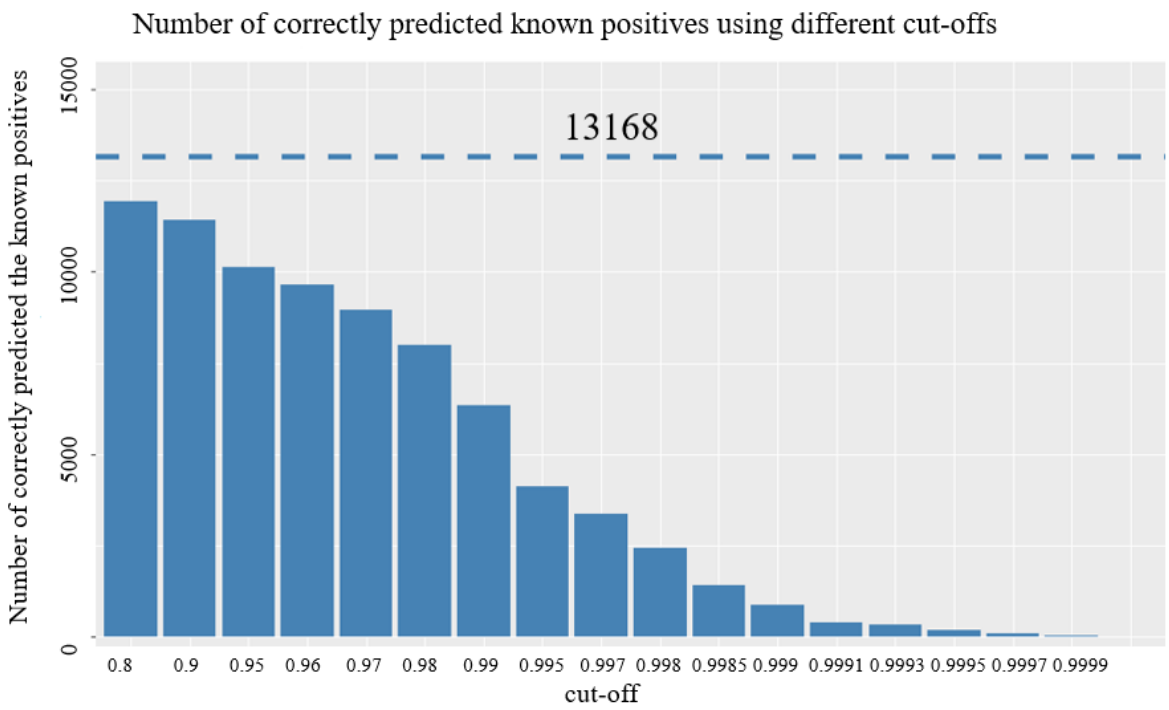


Figure 13 - Number of correctly categorized positive samples using trained DNN model verses different cut-offs.

3.6.2.3 Prediction results of the trained deep neural network model

Using the trained DNN model, we predicted all possible 18,138,422 DTIs combined from the 5,134 drugs and the 3,533 proteins. The prediction results were summarised using various cut-offs as shown in **Fig. 14**. Starting at 0.8, more than 2,000,000 drug-target pairs are predicted as positive samples and a decrease pattern is observed. The slope of decay slows down from 0.8 to 0.98. However, a large reduction is seen after 0.98. A relatively stable trend could be observed if a cut-off is set larger than 0.995.

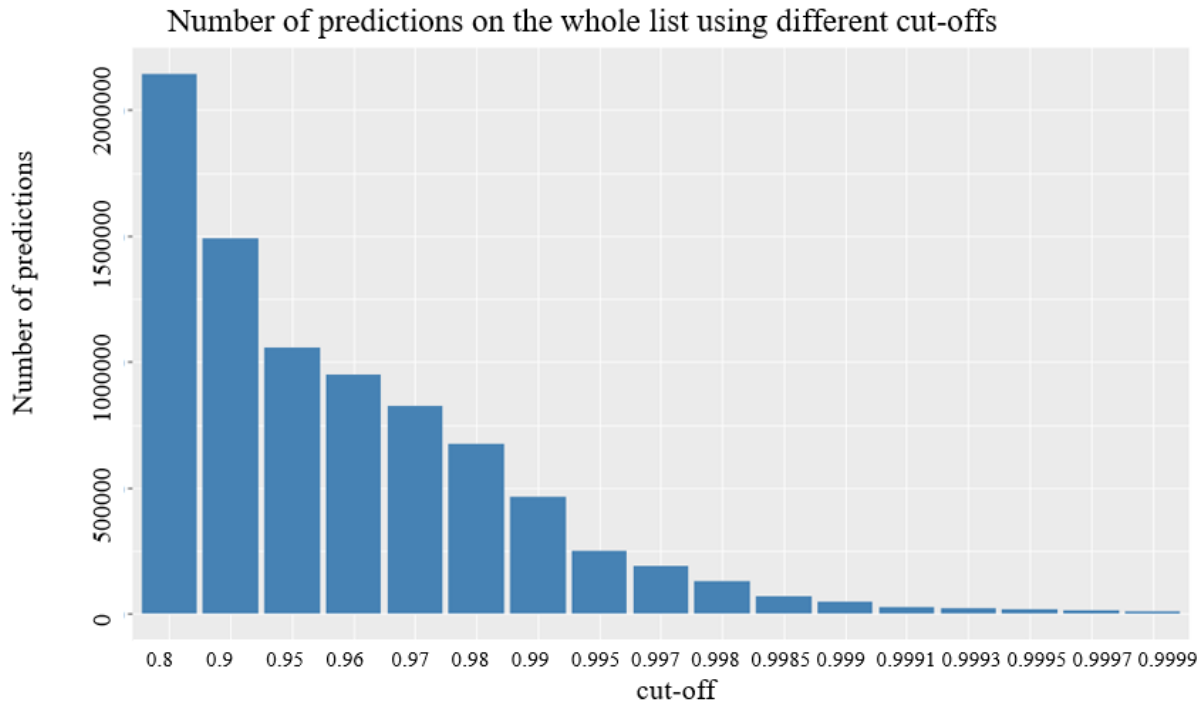


Figure 14 - Number of predictions on whole DTIs at different confidence levels.

The percentage trends on the predicted pairs provide another perspective of the prediction quality. **Fig. 15** shows the proportion of correctly predicted known positives and predicted candidate drug-target interactions using various cut-offs. As we can observe from the red line, at least half of the known positive pairs can still be predicted as positive interactions using cut off larger than 0.99, and the number of correctly predicted known interacted pairs drops rapidly if the cut-off is set larger than 0.9985. Also, it is not hard to find that very few number of known pairs are predicted as interaction given a relatively large cut-off, which also is supported by the fact that the number of non-interacting is far more than the number interacting. It is noticeable that the percentage decreases to somewhere below 0.1% when cut off is larger than 0.9993.

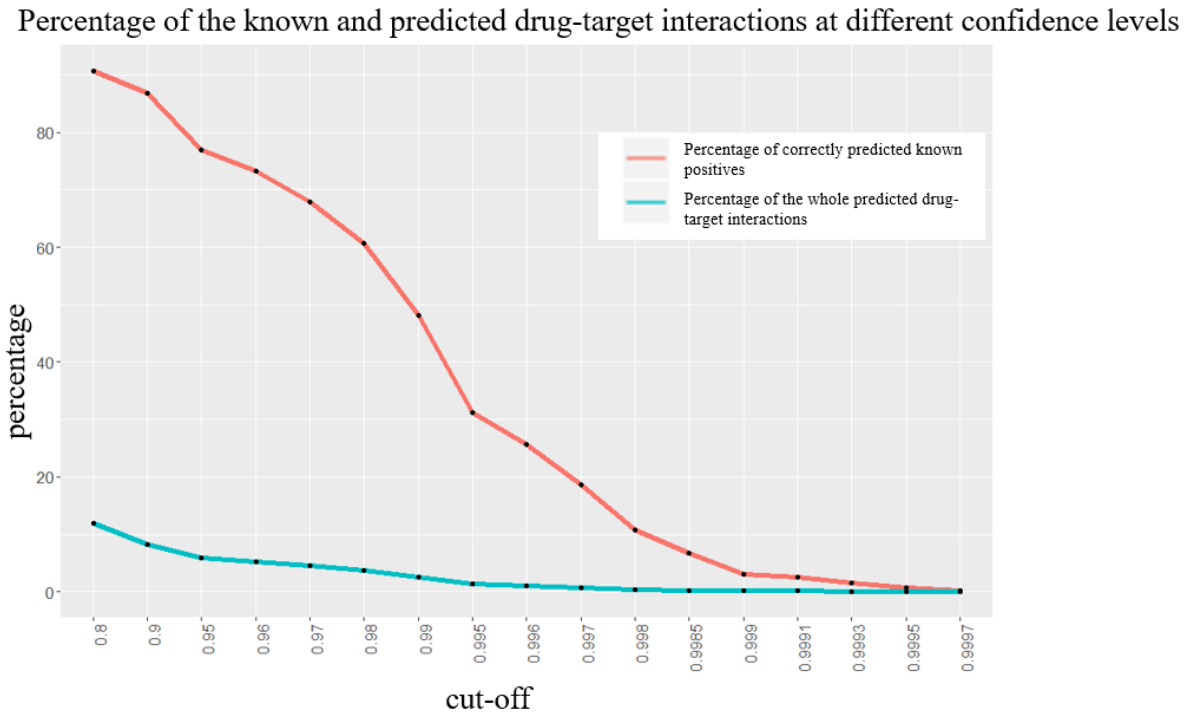


Figure 15 - Percentage of the known/predicted interaction pairs at various cut-offs. Most of the known positive pairs (60% of positive samples) can be correctly predicted using a cut off larger than 0.98. The results are as expected since the number of non-interacting pairs should be far more than the number of interacting pairs.

3.7 Conclusion and discussion

In this experiment, we first applied SLR and LASSO regression to predict DTIs. Our experiment showed that the LASSO-based regularized linear classification models achieved better results than the SLR models for predicting DTIs. We also demonstrated that the LASSO models can work well for data sets with a very large number of features, which often result in overfitting in the SLR models. We explored the prediction power of different protein feature sets and showed that the best performance can be reached by integrating all feature sets. Furthermore, we proposed a deep neural network model based on features extracted from the LASSO regression models fitted using the protein-specific and drug-specific features respectively. We showed that the deep neural network model outperformed logistic regression and LASSO models.

We predicted all the whole combined drug-target pairs from drug and protein list using the trained DNN model, the results confirmed that the number of non-interacting pairs are far more than interacting pairs.

Chapter 4 Repurposing drugs using the predicted DTI network

4.1 Introduction

DTI prediction is a vital component in drug repurposing. Based on the predicted interaction network, various methods could be applied to identify potential drug candidates. Some studies validated the promising candidates by experimental tests. For example, Cheng et.al tested 40 old drugs repurposed from predicted DTIs by in vitro assays. They tested the antiproliferative activities using one human breast cancer cell line. Drugs such as Itraconazole showed a higher antagonistic activity with IC50 on ERβ than tamoxifen. Simvastatin and ketoconazole also showed potent antiproliferative activities, making them promising treatments for breast cancer[14]. Additionally, Luo et.al proposed an integrated network-based approach to predict DTIs and the top-ranked predictions were also tested by experiments[45]. They validated the bioactivities of the COX inhibitors and the experimental results suggested 3 drugs, which were interacted with COX proteins, have shown positive effects on anti-inflammation.

Furthermore, some studies repurposed drugs from the DTI network by the rank of interaction probability. For example, Wen et al. applied DBN to identify DTIs and the predictions were ranked by the interaction probabilities [22]. They found 4 of the top 10 interactions supported by other literature and publications. Such as the drug Drospirenone which exhibits a low relative binding affinity to protein glucocorticoid receptor. The authors suggested their novel predictions add practical useful insights on drug repositioning. In this experiment, we first generated the top ranked DTI predictions and identified the most likely potential drugs for

IBD and breast cancer by the interaction frequency with risk targets, then we applied a clustering method to explore the most reliable predictions by identify the enriched clusters with disease-specific risk genes and drugs.

4.2 Data and material

4.2.1 Data collection

We used the top ranked drug-target interactions from the DNN prediction set with cut-off=0.9995, which included less than 0.1% prediction pairs. It contains 14,960 interacted drug-target pairs with 3,082 drugs and 3,514 proteins. This is a similar number of the interactions used in our positive set for the predictions. To decrease the effect of possible false positive pairs, we proposed a test to evaluate the proportion of interactions each drug and protein has in the prediction set as well as in the known positive samples, respectively. After excluding the potential false positive prediction pairs by the proportion-based testing, we collected 5,921 predicted interaction pairs with 1,633 unique drugs and 608 unique proteins for clustering analysis in this experiment.

4.2.2 GWAS risk genes of IBD and breast cancer

4.2.2.1 Inflammatory bowel disease (IBD)

Inflammatory bowel disease (IBD), including Crohn's disease (CD) and ulcerative colitis (UC), is a severe, incurable gastrointestinal illness that has chronic inflammation. These two most common forms of IBD may affect different parts of GI tract. Generally speaking, CD may

affect from mouth and anus to skin and liver while UC only have influence on the former. IBD greatly affects patients' quality of life involving severe abdominal diarrhea, pain, fatigue and gastrointestinal bleeding[54]. Approximately 1.5 million people have IBD in the United States and Canada, where the rates are among the highest in the world, and an estimated 17,000–93,000 new patients are diagnosed each year. Researchers found that inflammatory bowel disease is highly associated with some genes. Linkage studies have implicated several genomic regions as likely containing IBD susceptibility genes, with observed uniquely in Crohn's disease (CD) or ulcerative colitis (UC), and others common to both disorders such as CD susceptibility gene NOD2/CORD 15 can be observed on chromosome 16q[55]. To identify better treatments and individual patient care, it is critical to build computational models for IBD.

Katrina et al. identified 25 new susceptibility loci after 215 risk loci were found in previous GWAS [56]. Their study collected data on 25,395 individuals and conducted a meta-analysis with published data to perform a genome-wide association study. Three of newly identified loci contain integrin genes that encode proteins in pathways which are already recognized as significant targets in IBD. Their study has shown the potential of pursuing GWAS study and contributed promising targets for further therapeutics. For the 240 risk variants, Katrina et al. prioritized 811 IBD risk candidate genes [56].

4.2.2.2 Breast cancer

Breast cancer is the most common malignant cancer in women worldwide. It has been acknowledged that family history is one big factor causing breast cancer in women, however, the estimated of inherited components of breast cancer are not established. Approximately 100

breast cancer risk loci were identified in GWAS studies to date. In order to translate these findings into a greater understanding of the mechanisms that influence disease risk, it requires identification of the genes or non-coding RNAs that mediate these associations. Baxter et al.[57] used Capture Hi-C (CHi-C) to annotate 63 of the established breast cancer risk loci. They identified CHi-C interaction peaks involving 110 putative target genes mapping to 33 loci and demonstrated long-range interaction peaks some of which spanning megabase (Mb) distances and involving adjacent risk loci. Moreover, the breast cancer risk variants identified in genome-wide association studies explain only a small fraction of the familial relative risk, and the genes responsible for these associations remain largely unknown. To identify novel risk loci and likely causal genes, Wu et al. [58] performed a transcriptome-wide association study evaluating associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry. They used data from the Genotype-Tissue Expression Project [59] to establish genetic models and predict gene expression in breast tissue. They evaluated model performance using data from The Cancer Genome Atlas [60]. In total, 8,597 genes were evaluated, they identified 179 whose predicted expression was associated with breast cancer risk at $P < 1.05 \times 10\text{-}3$, a false discovery rate (FDR)-corrected significance level. In the experiment here, we combined the identified risk genes of Baxter's study [57] and Wu's study [58], for a total number of 288 risk genes of breast cancer which was then used for drug repurposing.

4.3 Drug repurposing using the predicted DTIs

Using the predicted DTI network, we repurposed the potential candidate drugs for IBD and breast cancer directly. We extracted the predicted drug-target pairs from the predicted

network that contained the IBD and breast cancer risk genes respectively. We calculated the frequency of each drug interaction with the IBD/breast cancer risk genes based on these extracted DTIs. We repurposed the top drugs ranked by frequency for both diseases.

4.4 Clustering DTIs and identifying disease causal gene-specific clusters

Using the predicted DTIs obtained by the trained DNN model, we performed clustering analysis. The drugs and proteins were treated as two types of nodes while the interactions were presented as edges. We identified the enriched clusters for certain risk gene groups of IBD and breast cancer using a bipartite clustering algorithm.

4.4.1 Introduction

Clustering analysis explores data from an intuitive perspective. It assigns massive nodes into groups in a network. Generally speaking, the aim of clustering analysis is to identify several clusters based on a known dataset so that within each identified cluster, the similarity of data objects is more significant than that of other clusters. Cluster models are widely applied in different fields such as bioinformatics and image analysis.

Network-based methods are applied widely in repurposing drug candidates owing to their simplicity and effectivity. For instance, a study in 2013 [61] reported a computational method based on heterogeneous network clustering, and the network was built between drugs and diseases based on interaction information of disease-gene and drug-target from KEGG[62]. They considered genes, biology process, pathway and phenotype or combinations of these as

features for drugs and diseases respectively. In this network, drugs and diseases were represented as nodes and the shared features denoted as edges. The similarity between two nodes was measured by Jaccard score and they applied two graph clustering algorithms ClusterONE and Louvain's modularity to this weighted heterogenous network. Ninety-eight clusters with repurposed 11,160 disease-drug pairs were generated from Louvain clustering algorithm while 110 clusters with 2,518 repositioned drug candidates were identified using ClusterONE. Newly extracted drug-disease pairs were then matched to case studies in clinical trials and the literature.

4.4.2 Bipartite clustering

A bipartite network is a term when nodes represent two distinct types, such that interactions can only occur between nodes of different types [63]. Examples of bipartite networks are plant-pollination interaction networks, movie-actor networks in social networks, networks between text terms and book category and drug-target interaction networks. It has been observed that some science networks in respect of society and computation are divided naturally into modules [64]. It has been realized that to categorize a bipartite network into small communities, also denoted modularity, is a challenging issue. However, bipartite networks are quite random like non-divisible networks, that is to say, the structure of bipartite networks relies heavily on sorting pairs of nodes of different types. For example, the plant-pollinator network, as illustrated in MATLAB package BiMat [65], was showed to have the propertires of (i) modular, subsets of nodes highly connect to each other compared to other nodes [66]; (ii) nested, the interaction between nodes usually could also be considered as subsets of each other [67]; (iii) multi-scale, whether the whole network is used results in different conclusions of structural property-based on different sorting of the same network [68].

4.4.3 Algorithms for clustering DTIs

In this study, we used the MATLAB version of BiMat package [65] to analyze the clusters in our DTI network. The package is mainly used to tackle the issues of bipartite network modularity and nestedness. Its features include the calculation and evaluation of the modularity of a network, the statistics behind the clusters, the statistics of internal modules and visualization of the network layout.

The modularity of a cluster represents the density of connected nodes within the network. From which, we could select a partition of the interactions in the network into clusters. There is an abundance of ways to define the modularity of a bipartite network. For example, for a drug-target network B with two node sets D (drug nodes) and T (target nodes), a corresponding bipartite adjacency matrix can be created to represent the edges between the two node sets with size m*n, where m is number of rows (drugs) and n is number of columns (targets), the number of interactions for each node is then presented as d_i for drug i and t_j for target i respectively in **Equation 12**:

$$d_i = \sum_i B_{ij} \qquad t_j = \sum_j B_{ij} \tag{12}$$

The whole network interaction can be defined as **Equation 13**:

$$\mathrm{E} = \sum_{ij} B_{ij} \tag{13}$$

Here, B_{ij} indicates the entry (i, j) of the bipartite adjacency matrix. Then the standard measure of modularity by Barber [69] is defined as **Equation 14**:

$$Q_b = \frac{1}{E}\sum_{ij}(B_{ij} - \frac{d_i t_j}{E})\delta(g_i, h_j) \quad (14)$$

Where the indexes of node *i* (drug) and j (target) are expressed as g_i and h_j. To maximize the modularity, BiMAT library provides three options for maximizing the equation, AdaptiveBrim, LPBrim and LeadingEigenvector. Here, we will discuss the default algorithm-AdaptiveBrim. The standard BRIM (Bipartite Recursively Induced Modules) uses the above equation in a matrix form as in **Equation 15**:

$$Q_b = \frac{1}{E}\, TrR^T \widetilde{\mathrm{B}} T \quad (15)$$

Here, the modularity matrix $\widetilde{\mathrm{B}} = B_{ij} - \frac{d_i t_j}{E}$, and the R(m*c) and T(n*c) are drug nodes and target nodes index matrix $\mathrm{R} = [d_1|d_2|\ldots|d_m]$, $\mathrm{T} = [t_1|t_2|\ldots|t_n]$, where c donates the number of modules and d_i and t_j are ones and zeros indicating if the node belongs to the specific module. Using a pre-set value c, this algorithm assigns nodes of one single type (say drug nodes) to maximize Q_b and the iteration stops when hits a local maximum. We will start with c equals to 1 (meanwhile $Q_b(1) = 0$ since the whole network is the one and only module) and increase c by doubling for every iteration till $Q_b(2c) < Q_b(c)$. Then C is interpolated in an interval $(\frac{c}{2}, 2c)$.

The modularity score is measured as Qb in the algorithm and the standard value is between 0 and 1. The closer it gets to the maximum, the higher modularity is indicted by the algorithm. Qb is determined by the network size and the link density [70]. Qb, as a structure value, the statistical significance could be tested by performing a statistical analysis. The algorithm

analyzes the statistical significance by comparing the value within null models. So, replicated null models are created for this purpose and the library currently offers 5 kinds of null modules including Equiprobable, Average, Average_cols, Average_rows and Fixed. The detailed principles of those 5 null modules are shown in **Table 4**. Then the statistical significance P value is calculated as the likelihood that the output structure value of the tested network is greater or equal than the randomly chosen networks.

Table 4. **Different features of the marginal distribution of original network preserved by different null models.**

Aspects of consideration / Null modules kind	Connectance of network	Number of interactions in which each node is involved	Number of interactions of row nodes	Number of interactions of col nodes	Total sums of each row and column of bipartite adjacency matrix
Equiprobable	Yes				
Average	Yes	Yes			
Average_col	Yes		Yes		
Average_rows	Yes			Yes	
Fixed					Yes

4.4.4 Methods for enrichment analysis

4.4.4.1 Enrichment analysis

To test the significance of the disease-specific risk genes enriched in the clusters or modules, we performed enrichment analysis for each module obtained from BiMat library. For each module, a 2 by 2 contingence table was made depicting the experimental results. The contingence table illustrates the distribution of two feature in rows and columns respectively. For the first column, we classify whether the risk genes are in the specific module or not. While the

second column specifies whether the non-risk proteins are in the specific module or not. The **Table 5** shows the graphic expression:

Table 5. **The distribution of classifications of the two variables.** A1: number of risk proteins with the specific module. B1: number of non-risk proteins with the specific module. A2: number of risk proteins not in the specific module but in the network. B2: number of non-risk proteins not in the specific module but in the network.

	Risk proteins	Non-risk proteins
In module	A1	B1
Not in module but in the network	A2	B2

4.4.4.2 Fisher's exact test

Statistical significance of the contingency table can be measured by a Fisher's exact test. The Fisher's exact test aims to explore the association between the two features (risk genes and modules) and the p-value from the test demonstrates the hypergeometric likelihood of the observed data. The null hypothesis is the independence to a hypergeometric distribution of the numbers in the cells of the contingence matrix. The formula of p is given as **Equation 16**:

$$p = \frac{(A1+B1)!(A2+B2)!(A1+A2)!(B1+B2)!}{A1!B1!A2!B2!n!} \quad (16)$$

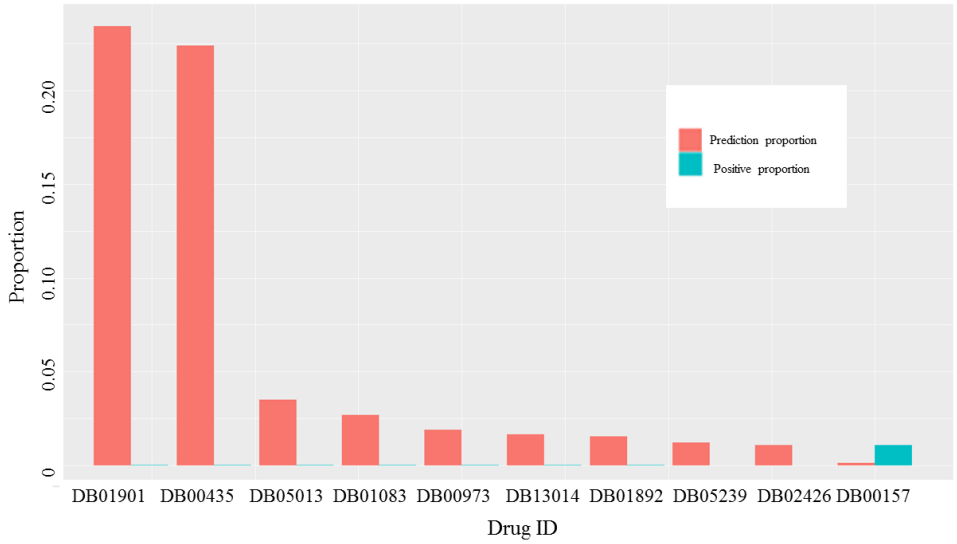
Here, n=A1+B1+A2+B2. Smaller p values reflect greater evidence in rejecting null hypothesis. A significance level of P<=0.005 is used is this experiment after multiple testing correction.

4.5 Results

4.5.1 Generation of the predicted drug-target interaction network

Using the trained DNN model, we predicted all possible 18,138,422 DTIs combined from the 5,134 drugs and the 3,533 proteins. We built a drug-target interaction network based on the top 14,960 interactions from the predicted DTIs with a probability score cut-off 0.9995, which is a similar number of DTIs in Drugbank (14,792, see **Fig. 2**). The network included 3,082 drugs and 3,514 proteins. For the predicted and known drug-target interaction networks, we calculated the interaction frequency of each drug interaction (**Fig. 16A**) and the interaction frequency of each protein interaction (**Fig. 16B**). It is obvious that DB01901 and DB00435 interacted with more than 20% of the interactions in the predicted network, and P07383 interacted with more than 15% of the interactions in the predicted network. Hence, we excluded the interactions containing DB01901, DB00435 or P07383 in the predicted network for the further exploration. The final predicted network included 5,921 drug-target interactions with 1,633 unique drugs and 608 unique proteins.

A:



B:

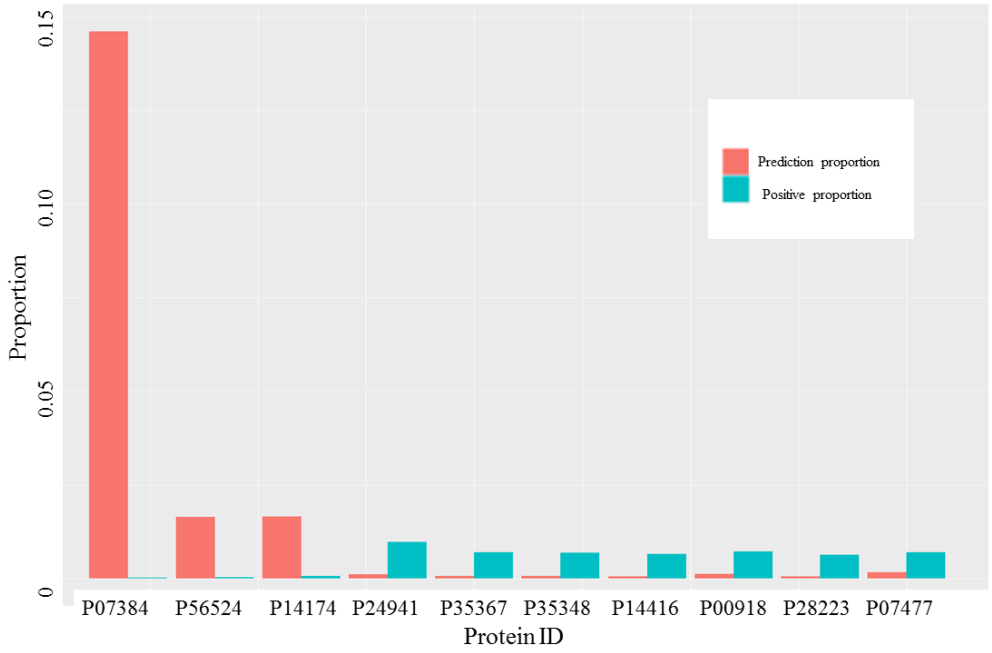


Figure 16 - Top 10 drugs (A) and proteins (B) with higher frequency of the interactions with other proteins and drugs respectively. A: DB01901 and DB00435 showed significantly higher proportions of the interactions with the proteins in the predicted drug-target interaction network. B: P07383 showed significantly higher proportions of the interactions with the drugs in the predicted drug-target interaction network.

4.5.2 Drug repurpose results using extracted predicted DTIs

For the 288 breast cancer risk genes and the final predicted network, we extracted a subnetwork with 138 DTIs, including only 6 of the 288 breast cancer risk genes and 57 drugs. These 57 drugs had at least two predicted interactions with the 6 breast cancer risk genes as shown in **Fig. 17**. **Table 6** shows the potentially repurposed drugs based on the subnetwork, which had interacted with at least 4 breast cancer risk genes. Overall, all these 5 drugs have shown some evidences related to the treatment of breast cancer.

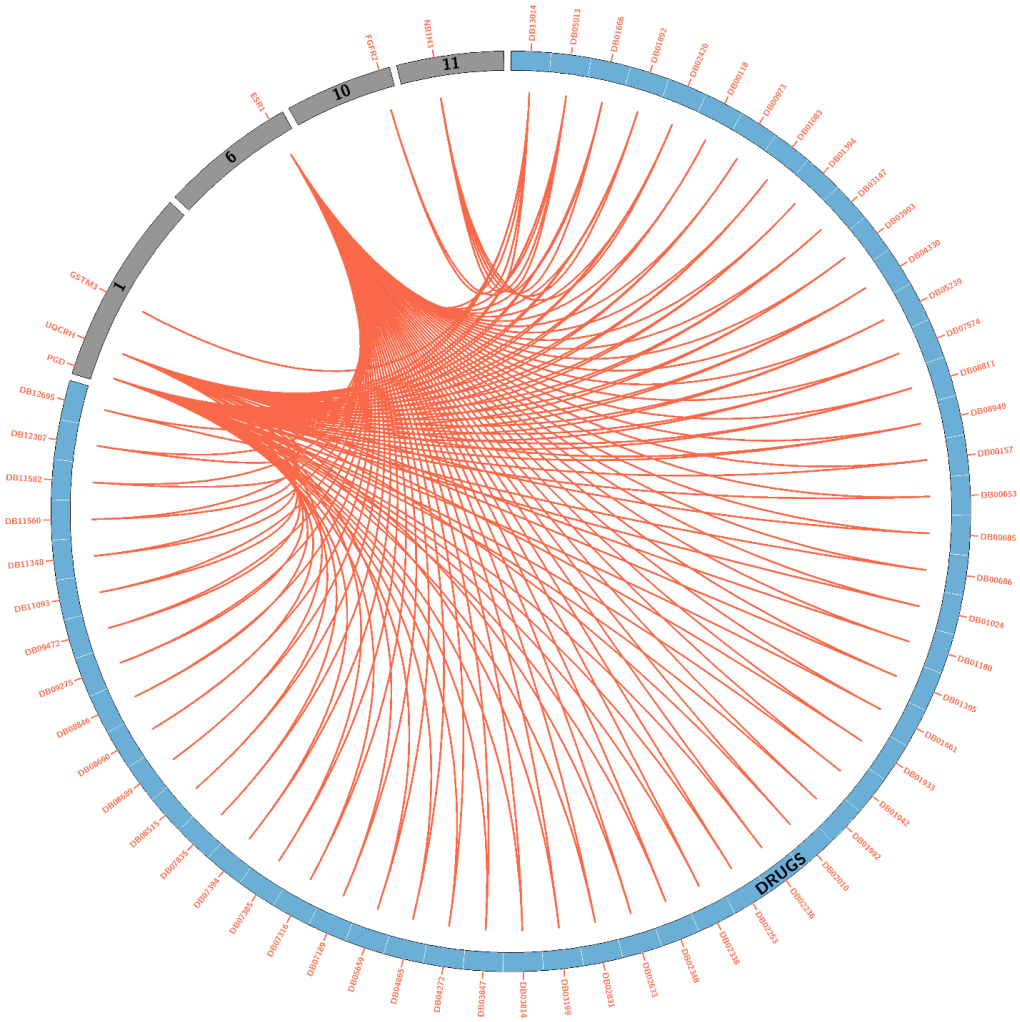


Figure 17 - 57 drugs interacted with 6 breast cancer risk genes in the predicted DTIs. Blue band locates drugs and gray band stands for the chromosomes.

Table 6. **Top drugs repurposed for breast cancer**

Drug ID (Name)	Interacted genes	Diseases treated by the drugs	Drug-breast cancer association
DB13014 (Hypericin)	ESR1,UQCRH, GSTM3,FGFR2, PGD,NR1H3	Used for Human Immunodeficiency Virus (HIV) Infections and Cutaneous T-Cell Lymphoma (CTCL)[39].	Hypericum perforatum has shown cytotoxic and apoptogenic effect on the MCF-7 human breast cancer cell line[71].
DB05013 (Ingenol ebutate)	ESR1,UQCRH,FGFR2,PGD,NR1H3	Used for the topical treatment of actinic keratosis[39].	Shown to be a potent anti-cancer drug and therapeutically effective in microgram quantities. Making it a promising candidate for use in breast cancer patients(Ogbourne, S. M. Crystalline Ingenol Mebutate. International Patent WO 2011128780A1, 2011).
DB01666 (D-Myo-Inositol-Hexasulphate)	ESR1,UQCRH,PGD,NR1H3	Myo-inositol hexasulfate is a potent inhibitor of Aspergillus ficuum phytase[72].	Experiment data of breast cancer cell lines treated in vitro with myo-Ins indicated that myo-Ins inhibits the principal molecular pathway supporting EMT in cancer cells[73].
DB01892 (Hyperforin)	ESR1,UQCRH,PGD,NR1H3	Hyperforin is a possible antidepressant component[74]. It also induces hepatic drug metabolism through activation of the pregnane X receptor[75].	Hyperforin has shown anti-inflammatory and anti-carcinogenic properties[76].
DB02426 (Carboxyatractylosid e)	ESR1,UQCRH,PGD,NR1H3	Carboxyatractyloside belongs to the class of organic compounds known as diterpene glycosides[39].	Carboxyatractyloside is a specific inhibitors of ANT ,and the expression of ANT isoforms is closely related to the energetic metabolic properties of tumoral cells[77].

A subnetwork of 112 DTIs were extracted to include the 22 IBD risk genes and 29 drugs. These 29 drugs had at least one predicted interaction with the 22 IBD risk genes as shown in **Fig. 18.** We reported these 29 repurposed drugs for IBD treatment.

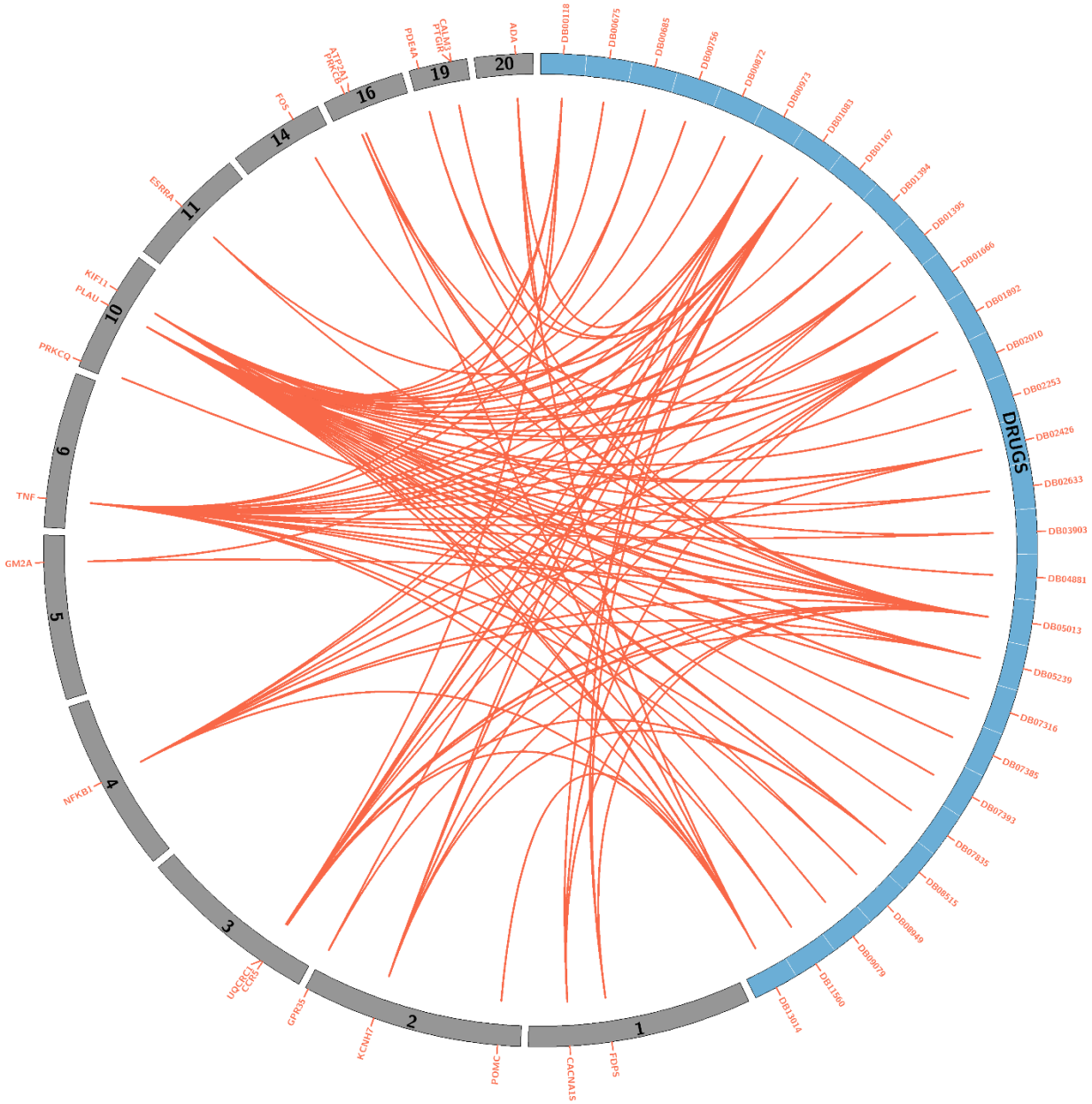


Figure 18 - Predicted interactions of 29 repurposed drugs and 22 risk IBD genes. Blue band locates drugs and grey band stands for the chromosomes.

4.5.3 Drug repurposed results based on clustering analysis

4.5.3.1 Results of identified clusters

14 clusters were generated through BiMat using 5,921 predicted DTIs with 1,633 unique drugs and 608 unique proteins, achieving the modularity score Qb=0.47 and a posteriori measure Qr=0.35. Qr describes the ratio of interactions built within the same module vs. between different modules [78]. Without isolated nodes within the clusters, around 65% of the interactions exist between different clusters. **Fig. 19A** and **Fig. 19B** illustrate the drug proportion and the protein distribution in each module respectively. **Table 7** demonstrates the number of drugs and proteins in each module.

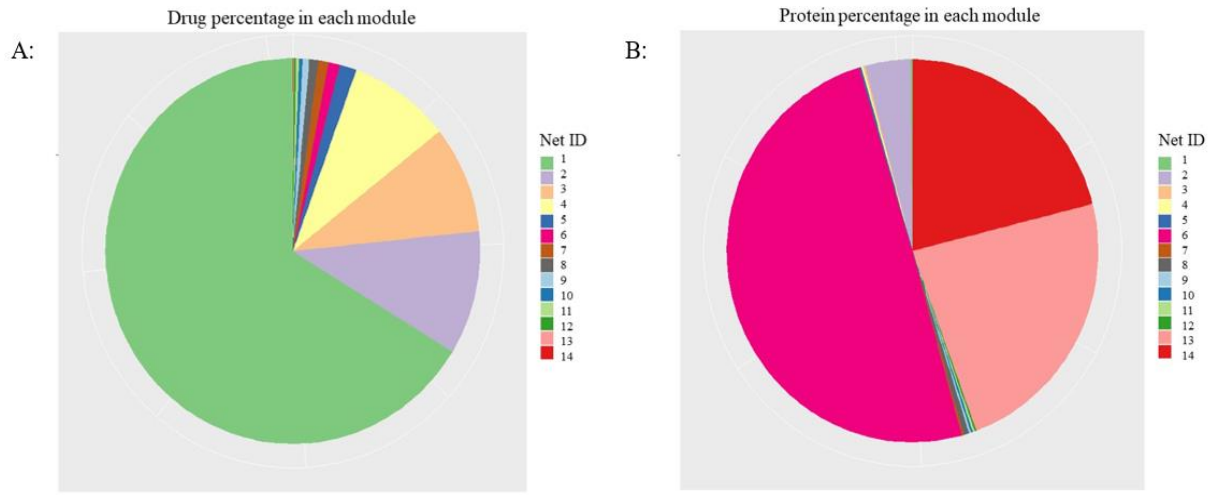


Figure 19 - Drug and protein percentages in each module. A: Drug percentage B: Protein percentage.

Table 7. **Number of drugs and proteins in each module.**

Net ID	1	2	3	4	5	6	7
Drug number	1082	170	147	144	24	16	14
Protein number	1	24	1	1	1	302	1
Net ID	8	9	10	11	12	13	14
Drug number	13	9	5	4	3	1	1
Protein number	3	1	1	1	1	142	128

4.5.3.2 Results of enrichment analysis

Using the risk genes of breast cancer and IBD collected from the above GWAS studies, 22 IBD risk genes and 6 breast cancer risk genes exist in the network. The Fisher's exact test showed significance of the enrichment of the breast cancer risk genes in clusters 5 and 7, and the enrichment of the IBD risk genes in clusters 6 and 13. **Fig. 20** showed the risk genes distribution in each module for IBD and breast cancer. **Table 8** summarised statistics of these enriched clusters. In total, 38 predicted DTIs from the enriched clusters 5 and 7 were collected for drug repurposing with one breast cancer risk genes in each of the two clusters. Meanwhile, 18 IBD risk genes were overrepresented in clusters 6 and 13 with total 53 predicted interactions.

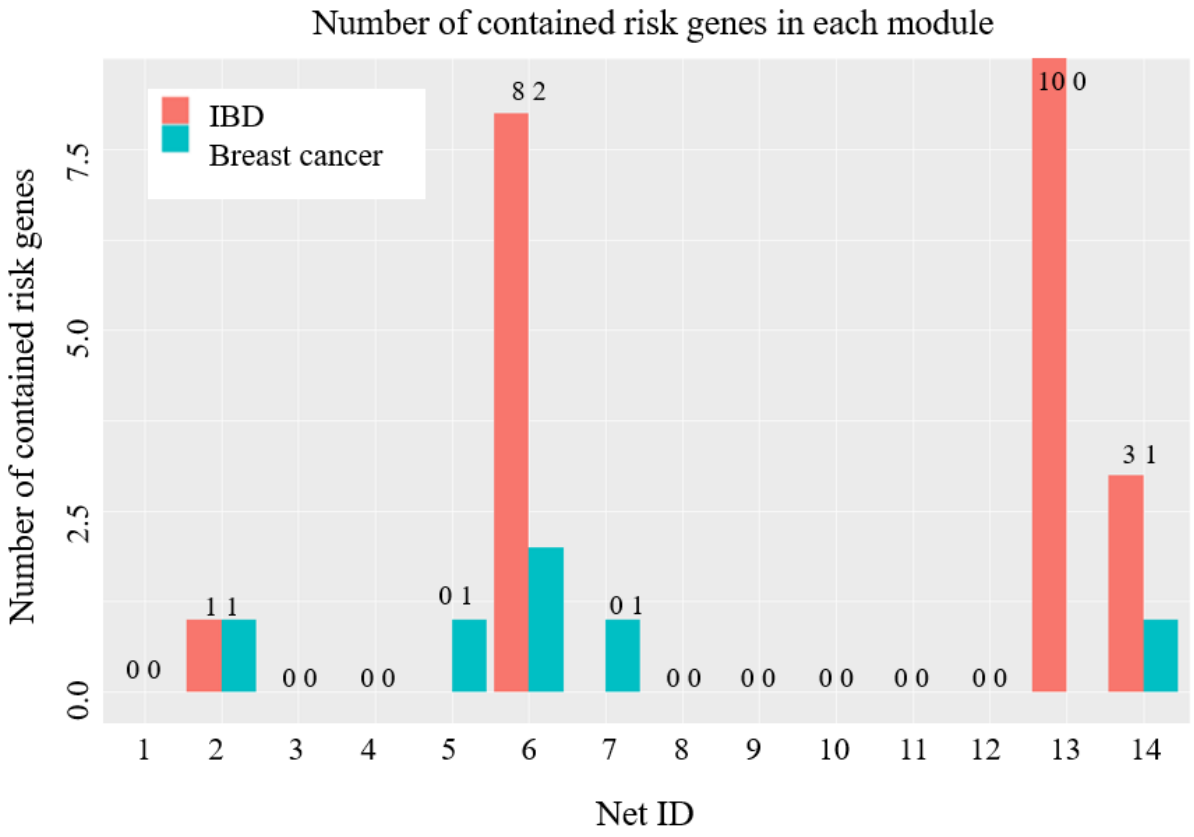


Figure 20 - Number of risk IBD/breast cancer genes in each module.

Table 8. **Summary statistics of the enriched clusters.** 38 (24+14) the predicted DTIs with 2 overrepresented breast cancer genes were reported for drug repurposing use. 53 (43+10) predicted DTIs with 18(8+10) over presented IBD genes were reported for IBD drug repurposing.

	Number of drugs	Number of proteins	Interactions	Breast cancer(BC) gene number	Interactions contain BC risk genes
Cluster 5	24	1	24	1	24
Cluster 7	14	1	14	1	14
	Number of drugs	Number of proteins	Interactions	IBD gene number	Interactions contain IBD risk genes
Cluster 6	16	302	1340	8	43
Cluster 13	1	142	142	10	10

4.5.3.3 Drug repurpose for IBD and breast cancer based on enriched clusters

4.5.3.3.1 Drug repurposing for breast cancer

Twenty-four and 14 newly identified DTIs were generated from the enriched clusters 5 and 7 with breast cancer risk genes respectively. In total, 38 drugs were repurposed for breast cancer in this experiment, containing that 24 drug candidates interact with PGD and 14 candidates interact with UQCRH. **Fig. 21** showed the predicted DTIs in cluster 5 while **Fig. 22** illustrated the predicted DTIs in the enriched cluster 7.

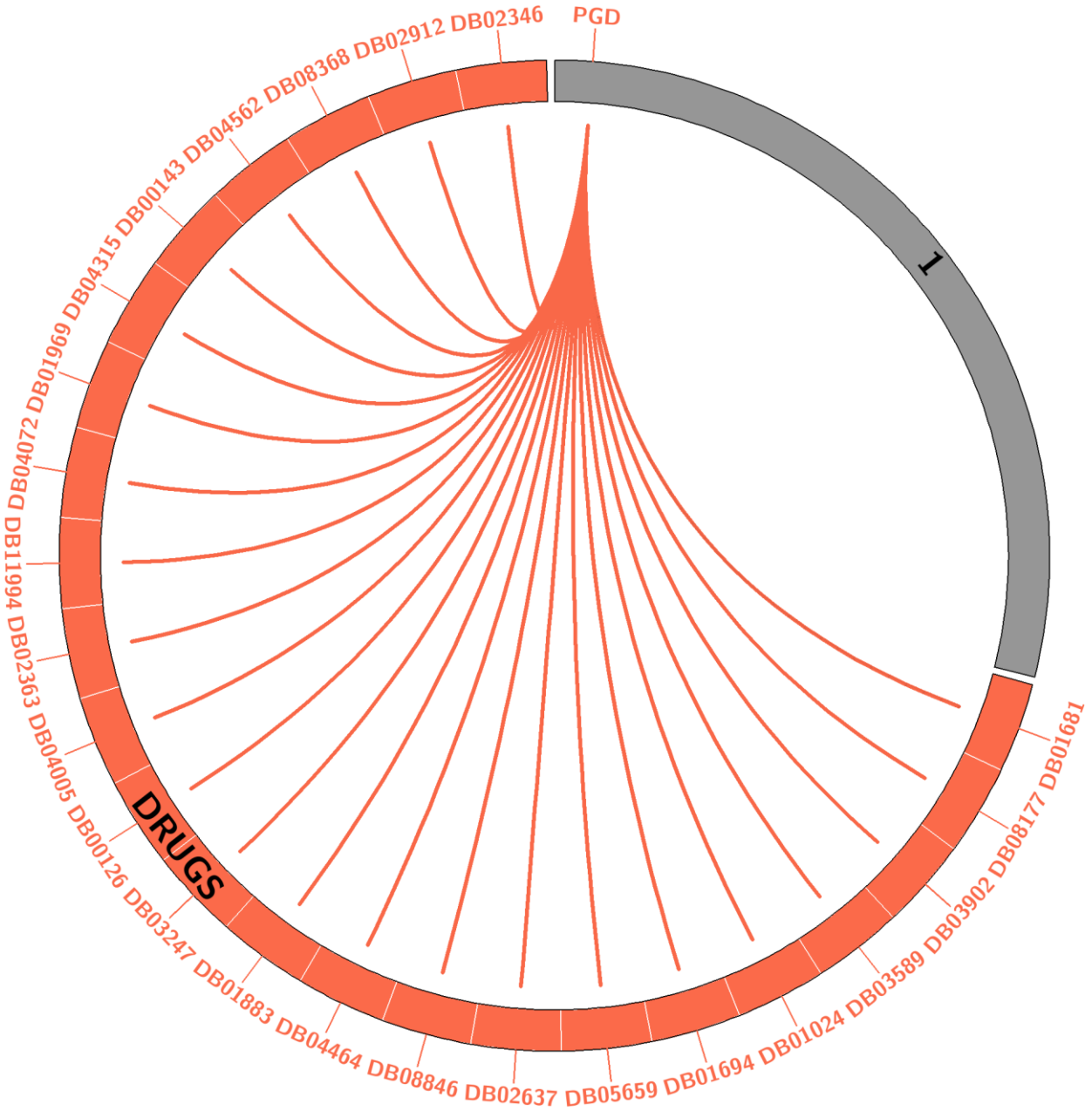


Figure 21 - Newly predicted DTIs in enriched cluster 5 for breast cancer. Red band locates drugs and grey band stands for the chromosomes. The only gene in this cluster PGD connected with 24 drugs. Since PGD is a risk gene of breast cancer, 24 drugs were repurposed for disease treatment.

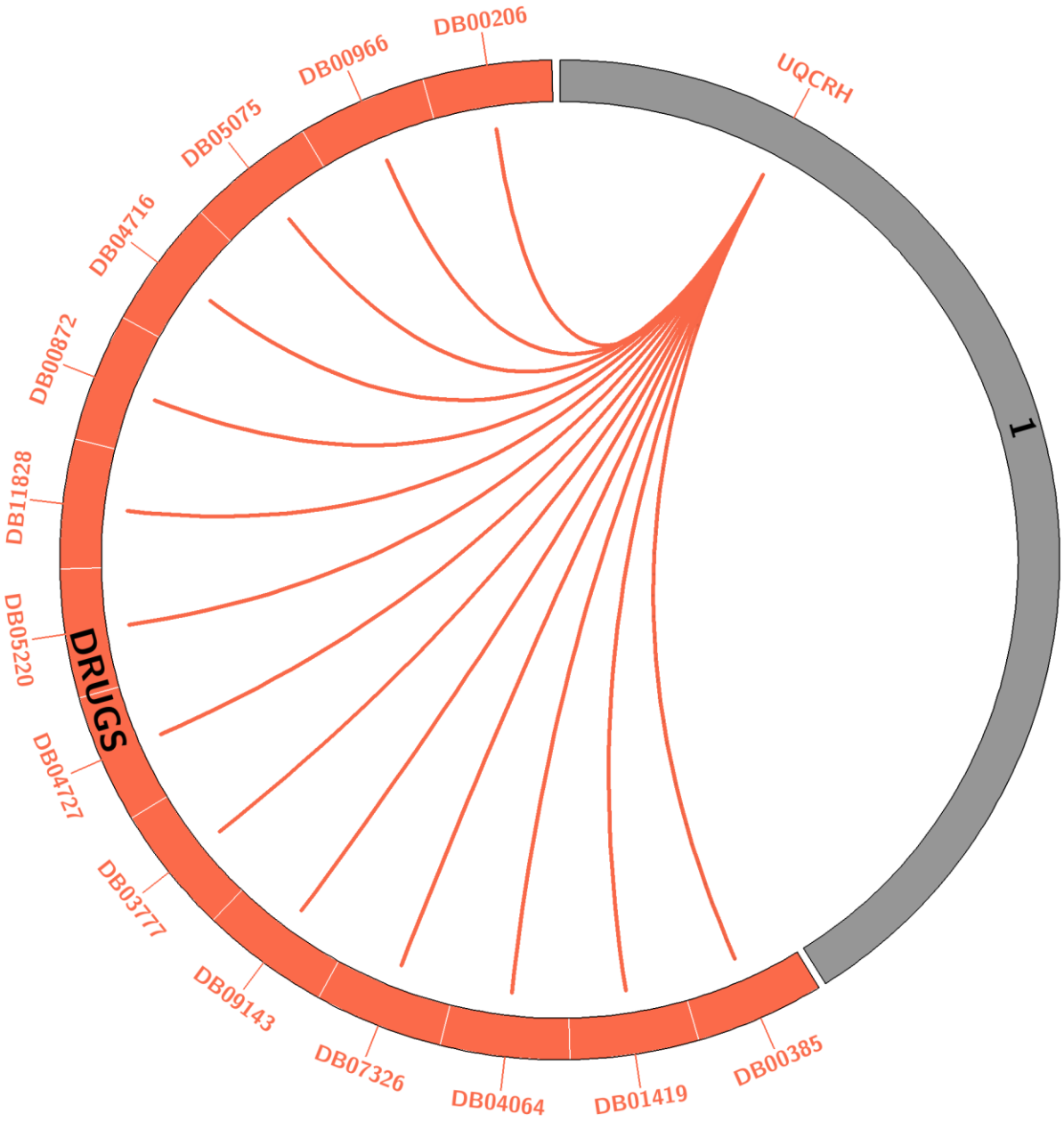


Figure 22 - Newly predicted DTIs in enriched cluster 7 for breast cancer. Red band locates drugs and grey band stands for the chromosomes. The only gene in this cluster UQCRH connected with 14 drugs. Since UQCRH is a risk gene of breast cancer, 14 drugs were repurposed for disease treatment.

4.5.3.3.2 Drug repurposing for IBD

The Cluster 6 and Cluster 13 showed statistical significance of enrichment of IBD risk genes. As shown in **Fig. 23**, the enriched cluster 6 included 1,340 interactions with 16 unique drugs and 302 unique proteins, 8 of which were IBD risk genes. 43 interactions marked in red contained these 8 risk genes and 13 unique drugs were repurposed for IBD treatment. **Fig. 24** showed the only candidate DB01083 in the enriched cluster 13, which interacted with 124 proteins. 8 of which were IBD risk genes: FDPS, CALM3, ESRRA, GM2A, PDE4A, PRDX5, PTGIR, CCR5, CACNA1S and GPR35. In total, 14(13+1) drug candidates DB01395, DB11560, DB02426, DB01666, DB13014, DB02633, DB05239, DB07316, DB08515, DB08949, DB00118, DB00973, and DB01083 were repositioned as novel drugs for IBD.

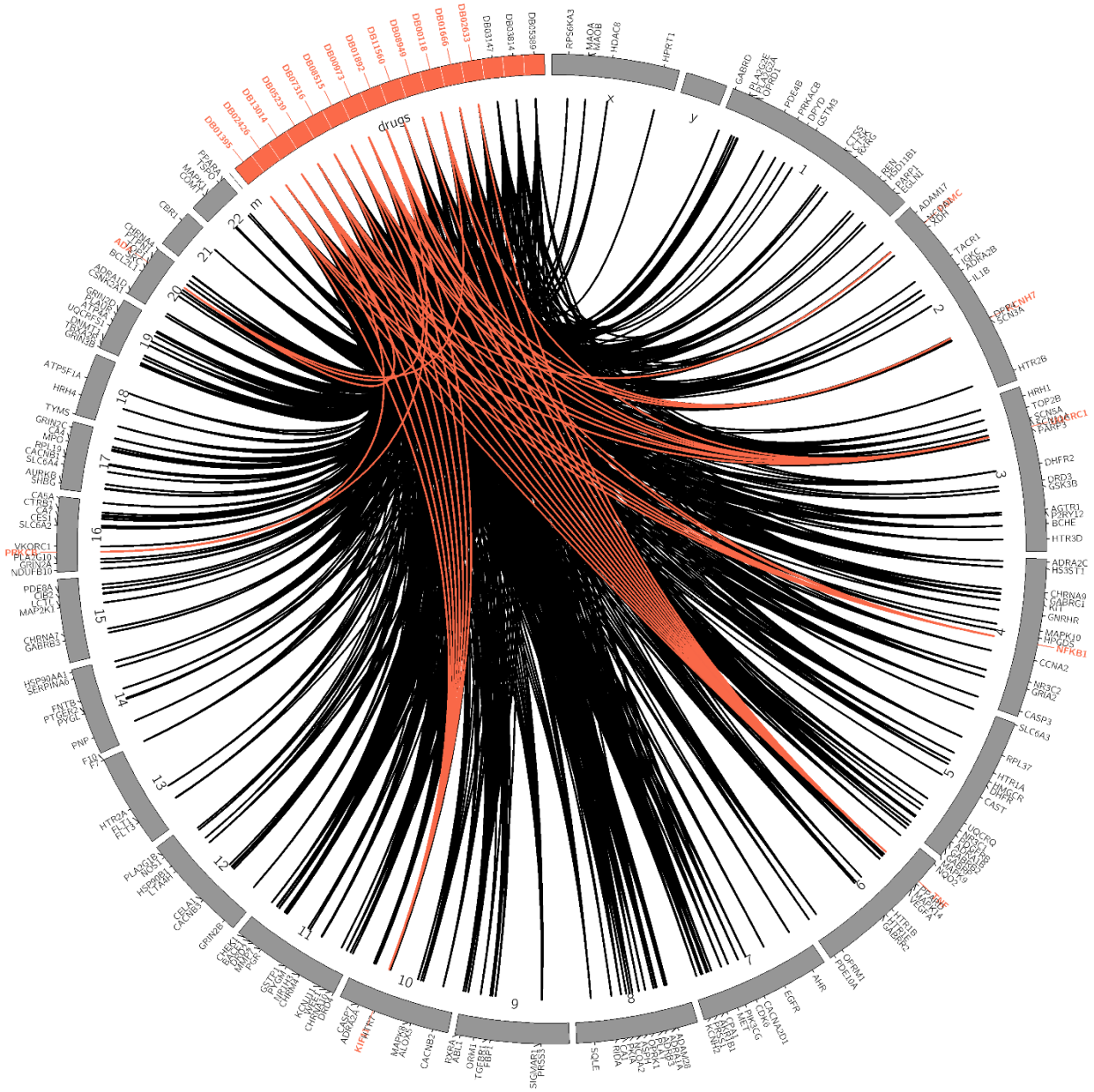


Figure 23 - Newly predicted DTIs in the enriched cluster 6 for IBD. Red band locates drugs and grey band stands for the chromosomes. Red notation represents IBD related drugs, proteins, or interactions. 13 drugs DB01395, DB11560, DB02426, DB01666, DB13014, DB02633, DB05239, DB07316, DB08515, DB08949, DB00118, and DB00973 were predicted to interact with IBD risk genes: TNF, NFKB1, UQCRC1, KIF11, ADA, POMC, PRKCB, KCNH7.

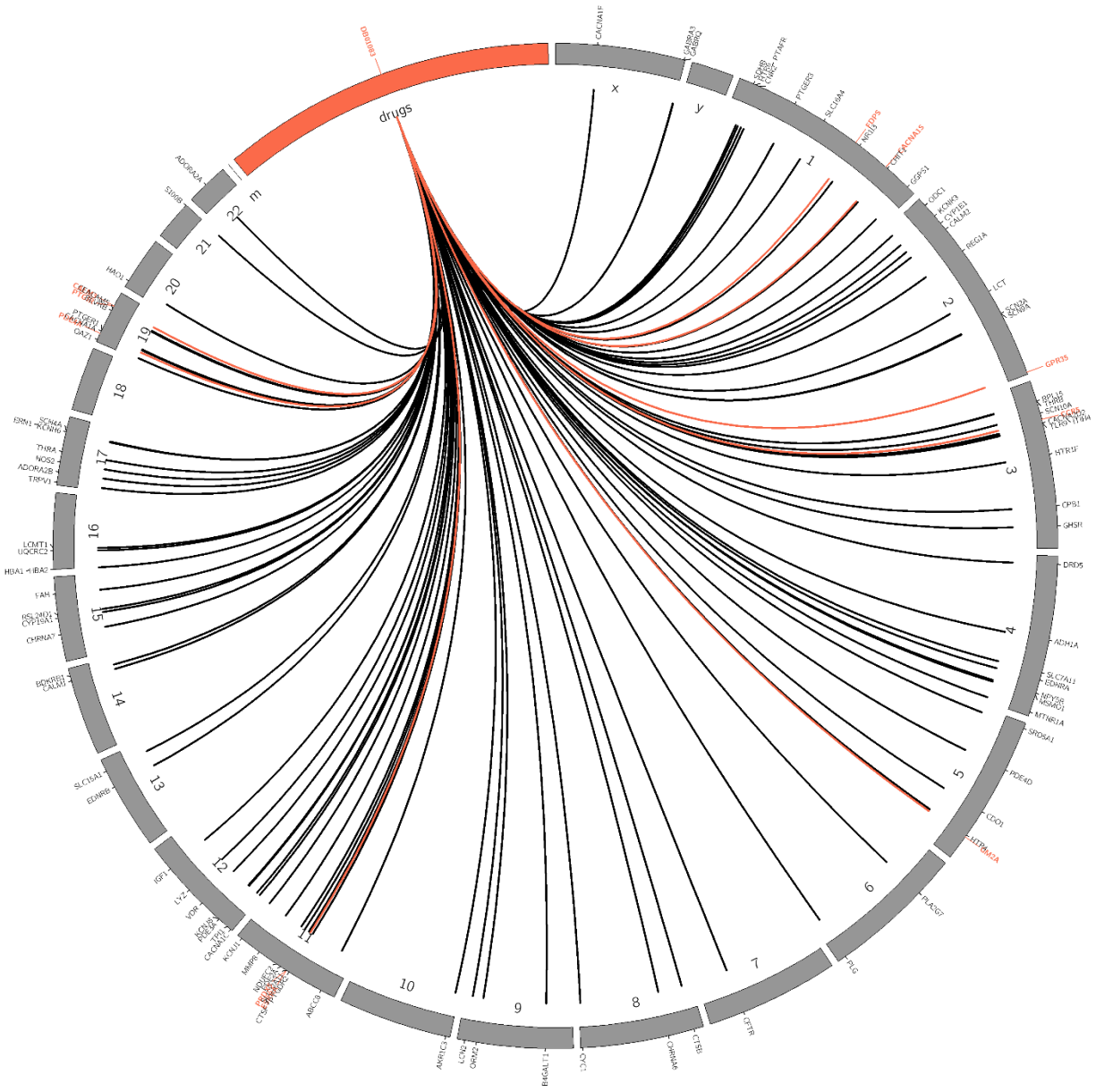


Figure 24 - Newly predicted DTIs in the enriched cluster 13 for IBD. Red band locates drugs and grey band stands for the chromosomes. Red notation represents IBD related drugs, proteins, or interactions. Drug DB01083 was predicted to interact with 124 IBD genes including 10 IBD risk genes: FDPS, CALM3, ESRRA, GM2A, PDE4A, PRDX5, PTGIR, CCR5, CACNA1S, and GPR35.

4.6 Conclusions

In this experiment, we first extracted subnetwork of predicted DTIs and then repurposed potential drug candidates for IBD and breast cancer. The repurposed drugs showed evidence related to the diseases, making them promising drugs for repurposing use. We also applied clustering analysis to the top ranked DTIs predicted from the trained DDN model, which contains drug-target pairs that are considered to be potentially interacted. The clustering analysis has advantages since it groups nodes with similarity together, which helps to recognize an underlying pattern beneath data. We showed that the network is modular to some extent and the clustering algorithm has the capability to categorize the network into modules. Given the disease-specific risk genes provided by GWAS, we also explored the enrichment of the clusters for IBD risk genes and breast cancer risk genes. The results showed that some modules were identified as enriched clusters for either IBD or breast cancer risk genes. Several risk genes of IBD and breast cancer showed significant enrichment in specific clusters and the corresponding interacting drugs were reported for repurposing use.

Chapter 5 Conclusion and further work

Drug repositioning is a promising technology in drug discovery. It simplifies the procedure before entering market and reduces the safety threats to develop new compounds. However, identifying drug-target interactions challenges the reliability and practicability of drug repurposing. Traditional machine learning approaches such as SVM and logistic regression have limitations in modelling high-dimensional biology data. Recently, the trend is observed in applying deep learning in drug-target interaction prediction. The strong traits of deep learning in handling large data offer an encouraging alternative for the interaction prediction. Several studies were investigated in drug repurposing using different deep learning architectures such as DBN[22].

In this thesis, we collected drug descriptors, protein sequence data from Drugbank and protein domain information from NCBI. We proposed SLR models, LASSO regression models as well as a deep neural network model for drug-target interaction prediction using integrated drug descriptors and protein sequence features from multiple sources. In the baseline SLR models, we used the top integrated features selected by Wilcoxon's test. We used different combinations of feature sets in LASSO models to explore the prediction power and showed that the best performance can be reached by integrating all feature sets. We demonstrated that the LASSO models can work well for data sets with a very large number of features, which often results in overfitting in the SLR models. Results showed that the LASSO-based regularized linear classification models achieved better results than the SLR models for predicting drug-target interactions. We also introduced a DNN model which takes features generated through LASSO regression as input for drug protein interaction prediction. Experimental results show

that the DNN model outperformed other traditional machine learning methods. In addition, deep learning has shown superiority in extracting informative features automatically and efficiently. However, the models in this study are sensitive to the selection of negative data since there is no existing gold standard negative data (not interacted drug-target pairs). Our models may vary when different cut-offs are applied in selection of potential positive DTIs or a different negative data generation method is used. Given this consideration, other machine learning approaches, such as collaborative filtering, a widely used method for user recommendations in social medias, which gives the recommendations for a user based on the large database of preferences. Additionally, the imbalance fact that there are many more negative drug-target pairs than the truly interacted DTIs leads to the bias issue in model training, which is often claimed as a between-class imbalance. This makes the prediction tend to predict more samples to the majority class (non-interacted samples). Therefore, accuracy is not a good performance measure in this case. We followed the most commonly accepted principle to have the same number of positive DTIs as negative DTIs to reduce the bias problem, but it results in discarding information in the majority group unavoidably [79]. Furthermore, we applied DNN to predict drug-target interaction networks in clustering analysis and demonstrated that various drug-target interactions in enriched clusters can be potentially repurposed for IBD and breast cancer. Thirty-eight drugs were repurposed for breast cancer through the enrichment analysis and 14 drugs were repurposed for potential IBD clinical treatment. Risk genes used in this study were recently published by GWAS. However, the existing drugs on market for specific diseases have been discovered for decades, which could be potential reasons why the existing drugs for specific diseases are not found in our repurposed drugs for the diseases.

Our study demonstrates that more drug and protein related features are helpful for drug-target interaction network prediction. In the future, we will include other types of genomic features, such as gene expression profiles, into the DNN model to improve the prediction performance. Our study shows the genetic information is helpful for drug repositioning. However, the collection of a set of experimental validated risk genes is still a challenging task. In the future, we will explore to use text mining technology to curate the robust risk gene list for complex disease drug repositioning.

Reference

[1] R. on T. G.-B. R. for Health, B. on H. S. Policy, and I. of Medicine, *Drug Repurposing and Repositioning: Workshop Summary*. 2014.

[2] T. T. Ashburn and K. B. Thor, "Drug repositioning: identifying and developing new uses for existing drugs," *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, 2004.

[3] D. F. Horrobin, "Realism in drug discovery - Could cassandra be right?," *Nature Biotechnology*, vol. 19, no. 12. pp. 1099–1100, 2001.

[4] J. Gilbert, P. Henske, and A. Singh, "Rebuilding big pharma's business model," *Vivo, Bus. Med. Rep.*, vol. 21, no. 10, pp. 73–80, 2003.

[5] M. J. Barratt and D. E. Frail, *Drug Repositioning: Bringing New Life to Shelved Assets and Existing Drugs*. 2012.

[6] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1998, vol. 1398, pp. 137–142.

[7] D. Yu and L. Deng, "Deep Learning and Its Applications to Signal and Information Processing [Exploratory DSP]," *Signal Process. Mag. IEEE*, vol. 28, no. 1, pp. 145–154, 2011.

[8] X. Chen *et al.*, "Drug-target interaction prediction: Databases, web servers and computational models," *Brief. Bioinform.*, vol. 17, no. 4, pp. 696–712, 2016.

[9] A. Ezzat, M. Wu, X.-L. Li, and C.-K. Kwoh, "Drug-target interaction prediction using ensemble learning and dimensionality reduction," *Methods*, 2017.

[10] K. Zhao and H.-C. So, "A machine learning approach to drug repositioning based on drug

expression profiles: Applications to schizophrenia and depression/anxiety disorders," pp. 1–18, 2017.

[11] Y. Yamanishi, M. Kotera, Y. Moriya, R. Sawada, M. Kanehisa, and S. Goto, "DINIES: Drug-target interaction network inference engine based on supervised analysis," *Nucleic Acids Res.*, vol. 42, no. W1, 2014.

[12] A. P. Chiang and A. J. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses," *Clin. Pharmacol. Ther.*, vol. 86, no. 5, pp. 507–510, 2009.

[13] L. Yang and P. Agarwal, "Systematic drug repositioning based on clinical side-effects," *PLoS One*, vol. 6, no. 12, 2011.

[14] F. Cheng *et al.*, "Prediction of drug-target interactions and drug repositioning via network-based inference," *PLoS Comput. Biol.*, vol. 8, no. 5, 2012.

[15] F. Napolitano *et al.*, "Drug repositioning: A machine-learning approach through data integration," *J. Cheminform.*, vol. 5, no. 6, 2013.

[16] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 1, pp. 30–42, 2012.

[17] A. Krizhevsky, Ii. Sulskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Adv. Neural Inf. Process. Syst.*, vol. 60, no. 6, pp. 84–90, 2012.

[18] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.

[19] A. Mohamed and G. Hinton, "Phone Recognition Using Restricted Boltzmann Machines,"

Icassp, no. 3, pp. 4354–4357, 2010.

[20] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2014, pp. 5337–5341.

[21] S. Zhai, K. Chang, R. Zhang, and Z. M. Zhang, "DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016, pp. 1295–1304.

[22] M. Wen *et al.*, "Deep-Learning-Based Drug-Target Interaction Prediction," *J. Proteome Res.*, vol. 16, no. 4, pp. 1401–1409, 2017.

[23] S. Min, B. Lee, and S. Yoon, "Deep learning in bioinformatics," *Briefings in bioinformatics*, vol. 18, no. 5. pp. 851–869, 2017.

[24] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[25] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine Learning in Genomic Medicine: A Review of Computational Problems and Data Sets," *Proc. IEEE*, vol. 104, no. 1, pp. 176–197, 2016.

[26] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[27] D. R. Kelley and Y. A. Reshef, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks," *bioRxiv*, 2017.

[28] J. Zhou and O. G. Troyanskaya, “Predicting effects of noncoding variants with deep learning-based sequence model,” *Nat. Methods*, vol. 12, no. 10, pp. 931–934, 2015.

[29] J. L. Haines *et al.*, “Complement factor H variant increases the risk of age-related macular degeneration.,” *Science*, vol. 308, no. 5720, pp. 419–21, 2005.

[30] T. Wellcome, T. Case, and C. Consortium, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.,” *Nature*, vol. 447, no. 7145, pp. 661–78, 2007.

[31] G. B. Ehret *et al.*, “Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk,” *Nature*, vol. 478, no. 7367, pp. 103–109, 2011.

[32] J. MacArthur *et al.*, “The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog),” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D896–D901, 2017.

[33] J. L. E. Pritchard, T. A. O’Mara, and D. M. Glubb, “Enhancing the promise of drug repositioning through genetics,” *Frontiers in Pharmacology*, vol. 8, no. DEC. 2017.

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

[35] L. Wang, H. Liu, C. G. Chute, and Q. Zhu, “Cancer based pharmacogenomics network supported with scientific evidences: From the view of drug repurposing,” *BioData Min.*, vol. 8, no. 1, 2015.

[36] J. Z. Liu *et al.*, “Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis,” *Nat Genet*, vol. 44, no. 10, pp. 1137–1141, 2012.

[37] V. Collij, E. A. M. Festen, R. Alberts, and R. K. Weersma, “Drug Repositioning in Inflammatory Bowel Disease Based on Genetic Information,” *Inflamm. Bowel Dis.*, vol. 22, no. 11, pp. 2562–2570, 2016.

[38] S. J. Connor *et al.*, “CCR2 expressing CD4+ T lymphocytes are preferentially recruited to the ileum in Crohn’s disease,” *Gut*, vol. 53, no. 9, pp. 1287–1294, 2004.

[39] D. S. Wishart *et al.*, “DrugBank 5.0: A major update to the DrugBank database for 2018,” *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, 2018.

[40] S. A. R. P. Jagarlapudi and K. V. R. Kishan, “Chemogenomics,” vol. 575, pp. 159–172, 2009.

[41] D. S. Cao, N. Xiao, Q. S. Xu, and A. F. Chen, “Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions,” *Bioinformatics*, vol. 31, no. 2, pp. 279–281, 2015.

[42] I. Sushko *et al.*, “Online chemical modeling environment (OCHEM): Web platform for data storage, model development and publishing of chemical information,” *J. Comput. Aided. Mol. Des.*, vol. 25, no. 6, pp. 533–554, 2011.

[43] Z. He *et al.*, “Predicting drug-target interaction networks based on functional groups and biological features,” *PLoS One*, vol. 5, p. e9603, 2010.

[44] D. L. Wheeler *et al.*, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, 2007.

[45] Y. Luo *et al.*, “A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information,” *Nat. Commun.*, vol. 8, no. 1, 2017.

[46] D. Bajusz, A. Rácz, and K. Héberger, “Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?,” *J. Cheminform.*, vol. 7, no. 1, 2015.

[47] N. Xiao, D. S. Cao, M. F. Zhu, and Q. S. Xu, “Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences,” in

Bioinformatics, 2015, vol. 31, no. 11, pp. 1857–1859.

[48] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization Paths for Generalized Linear Models via Coordinate Descent," *J. Stat. Softw.*, vol. 33, no. 1, 2010.

[49] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "Misc functions of the Department of Statistics (e1071), TU Wien," *R package version 1.6-2*. p. http://cran.r--project.org/package=e1071, 2014.

[50] D. P. Kingma and J. L. Ba, "Adam: a Method for Stochastic Optimization," *Int. Conf. Learn. Represent. 2015*, pp. 1–15, 2015.

[51] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, 2014.

[52] M. Hamanaka *et al.*, "CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning," *Mol. Inform.*, vol. 36, no. 1, 2017.

[53] M. Abadi *et al.*, "TensorFlow: A System for Large-Scale Machine Learning TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)*, 2016, pp. 265–284.

[54] R. H. Duerr *et al.*, "A genome-wide association study identifies IL23R as an inflammatory bowel disease gene.," *Science*, vol. 314, no. 5804, pp. 1461–3, 2006.

[55] D. K. Bonen and J. H. Cho, "The genetics of inflammatory bowel disease," *Gastroenterology*, vol. 124, no. 2, pp. 521–536, Feb. 2003.

[56] K. M. De Lange *et al.*, "Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease," *Nat. Genet.*, vol. 49, no. 2, pp. 256–261, 2017.

[57] J. S. Baxter *et al.*, “Capture Hi-C identifies putative target genes at 33 breast cancer risk loci,” *Nat. Commun.*, vol. 9, no. 1, 2018.

[58] L. Wu *et al.*, “A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer,” *Nat. Genet.*, vol. 50, no. July, p. 1, 2018.

[59] J. Lonsdale *et al.*, “The Genotype-Tissue Expression (GTEx) project,” *Nature Genetics*, vol. 45, no. 6. pp. 580–585, 2013.

[60] T. C. G. A. R. Network, “The Cancer Genome Atlas Pan-Cancer analysis project.,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–1120, 2013.

[61] C. Wu, R. C. Gudivada, B. J. Aronow, and A. G. Jegga, “Computational drug repositioning through heterogeneous network clustering,” *BMC Syst. Biol.*, vol. 7, no. SUPPL 5, 2013.

[62] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto encyclopedia of genes and genomes,” *Nucleic Acids Research*, vol. 27, no. 1. pp. 29–34, 1999.

[63] C. C. New, “Introductory Graph Theory,” *Journal of the Operational Research Society*, vol. 29, no. 12. pp. 1245–1245, 1978.

[64] M. E. J. Newman, “Modularity and community structure in networks,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, 2006.

[65] C. O. Flores, T. Poisot, S. Valverde, and J. S. Weitz, “BiMat: A MATLAB package to facilitate the analysis of bipartite networks,” *Methods Ecol. Evol.*, vol. 7, no. 1, pp. 127–132, 2016.

[66] J. M. Olesen, J. Bascompte, Y. L. Dupont, and P. Jordano, “The modularity of pollination networks,” *Proc. Natl. Acad. Sci.*, vol. 104, no. 50, pp. 19891–19896, 2007.

[67] J. Bascompte, P. Jordano, C. J. Melian, and J. M. Olesen, “The nested assembly of plant-animal mutualistic networks,” *Proc. Natl. Acad. Sci.*, vol. 100, no. 16, pp. 9383–9387, 2003.

[68] C. O. Flores, S. Valverde, and J. S. Weitz, “Multi-scale structure and geographic drivers of cross-infection within marine bacteria and phages,” *ISME J.*, vol. 7, no. 3, pp. 520–532, 2013.

[69] M. J. Barber, “Modularity and community detection in bipartite networks,” *Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys.*, vol. 76, no. 6, 2007.

[70] C. F. Dormann, J. Frund, N. Bluthgen, and B. Gruber, “Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks,” *Open Ecol. J.*, vol. 2, no. 1, pp. 7–24, 2009.

[71] S. A. Mirmalek *et al.*, “Cytotoxic and apoptogenic effect of hypericin, the bioactive component of Hypericum perforatum on the MCF-7 human breast cancer cell line,” *Cancer Cell Int.*, vol. 16, no. 1, 2016.

[72] A. H. J. Ullah and K. Sethumadhavan, “Myo-inositol Hexasulfate Is a Potent Inhibitor ofAspergillus ficuumPhytase,” *Biochem. Biophys. Res. Commun.*, vol. 251, no. 1, pp. 260–263, 1998.

[73] M. Bizzarri, S. Dinicola, A. Bevilacqua, and A. Cucina, “Broad Spectrum Anticancer Activity of Myo-Inositol and Inositol Hexakisphosphate,” *International Journal of Endocrinology*, vol. 2016. 2016.

[74] S. S. Chatterjee, S. K. Bhattacharya, M. Wonnemann, A. Singer, and W. E. Müller, “Hyperforin as a possible antidepressant component of hypericum extracts,” *Life Sci.*, vol. 63, no. 6, pp. 499–510, 1998.

[75] L. B. Moore *et al.*, “St. John’s wort induces hepatic drug metabolism through activation of

the pregnane X receptor.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 13, pp. 7500–7502, 2000.

[76] A. Koeberle *et al.*, "Hyperforin, an Anti-Inflammatory Constituent from St. John's Wort, Inhibits Microsomal Prostaglandin E2 Synthase-1 and Suppresses Prostaglandin E2 Formation in vivo," *Front. Pharmacol.*, vol. 2, 2011.

[77] A. Chevrollier, D. Loiseau, P. Reynier, and G. Stepien, "Adenine nucleotide translocase 2 is a key mitochondrial protein in cancer metabolism," *Biochimica et Biophysica Acta - Bioenergetics*, vol. 1807, no. 6. pp. 562–567, 2011.

[78] T. Poisot, "An a posteriori measure of network modularity," *F1000Research*, 2013.

[79] A. Ezzat, M. Wu, X. L. Li, and C. K. Kwoh, "Drug-target interaction prediction via class imbalance-aware ensemble learning," *BMC Bioinformatics*, vol. 17, 2016.