Faculty of Engineering and Information Technology
University of Technology Sydney

Text and Data Mining for Human Drug Understanding

A thesis submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

by

Yi Zheng

October 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Yi Zheng declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Advanced Analytics Institute, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:
Signature removed
prior to publication.

SIGNATURE OF CANDIDATE: _____

[Yi Zheng]

DATE: 13th June, 2019

PLACE: Sydney, Australia

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Prof. Jinyan Li for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me a lot in all the time of my research and writing of this thesis. With his help, my research skills such as scientific writing, academic communication, and presentation have been improved to a new stage. Without his guidance and persistent help, this thesis and related research work would not be possible.

I would like to thank my co-supervisor Prof. Longbing Cao for his help on the enrollment of UTS and valuable suggestions for my research work. I also thank my Master's supervisor Quanyuan Wu, for his strong support when I apply for the CSC (China scholarship council) scholarship. Many thanks to Dr. Jie Yin and Dr. Xiaoying Gao, two of my research partners, for their insightful suggestions and patience to discuss the research problems with me.

My sincere thanks also go to my research team members Hui Peng, Zhixun Zhao, Xiaocai Zhang, Chaowang Lan, and Yuansheng Liu for their help in both my research and daily life. I am grateful to three former team members Dr. Jing Ren, Dr. Shameek Ghosh, and Dr. Renhua Song, for their kind help at the beginning of my Ph.D. study. I also want to thank Tao Tang and Xuan Zhang who join us recently. It's my honor to be one member of my research group. Thank you all for bringing me the feeling of home in Australia. I will memorize the unforgettable experiences forever.

In addition, I am grateful to acknowledge the funding sources of my study and conference travel, including the tuition fee and living expenses provided by the China Scholarship Council and Graduate Research School, travel funds provided by Faculty of Engineering and Information Technologies. Thanks to all staffs of Advanced Analytics Institute and School of Software who provide services and conveniences to my study and research in UTS.

At last, I really appreciate my wife Chengcheng Sun and my parents for their strong support during my overseas study. Especially my wife, she sacrificed her time and opportunities to support my study. A thank you to my relatives and friends in China, for your concern about my life and study abroad.

Yi Zheng
June 2019 @ UTS

Contents

List of Figures

List of Tables

List of Publications

The journal and conference papers published during my PhD study are listed as follows:

Related to the Thesis :

1. **Y. Zheng**, H. Peng, J. Li, et al. Inverse similarity and reliable negative samples for drug side-effect prediction [J], BMC Bioinformatics, 2019, 19(13): 554.

2. **Y. Zheng**, H. Peng, J. Li, et al. Predicting adverse drug reactions of combined medication from heterogeneous pharmacologic databases [J], BMC Bioinformatics, 2018, 19(S19).

3. **Y. Zheng**, H. Peng, J. Li, et al. Predicting Drug Targets from Heterogeneous Spaces using Anchor Graph Hashing and Ensemble Learning [C], 2018 International Joint Conference on Neural Networks. IEEE, 2018: 1-7.

4. **Y. Zheng**, S. Ghosh, J. Li, et al. An Optimized Drug Similarity Framework for Side-effect Prediction [C], Computing in Cardiology. IEEE, 2017: 1-4.

5. **Y. Zheng**, C. Lan, H. Peng, J. Li, et al. Using constrained information entropy to detect rare adverse drug reactions from medical forums [C], 2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC). IEEE, 2016: 2460-2463.

6. **Y. Zheng**, H. Peng, J. Li, et al. Old drug repositioning and new drug discovery through similarity learning from drug-target joint feature spaces [J], The 30th International Conference on Genome Informatics 2019 (transferred to BMC Bioinformatics). **(conference accepted, journal potentially accepted)**.

7. **Y. Zheng**, H. Peng, J. Li, et al. DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions [J], International Conference on Bioinformatics 2019 (transferred to BMC Bioinformatics). **(conference accepted, journal potentially accepted)**.

Others :

8. **Y. Zheng**, S. Ghosh, T. Lammers, J. Li, et al. Deriving Public Sector Workforce Insights: A Case Study using Australian Public Sector Employment Profiles [C], 12th International Conference on Advanced Data Mining and Applications (ADMA 2016), Springer, 2016: 764-774. **(co-first author)**

9. **Y. Zheng**, C. Sun, C. Zhu, et al. LWCS: A large-scale web page classification system based on anchor graph hashing [C], 2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2015: 90-94.

10. H. Peng, **Y. Zheng**, J. Li, et al. CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling [J]. Bioinformatics, 2018, 34 (18), 3069-3077.

11. H. Peng, **Y. Zheng**, J. Li et al. Recognition of CRISPR/Cas9 off-target sites through ensemble learning of uneven mismatch distributions [J]. Bioinformatics, 2018, 34 (17), i757-i765.

12. H. Peng, C. Lan, **Y. Zheng**, J. Li, et al. Cross disease analysis of co-functional microRNA pairs on a reconstructed network of disease-gene-microRNA tripartite [J]. BMC bioinformatics, 2017, 18(1): 193.

13. X. Zhang, Z. Zhao, **Y. Zheng**, and J Li. Prediction of Taxi Destinations Using a Novel Data Embedding Method and Ensemble Learning [J]. IEEE Transactions on Intelligent Transportation Systems, 2019.

14. X. Zhang, Y. Liu, **Y. Zheng**, Z. Zhao, J Li et al. Distinction Between Ships and Icebergs in SAR Images Using Ensemble Loss Trained Convolutional Neural Networks [C]. Australasian Joint Conference on Artificial Intelligence. Springer, 2018: 216-223.

15. Z. Zhao, H. Peng, C. Lan, **Y. Zheng**, J. Li et al. Imbalance learning for the prediction of N 6-Methylation sites in mRNAs [J]. BMC genomics 19 (1), 574.

Abstract

This research employs text and data mining methods to gain valuable knowledge for human drugs. Specifically, computational methods are developed for three topics, namely drug-side-effect prediction, drug-target identification, and drug-drug-interaction detection. The key innovations of the proposed methods lie in the feature space construction using medical domain knowledge, generation of reliable negative samples, and successful application of machine learning algorithms.

The drug-side-effect prediction problems are studied in Chapters 3-5. Side-effects are secondary phenotypic responses of human organisms to drug treatments. Side-effect prediction is an important topic for drugs especially in post-marketing surveillance because they cause significant fatality and severe morbidity. To overcome the limitations of existing computational methods such as lack of proper drug representation and reliable negative samples, this thesis presents three novel methods.

The first method is to predict side-effects for single drug medication as described in Chapter 3. A comprehensive drug similarity framework is developed by integrating several types of similarities measured by representative features of drugs first. Then reliable negative samples are generated through analyzing the comprehensive drug similarities. Trained with generated reliable negatives, the prediction performance of four classical classifiers are improved significantly, outperforming those state-of-the-art methods. Chapter 4 describes the method proposed to predict side-effects for combined medication of multi-drugs. A scoring method on a drug-disease-gene

tripartite network is developed to prioritize interacting drugs, paving a way to generate credible negative samples for side-effect prediction of combined medication. It creatively characterized a drug with its chemical structures, target proteins, substituents, and enriched pathways. The drug-drug pairs are represented as novel feature vectors to train binary classifiers for prediction. This novel representation and the inferred negative samples contribute to the superior performance of the proposed method in drug-drug-side-effect association prediction. Chapter 5 introduces the last method for detecting adverse drug reactions (ADRs, i.e., side-effects) from medical forums. It filters the cause-result relationship between drugs and ADRs using a self-built dictionary and detects drug-ADRs associations by information entropy. Compared with conventional co-occurrence based methods, the proposed method captures both high-frequency and low-frequency ADRs simultaneously. Besides, it returns drug-related ADRs only owing to the self-built relation dictionary.

Drug-target identification plays a crucial role in drug discovery. Existing computational methods have achieved remarkable prediction accuracy, however, usually obtain poor prediction efficiency due to computational problems. Chapter 6 presents a method to improve the prediction efficiency using an advanced technique named anchor graph hashing (AGH). AGH embeds data into low-dimensional Hamming space while maintaining the neighbourship. It turns the drug-target identification problem into a binary classification task where inputs are AGH-embedded vectors of drug-target pairs, and labels are judgments of their associations. Ensemble learning with random forest and XGBoost is employed to learn a good decision boundary. The proposed method is demonstrated to be the most efficient method and achieves comparable prediction accuracy with the best literature method.

Chapter 7 introduces a novel positive-unlabeled learning method named DDI-PULearn for large-scale detection of drug-drug interactions (DDIs). DDI-PULearn first generates seeds of reliable negatives via OCSVM (one-class support vector machine) under a high-recall constraint and via the

cosine-similarity based KNN (k-nearest neighbors) as well. Then trained with all the labeled positives (i.e., validated DDIs) and the generated seed negatives, DDI-PULearn employs an iterative SVM to identify the set of entire reliable negatives from the unlabeled samples. The identified negatives and validated positives are represented as vectors using the bit-wise similarity of corresponding drug pairs to train random forest for prediction. Its excellent performance is confirmed by comparing with two baseline methods and five state-of-the-art methods.

Chapter 1

Introduction

This chapter describes the background, research topics, contributions and structure of the thesis. Specifically, the studied drug properties including drug side-effects, drug targets and drug interactions, the data sources, and the researched techniques including text mining and data mining are presented in Section 1.1. The three topics researched in this thesis are described in Section 1.2. The contributions and structure of the thesis are detailed in Section 1.3 and Section 1.4 respectively.

1.1 Background

This thesis presents three topics of my Ph.D. research related to drug side-effects, drug targets, and drug-drug interactions. In this section, background knowledge of the researched drug properties, the data sources, and the researched text and data mining techniques are described.

1.1.1 Side-effects, targets, and drug-drug interactions

Our research focuses on studying three main drug properties including drug side-effects, drug targets, and drug-drug interactions.

Drug side-effects

Drug side-effects or adverse drug reactions (ADRs) are phenotypic responses to certain medicinal products which are unintended and noxious for the normal use of medication at normal doses (Karimi & Wang 2015, Edwards & Aronson 2000). Some side-effects are predictable consequences of the known mechanism of a drug (Niu & Xin 2015). Other side-effects are not predicted, caused by unintended activity at off-targets. Drug side-effects have gained public attention because of the significant morbidity and mortality they cause (Zheng & Ghosh 2017). For example, in America, drug side-effects are the fourth leading cause of deaths, resulting in 100,000 deaths every year (Giacomini, Krauss, Roden, Eichelbaum, Hayden & Nakamura 2007). A considerable amount of material resources and financial resources have spent to handle accidents caused by side-effects as well. Taking Australia as an example, 660 million dollars were spent on medicine-related hospitalizations including drug side-effects and medication errors in 2009 (Roughead & Semple 2009). In addition, side-effects are one of the major causes of drug failure in the drug development process (Liu & Wu 2012). Serious drug side-effects even lead to drug withdrawal from the market (Niu & Xin 2015). Consequently, predicting potential side-effects helps to reduce patients' risk and cost on society. Besides, the drug discovery process could benefit from the predictions as well.

Drug targets

Drug targets are molecular structures which interact with drugs and to which drugs are bound (Imming, Sinning & Meyer 2006). Drug targets can be mainly classified into two categories, namely protein drug targets and non-protein drug targets (Palmer, Chan, Dieckmann & Honek 2013). Most drug targets belong to the protein category, including enzymes, ion channel, G protein-coupled receptors (GPCRs), nuclear receptors, membrane transporters, and cytoskeletal proteins. Typical examples of non-protein drug targets are DNA, RNA and lipid membranes. Though the advance of

medical techniques, a large number of drug targets still remain unidentified. For example, PubChem (Kim & Thiessen 2015) contains about 35 million compounds, however, less than 7000 compounds have target information (Ding & Takigawa 2013). Identification of these potential targets for a given drug shows bright prospects in drug repositioning which aims to discover new therapeutic indications of existing drugs. Identification of interactions between drugs and targets could also facilitate the process of new drug discovery (Hopkins 2009) and the drug side-effect prediction (Brouwers, Iskar, Zeller, Van Noort & Bork 2011).

Drug-drug interactions

Drug-drug interactions (DDIs) refer to the efficacy change of one drug caused by the co-administration of another drug (Park, Kim, Ha & Lee 2015). DDIs may occur when two or more drugs are taken together or concomitantly. They can be categorized into three types: pharmaceutical, pharmacokinetic, and pharmacodynamic (Beijnen & Schellens 2004, Huang, Niu, Green, Yang, Mei & Han 2013). Pharmaceutical DDIs occur due to a physical or chemical incompatibility. Pharmacokinetic DDIs occur when one drug alters the absorption, distribution, metabolism, or elimination of another drug, resulting in changes in the concentration of the affected drugs (Cami & Manzi 2013). Pharmacodynamic DDIs occur if a drug interferes another drug at the target site, causing an antagonistic, additive, synergistic or indirect pharmacologic effect on the affected drug. In terms of severity, DDIs are categorized into minor, moderate and severe (Cami & Manzi 2013). Minor DDIs only call for routine patient monitoring, while moderate DDIs may require dosage changes and closer monitoring. Severe DDIs are of the most important clinical significance, possibly leading to serious side-effects and should be avoided.

DDIs account for around one-third of all adverse drug reactions (Strandell, Bate, Lindquist, Edwards & Swedish 2008, Huang, Temple, Throckmorton & Lesko 2007, Zheng, Peng, Zhang, Gao & Li 2018), leading to significant

morbidity and mortality worldwide (Vilar & Uriarte 2013). However, the use of multiple drugs is common in our daily life. One study (Bushardt, Massey, Simpson, Ariail & Simpson 2008) estimates that 29.4% of elderly patients in America are taking six or more drugs. When multi drugs are administered together, we are at the risk of adverse drug reactions as one drug can change the effect of other drugs. In addition, various factors could affect DDI interactions, such as activity of the liver enzyme system, and intake of certain food. Currently a few DDIs are identified via wet-lab experiments, however, a large number of DDIs remain unknown (Cheng & Zhao 2014). Thus, there is an urgent need to detect potential DDIs to reduce patients' risks and economic costs.

1.1.2 Pharmacologic databases and social media

We employ two types of data sources for research, namely online pharmacologic databases and social media.

Pharmacologic databases

With the development of information technology, more and more pharmacologic data are collected, stored, and available in public online databases. These pharmacologic data are usually processed and reviewed manually, thus they are of high data quality. The pharmacologic databases make it possible to mine potential knowledge from existing pharmacologic data. A brief introduction of the pharmacologic databases used in this thesis is described in Table 1.1.

DrugBank is a comprehensive, freely accessible online drug database which stores annotated bioinformatic and cheminformatic information of drugs (Wishart & Feunang 2017). DrugBank is updated frequently. Up to now, 12,666 drugs including 3,878 approved drugs and over 6,310 experimental drugs are available (version 5.1.3, released 2019-04-02). In our work, we extract drug information including drug chemical structures, substituents, indications, targets, and ATC codes as well as drug-drug

Table 1.1: **Pharmacologic databases used in this thesis.**

Database	Data type	URL
DrugBank	comprehensive drug information	https://www.drugbank.ca
SIDER	drug-side-effect associations	http://sideeffects.embl.de
EMBL-EBI	protein-GO associations	https://www.ebi.ac.uk
DrugCentral	drug-target associations	http://drugcentral.org
CTD	drug-gene, drug-pathway, drug-disease, disease-gene	http://ctdbase.org
Twosides	drug-drug-side-effect associations, drug-drug interactions	http://tatonettilab.org/resources/tatonetti-stm.html
PubChem	drug-chemical-substructure associations	https://pubchem.ncbi.nlm.nih.gov

interactions from DrugBank. **SIDER** is a side-effect database which contains information on marketed medicines and their recorded side-effects (Kuhn & Letunic 2015). The information is collected from package inserts and public documents. Information of side-effects including their frequencies, associated drugs, categories is provided. **EMBL-EBI** is a data warehouse which provides big data in biology (McWilliam, Li, Uludag, Squizzato, Park, Buso, Cowley & Lopez 2013). A wide range of services, tools, and databases are made freely available by EMBL-EBI. We extract the protein-GO (gene ontology) associations from EMBL-EBI. **DrugCentral** an open-access online drug compendium. It integrates structures, bio-activities, pharmacologic actions and indications for drugs approved by regulatory agencies, such as FDA (Food and Drug Administration) (Ursu & Holmes 2016). Drug-target associations from DrugCentral are used as supplement data for the drug-target research. **CTD** (comparative toxicogenomics database) provides information about interactions between chemicals and gene products, and

their relationships with diseases (Davis & Grondin 2016). The data are manually curated from the literature. Associations including drug-gene, drug-pathway, drug-disease, and disease-gene are extracted from CTD for our research. **Twosides** is a resource of polypharmacy side effects for pairs of drugs (Tatonetti & Patrick 2012). 868,221 associations between 59,220 drug pairs and 1301 side-effects which cannot be clearly attributed to either drug alone are stored in Twosides. **PubChem** (Kim & Thiessen 2015) is a public repository for information on chemical substances and their biological activities. It defines 881 chemical substructures to represent each chemical substance (e.g., drug) as an 881-bit fingerprint. In our research, we utilize the PubChem fingerprints as the chemical representation of drugs.

Social media

Web 2.0 has changed the manner we interact, even our lifestyles. People share their feelings and opinions including their medical situations (e.g., drugs they are taking, their reactions) on social media, such as Twitter, Facebook or social media websites. A survey shows that 61% of American adults searched for health information online, and 41% had read others' experience or commentary on social media (Fox 2011). Social media become an important data source for post-marketing drug surveillance. Compared with data from pharmacologic databases, social media data is of low data-quality due to much noisy information. However, social media have two advantages: (1) Information about drug-side-effect associations or drug-drug interactions may appear on social media a long time before it reaches other data sources. (2) Social media provide rich data automatically which largely reduce the cost of data collection for researchers.

Currently, there are still a few challenges to mine knowledge from the social media data source. The first challenge is to process unstructured data, free text in most cases. The commonly-used language on social media is the colloquial language which is full of nonstandard abbreviations, a wide variation of syntax, spelling and vocabulary (Leaman & Wojtulewicz 2010).

The second challenge lies in considerable noisy information in social media. To obtain useful information from such noisy data is challenging. The last challenge is that the mining results may lack interpretability. It is due to the way how social media data are produced.

1.1.3 Text mining and data mining

We utilize two techniques including text mining and data mining to mine knowledge for human drug understanding. Their definitions, mining procedures, and applications are presented as follows.

Text Mining

Text mining is the process of deriving potentially useful knowledge from unstructured textual data or documents, e.g. newspaper, blogs and articles. Different from data mining which usually mines information from structured and well-organized databases, text mining focuses on mining unstructured and free-format text. Its overarching goal is to translate those free texts into structured data using natural language processing techniques for further analysis. Consequently, text mining typically includes sub-tasks like information retrieval, name entity recognition and sentiment analysis, which have strong associations with natural language processing techniques. In general, the text mining process starts with a collection of documents, which are repeatedly fed for information extraction (Vidhya & Aghila 2010). Then these documents are parsed and pre-processed to remove noise data. Following that, features which are capable to represent documents are generated and part of them are selected as a feature set to vectorize each document. Finally, text mining techniques, for example, text clustering and classification tools, can be applied to these document vectors for further analysis.

Data Mining

Data mining is an interdisciplinary subfield of computer science (John 2010). Its goal usually lies in discovering significant and understandable information (e.g. patterns and trends) from a data set. To discover such information, data mining commonly involves six types of tasks, namely anomaly detection, association rule learning, clustering, classification, regression and summarization. Different task involves different mining process. However, data mining generally can be viewed an iterative process involves the following phases: problem definition, data exploration, data preparation, modeling, evaluation and deployment (Kurgan & Musilek 2006, Center 2013). To start with, the mining objective should be translated into a proper data mining problem definition. Then simple traditional data analysis tools (e.g. statistics) are leveraged to explore the data and obtain basic understanding of the data. After data exploration, the data will be preprocessed to enhance data reliability. During this phase, missing values and noise data are processed properly. Then follows the most important part of data mining, modeling, which requires careful selection of mining algorithms or tools according to the type of mining problems and data. Once the model is built, the evaluation phase will be activated. The model will be rebuilt by optimizing its parameters until it meets the expectations. Consequently, the modeling phase and model evaluation phase may iterate several times. Only after a satisfactory model is obtained, the model can be deployed for practical tasks.

Text and data mining in drug research

Recent advances in information technology have led to the exponential growth of different medical data sources, such as pharmacologic databases, electronic health records, medical literature, and social media. Text and data mining techniques can be used to address a variety of research questions on drugs using these unstructured and structured data sources. For example, text mining has been successfully applied to detect adverse drug

events (Leaman & Wojtulewicz 2010, Nikfarjam & Gonzalez 2011, Ginn, Pimpalkhute, Nikfarjam, Patki, O'Connor, Sarker, Smith & Gonzalez 2014, Patki, Sarker, Pimpalkhute, Nikfarjam, Ginn, O'Connor, Smith & Gonzalez 2014), screen drug interactions (Tari, Anwar, Liang, Cai & Baral 2010, Percha, Garten & Altman 2012, Iyer & Harpaz 2013) and guide drug discovery (Agarwal & Searls 2008, Andronis, Sharma, Virvilis, Deftereos & Persidis 2011, Zhou, Peng & Liu 2010). Text mining techniques make it possible to extract and relate information from large-scale unstructured text data automatically. Data mining shows its advantages over experimental approaches in drug side-effect prediction (Pauwels & Yamanishi 2011, Liu & Wu 2012, Zhang & Zou 2016, Zhang & Chen 2016), drug-target identification (Campillos & Kuhn 2008, Kuhn & Campillos 2008, Yamanishi, Kotera, Kanehisa & Goto 2010, Emig, Ivliev, Pustovalova, Lancashire, Bureeva, Nikolsky & Bessarabova 2013), drug-drug interaction detection (Vilar & Harpaz 2012, Vilar & Uriarte 2013, Zhang, Wang, Hu & Sorrentino 2015) and drug repositioning (Dudley, Deshpande & Butte 2011, Wu, Wang & Chen 2013, Luo & Wang 2016). Comparing with experimental methods which are time-consuming and expensive, data mining methods are much more efficient and with low-cost. Besides, tasks unfeasible using experimental methods become possible using data mining techniques. For instance, it's unfeasible to test all possible interactions between a new drug and existing approved drugs using experimental methods. However, it is not difficult for data mining methods (see examples in Chapter 7). We just need to collect the drug information which is available in most cases and choose a proper algorithm for mining. All in all, text and data mining provide us useful and powerful tools to gain valuable knowledge for human drugs from the big medical data.

1.2 Research topics

Three topics are researched in this thesis, namely side-effect prediction, drug-target association prediction, and drug-drug interaction identification. There are several specific questions to be solved for the above three topics. The detailed research questions and their formulations are listed in below Q1 to Q4.

Q1: Drug side-effect prediction from pharmacologic databases

There are many factors which can lead to side-effects of drugs, such as off-target activities or drug interactions. In this thesis, we distinguish side-effects caused by single drug medication with side-effects caused by combined medication of multi-drugs.

Most post-marketing surveillance studies focus on predicting side-effects for single drug medication only (Leaman & Wojtulewicz 2010, Hammann & Gutmann 2010, Yamanishi & Pauwels 2012, Zhang & Chen 2016, Zheng & Ghosh 2017). For a given side-effect s and a drug d_i, this research problem can be formalized into the following formula:

$$f(d_i, s) = \begin{cases} +1 & \text{if } d_i \text{ causes } s, \\ -1 & \text{else} \end{cases} \tag{1.1}$$

where $f(d_i|s)$ is the well-trained binary classifier to classify the drug d_i as positive (+1, if the drug causes the side-effect) or negative (-1, otherwise). A binary classifier is built and trained for each side-effect, where inputs are feature vectors of drugs, labels are the judgments of their associations (i.e., drugs cause the side-effect are labeled as positive, otherwise negative). To complete the task, we need to address the following problems: (1) Proper drug representation. Drugs have a variety of properties, we have to decide which property to use for their representations. (2) Lack of negative samples. Existing pharmacologic databases only collect validated drug-side-

effect associations (i.e. positives in this research question) but ignore the non-drug-side-effect associations (i.e., negatives). (3) Learning from unbalanced data. Some side-effects are associated with most drugs, however, others are associated with only a few drugs. Consequently, the learning has to be performed on the unbalanced data. (4) The remarkable performance of state-of-the-art methods. Several computational methods have been proposed to predict side-effects for single drug medication and achieved remarkable performance. It is a challenge to overcome their limitations and improve the prediction performance.

In this thesis, we also research on predicting side-effects caused by combined medication of multi-drugs. These side-effects are those cannot be clearly attributed to medication of a single drug alone. They are caused by interactions between drugs in most cases (Tatonetti & Patrick 2012). Specifically, we focus on predicting side-effects for drug pairs in this thesis. Compared with side-effect prediction for single drug medication, the side-effect prediction for combined medication is more complex: (1) Data collection. Conventional pharmacologic databases record the simple drug-side-effect associations alone. Validated associations between a drug pair and their side-effects are hard to get. (2) The representation of drug-pairs is more complex as one more drug is involved. (3) Lack of negative samples. The space of potential negatives is larger than side-effect prediction for single drug medication. (4) The validation of predicted results. We cannot perform wet-lab experiments to validate the predicted drug-drug-side-effect associations (DDSAs).

Q2: Adverse drug reaction detection from social media

Social media provide a new data source to identify drug-side-effect associations. In this thesis, we focus on detecting adverse drug reactions from medical forums, a typical type of social media. Medical forums are online sites where people discuss medical products and their health concerns via posting messages (Karimi & Wang 2015). The medical forums are usually

hierarchical: a medical forum contains several categories; each category has several topics; and within each topic, there are several discussions (i.e., threads) involving many users. The users submit their messages which may contain the time and their personal information to the discussion section. An advantage of medical forums over other types of social media such as Twitter and Facebook, is that they are highly health related. However, it is still not easy to detect drug-side-effect associations from the sea of information. The challenges for addressing this question are as follows: (1) Crawling web pages from medical forums. A specialized web-crawler has to be coded to crawl the whole web pages (HTML files) from the online forms. (2) Parsing, denoising and extracting post contents. The semi-structured HTML files should be parsed first, then the post contents are extracted from them. (3) Recognition of entity. Entities including drug names and the ADRs have to be identified from the post contents written in the colloquial language which is full of nonstandard abbreviations, a wide variation of syntax, spelling, and vocabulary. (4) Detection of relations. Relations between drugs and side-effects are detected using statistical methods (e.g., information entropy).

Q3: Drug-target association prediction

For a given drug d and a target t, the drug-target association prediction problem can be expressed as the following formula:

$$f(d,t) = \begin{cases} +1 & \text{if } t \text{ is targeted by } d, \\ -1 & \text{else} \end{cases} \tag{1.2}$$

where $f(d,t)$ is a well-learned binary classifier which classifies the association between drug d and target t as positive (+1, if t is targeted by d) or negative (-1, otherwise). The inputs of the classifier are combined feature vectors of drugs and targets, and the labels are the judgments of their associations. The following problems have to be solved: (1) Proper representation of drug-target associations. How to determine which properties of drugs and targets to use and how to combine them to represent their associations are

challenging. (2) Lacking validated negatives. Non-interacting associations between drugs and targets don't exist in existing databases. (3) Dimension reduction of the feature space. The dimension of association feature vectors is quite high which incurs a huge computational cost. (4) The prediction efficiency needs to be improved. There is enormous space to improve the prediction efficiency of state-of-the-art methods.

Q4: Drug-drug interaction identification

For a drug pair d_i-d_j, the identification of their interaction is performed with the following model:

$$f(d_i, d_j) = \begin{cases} +1 & \text{if } d_i \text{ interacts with } d_j, \\ -1 & \text{else.} \end{cases} \tag{1.3}$$

The function f is a trained classifier to determine whether the two drugs interact or not. The difficulties in this question are: (1) Transforming identification of drug-drug interactions into a binary classification task. Inputs are feature vectors of drugs and labels are judgments of their interactions. (2) Proper representation of drug pairs which can characterize them accurately. (3) Absence of non-interacted drugs. All unobserved DDIs which don't exist in validated drug-drug-interactions have considerable possibilities to be real DDIs. (4) Imbalanced learning. The number of potential negatives (i.e., unobserved DDIs) is much more than validated DDIs.

1.3 Research contributions

To address the above four research questions, we have proposed five novel methods as described from Chapter 3 to 7. Our contributions are summarized as follows (C1 to C5).

C1: Inverse similarity and reliable negative samples for drug side-effect prediction

In chapter 3, we present a novel method to predict side-effects for single drug medication with reliable samples generated by inverse drug similarity. Our contributions are: (1) Development of a drug similarity integration framework which integrates drug chemical structure similarity, drug target protein similarity, drug substituent similarity, and drug therapeutic similarity into a unified comprehensive similarity. (2) Selection of reliable negative drug samples for each side-effect from the candidate negative drugs (i.e., unlabeled drugs) based on their comprehensive similarities with validated positive drugs. (3) Investigation on impacts of the imbalance between negative samples and positive samples in the training set by comparing with a self-built balanced training set. (4) Use of machine learning methods to predict potential drug side-effects based on validated positive drug samples and the selected reliable negative drug samples, and compare the results with the existing predictive methods. (5) Extensive comparison experiments demonstrate that side-effect prediction using reliable negative drug samples selected based on the proposed drug similarity integration framework can achieve the best performance.

C2: Predicting side-effects of combined medication from heterogeneous databases

Chapter 4 describes a method to predict side-effects for combined medication of multi-drugs via integrating data from heterogeneous databases. The contributions of this work are three folds: (1) We represent drugs as multi-dimensional vectors according to their abundant features collected from heterogeneous databases. This makes it possible to employ advanced machine learning methods for predictions. (2) We design a scoring method on a drug-disease-gene tripartite network to prioritize interacting drugs, paving a way to select credible negative samples for drug-drug-side-effect prediction. (3) We

learn a high-performance machine learning model for drug-drug-side-effect prediction and demonstrate its superior prediction performance by comparing it with three baseline methods.

C3: Constrained information entropy for detecting adverse drug reactions from medical forums

In Chapter 5, a new method to detect adverse drug reactions (ADRs, i.e., side-effects) from medical forums via constrained information entropy (CIE) is introduced. Its contributions are in the following aspects: (1) A novel CIE-based method is proposed to detect ADRs from medical forums. (2) This method captures both high-frequency and low-frequency ADRs simultaneously. (3) This method only returns pure ADRs, which can reduce the false reaction rate from the detection results. (4) Extended experiments demonstrate the proposed method outperforms other state-of-the-art detection methods regarding to the completeness and accuracy of detection results.

C4: Predicting drug targets using anchor graph hashing and ensemble learning

Chapter 6 presents an improved method to predict drug targets using anchor graph hashing and ensemble learning. Our contributions include: (1) We integrate chemical structures, side-effects, substituents of drugs, and GO terms of targets in a unified framework for drug-target representation. (2) We are the first one to use the advanced technique "anchor graph hashing" in the drug-target prediction tasks which successfully improves the prediction efficiency significantly. (3) We employ ensemble learning of random forest and XGBoost for the final predictions which guarantees the prediction accuracy. (4) Related comparison experiments demonstrate that the proposed method has the highest efficiency while achieving comparable prediction accuracy with the best literature method.

C5: DDI-PULearn: a novel PU learning method for prediction of drug-drug interaction

In chapter 7, we describe the proposed PU (positive-unlabeled) learning method to predict drug-drug interactions. Our contributions are as follows: (1) We develop a novel positive-unlabeled learning method named DDI-PULearn for large-scale drug-drug interaction predictions. DDI-PULearn first generates seeds of reliable negative samples via OCSVM (one-class support vector machine) under a high-recall constraint and via the cosine-similarity based KNN (k-nearest neighbors). Then trained with all the labeled positives (i.e., the validated DDIs) and the generated seed negatives, DDI-PULearn employs an iterative SVM to identify the set of entire reliable negatives from the unlabeled samples. (2) We design a similarity-based method using abundant drug properties to represent drug-drug interactions. (3) We validate the performance of DDI-PULearn on simulative prediction for 149,878 possible interactions between 548 drugs, comparing with two baseline methods and five state-of-the-art methods. (4) Related experiment results show that the proposed method for the representation of DDIs characterizes them accurately. DDI-PULearn achieves superior performance owing to the identified reliable negatives, outperforming all other methods significantly.

1.4 Thesis structure

The structure of this thesis is illustrated in Figure 1.1 and briefly introduced as follows:

Chapter 1 introduces the background of the whole thesis, the research topics and the corresponding contributions. The structure of the thesis is illustrated in the end for ease of reading. **Chapter 2** presents related work about the researched questions. The current research progress and the limitations of existing methods are included. **Chapter 3** to **Chapter 7** describe the proposed methods for solving the research questions including drug side-effect prediction, drug-target identification, and drug-

drug interaction detection. Details of the methods, experimental evaluation and comparison, and corresponding case studies are included. **Chapter 8** provides a summary of the work and discussions of my future directions.

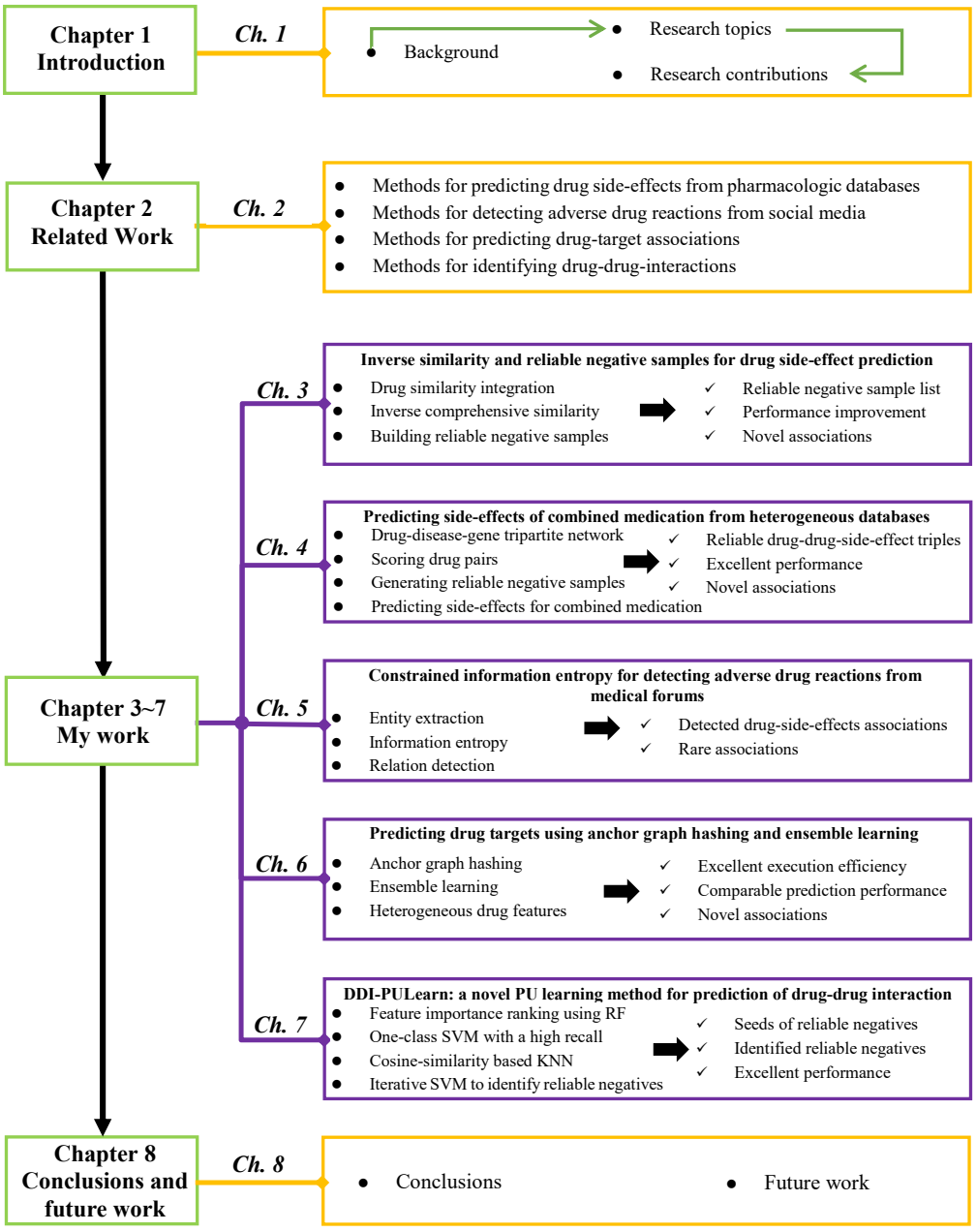


Figure 1.1: **Thesis Structure.** It consists of the following four parts: introduction, related work, my work, and conclusions and future work. Short introduction of each part is shown in the right side.

Chapter 2

Related work and literature review

This chapter describes the related work and literature review of the work in this thesis. The existing methods for predicting side-effects of single drug medication and combined medication of multi-drugs are surveyed in Section 2.1. Then, the state-of-the-art methods for detecting ADRs from social media are reviewed in Section 2.2. Following that, Section 2.3 introduces the published methods for identifying drug-target associations. In section 2.4, related studies about drug-drug-interaction detection and their limitations are presented. Finally, we briefly summarize the contents in this chapter.

2.1 Drug side-effect prediction from pharmacologic databases

2.1.1 Drug side-effect prediction for single drug medication

Conventional approaches for side-effect prediction of single drug medication are pharmacology assays such as *in vivo* assays and *in vitro* assays (Whitebread, Hamon, Bojanic & Urban 2005). However, such experimental

predictions are expensive, time-consuming, and tedious. Recently, several computational methods have been proposed to tackle the side-effect prediction problem based on drug profiles (Yamanishi & Pauwels 2012, Scheiber, Chen, Milik, Sukuru, Bender, Mikhailov, Whitebread, Hamon, Azzaoui & Urban 2009, Fukuzaki & Seki 2009, Bender & Scheiber 2007, Hammann & Gutmann 2010, Atias & Sharan 2011, Pauwels & Yamanishi 2011, Scheiber, Jenkins, Sukuru, Bender, Mikhailov, Milik, Azzaoui, Whitebread, Hamon & Urban 2009, Ma, Meng, Xiao, Li, Gao, Su & Zhang 2017, Lee, Huang, Chang, Lee & Lai 2017, Zhang & Zou 2016). These methods can be categorized into target protein-based methods and chemical structure-based methods.

The main idea of target protein-based methods is to relate drug side-effects to drug target proteins directly or indirectly. Previous studies on drug target identification by side-effects have demonstrated the strong associations between drug targets and drug side-effects (Campillos & Kuhn 2008, Mizutani, Pauwels, Stoven, Goto & Yamanishi 2012, Yamanishi, Kotera, Moriya, Sawada, Kanehisa & Goto 2014). In 2012, Yamanishi et al. (Yamanishi & Pauwels 2012) explored to predict drug side-effects by integrating target proteins and drug structures in a unified framework. Their experiments showed that the prediction accuracy can be improved owing to the integration of target protein information. Researchers developed prediction methods based on pathways which indirectly involve proteins targeted by drugs. Links between pathways and side-effects have been established by analyzing compounds which share the same toxic phenotypes and comparing these pathways with pathways modulated by nontoxic compounds (Scheiber, Chen, Milik, Sukuru, Bender, Mikhailov, Whitebread, Hamon, Azzaoui & Urban 2009). An efficient algorithm, named CoopeRativE Pathway Enumerator, was proposed to enumerate cooperative pathways which share common active conditions. Then these cooperative pathways were further adopted to predict side-effects and the method achieved satisfactory performance (Fukuzaki & Seki 2009).

The principle of chemical structure-based methods is to relate drug side-effects to drug chemical structures. Early in 2007, Bender et al. (Bender & Scheiber 2007) attempted to predict drug side-effects across hundreds of categories from their chemical structures alone and established correlations between the chemical space and side-effect prediction. Hammann et al. (Hammann & Gutmann 2010) used a decision tree to determine the chemical, physical and structural properties of drugs that predispose them to cause side-effects. Canonical correlation analysis (CCA) was employed for simultaneous prediction of multiple side-effects from chemical structures by Atias (Atias & Sharan 2011). Based on CCA, an improved method called sparse canonical correlation analysis (SCCA) was designed to predict potential drug side-effects from chemical substructures (Pauwels & Yamanishi 2011). Related comparison experiments with CCA have demonstrated that SCCA could provide more selective and informative correlation between drug chemical substructures and side-effects without losing performance. Schiber et al. (Scheiber, Jenkins, Sukuru, Bender, Mikhailov, Milik, Azzaoui, Whitebread, Hamon & Urban 2009) tried a global linkage analysis to map drug chemical features to side-effects on a large scale. However, the authors just aimed to relate chemical structures to side-effects instead of understanding the mechanism of relations. Moreover, they did not provide a framework to predict drug side-effects for drugs.

Usually for the side-effect prediction task, only positive samples which are composed of drugs known to have certain side-effects are validated. There is no definite prior knowledge that a drug is certain not to have a side-effect. Thus there are no validated negative samples in this task. Most existing side-effect prediction methods took labeled drugs as positive samples and unlabeled drugs as negative samples to perform the prediction directly (Pauwels & Yamanishi 2011, Yamanishi & Pauwels 2012, Bender & Scheiber 2007, Zhang & Chen 2016, Liu & Zhao 2016). Labeled and unlabeled drugs are drugs which are known and not known to have the side-effect respectively according to the prior knowledge. They solely depend on

the employment of known drug-side-effect relations (i.e., validated positive samples) to make predictions. However, the unlabeled drugs still have a considerable probability to have these side-effects. It means that the assumed negative samples may include a considerable number of real positive samples which are yet unknown. As a result, the quality of their negative samples cannot be guaranteed, and inaccurate selection of negative samples would largely degrade the prediction performance (Chen, Liu & Yan 2012).

2.1.2 Drug side-effect prediction for combined medication

Side-effects of combined medication occur when individually safe drugs interact pharmacokinetically or pharmacodynamically (Banda & Callahan 2016). Currently, two major approaches have been developed for detecting potential side-effects of combined medication: pre-marketing review and post-marketing surveillance. In the pre-marketing review process, *in vivo* and *in vitro* assays are employed to test new drugs with existing drugs to identify potential risks (Zhang & Huang 2009). However, it is experimentally unfeasible to test every possible interaction between a new drug and all existing drugs; let alone, some DDIs manifest only after multiple periods of exposure (Triaridis, Tsiropoulos, Rachovitsas, Psillas & Vital 2009, Zhang & Huang 2009). Therefore, post-marketing surveillance becomes important. Most post-marketing surveillance studies focus on predicting side-effects for single drug medication or identifying drug-drug interactions only (Liu & Zhao 2016, Zhang & Chen 2016, Yang & Jiang 2012, Liu & Chen 2013, Zheng & Ghosh 2017, Leaman & Wojtulewicz 2010, Yang & Jiang 2012, Yamanishi & Pauwels 2012, Hammann & Gutmann 2010). Post-marketing side-effects identification of combined medication has not been adequately studied.

Current post-marketing surveillance primarily relies on spontaneous reporting systems (SRSs). Harpaz et al. (Harpaz, Chase & Friedman 2010) applied association rule mining to detect side-effects of multi-drugs from 162,744 FAERS (Food and Drug Administration Adverse Event Reporting

System) reports. They successfully obtained 1,167 multi-drug side-effects associations, demonstrating the feasibility to detect side-effects of multi-drugs from SRSs using conventional data mining tools. Thakrar et al. (Thakrar, Grundschober & Doessegger 2007) investigated a multiplicative and an additive statistical model to detect side-effects of drug pairs from FAERS. They validated the two models on 4 known and 4 unknown drug-drug-side-effect associations (DDSAs), showing that all 8 DDSAs were successfully predicted by both models. Although SRSs are proved to be a useful data source to detect DDSAs. However, they suffer from common limitations such as high under-reporting ratio, duplicate reports, and delays in reporting, making them unable to detect side-effects of new or rare drugs.

With the increasing use of electronic health records (EHRs) in hospital and for research, researchers have also attempted to detect side-effects of combined medication from EHRs. For example, Banda et al. (Banda & Callahan 2016) investigated the feasibility of prioritizing DDSAs derived from EHRs using four different information sources. They first filtered out known DDSAs found in public databases such as Drugs.com (Wikipedia contributors 2018) and DrugBank (Law & Knox 2013) from the candidate DDSA list, and then derived a ranking score measured from the remaining sources. Finally, candidate DDSAs with the highest scores were predicted as identified associations and used for further validation. Iyer et al. (Iyer & Harpaz 2013) used standard methods that measure the disproportionality of the mention of adverse reactions to detect DDSAs from 50 million clinical notes. They showed that using clinical notes their method achieved as good performance as established methods on SRSs (i.e., FARES). In reality, EHRs often have restricted access, because they are always privately owned by health organizations (e.g., hospitals), available only to their cooperated research groups.

On the other hand, social media provides a new channel for people to share experiences and seek help online. A recent survey shows that 72% of Internet users went online to seek health information (White, Wang, Pant,

Harpaz, Shukla, Sun, DuMouchel & Horvitz 2016, Yang & Yang 2016). Several attempts have been made to detect DDSAs from social media. For example, White et al. (White, Tatonetti, Shah, Altman & Horvitz 2013) performed a large-scale study on Web search logs to detect a specific DDSA, i.e., paroxetine-pravastatin-hyperglycemia association. Yang et al. (Yang & Yang 2016) proposed to discover drug-drug interactions and DDSAs from MedHelp.org (MedHelp.org 2019, May 2), a popular online health community. They first built a heterogeneous healthcare network comprising entities (e.g., drugs, side-effects) extracted from the posts, and then learned a logistic regression classifier using features derived from nodes, links, and triads in the network to predict the occurrences of DDSAs, which achieved satisfactory results. However, social media text is often very noisy, phrased using colloquial language, and potentially contains inaccurate information, making it extremely difficult to extract mentions of drugs and side-effects and to establish their associations when context information (i.e., patient condition) is missing. This has significantly limited the accuracy and reliability of DDSA detection.

In summary, SRSs have the disadvantages of under-reporting and inability for new and rare drugs, EHRs are often difficult to access, and much noisy and unstructured information is contained in social media data. In addition, side-effects cannot be identified before happening using the above three data sources. Currently, more and more pharmacologic data have been collected and stored in public pharmacologic databases. Compared with the above three data sources, pharmacologic databases provide rich useful resources for identifying side-effects of combined medication, owing to the openness, high data quality, and coverage for both new and rare drugs. Nevertheless, the application of machine learning methods on pharmacologic databases is severely impeded by the lack of proper drug representation and credible negative samples. This is attributed to several reasons. First, each pharmacologic database contains different types of data and different aspects of information in relation to drugs, diseases, and side-effects.

Such scattered information populates over different databases, lacking an appropriate drug representation to establish the associations between drugs and side-effects. Second, pharmacologic databases, like Twosides databases (Tatonetti, Fernald & Altman 2011), define a set of positive samples, namely known drug-drug-side-effect associations included in the database. However, they often do not contain any explicit set of negative samples—drug-drug-side-effects that have no associations—and this poses a major difficulty in the use of machine learning methods to learn an optimal decision boundary.

2.2 Adverse drug reaction detection from social media

Several attempts to identify adverse drug reactions (ADRs) from social media have demonstrated promising results (Freifeld & Brownstein 2014, Leaman & Wojtulewicz 2010, Karimi & Cavedon 2011, Benton & Ungar 2011, Yang & Jiang 2012, Liu & Chen 2013, Yeleswarapu, Rao, Joseph, Saipradeep & Srinivasan 2014). Leaman et al. (Leaman & Wojtulewicz 2010) proposed to extract ADRs from health-related websites, specifically DailyStrength, with a combination lexicon. Following their work, Karimi et al. (Karimi & Cavedon 2011) presented to identify both beneficial and adverse drug reactions from patient forums. Benton et al. (Benton & Ungar 2011) extracted drug and reaction pairs from breast cancer message boards and evaluated specified 4 drugs by comparing the extracted ADRs with those on drug labels. Clark et al. (Freifeld & Brownstein 2014) mined ADRs from Twitter, aiming at evaluating the level of concordance between Twitter posts mentioning AE-like reactions and spontaneous reports received by a regulatory agency .

Different from other types of social media such as Facebook, medical forums are highly health related. Thus, ADRs detection from medical forums can produce more precise results. In 2012, Yang et al. (Yang & Jiang 2012) proposed a detection approach to mine the relationships between drugs and ADRs from MedHelp, a typical medical forum. Followed by that, Liu et al.

(Liu & Chen 2013) proposed an analytical framework for ADRs detection from patient forms. They developed the AZDrugMiner system to extract ADRs from the patient discussions and experiences. All researches above belong to the co-occurrence based methods, which reveals pattern according to the co-occurrences between drugs and ADRs. Though the co-occurrence based methods are able to capture high-frequency ADRs, however, it has several drawbacks: (1) It overlooks the rare ADRs. (2) It can't distinguish drug-related ADRs from unrelated ones. For instance, in the sentence 'We ask his lung doctor about an allergy medicine for mild allergy symptoms.', allergy medicine is prescribed to treat the allergy symptoms. However, allergy will still be recognized as an ADR of allergy medicine if their co-occurrence is larger than the pre-defined threshold.

2.3 Drug-target association identification

Though various biological assays are available nowadays, experimental identification of drug-target interaction remains challenging (Haggarty, Koeller, Wong, Butcher & Schreiber 2003). To guide and speed up the expensive and laborious experimental approaches, a variety of *in silico* prediction methods have been developed (Van & Nabuurs 2011). These approaches can be classified into two types, namely docking simulation and machine learning. Docking simulation is a powerful approach, however, it requires 3D structure information of the target proteins (Halperin, Ma, Wolfson & Nussinov 2002). Moreover, it is time-consuming, in that a large amount of computational resources are required (Ding & Takigawa 2013). Therefore, it's hard to apply this technique on a large scale. By comparison, machine learning is more efficient and capable in large-scale predictions.

Machine learning methods are based on the observation that similar drugs tend to interact with similar targets (Van & Nabuurs 2011). The prediction is performed based on information such as drug chemical structures (Martin, Kofron & Traphagen 2002, Nagamine & Sakakibara 2007), drug side-

effects (Campillos & Kuhn 2008), target protein sequences (Bleakley & Yamanishi 2009), and the known drug-target interactions. A straightforward approach is to use binary classification methods where the drug-target pairs are vectorized as input for binary classifiers, e.g., artificial neural network (ANN) and support vector machine (SVM) (Nagamine & Sakakibara 2007, Bleakley & Yamanishi 2009, Bock & Gough 2005, Faulon, Misra, Martin, Sale & Sapra 2007). Another encouraging approach named bipartite local model (BLM) (Bleakley & Yamanishi 2009) is to view drugs and targets as a bipartite graph. BLM builds one classifier from the drug side and one classifier from the target side for each drug-target interaction to be predicted. It can achieve remarkable accuracy, however, leading to a serious computational problem because a large number of classifiers have to be trained. The current state-of-the-art methods for predicting drug-target interactions using machine learning is to embed drug similarity measures and target similarity measures into kernels (Bleakley & Yamanishi 2009, Jacob & Vert 2008, Nagamine & Sakakibara 2007, Wassermann, Geppert & Bajorath 2009, Ding & Takigawa 2013). A good example is the pairwise kernel method (PKM) (Ding & Takigawa 2013) which treats drug-target interactions as instances and uses similarities between interactions to compute the kernel. By using kernels, multiple information sources can be integrated to achieve high prediction accuracy. However, one serious problem is the scale of the training set with "the number of training drugs times the number of training targets", bringing computational difficulties in large-scale applications.

As described above, existing machine learning methods either use vectors extracted from drug-target pairs or the similarity matrix as input for predictions. The dimension of the input vectors and the similarity matrix is quite high, which usually leads to serious computational problems. Reliable and high-performance compression methods are badly in need to tackle this problem.

2.4 Drug-drug interaction detection

Conducting experimental trials to detect potential interactions between a great number of drug pairs is unrealistic due to the huge time and monetary cost. Recently, several computational methods have been successfully applied to detect DDIs. Here, we categorize these methods roughly into three categories: similarity-based methods, knowledge-based methods, and classification-based methods.

The similarity-based methods assume that drugs with similar properties tend to interact with the same drug (Vilar & Harpaz 2012). Based on this assumption, different drug similarity measures have been designed employing various drug properties. Vilar et al. measured the drug similarity as the Tanimoto coefficient between molecular fingerprints (Vilar & Harpaz 2012) and between interaction profile fingerprints of drug pairs (Vilar & Uriarte 2013). Gottlieb et al. (Gottlieb, Stein, Oron, Ruppin & Sharan 2012) built their DDI predictive model by integrating seven drug similarity measures, namely chemical structure similarity, ligand similarity, side-effect similarity, annotation similarity, sequence similarity, closeness similarity in the protein-protein network, and Gene Ontology similarity. By using the drug-drug similarity indirectly, Zhang et al. (Zhang, Wang, Hu & Sorrentino 2015) designed a label propagation framework to predict DDIs based on drug chemical structures, labeled side-effects, and off-labeled side-effects. Similarity-based methods have achieved remarkable prediction performance, however, interactions for drugs lacking similarity information cannot be predicted. In addition, the assumption of similarity-based methods has one limit: dissimilar drugs may interact with the same drug.

The knowledge-based methods detect DDIs from scientific literature (He & Yang 2013), electronic medical records (Duke, Han, Wang, Subhadarshini, Karnik, Li, Hall, Jin, Callaghan & Overhage 2012), and the Food and Drug Administration Adverse Event Reporting System (FAERS) (Tatonetti, Denny, Murphy, Fernald, Krishnan, Castro, Yue, Tsau, Kohane & Roden 2011, Tatonetti, Fernald & Altman 2011). He et al. (He & Yang

2013) presented a Stacked generalization-based approach for automatic DDI extraction from biomedical literature. Tatonetti et al. (Tatonetti, Denny, Murphy, Fernald, Krishnan, Castro, Yue, Tsau, Kohane & Roden 2011) identified drug interactions and effects from FAERS using statistical methods. They found that interaction between paroxetine and pravastatin increased blood glucose levels. Knowledge-based methods rely on the accumulation of post-marketing clinical evidence. Consequently, they are incapable to detect all DDIs and cannot warn the public of the potentially dangerous DDIs before drugs reach the market.

Classification-based methods formulate DDI prediction as a binary classification task. Cami et al. (Cami & Manzi 2013) represented drug-drug pairs as feature vectors using three types of covariates from their constructed pharmacointeraction network. Then they defined the presence or absence of interactions as labels and finally built logistic regression models for predictions. Cheng et al. (Cheng & Zhao 2014) encoded each drug pair as a 4-dimensional vector of four different similarities, and employed five classical prediction algorithms for predictions. Compared with similarity-based methods and knowledge-based methods, classification-based methods don't have the assumption limitation or dependence on evidence accumulation. Nevertheless, two classes of data are required for classification methods: positive samples and negative samples. Existing classification-based methods used drug-pairs known to interact as positive samples, and other unlabeled drug-pairs as negative samples (Cami & Manzi 2013, Cheng & Zhao 2014). These unlabeled drug pairs may include a considerable number of real positive samples which can degrade the prediction performance.

From the above survey, it is understood that similarity-based methods and knowledge-based methods are limited to their application ranges, while classification-based methods are lack of reliable negative samples.

2.5 Summary

This chapter mainly reviews existing methods for addressing research problems including drug-side-effect prediction from pharmacologic databases, adverse drug reaction detection from social media, drug-target association prediction, and drug-drug interaction detection. Details of these methods and their limitations are introduced. In conclusion, the proper sample representation is important in designing the models. Combining domain-knowledge features and feature selection can improve the prediction performance significantly. Besides, the generation of reliable negative samples plays a key role to address these problems. The common challenge of the above problems lies in the lack of negative samples. There is no doubt that high-quality negative samples can improve the performance of the supervised models a lot.

Chapter 3

Inverse similarity and reliable negative samples for drug side-effect prediction

3.1 Introduction

As mentioned in Section 2.1.1, there are no validated negative samples in the side-effect prediction task. Existing side-effect prediction methods took unlabeled drugs as negative samples to perform the prediction directly. However, the unlabeled drugs still have considerable probabilities to cause these side-effects. It means that the assumed negative samples may include a considerable number of real positive samples which are yet unknown. As a result, the quality of their negative samples can not be guaranteed, and inaccurate selection of negative samples would largely degrade the prediction performance (Chen, Liu & Yan 2012). To improve the prediction performance, we need good methods to select reliable negative samples for the prediction task.

Existing side-effect prediction methods are essentially based on the assumption that similar drugs are inclined to share the same side-effects (Zhang & Chen 2016), which have generated remarkable performance. Thus,

it is rational to determine potential negative samples based on the inverse proposition that dissimilar drugs are less likely to share the same side-effects. Based on this hypothesis, we proposed a novel method to select highly-reliable negative samples based on the comprehensive drug similarity from the set of unlabeled drugs. The comprehensive drug similarity measurement integrates the chemical space of drug chemical structures, biological space of drug target proteins and other space of drug substituents, and therapeutic classification into a unified framework, to capture drug features from different perspectives. Candidate negative drugs which have lower similarity scores with validated positive drugs were preferentially selected to form the negative drug sample set. The performance of the proposed method was evaluated on simulative side-effect prediction of 917 DrugBank drugs, comparing with four machine-learning algorithms. Extensive experiments show that the drug similarity integration framework has superior capability in capturing drug features, achieving much better performance than those based on a single type of drug property. Besides, the four machine-learning algorithms achieved significant improvement in macro-averaging F1-score (e.g., SVM from 0.655 to 0.898), macro-averaging precision (e.g., RBF from 0.592 to 0.828) and macro-averaging recall (e.g., KNN from 0.651 to 0.772) complimentarily attributed to the highly-reliable negative samples selected by the proposed method.

3.2 Materials

3.2.1 Drug side-effect profiles

The side-effect data set was downloaded from SIDER (version 4.1), a comprehensive drug side-effect database (Kuhn & Letunic 2015). In this chapter, we focus on side-effects of drugs which are grouped as "Small Molecules" in the DrugBank database (Law & Knox 2014). Our basic idea lies in predicting drug side-effects according to drug similarities. Therefore, those drugs whose similarity information are not available were removed.

Finally, we obtained a data set of 917 drugs, 500 side-effect terms, and 78,855 drugs side-effect pairs (DSPs). Details of the dataset are described in both Table 3.1 and Figure 3.1. All the above data and the source codes are included in **Additional file 3**.

Table 3.1: **The drug side-effect dataset.**

Field	Value
Drug Group	Small Molecules
Number of Drugs	917
Number of Side-effects	500
Number of DSPs	78,855
Side-effects per Drug	86.0
Drugs per Side-effect	157.7

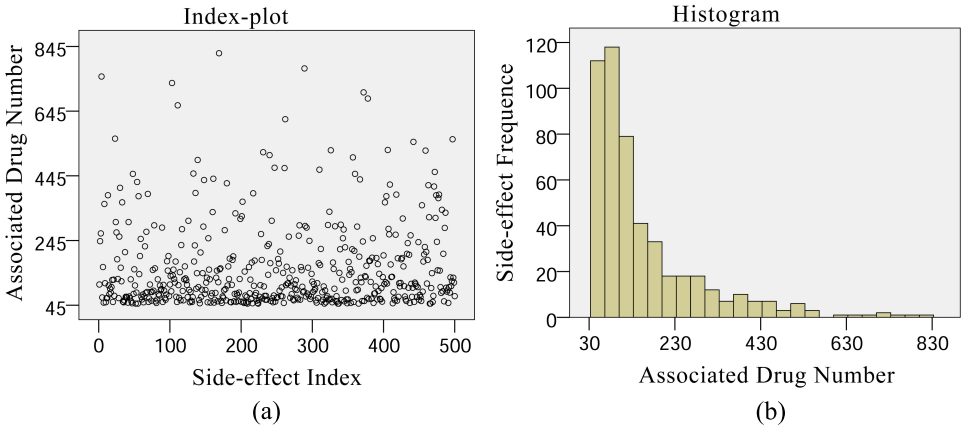


Figure 3.1: **Characteristics of side-effects and their associated drugs.** The left panel is the index-plot of the number of associated drugs for each side-effect and the right panel is the histogram of the associated drug number for the side-effects.

3.2.2 Drug similarities

Chemical structure similarity

Chemical structures of drug molecules were downloaded from DrugBank (stored in SMILES files) (Law & Knox 2014). The molecular fingerprints were retrieved from these SMILES files using an open source tool called Chemical Development Kit (CDK) (Steinbeck, Han, Kuhn, Horlacher, Luttmann & Willighagen 2003). The chemical structure similarity score between two drugs is calculated according to the Tanimoto 2D score between their molecular fingerprints. For a drug d, it can be represented by its hybridization fingerprint $f^d(f_i^d \in \{0,1\}, i \in \{1...1024\})$. Then the chemical similarity score between drug d_j and drug d_k is given by:

$$S_{chem}(d_j, d_k) = Tanimoto(f^j, f^k) = \frac{\sum_{l=1}^{1024}(f_l^j \wedge f_l^k)}{\sum_{l=1}^{1024}(f_l^j \vee f_l^k)} \tag{3.1}$$

where $\wedge$ and $\vee$ are bit-wise "and" and "or" operators respectively; f_l^j and f_l^k are the l^{th} bit of fingerprints of drug d_j and drug d_k respectively.

Drug target protein similarity

The similarity of two proteins is calculated based on the overlapping rate of their associated Gene Ontology (GO) terms. Suppose GO^m and GO^n are the GO term sets for protein p_m and protein p_n respectively, the similarity score between p_m and p_n is defined as

$$S_{go}(p_m, p_n) = \frac{GO^m \cap GO^n}{GO^m \cup GO^n} \tag{3.2}$$

where $\cap$ and $\cup$ are "intersection" and "union" operators respectively. The GO terms of target proteins were downloaded from the EMBL-EBI website (Consortium 2004, Lan, Chen & Li 2016). Since drugs are expected to interact with proteins which have similar cellular components, or share similar molecular functions, or go through similar biological processes. Therefore, all the three types of ontologies were utilized in the similarity definition. Then the drug target protein similarity between each pair of drugs

was calculated by integrating protein similarities of their target proteins. The target protein similarity score between drug d_j and drug d_k is calculated as follows:

$$S_{tar}(d_j, d_k) = \frac{\sum_{m=1}^{N_j} \sum_{n=1}^{N_k} S_{go}(p_m, p_n)}{N_j * N_k} \tag{3.3}$$

where N_j and N_k are the total number of proteins in the interacted protein sets of drug d_j and drug d_k respectively.

Drug substituent similarity

Substituents are atoms which replace hydrogen atoms on the parent chain of hydrocarbon. Its subsets, functional groups, are responsible for the characteristic chemical reactions of molecules. Therefore, it's reasonable to measure similarity between drugs via their substituents. The drug substituent similarity between drug d_j and drug d_k is calculated via Jaccard score which is given by:

$$S_{sub}(d_j, d_k) = \frac{SUB_j \cap SUB_k}{SUB_j \cup SUB_k} \tag{3.4}$$

where SUB_j and SUB_k are the substituent sets of drug d_j and d_k respectively; $\cap$ and $\cup$ are intersection and union operators respectively.

Drug therapeutic similarity

The Anatomical Therapeutic Chemical (ATC) codes of drugs are assigned according to their therapeutic, pharmacological and chemical properties (Chen, Zeng, Cai, Feng & Chou 2012). The ATC codes have been demonstrated to be useful in predicting the drug poly-pharmacological profiles (Cheng, Li, Wu, Wang, Zhang, Li, Liu & Tang 2013). Hence, we take the ATC codes as one part of the drug similarity measurement. The ATC codes used in this chapter were extracted from the DrugBank database. There are 5 levels in the ATC code. First, we calculate the drug therapeutic similarity at each level separately. The l^{th} level drug therapeutic similarity

S_l between the drug d_j and d_k is defined as:

$$S_l(d_j, d_k) = \frac{ATC_l(d_j) \cap ATC_l(d_k)}{ATC_l(d_j) \cup ATC_l(d_k)} \tag{3.5}$$

where $ATC_l(d_j)$ denotes the l^{th} level ATC code for drug d_j; $\cap$ and $\cup$ are intersection and union operator respectively. The average value of the five-level similarity scores is used as the therapeutic similarity of a drug pair:

$$S_{thera}(d_j, d_k) = \frac{\sum_{l=1}^{n} S_l(d_j, d_k)}{n} \tag{3.6}$$

where $n = 5$, is the total number of ATC code levels.

3.3 Methods

We transformed the side-effect-prediction problem into a set of independent binary classification problems, where each drug causes or does not cause a given side-effect. For each side-effect, we built a classifier using its validated positive drugs and selected reliable negative drugs. In this section, the framework to integrate drug similarity is introduced first. Then, processes to select reliable negative samples based on integrated drug similarities are detailed. And finally, the way to build classifiers is presented.

3.3.1 Drug similarity integration framework

We integrate the above four measurements of drug similarities into a single comprehensive similarity using three consensus similarity inference methods: maximum, mean and geometric mean.

(1) Maximum

$$\begin{aligned} S_{max}(d_j, d_k) = max\{&S_{chem}(d_j, d_k), S_{tar}(d_j, d_k) \\ &, S_{sub}(d_j, d_k), S_{thera}(d_j, d_k)\} \end{aligned} \tag{3.7}$$

(2) Mean

$$\begin{aligned} S_{mean}(d_j, d_k) = [&S_{chem}(d_j, d_k) + S_{tar}(d_j, d_k) \\ &+S_{sub}(d_j, d_k) + S_{thera}(d_j, d_k)]/4 \end{aligned} \tag{3.8}$$

(3) Geometric Mean

$$S_{GM}(d_j, d_k) = \sqrt[4]{S_{chem}(d_j, d_k) * S_{tar}(d_j, d_k) * S_{sub}(d_j, d_k) * S_{thera}(d_j, d_k)} \tag{3.9}$$

where $S_{chem}(d_j, d_k)$, $S_{tar}(d_j, d_k)$, $S_{sub}(d_j, d_k)$ and $S_{thera}(d_j, d_k)$ are chemical similarity, target protein similarity, drug substituent similarity and drug therapeutic similarity between drug d_j and d_k respectively; $S_{max}(d_j, d_k)$, $S_{mean}(d_j, d_k)$ and $S_{GM}(d_j, d_k)$ are the three combined comprehensive similarities for the drug pair drug d_j and d_k.

3.3.2 Negative sample selection based on comprehensive drug similarities for side-effect predictions

Most existing prediction methods assume that similar drugs tend to share the same side-effects. We adopt not only this assumption, but also its inverse proposition to make predictions. Particularly, we adopt its inverse proposition, i.e., dissimilar drugs are less likely to share the same side-effects, to select highly-reliable negative samples. Figure 3.2 illustrates the flow diagram of the proposed method to select negative samples for predictions.

Starting with the preprocessing procedure, side-effects and drugs which do not meet the requirements are removed (see Section 3.2.1). Then we compute the comprehensive similarity between every two drugs from the 917 drugs. Following that, we build the positive drug set and the negative drug set for each side-effect. Here we take the side-effect *se* as an example to describe the process.

- Build the positive drug set and the candidate negative drug set for *se*. Specifically, the positive drug set and the candidate negative drug set are formed by drugs which are known and unknown to cause *se* in the prior knowledge respectively.

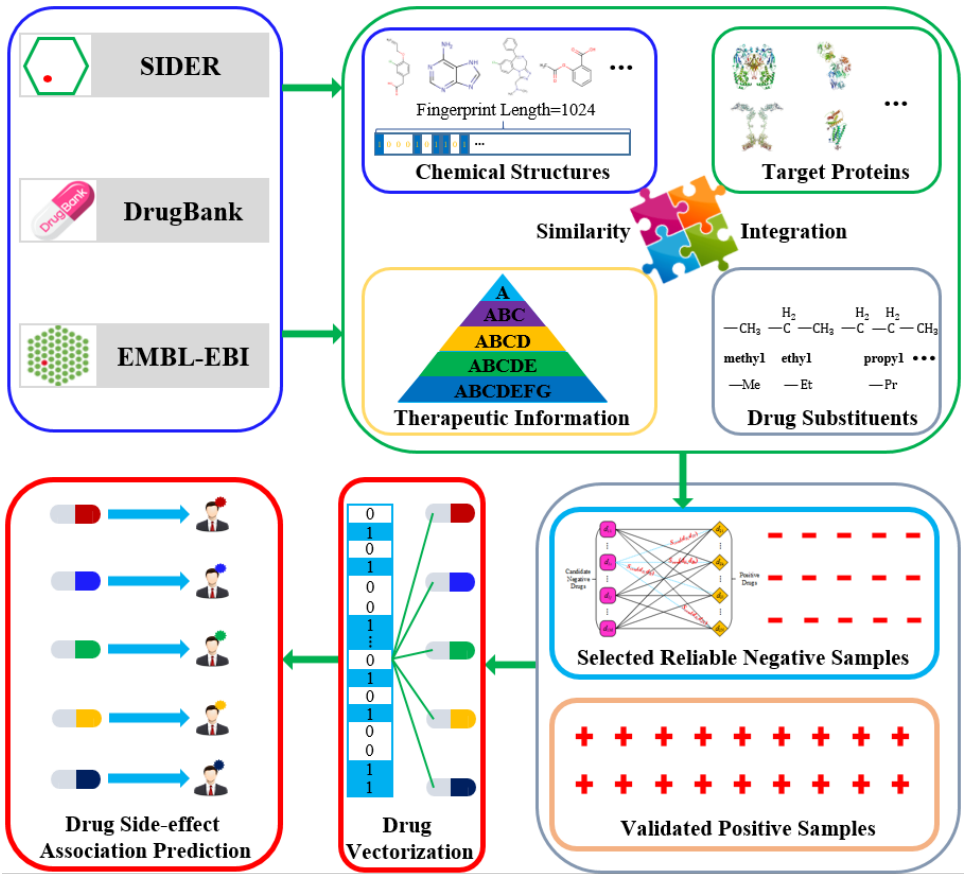


Figure 3.2: **Flow diagram of drug side-effect prediction with the proposed negative sample selection method using the comprehensive drug similarity.**

- Calculate the accumulative similarity score for each drug in the candidate negative drug set using the comprehensive drug similarity.
  Example. The accumulative similarity score of a candidate negative drug d_{1i} equals the sum of similarities between d_{1i} and each drug in the positive drug set.

$$Score_{d_{1i}} = S_{com}(d_{1i}, d_{21}) + ... + S_{com}(d_{1i}, d_{2j}) + ... + S_{com}(d_{1i}, d_{2N}) \tag{3.10}$$

  where $S_{com} \in \{S_{max}, S_{mean}, S_{GM}\}$, is the comprehensive similarity; d_{2j} is the jth drug in the positive drug set $(1 \leq j \leq N)$; N is the total

drug number in the positive drug set.

- Rank all drugs in the candidate negative drug set in an ascending order of their accumulative similarity scores, and those with lower scores are preferably selected to form the negative drug set. The threshold score depends on the number of negative drugs should be selected.

After we get the validated positive drug set and selected negative drug set for all side-effects, we vectorize each drug as a 917-dimensional feature vector using comprehensive drug similarities. For example, drug d_i is represented as $d_i = \{S_{com}(d_i, d_1), ..., S_{com}(d_i, d_j), ..., S_{com}(d_i, d_{917})\}^T$, where each element encodes the comprehensive drug similarity between d_i and each drug from the whole drug set. Finally, we construct one classifier for each side-effect, optimize related parameters via cross-validations, and finally use the optimized classifier to predict potential drug-side-effect associations. The inputs are vectors of the validated positive drug samples and the selected negative drug samples.

3.4 Results and discussions

The prediction results under different similarity measurements and different similarity integration methods are presented. The prediction performances on balanced and imbalanced training sets are also compared through a series of experiments. Finally, we demonstrate the excellent prediction performances of approaches achieved by using the highly-reliable negative samples selected based on the inverse similarity hypothesis and the similarity integration framework.

3.4.1 Evaluation on drug similarity integration framework

To demonstrate advantages of the proposed drug similarity integration framework, we report prediction performances under different similarity

measurements and different similarity integration methods. We tested eight situations derived from the four similarity measurements and three similarity integration methods: (1) Chem, (2) Tar, (3) ChemTarMax, (4) ChemTarMean, (5) ChemTarGM, (6) ComMax, (7) ComMean and (8) ComGM, where Chem, Tar, ChemTar and Com denote the four similarity measurements, namely the chemical, target and chemical-target and comprehensive similarity respectively; Max, Mean and GM denote the three similarity integration approaches, namely maximum, mean and geometric mean respectively. A typical classifier, k-nearest neighbors (KNN), was employed to perform these tasks. The k parameter of KNN was set as 50 based on 5-fold cross validation optimization. In addition, all drugs known and unknown to cause the side-effect were directly used as positive and negative samples for prediction. Detailed results of the eight situations are illustrated in Figure 3.3.

It can be seen that for situations based on chemical-target similarity as well as the comprehensive similarity, the integration method "Mean" and "Maximum" perform much better than "Geometric Mean", and "Maximum" outperform "Mean" a bit (detailed in Figure 3.4 (a), Figure 3.4 (b) and Figure 3.4 (c)). Therefore, we leveraged "Maximum" as the integration method in the subsequent experiments. As for performances of different similarity measurements, chemical-target based situations achieved better performance than chemical-structure based situations and target-protein based situations (see Figure 3.4 (e) and Figure 3.4 (f)). Situations based on the proposed comprehensive similarity outperformed the chemical-target based situation (see Figure 3.4 (d)), producing the best results. Examples of side-effects with low $F1$-scores are "unspecified visual loss", "hyperphosphataemia", "corneal opacity", "red blood cell sedimentation rate increased", and "exacerbation of asthma". Examples of high $F1$-score side-effects are "pseudomembranous colitis", "nausea", "febrile neutropenia", "headache", and "vomiting". Please refer to the specific $F1$-scores of each situation in Table S3 in **Additional file 2**.

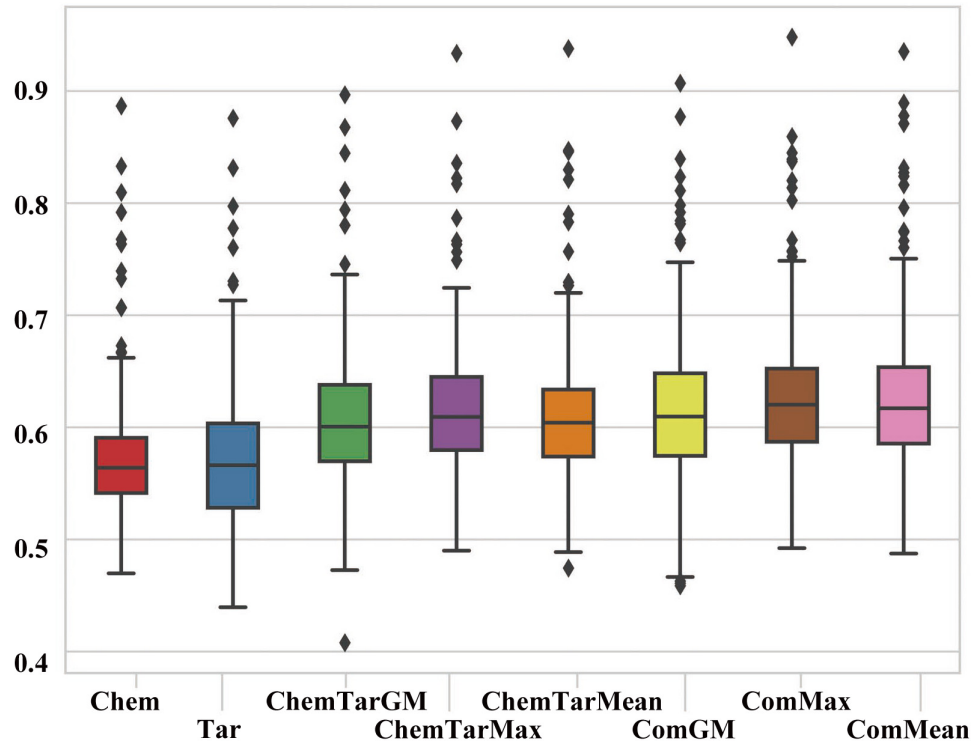


Figure 3.3: **Boxplots of the *F*1-scores for different similarity measurements and different similarity integration methods using the KNN classifier.**

3.4.2 Evaluation on balanced and imbalanced training set

It can be seen from Figure 3.1 that the number of drugs, with which a side-effect is associated, varies a lot. It means the labeled drugs and unlabeled drugs for most side-effects are imbalanced. In fact, 479 (95.8%) side-effects have more unlabeled drugs than labeled drugs. If all labeled and unlabeled drugs were directly taken as the positive and negative samples respectively to perform the side-effect prediction, the imbalance between them would become a problem that is challenging in the field of machine learning (Tan 2005). We investigated the impact of imbalance between positive samples and negative samples on side-effect prediction performance, before reporting results of the proposed negative sample selection method.

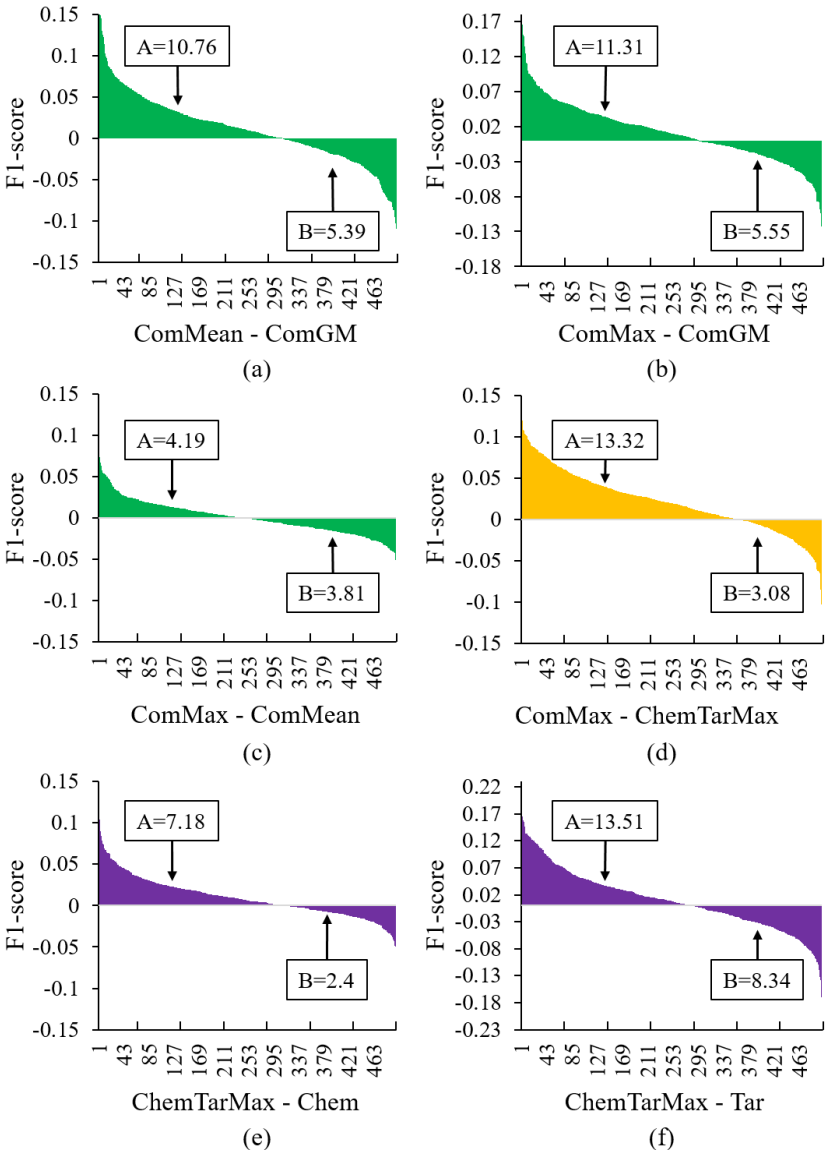


Figure 3.4: **Differences among similarity measurements and similarity integration methods.** In each panel, the x-axis denotes the index of each side-effect and the y-axis denotes the $F1$-score difference between two methods. For instance, Figure 3.4 (a) describes the differences between "ComMean" and "ComGM" using the $F1$-score of each side-effect from ComMean minus that from ComGM (i.e. difference=$F1$-score(ComMean)$-F1$-score(ComGM)). Thus we can identify which method performs better by comparing the area under the curve above zero (i.e., area A) with the area above the curve under zero (i.e., area B).

To simply compare with the original imbalanced training set, we developed a balanced training set in the following way: (1) the smaller number n_s, between the labeled drug number and the unlabeled drug number was obtained; (2) n_s labeled drugs were selected to form the positive drug sample set; (3) n_s unlabeled drugs were selected to form the negative drug sample set. More efficient classifiers, including extreme learning machine (ELM), support vector machine (SVM), and radial basis function (RBF) networks were employed in this task. We tested eight situations: (1) KNNComBal, (2) KNNComUnbal, (3) ELMComBal, (4) ELMComUnbal, (5) SVMComBal, (6) SVMComUnbal, (7) RBFComBal and (8) RBFComUnbal. Since the best prediction performance obtained among the three drug similarity integration methods was via "Maximum", we used it again as the integration method for the eight situations. These situations are named according to the classifier involved, similarity measurements, and balance or not. Situations ended with "Bal" are based on the self-built balanced training set and those ended with "Unbal" are based on the original imbalanced training set. For example, SVMComBal refers to the situation which is performed by SVM based on the self-built balanced training dataset using the comprehensive similarity.

To avoid bias, situations based on the self-built balanced training set were repeated 5 times. Each time, the n_s positive drugs and the n_s negative drugs were selected randomly. Average values of the performance evaluation metrics were used for comparison. The scatter plots of prediction results are shown in Figure 3.5. From it, we can see that most dots are located at the upper side of the reference line for all the classifiers. It reveals that situations based on the balanced training set outperform those on the imbalanced training set for most side-effects.

In Figure 3.5 (b) and Figure 3.5 (c), a few dots (i.e., side-effects) are concentrated in the left side of the chart and are close to the y-axis, where $F1$-scores of the imbalanced situations are close to 0. It is caused by the large degree of imbalance between the positive and negative samples in the training

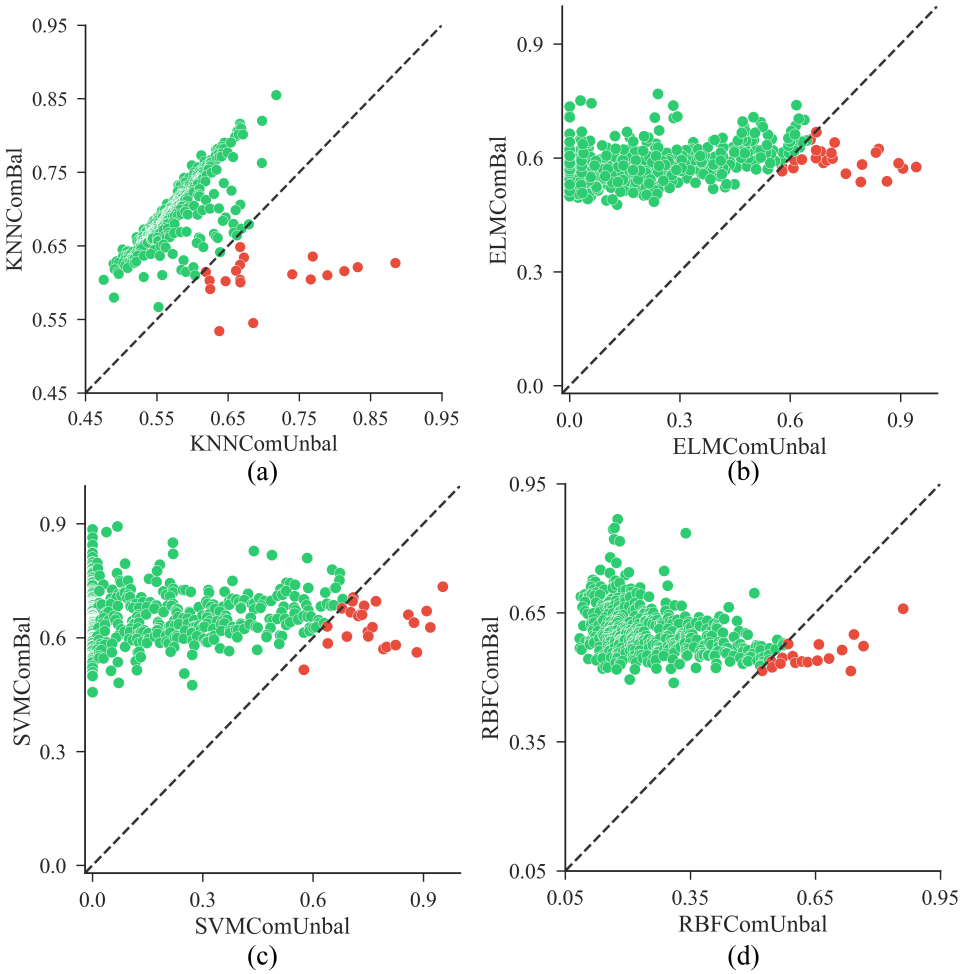


Figure 3.5: **Scatter plots of *F*1-scores for different classifiers using the comprehensive similarity on balanced and imbalanced training sets.** The x-axis denotes $F1$-scores of results based on imbalanced training sets, and the y-axis for balanced training sets. The line "$y = x$" on which $F1$-scores are equal, is the reference line to better visualize the results. Dots above the reference line are colored green while dots below the reference line are colored red.

set. Moreover, ELM and SVM are sensitive to the imbalanced training set. We examined the ratios of positive and negative samples in the imbalanced training sets of these side-effects. For most low $F1$-score side-effects, the number of negative samples in the training set is several times more than

positive samples. For example, in terms of side-effects with $F1\text{-score} \leq 0.1$ using ELMComUnBal, the negative samples are 10.75 times that of positive samples on average. A large degree of imbalance led to the bias of the classification decision boundary against the positive samples, therefore very few samples were predicted as positive drugs. It means the true positive rate is low, resulting in a low $F1$-score.

In Figure 3.5 (b), Figure 3.5 (c) and Figure 3.5 (d), a majority of side-effects for which the imbalanced situations outperformed the balanced ones, have relatively high $F1$-scores. Analogously, we investigated the numbers of positive and negative samples for these side-effects. It was found that most of these side-effects have more positive samples in the training set. Taking results from SVMComUnBal as an example, 21 out of 23 side-effects which own higher $F1$-scores than SVMComBal, have more positive samples in the training set.

With the above analyses, it is suggestive that the ratios of positive and negative samples in the training set can influence the prediction performance a lot. Using imbalanced training sets, classifiers are easy to fall into bias. Therefore, situations with balanced training sets can perform better than those situations with imbalanced training sets on most side-effects. This observation can also be confirmed by prediction results using the similarity ChemTar (Figure S1 in **Additional file 1**).

3.4.3 Performance improvement brought by the selection of highly-reliable negative samples

From analyses presented in the previous sections, it has been understood that different similarity measurements, different similarity integration methods, and the balance ratios between positive and negative samples have heavy influence on the prediction performance. It is confirmed that approaches based on self-built balanced training sets achieved better performance than those based on original imbalanced training sets. With the best options of similarity measurements, integration methods and balance strategy, we

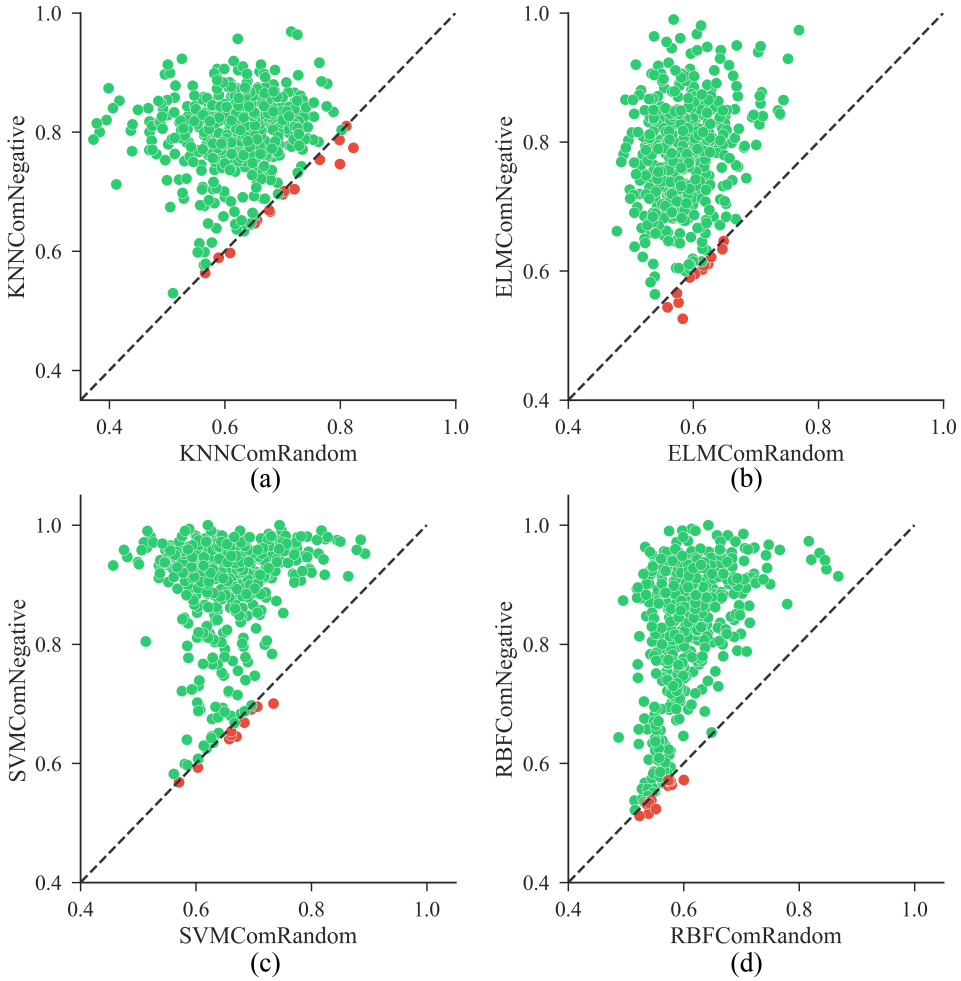


Figure 3.6: **Comparison results using the proposed negative sample selection method and random sample selection method.** The x-axis denotes $F1$-scores of results based on negative samples selected randomly, and the y-axis denotes $F1$-scores of results based on negative samples selected by the proposed method. The line "$y = x$" on which $F1$-scores are equal, is used as the reference line. Dots above the reference line are colored green while dots below the reference line are colored red.

evaluated the performance of side-effect prediction when negative samples are selected by the proposed method. As a comparison, we randomly selected negative samples for each side-effect from its unlabeled drugs. We treat this method as the "Baseline". We tested the following situations:

(1) KNNComNegative, (2) KNNComRandom, (3) ELMComNegative, (4) ELMComRandom, (5) SVMComNegative, (6) SVMComRandom, (7) RBF -ComNegative and (8) RBFComRandom, where situations end with "Negative" stand for the negative samples are selected by the proposed negative sample selection method, while those ended with "Random" denote the negative samples are selected randomly. To avoid bias, situations based on randomly selected negative samples were repeated 5 times. Note that all the above situations are based on the comprehensive similarity, the similarity integration method "Maximum", and the proposed strategy to build balanced training sets (see Section 3.4.2). Related results are illustrated in Figure 3.6. As shown in all sub-graphs of Figure 3.6, a majority of dots are located at the upper side of each reference line. This suggests that performances of all classifiers were significantly improved owing to the proposed negative sample selection method.

We further investigated the improvement on macro-averaging values of the three performance measurement indices (see Table 3.2). They were calculated as the average values of all side-effects using formula (A.13). The highest performance of each classifier was highlighted via bold numbers. Clearly, situations using the proposed negative sample selection method based on the comprehensive similarity (ComNegative) achieved significantly higher performance than those using randomly selected negative samples based on comprehensive similarity (ComRandom) or chemical-target similarity (ChemTarRandom, see Figure S2 in **Additional file 1**). For example, for the four classifiers from KNN to RBF, the macro-averaging $F1$-score improvement over "ComRandom" is 18.1%, 18.8%, 24.3%, 21.6% respectively, over "ChemTarRandom" is 21.9%, 20.2%, 29.2%, 25.1% respectively.

Both the results from Figure 3.6 and Table 3.2 confirm the performance improvement brought by the selected highly-reliable negative samples. The proposed method is based on the assumption that drugs dissimilar to any drugs known to cause a given side-effect are less likely to cause the side-

Table 3.2: **Macro-averaging *F*1-score, precision, and recall of four typical classifiers based on negative samples selected by the proposed negative sample selection method and randomly selected negative samples.**

Measure	Macro_$F1$	Macro_P	Macro_R
KNN & ComNegative	**0.800**	**0.728**	**0.772**
KNN & ComRandom	0.619	0.598	0.651
KNN & ChemTarRandom	0.581	0.553	0.626
ELM & ComNegative	**0.774**	**0.761**	**0.604**
ELM & ComRandom	0.586	0.572	0.601
ELM & ChemTarRandom	0.572	0.561	0.585
SVM & ComNegative	**0.898**	**0.938**	**0.861**
SVM & ComRandom	0.655	0.670	0.642
SVM & ChemTarRandom	0.606	0.622	0.598
RBF & ComNegative	**0.822**	**0.828**	**0.818**
RBF & ComRandom	0.606	0.592	0.622
RBF & ChemTarRandom	0.571	0.561	0.583

effect. Drugs that locate far from all positive samples in the chemobiological space are used as negative samples, which really contributed to improving the prediction performance for different classifiers. The selected high-reliable samples help to learn an optimal classification decision boundary, better differentiating positive samples and negative samples.

3.4.4 Comparison with other methods

To demonstrate the superior performance of the proposed method, we compared it with several state-of-the-art methods on a widely-used benchmarking dataset (i.e., Liu's dataset) (Liu & Wu 2012, Zhang, Liu, Luo

& Zhang 2015, Zhang & Chen 2016, Muñoz & Nováček 2017). There are 832 drugs and 1385 side-effects in this dataset. Since the ATC codes and substituents of some drugs are not available, we used drug chemical substructures and drug targets as features, and "Max" as the similarity integration method to measure drug similarities. We compared the performance of our method with the state-of-the-art methods reported in (Muñoz & Nováček 2017). As our method is binary-classification based, we just report metrics which are designed for binary classification. The results are listed in Table 3.3 and the best performance of a given metric is highlighted in bold values. Table 3.3 shows that our method outperformed all other methods in terms of AUC-PR (area under the precision-recall curve). In addition, our method achieved comparable average precision and AUC-ROC (area under the receiver operating characteristic curve) with the best performance (0.5439 vs 0.5476 in average precision, 0.9086 vs 0.9091 in AUC-ROC). The results further confirm the predictive power of the proposed method.

3.4.5 Drugs that have the predicted side-effect "drug eruption": a case study

This section presents a list of drugs which are predicted to have the side-effect "drug eruption" by the proposed method. Drug eruption is a side-effect on skin (Wikipedia 2018, April 1*b*). Most drug-induced eruptions are mild and they can disappear after stop taking the drugs. However, serious drug eruptions sometimes are associated with organ injuries like kidney and liver damage. It has been estimated that every 2-3 in 100 hospitalized patients have been suffering a drug eruption, and serious drug eruptions occur in around 1 in 1000 patients (Roujeau & Stern 1994). Consequently, to predict potential drugs which could possibly lead to drug eruptions is of great interests.

We performed the prediction using KNN ($k = 50$) with validated positive samples and reliable negative samples selected by the proposed negative

Table 3.3: **Performance of the proposed method and state-of-the-art-methods using 5-fold cross-validation on Liu's data set.**

Measure	Average Precision	AUC-ROC	AUC-PR
Our method	0.5439	0.9086	**0.5424**
Liu's method (Liu & Wu 2012)	0.2610	0.8850	0.2514
FS-MLKNN (Zhang, Liu, Luo & Zhang 2015)	0.5134	0.9034	0.4802
LNSM-SMI (Zhang, Liu, Luo & Zhang 2015)	**0.5476**	0.8986	0.5053
LNSM-CMI (Zhang, Liu, Luo & Zhang 2015)	0.5329	**0.9091**	0.4909
KG-SIM-PROP (Muñoz, Nováček & Vandenbussche 2017)	0.4895	0.8860	0.4295

sample selection method. Like other data mining results, it is unrealistic to expect every predicted drug is of value to domain experts (Jin & Chen 2008). Therefore, we shortlist the 50 top-ranked drugs in terms of their prediction scores in Figure 3.7. The circle in the center of the figure is the side-effect "drug eruption" and the other circles are the top 50 drugs. Among the top 50 drugs, the larger the prediction score is, the larger its circle is. Besides, the labels on the edges show the ranking positions and evidence types of the predicted associations. The symbols "#" and "$" denote that the corresponding associations can be validated by records from the side-effect database SIDER (colored green) and related literature (colored red) respectively. The symbol "?" means the predicted associations cannot be validated to the best of our knowledge (colored orange). Overall, 49 of the top 50 predicted associations can be verified by the SIDER and other literature

(SIDER: 45, other literature: 4).

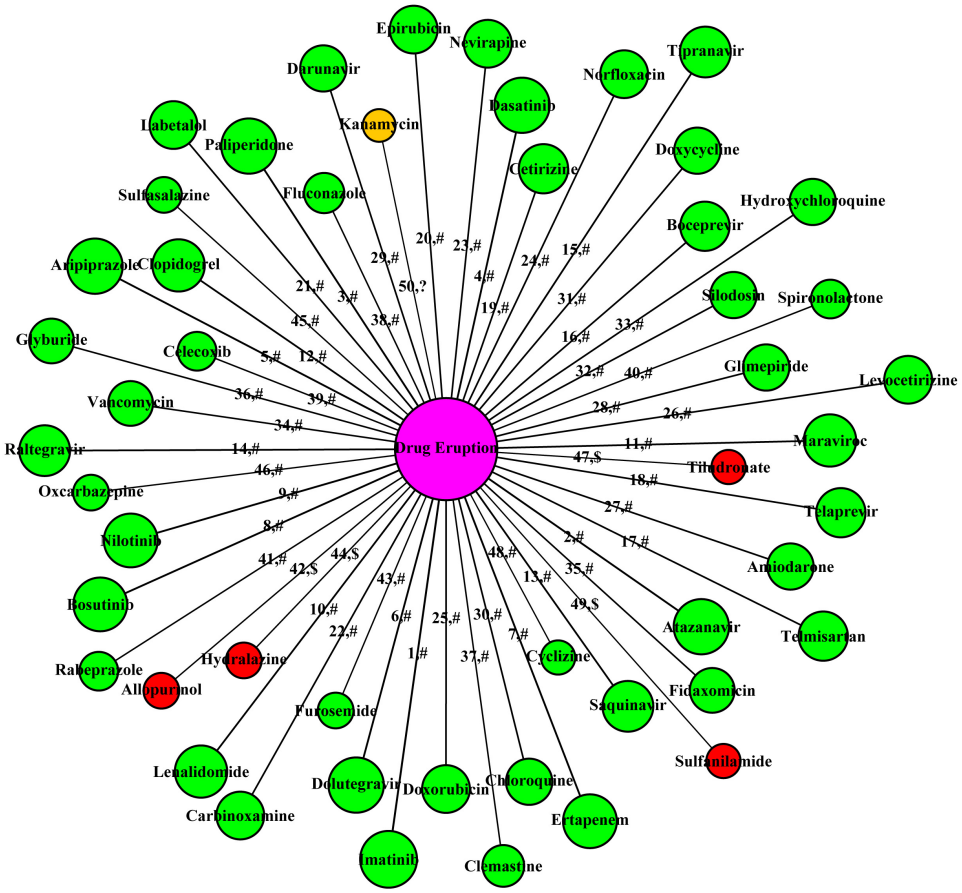


Figure 3.7: **The top 50 drugs which are predicted to have the side-effect "drug eruption".** Labels on the edges illustrate the ranks of predicted associations and the confirmation types. The symbols "#" and "$" denote that the corresponding associations can be validated by records from the side-effect database SIDER (colored green) and related literature (colored red) respectively. The symbol "?" means the predicted associations cannot be validated to the best of our knowledge (colored orange).

Drugs including allopurinol, hydralazine, sulfanilamide, tiludronate and kanamycin are newly predicted (i.e., not stored in SIDER) to be associated with "drug eruption". 4 of the above 5 novel associations can be confirmed by the literature work. Allopurinol (Wikipedia 2018, April 1*a*) is a medication

used to decrease high blood uric acid levels. In 2012, Kim et al. (Kim & Shim 2012) studied its relationship with fixed drug eruptions using a lymphocyte transformation test. Fixed drug eruption is a distinct type of drug eruption which occurs in the same skin area each time when the patients take the drug (Wikipedia 2018, April 1*b*, Shiohara 2009). They finally confirmed that allopurinol is one of the causative drugs that induced fixed drug eruption (Kim & Shim 2012). Hydralazine, a well-known antihypertensive drug, has been in vogue for the last three decades. It was reported to induce fixed drug eruption in literature (Sehgal & Gangwaani 1986). Sulfanilamide, a sulfonamide antibacterial, was widely used by the Allies in World War II to reduce infection rates. It contributed a lot to reducing the mortality rates. Early in 1939, Loveman et al. (Loveman & Simon 1939) reported fixed drug eruptions and stomatitis due to sulfanilamide. Tiludronate is a bisphosphonate used for treatment of Paget's disease of bone. Its association with drug eruption can be found in the 22nd edition of Litt's Drug Eruption and Reaction Manual (Wiffen 2016).

To further demonstrate the capacity of the proposed method in predicting new DSPs, we also investigated those drugs which are ranked from top-51 to top-60. Among them, triclosan, nitric oxide, carbimazole, propylthiouracil, neomycin and dapsone are newly predicted drugs to cause "drug eruptions" and require further validation. Such newly predicted associations may provide interesting information for domain experts. In summary, the above successful prediction instances further demonstrate that our method has the capacity to predict both existing and novel drug-side-effect associations.

Chapter 4

Predicting side-effects of combined medication from heterogeneous databases

4.1 Introduction

As mentioned in Section 2.1.2, existing methods work on three types of data sources namely spontaneous reporting systems (SRSs), electronic health records (EHRs) and social media. They have different limitations and shortcomings. To overcome these limitations, we explore the use of multiple pharmacologic databases as a new data source to identify side-effects of combined medication. However, the application of machine learning methods on pharmacologic databases is severely impeded by the lack of proper drug representation and credible negative samples.

To solve the above problems, we propose a machine learning method to predict side-effects of combined medication from pharmacologic databases by building up highly-credible negative samples (HCNS). We fuse heterogeneous information from different databases and represent each drug as a multi-dimensional vector according to its chemical substructures, target proteins, substituents, and related pathways first. Then, a drug-pair vector is obtained

by appending the vector of one drug to the other. Next, we construct a drug-disease-gene network and devise a scoring method to measure the interaction probability of every drug pair via network analysis. Drug pairs with lower interaction probability are preferentially selected as negative samples (see sorted drug pairs in **Additional file 4**). Following that, the validated positive samples and the selected credible negative samples are projected into a lower-dimensional space using the principal component analysis. Finally, a classifier is built for each side-effect using its positive and negative samples with reduced dimensions. The performance of the proposed method is evaluated on simulative prediction for 1,276 side-effects and 1,048 drugs, comparing using four machine learning algorithms and with two baseline approaches. Extensive experiments show that the proposed way to represent drugs characterizes drugs accurately. With highly-credible negative samples selected by HCNS, the four machine learning algorithms achieve significant performance improvements. HCNS is also shown to be able to predict both known and novel drug-drug-side-effect associations, outperforming two other baseline approaches significantly.

4.2 Methods

4.2.1 Data resources

We focus on drugs from DrugBank (Wishart & Feunang 2017), a comprehensive drug database. The drug chemical structures, drug target proteins, and drug substituents are extracted from DrugBank. Since the validated drug target proteins in DrugBank are quite sparse, we also integrate the target information from DrugCentral (Ursu & Holmes 2017), an open-access online drug compendium. The association data, including drug-gene (Homo sapiens), drug-disease, drug-pathway, and disease-gene associations are retrieved from the CTD (comparative toxicogenomics database) (Davis & Grondin 2016). The drug-drug-side-effect associations are downloaded from the Tatonetti Lab (Tatonetti & Patrick 2012) (Twosides databases). The

integration of the above data finally produces 1,048 drugs, 1,276 side-effects, and 1,155,754 drug-drug side-effect associations (DDSAs). Information of all researched drugs and side-effects is available in **Additional file 5**. The disease-gene associations and drug-drug-side-effect associations are included in **Additional file 6** and **Additional file 7** respectively. All the above data and the source codes are included in **Additional file 8**.

4.2.2 Proposed method

We cast the prediction of drug-drug-side-effect associations into a binary classification task, where inputs are vectors of drug pairs and labels are side-effects. We represent each drug as a multi-dimensional vector using their heterogeneous features. To address the lack of negative samples in the binary classification task, we design a method to select credible negative samples based on a drug-disease-gene tripartite network constructed in this chapter. The specific process is detailed as pseudo codes in Algorithm 4.1 and the framework is outlined in Figure 4.1. The framework consists of three components, namely drug representation, credible negative sample generation, and drug-drug-side-effect association prediction.

Algorithm 4.1 Predicting adverse drug reactions of combined medication by building up highly-credible negative samples.

Input drug set V_r, disease set V_d, target gene set V_g, side-effect set V_s, drug chemical structure vectors f_{cs}, drug target protein vectors f_{tp}, drug substituent vectors f_{sub}, drug pathway vectors f_{pat}, and negative sample ratio nsr.

for $R \in V_r$ **do**

$R_{cs} = f_{cs}[R]$;

$R_{tp} = f_{tp}[R]$;

$R_{sub} = f_{sub}[R]$;

$R_{pat} = f_{pat}[R]$;

$R_v =< R_{cs}, R_{tp}, R_{sub}, R_{pat} >$;

end for

for $i = 1$ **to** $(|V_r| - 1)$ **do**
  for $j = (i + 1)$ **to** $|V_r|$ **do**
    $R_1 = V_r[i]$;
    $R_2 = V_r[j]$;
    $D_1 = V_d[R_1]$;
    $D_2 = V_d[R_2]$;
    $G_1 = V_g[R_1]$;
    $G_2 = V_g[R_2]$;
    $inter_g(R_1, R_2) = (G_1 \cap G_2)/(G_1 \cup G_2)$;
    $inter_d(R_1, R_2) = ((D_1 \cap D_2)/(D_1 \cup D_2)) * ((D_1 \cap D_2)/|V_d|)$;
    Obtain the shared disease set V_{sd} between R_1 and R_2, let G_k^d denote the gene set of the k^{th} shared disease d_k;
    $inter_{gd}(R_1, R_2) = (\cup_{k=1}^{|V_{sd}|}((G_1 \cap G_2) \cap G_k^d))/(G_1 \cup G_2)$;
    $inter_{score}(R_1, R_2) = inter_g(R_1, R_2) + inter_d(R_1, R_2) + inter_{gd}(R_1, R_2)$
  end for
end for
Rank all drug pairs by their interaction scores (i.e., $inter_{score}$) ascendingly and obtain the candidate negative drug-pair list *candidate_list*;
for $S_i \in V_s$ **do**
  positive_set $\leftarrow$ drug pairs known to cause S_i;
  negative_set $\leftarrow$ first $(nsr * |positive\_set|)$ drug pairs from the *candidate_list* (exclude drug pairs in *positive_set*);
  Separate *positive_set* and *negative_set* into *training_set* and *test_set*;
  Obtain vectors of drug pairs from the *training_set* and *test_set* by appending vector of one drug to the other;
  Fit PCA using the *training_vectors*;
  Transform the *training_vectors* and *test_vectors* into lower dimensional vectors *comp_training_vectors* and *comp_test_vectors* using PCA;
  Train a binary classifier using *comp_training_vectors*;
  Predict labels for the *comp_test_vectors* using the trained classifier;
end for

Evaluate the performance of all predictions;

Output Predicted drug-drug-side-effect associations, prediction performances, and the candidate negative drug-pair list.

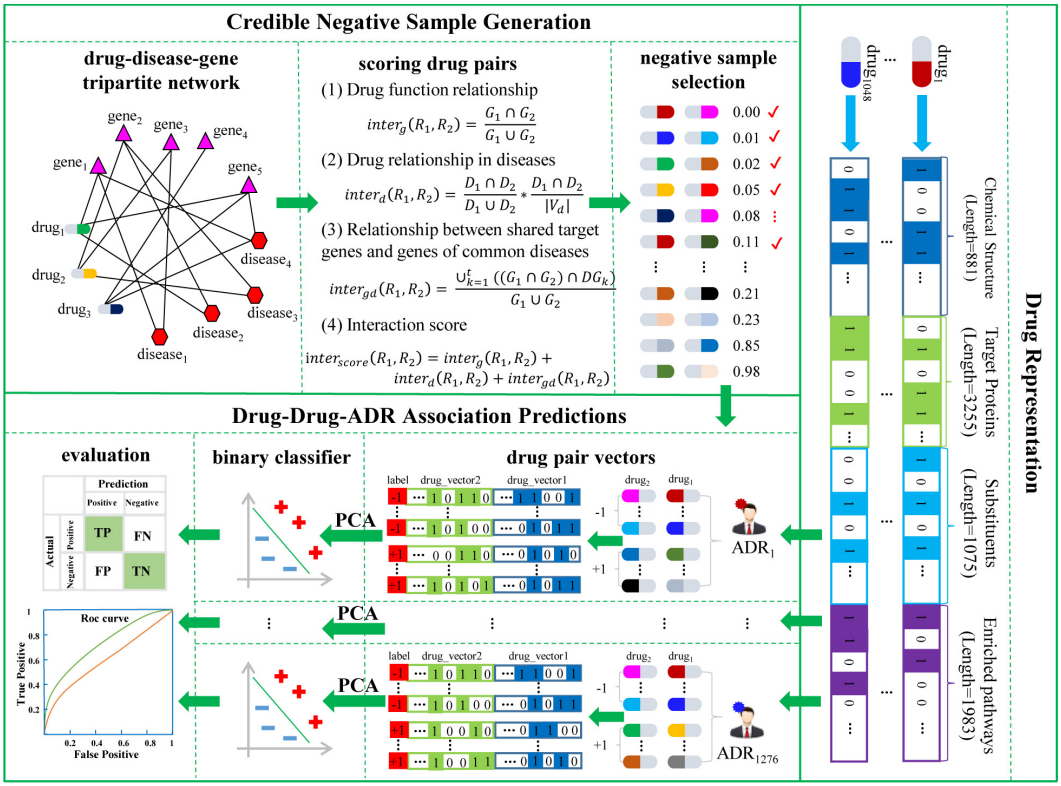


Figure 4.1: **The framework of HCNS.** It consists of three components: drug representation, credible negative sample generation, and drug-drug-side-effect association prediction.

Drug Representation

We represent each drug as a feature vector using its chemical structures, target proteins, substituents, and enriched pathways. Specifically, we use the PubChem fingerprint, which corresponds to 881 substructures defined in the PubChem database (Chen & Wild 2009), to encode the drug chemical structure. Each drug is represented as an 881-bit binary profile,

whose elements denote the absence/presence of the corresponding PubChem substructure by 0/1. We obtain 3,255 unique drug target proteins by merging target proteins of all drugs. Then each drug can be represented as a 3,255 dimensional vector whose elements encode for the absence/presence of the corresponding target proteins by 0/1. Analogously, each drug is represented as a 1,075 and 1,983 dimensional vector according to its substituents and enriched pathways. Finally, we represent each drug as a single vector by concatenating its four separate vectors. This produces a 7,194 (881+3,255+1,075+1,983) dimensional vector for each drug.

Credible Negative Sample Generation

In this chapter, we are interested in predicting adverse drug reactions of combined medication (pair-wise drug medication in particular). These adverse drug reactions can not be attributed to either drug alone, occurring only when two drugs are taken together or concomitantly. Known drug-drug-side-effect associations can serve as positive samples for prediction. However, there is a lack of explicit negative samples, namely drug-drug-side-effects that are confirmed to have no associations, making it difficult to directly apply machine learning methods to solve the prediction problem.

Since side-effects of combined medication are caused by pharmacokinetic or pharmacodynamic interactions between drugs (Banda & Callahan 2016), it is reasonable to hypothesize a drug pair with lower interaction probabilities are less likely to cause any side-effects. Base on this hypothesis, we explore to use drug pairs with lower interaction probabilities as negative samples for building a classifier. We propose a method to measure the interaction probability of a drug pair according to the following observations: (i) interacting drugs prefer to share the same target genes; (ii) interacting drugs are more likely to associate with a same set of diseases; (iii) the common target genes of two drugs tend to be the common disease genes of their associated diseases.

To explicitly quantify the interaction probability of a drug pair, we first

construct a drug-disease-gene (RDG) tripartite network $rdg = (V_r \cup V_d \cup V_g, E)$, where V_r, V_d, V_g is a set of drugs, diseases, and genes, respectively; E is the set of associations among them. Given a drug pair R_1 and R_2 $(R_1, R_2 \in V_r)$, we denote their gene sets as $G_1 = \{g_{11}, g_{12}, ..., g_{1i}, ..., g_{1x}\}$ and $G_2 = \{g_{21}, g_{22}, ..., g_{2j}, ..., g_{2y}\}$, respectively $(g_{1i}, g_{2j} \in V_g; (R1, g_{1i}), (R2, g_{2j}) \in E))$. Analogously, we represent the disease sets of R_1 and R_2 as $D_1 = \{d_{11}, d_{12}, ..., d_{1p}, ..., d_{1m}\}$ and $D_2 = \{d_{21}, d_{22}, ..., d_{2q}, ..., d_{2n}\}$ respectively $(d_{1p}, d_{2q} \in V_g; (D1, d_{1p}), (D2, d_{2q}) \in E))$. Then the associated genes of a disease d_s can be denoted as $DG_s = \{g_1, g_2, ..., g_k, ..., g_l\}$.

We quantify (i) the function relationship between a drug pair; (ii) the drug therapeutic relationship in different diseases; (iii) the relationship between the shared genes of two drugs and the common disease genes of their associated diseases:

- Drug function relationship. The function relationship between drug R_1 and drug R_2 is measured as the proportion of their shared target genes.

$$inter_g(R_1, R_2) = \frac{G_1 \cap G_2}{G_1 \cup G_2}. \tag{4.1}$$

- Drug therapeutic relationship in diseases. The idea is that drugs have higher overlapping ratio in associated diseases are more likely to interact. Besides, the more diseases two drugs share, the more likely they will interact. The overlapping ratio of associated diseases is measured by dividing the shared disease number by the number of their union diseases. The shared diseases are measured as the percentage of the shared diseases comparing to all diseases in RDG. The two factors are combined to measure drug therapeutic relationship as follows:

$$inter_d(R_1, R_2) = \frac{D_1 \cap D_2}{D_1 \cup D_2} * \frac{D_1 \cap D_2}{|V_d|}. \tag{4.2}$$

- Relationship between shared target genes of drug R_1 and drug R_2 and genes of their common diseases. The idea is that interacted drugs

always associate with common disease genes which contribute to the disease development.

$$inter_{gd}(R_1, R_2) = \frac{\cup_{k=1}^{t}((G_1 \cap G_2) \cap DG_k)}{G_1 \cap G_2}, \tag{4.3}$$

where DG_k is the gene set of the k^{th} shared disease d_k, and t is the total number of common diseases between R_1 and R_2.

By integrating the three utilities above, we define the score for calculating the interaction probability of the drug pair R_1 and R_2 as follows:

$$\begin{aligned} inter_{score}(R_1, R_2) = inter_g(R_1, R_2) + \\ inter_d(R_1, R_2) + inter_{gd}(R_1, R_2). \end{aligned} \tag{4.4}$$

We rank all candidate drug pairs (548,628) in an ascending order of their interaction scores (i.e., $inter_{score}$). Drugs from a drug pair with a lower position in the rank are less likely to interact with each other. Therefore, we selectively choose drug pairs with lower positions as negative samples for DDSA prediction.

Drug-drug-side-effect association prediction

We build a binary classifier for each side-effect. Drug pairs known to cause the side-effect are used as positive samples directly. A corresponding number of negative samples is selected from the above ranked drug-pair list. The specific number is determined by the negative sample ratio, which will be discussed in the result section. Each drug pair is represented as a vector by concatenating two individual drug vectors. A drug-pair vector has a high dimensionality of 14,388 (7,194+7,194), which incurs a high computational cost to train a classifier. To speed up training, we employ PCA to perform dimension reduction on the drug-pair vectors. Specifically, all training drug-pair vectors are used to fit the PCA first. Then the fitted PCA is used to transform the training and test drug-pair vectors into lower-dimensional vectors. Finally, the resulted vectors of drug pairs are used as inputs to train and validate the binary classifier.

4.3 Results and discussions

5-fold cross-validation is performed to evaluate the prediction performance. The macro-averaging value of precision, recall, accuracy, F1-score, and AUC (area under the receiver operating characteristic curve) are used as evaluation metrics.

4.3.1 Parameter optimization

The key factors of HCNS are the negative sample ratio (NSR) and the PCA component number (PCN). NSR refers to the ratio of negative samples relative to positive samples. All the 1,155,754 validated DDSAs are served as positive samples. We first perform experiments to obtain the best settings for NSR and PCN. We employed SVM (support vector machine) in the Python sklearn package with default settings as the classifier. We experiment with the following settings: $NSR \in \{0.5, 1, 2, 3\}$ and $PCN \in \{10, 20, 30, 40, 50, 80, 100, 150, 200, 300, 400, 500, 600, 700, 800,$
$900, 1000\}$.

Figure 4.2 illustrates the macro-averaging F1-scores for different combinations of NSR and PCN. It can be observed that, when $PCN <=$ 300, the macro-F1 increases dramatically with PCN for $NSR = 2$ (green) and $NSR = 3$ (yellow). By contrast, the macro-F1 of $NSR = 0.5$ (blue) and $NSR = 1$ (red) increases very slightly with PCN. And all macro-F1 values plateau when the PCA component number is larger than 300. Similar conclusions can be drawn from the macro-AUC results, as shown in Figure S3 in **Additional file 9**. Besides, $NSR = 1$ slightly outperforms $NSR = 0.5$, and both of them significantly outperform $NSR = 2$ and $NSR = 3$. Based on the above observation and considering the time-cost (computational time increases with PCN), we set $NSR = 1$ and $PCN = 300$ for HCNS in the following experiments.

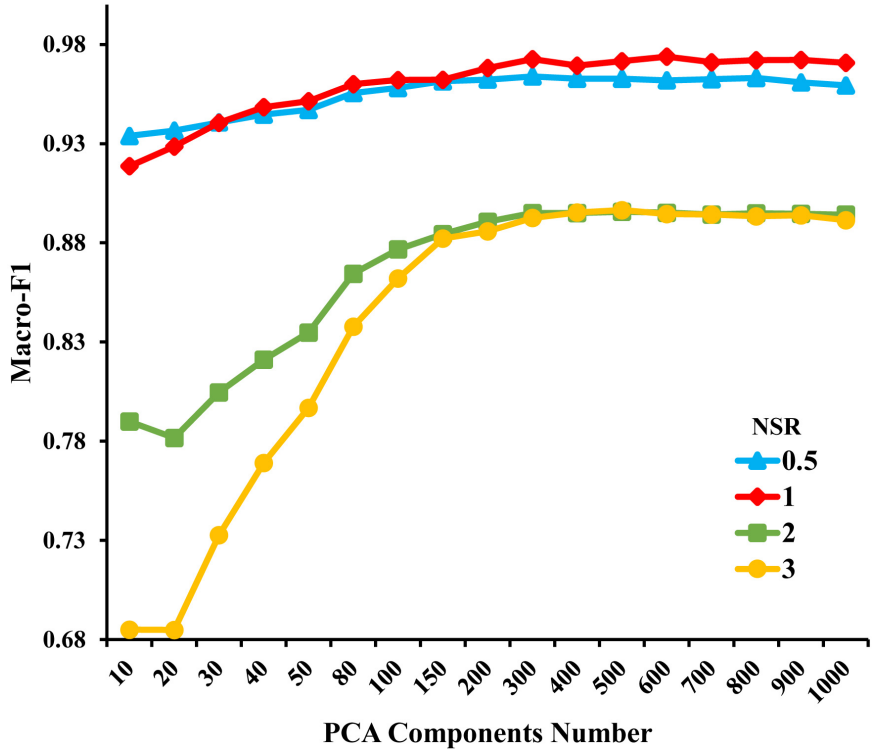


Figure 4.2: **The macro-averaging *F*1-scores with different PCA component numbers and different negative sample ratios.**

4.3.2 Evaluation on classic classifiers

To demonstrate the superior performance of HCNS and its efficacy of credible negative sample generation, we first conduct experiments to compare four classic classifiers learned based on randomly generated negative samples (RGNS). The randomly generated samples are produced by randomly sampling drug pairs which are not in the positive samples. Except for the way to generate negative samples, other details of RGNS are the same as HCNS. To avoid bias, RGNS is repeated 5 times and the average results are used for the final evaluation. The four classic classifiers are KNN (k-nearest neighbors), SVM, RF (Random Forest) and LR (Logistic Regression). All classifiers are run using Python 2.7.13 (sklearn) with the default settings.

Table 4.1 shows the evaluation indices of the above four classifiers. We can see that all classifiers achieve significant performance gains using negative

Table 4.1: **Macro-averaging AUC, *F*1-score, precision, recall and accuracy of four typical classifiers based on negative samples selected by HCNS and RGNS.**

Classifier	Negative Samples	Macro AUC	Macro F1	Macro Precision	Macro Recall	Macro Accuracy
SVM	HCNS	0.994	0.973	0.985	0.963	0.975
SVM	RGNS	0.946	0.887	0.896	0.888	0.893
LR	HCNS	0.998	0.980	0.991	0.971	0.981
LR	RGNS	0.963	0.903	0.898	0.913	0.905
KNN	HCNS	0.983	0.920	0.972	0.883	0.936
KNN	RGNS	0.923	0.859	0.850	0.877	0.862
RF	HCNS	0.943	0.840	0.928	0.781	0.861
RF	RGNS	0.787	0.713	0.753	0.700	0.717

samples generated by HCNS in comparison to their RGNS counterparts using randomly generated samples. For example, as compared to RNGS, HCNS achieves an improvement of 5.07%, 3.70%, 6.49% and 19.75% on the macro-averaging AUC, and 9.65%, 8.48%, 7.09% and 17.79% on the macro-averaging F1, for SVM, Logistic Regression, KNN, and Random Forest, respectively.

4.3.3 Comparison with baseline methods

Lots of methods have been developed to predict side-effects, however, they are for single drug medication only. To our knowledge, HCNS is the first method proposed to predict side-effects for combined medication. To further confirm the superiority of HCNS, we compare it with two baseline methods, namely random assignment and one-class SVM.

Random assignment (Random). To show how difficult the prediction problem is, we use a random assignment method as one baseline, where the label is randomly assigned to a drug pair according to the probability of their

ratios. For instance, if the ratio of label "+1" in the training set is 50%, we randomly assign "+1" to 50% samples in the test set. The other 50% samples in the test set are then assigned with "-1". For this method, the prediction is made at random.

One-class SVM (OCSVM). One-class SVM is a one-class learning technique which is widely used in the scenario where there are only positive samples, while negative samples are hard to acquire (Erfani, Rajasegarar, Karunasekera & Leckie 2016). It has achieved remarkable performance in a few fields, such as image retrieval, and anomaly detection (Chen, Zhou & Huang 2001, Li, Huang, Tian & Xu 2003). One-class SVM is trained with validated positive samples only. In this chapter, we report evaluation results of OCSVM on two sets of test data: one with negative samples selected randomly (OCSVM_Random), and the other with negative samples generated according to our screen list (OCSVM_Screen). The negative sample ratios of both methods are set as 1.

To avoid bias, the method OCSVM_Random and Random are repeated 5 times and the average results are used for the final evaluation. The performances achieved by these predictive methods are shown in Figure 4.3. Clearly, HCNS remarkably outperform the other three methods on all evaluation indices. While OCSVM_Screen and OCSVM_Random perform slightly better than the random assignment method, they have comparable prediction results on two test data. It suggests that, with positive samples only, OCSVM is unable to learn an accurate decision boundary for the drug-drug-side-effects association prediction problem. This necessitates the generation of credible negative samples by the proposed method to achieve satisfactory prediction results.

4.3.4 Predicted adverse drug reactions for the drug pair "Albuterol-Zolpidem": a case study

After confirming the superior performance of the proposed method, we build 1,276 SVM classifiers using the corresponding validated positive samples and

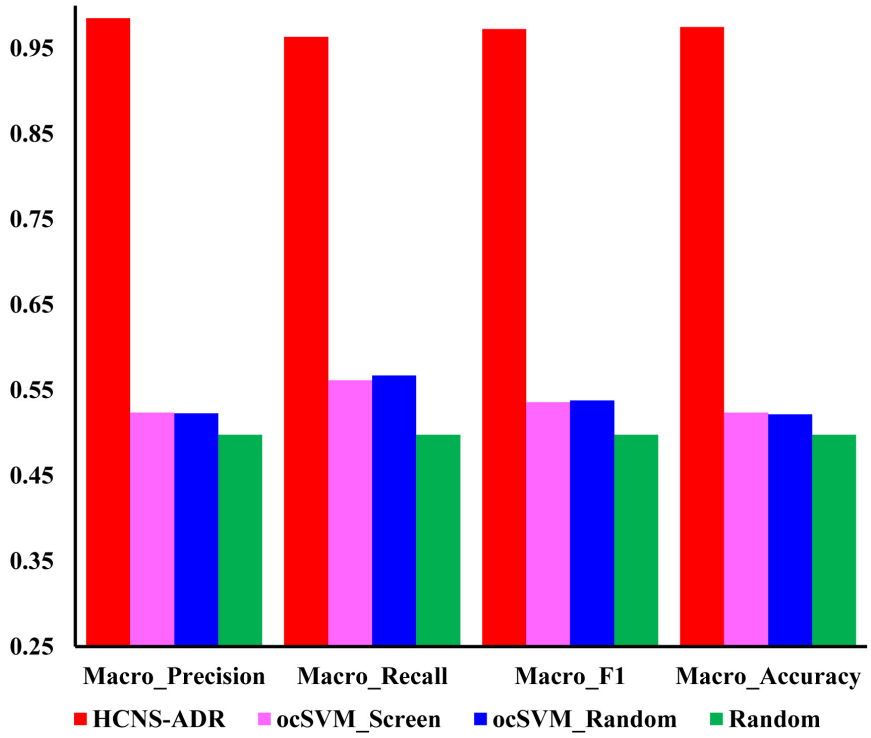


Figure 4.3: **The macro-averaging precision, recall, *F*1-score and accuracy of HCNS and other three comparison methods.**

selected negative samples to predict potential side-effects for any drug-pairs. The negative sample ratio is set as 1. Here, we report the prediction results for the drug pair "Albuterol-Zolpidem" as a case study.

Like other data mining results, it is unrealistic to expect every highly ranked side-effects valuable to domain experts (Zheng, Lan, Peng & Li 2016). Therefore, we shortlist the top 40 side-effects according to their prediction scores, as shown in Figure 4.4 . The UMLS names of side-effects are labeled on the circle, and their ranks and confirmation types are labeled on the edges. "#" denotes the relation is known in the Tatonetti Lab dataset, "$" means the relation is the common side-effects of the drug pair, and "?" indicates there are no evidences for the relation.

It can be observed that 36 out of the top 40 side-effects are known in the Tatonetti Lab dataset. The other 4 side-effects are newly predicted side-effects for the combined medication "Albuterol-Zolpidem". We further

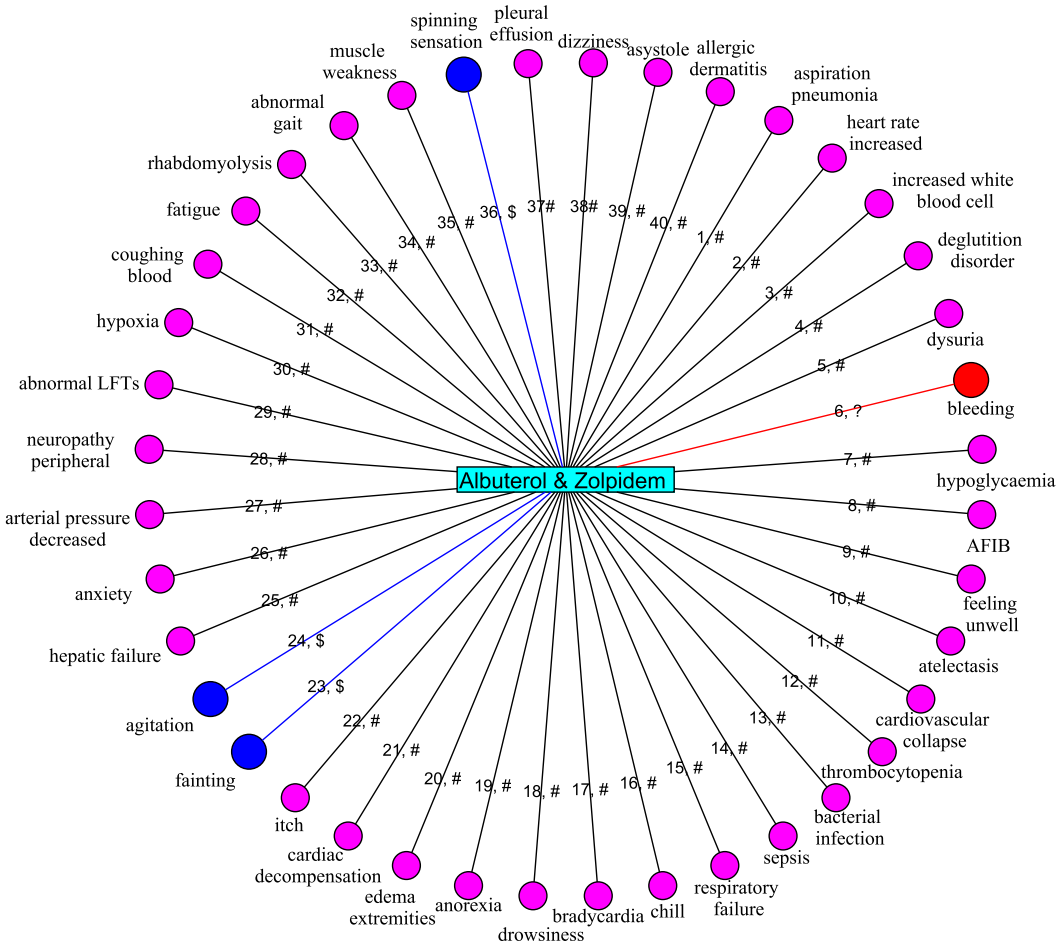


Figure 4.4: **The top 40 side-effects which are predicted to be associated with the drug pair "Albuterol-Zolpidem".** Labels on the edges illustrate the ranks of predicted associations and the confirmation types. "#" denotes the relation is known in the Tatonetti Lab dataset, "$" means the relation is the common side-effects of the drug pair, "?" indicates there are no evidence for the relation.

confirm the newly predicted associations from other data sources; 3 out of them are indeed confirmed as common side-effects of Albuterol and Zolpidem. More specifically, fainting and spinning sensation are found to be two common side-effects of Albuterol and Zolpidem in Drugs.com (Wikipedia

contributors 2018), and agitation is found in SIDER (Kuhn & Letunic 2015), a comprehensive drug side-effect database.

To further demonstrate HCNS's ability to predict new DDSAs, we also investigate the side-effect ranked list from top-41 to top-50. Among others, acute respiratory distress syndrome, hive, arterial pressure NOS increased, loss of consciousness, ascites and pulmonary arrest are newley predicted side-effects of Albuterol and Zolpidem, requiring further validation. These newly predicted associations provide valuable information to domain experts. In summary, this case study confirms that HCNS is able to predict both known and novel drug-drug-side-effect associations.

Chapter 5

Constrained information entropy for detecting adverse drug reactions from medical forums

5.1 Introduction

As we figured out in Section 2.2, existing methods for detecting adverse drug reactions from medical forums are mostly co-occurrence based methods and face several issues, in particularly, leaving out the rare ADRs and unable to distinguish irrelevant ADRs. We introduce a constrained information entropy (CIE) method to solve these problems. CIE first recognizes the drug-related adverse reactions using a predefined keyword dictionary and then captures high- and low-frequency (rare) ADRs by information entropy. Extensive experiments on medical forums dataset demonstrate that CIE outperforms the state-of-the-art co-occurrence based methods, especially in rare ADRs detection.

5.2 Methods

5.2.1 Framework

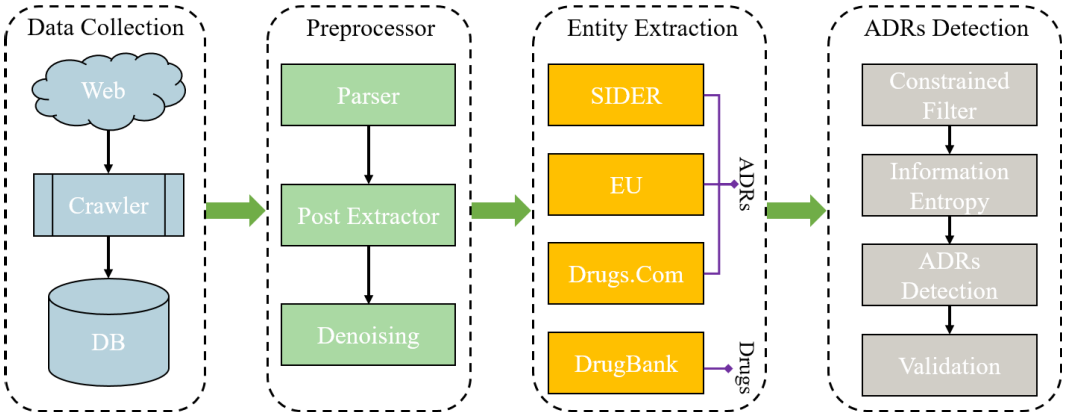


Figure 5.1: **Framework of CIE.** It consists of four components namely data collection, preprocessing, entity extraction and ADRs detection.

The overall framework of the proposed method is illustrated in Figure 5.1. First, the specialized crawler crawls medical web pages from two specified medical forums and store them into the database. Then, the crawled semi-structured HTML (hypertext markup language) web pages are parsed to extract post contents. After denoising the post contents, we extract all entities (i.e., ADRs and drugs) using two self-built dictionaries. Finally, drug-ADRs associations are filtered, detected, and validated by the ADRs detection module. Details of the proposed method are provided in the following subsections.

5.2.2 Data collection

We develop a specialized crawler to collect data from two medical forums including SteadyHealth.Com (Steady Health Comunity 2019) and MedHelp.Org (MedHelp.org 2019). The crawler crawls the whole site of the

two forums and obtains all HTML pages. Then, the crawled pages are stored into the database for further processing.

5.2.3 Preprocessing

We need to preprocess the crawled pages before further analysis. Since all crawled pages are semi-structured HTML files, we use HtmlParser a high-performance Java-based tool as our parser to parse all the pages. Following that, post contents are extracted from these parsed pages. Due to inevitable noise data among posts, such as substandard HTML tags and duplicated punctuations, we formulate rules using regular expressions to remove such noise data from the extracted post contents.

5.2.4 Entity extraction

Previous study (Liu & Chen 2013) suggests that the dictionary-based method is the most appropriate approach to extract medical entities from patient-generated contents. In this study, we explore to extract entities from the post contents using the dictionary built from multi-data sources. Specifically, we build two dictionaries: the ADRs dictionary consists of 10,605 ADRs merged from SIDER (Kuhn & Letunic 2015), PROTECT (European Medicines Agency 2019) and Drugs.Com (Wikipedia contributors 2018); the drug dictionary includes 23,300 drug names (i.e., generic names and brand names) obtained from DrugBank (Wishart & Feunang 2017), a comprehensive drug database.

5.2.5 ADRs detection

To distinguish drug-related adverse drug reactions from unrelated reactions, we construct a keyword-based dictionary which contains keywords of the causality as follows: (1) We conduct an analysis of n-grams frequency on the whole text corpus first. High-frequency n-grams ($0 < n < 4$) which appear more than 20 times are identified. (2) We get the basic forms of

the obtained n-grams using Stanford CoreNLP (Manning, Surdeanu, Bauer, Finkel, Bethard & McClosky 2014). (3) We review and deduplicate the n-grams of basic forms. n-grams which don't denote the cause-and-effect relationship are removed and duplicate n-grams are removed. Performing the above steps, we finally obtain 47 n-grams (i.e., phrases), as listed in Table 5.1.

Table 5.1: **A set of keywords or phrases denote the causality relationship between drugs and ADRs.**

attack	because	because of	cause
cause of	develop	disorder	drive
due to	effect	effect of	effect side
experience	feel	find out	find that
for	gain	get	have
have a problem	horrible	hurt	increase
induce	infect	injury	lead
lose	notice	prevent	problem with
react to	result	result of	risk
serious	severe	side effect	since
start get	stop	stop take	suffer
terrible	with	with problem	

The keyword-based dictionary is utilized as a filter to filter out drug-ADRs pairs which don't have any keywords between the drug and ADRs within 10-word space. Being filtered, those drug-ADRs pairs having no cause-and-effect associations are removed and the remained drug-ADRs pairs are preserved for further analysis.

As we pointed out in the beginning that the co-occurrence based methods focus on the high co-occurrence frequency pairs. They are incapable to identify those relatively low-frequency but important drug-ADRs pairs. To detect such ignored rare drug-ADRs pairs, we propose to use information entropy instead of co-occurrence frequency to detect the ADRs. According

to information entropy theory (Harte & Newman 2014), the more even the distribution of ADRs caused by a drug is, the larger information entropy the drug has and harder to identify the ADRs. Adversely, the less even the distribution of ADRs is, the smaller information entropy the drug has and easier to identify the ADRs. In addition, the most frequent ADRs of a drug carry the most information of the drug. Thus, it's reasonable to preferentially select the most frequent ADRs of a certain drug with smaller information entropy as detected ADRs. The definition of information entropy is defined as follows:

$$H(d) = -\sum_{i=1}^{N_d} p(a_i) log(p(a_i)), \tag{5.1}$$

$$p(a_i) = \frac{f(a_i)}{\sum_{j=1}^{N_d} f(a_j)} \tag{5.2}$$

Where $H(d)$ is the information entropy of drug d, N_d is the number of ADRs caused by drug d, $p(a_i)$ and $f(a_i)$ are the probability and frequency of the i^{th} ADRs respectively. ADR which occurs the most frequently (frequency>2) in each drug is identified as the adverse reaction of the drug. The identified drug-ADRs pairs are sorted in ascending order of their information entropy.

5.3 Results and discussions

5.3.1 Dataset and evaluation metrics

We crawled the posts from SteadyHealth.Com (Steady Health Comunity 2019) and MedHelp.Org (MedHelp.org 2019), two well-known health-related forums, as the experiment data for evaluation. The statistical information of the dataset is shown in Table 5.2 and Table 5.3.

The proportion of known drug-ADRs pairs to all detected drug-ADRs pairs is used as the evaluation metric.

$$precision = \frac{\text{known pair number}}{\text{total detected pair number}} \tag{5.3}$$

Table 5.2: **Dataset summary of SteadyHealth.Com**

Field Name	**Value**
Forum Name	www.steadyhealth.com
Post Number	56,866
Topic Number	14,939
Total Number of Sentences	438,826
Average Number of Sentences Per Post	7.72

Table 5.3: **Dataset summary of MedHelp.Org**

Field Name	**Value**
Forum Name	www.medhelp.org
Post Number	15,150
Topic Number	773
Total Number of Sentences	142,083
Average Number of Sentences Per Post	9.38

The known drug-ADRs pairs are those hit the associations in the SIDER, PROTECT or Drugs.com.

5.3.2 Adverse drug reaction detection

To evaluate the efficiency of the keyword-based dictionary filter, we performed comparison experiments using the basic co-occurrence based method (BCM) (Yang & Jiang 2012) to detect ADRs from SteadyHealth.Com with and without the filter. The comparison results are shown in Figure 5.2. The x-axis stands for the top x pairs having the highest co-occurrence frequencies and the y-axis is the precision. It can be observed from Figure 5.2 that BCM with the filter contributes to an increased precision than that without the filter. It indicates that the keyword filter can filter out those irrelevant drug-ADRs pairs. The irrelevant pairs mainly consist of pairs which have no causality relationship, and pairs in which drugs are used to cure the adverse drug reaction symptoms. Such irrelevant drug-ADRs

pairs bring lots of noise data to the analysis, inevitably reduce the detection performance.

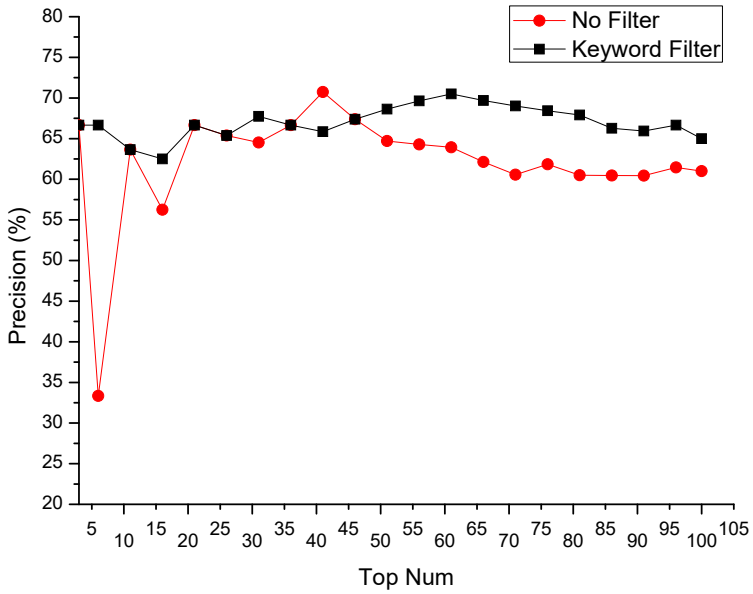


Figure 5.2: **ADRs detection performance from SteadyHealth.Com using the co-occurrence based method with and without the keyword filter**.

To demonstrate the superior performance of the proposed method CIE, we conducted comparison experiments with BCM. The overall precisions of CIE are 73.46% in SteadyHealth.Com and 34.42% in Medhelp.Org respectively, better than 49.06% and 26.73% of BCM. Like other data mining results, it is unrealistic to expect every highly ranked drug-ADRs pair to be of value to domain experts (Jin & Chen 2008). Therefore, we shortlist 20 highest ranked true pairs from our method in Table 5.4 and Table 5.5. As shown in them, most of the drug-ADRs pairs detected by our method are of relatively low-frequency. The ranks of these low-frequency pairs improve a lot in our method. For example, the rank of the pair topamax-migraine improves from

68 to 2 and abilify-polyuria's rank improves from 79 to 3 as shown in Table 5.4 and Table 5.5 respectively. These ranks are much higher than their ranks by BCM, which shows that our method is able to detect relatively low-frequency ADRs. Such rare ADRs will not be detected by BCM if the threshold is larger than their frequencies. Note that the drug-ADRs pairs which are marked 'newfound' in column 'Known or Newfound' are the newfound pairs, requiring further verification by domain experts.

Table 5.4: **Top 20 identified ADRs from SteadyHealth.Com**

CIE Rank	**Drug**	**ADRs**	**Known or Newfound**	**Frequency Rank**
1	norepinephrine	anxiety	known	21
2	**topamax**	**migraine**	**known**	**68**
3	arthritis pain	pain	newfound	39
4	advil	headache	known	17
5	lithium	pain	known	42
6	neurontin	pain	known	66
7	azithromycin	bronchitis	newfound	22
8	methadose	pain	known	28
9	spironolactone	acne	newfound	72
10	enbrel	psoriasis	known	30
11	naproxen sodium	pain	known	38
12	lyrica	pain	known	26
13	clindamycin	acne	known	24
14	avelox	sinusitis	known	59
15	acetaminophen	pain	known	54
16	lexapro	anxiety	known	40
17	hepsera	hepatitis b	known	49
18	ultram	pain	known	18
19	diflucan	vaginal yeast	known	53

20	suboxone	surgery	newfound	8

Table 5.5: **Top 20 identified ADRs from Medhelp.Org**

CIE Rank	**Drug**	**ADRs**	**Known or Newfound**	**Frequency Rank**
1	cortisone	coccydynia	newfound	4
2	analgesic	pain	newfound	38
3	**abilify**	**polyuria**	**known**	**79**
4	antibacterial	infection	newfound	26
5	eye drop	surgery	newfound	83
6	provigil	excessive daytime sleepiness	known	98
7	advil	pain	known	80
8	benadryl	erectile dysfunction	newfound	27
9	compazine	nausea	newfound	64
10	navelbine	pain	known	60
11	oxygen	pneumonia	newfound	67
12	tambocor	stress	newfound	82
13	cough syrup	asthma	newfound	53
14	tramadol	dependence	newfound	97
15	epinephrine	overdose	newfound	85
16	celebrex	arthritis	newfound	29
17	avelox	bronchitis	known	42
18	acetaminophen	pain	known	24
19	novasen	clot	newfound	11
20	piperacillin	hypersensitivity	known	88

Chapter 6

Predicting drug targets using anchor graph hashing and ensemble learning

6.1 Introduction

As we analyzed in section 2.3, existing machine learning methods may come across computational problems. To reduce the computational cost and improve the prediction efficiency, we employ anchor graph hashing (AGH), an excellent hashing method in terms of execution speed and accuracy (Zheng, Sun, Zhu, Lan, Fu & Han 2015) as the compression method. It can embed the data into Hamming space so that the neighbors in the original data space remain neighbors in the Hamming space (Liu & Wang 2011). Therefore, AGH is an ideal technique for dimension reduction. To maintain the prediction accuracy, we employ ensemble learning as the learning method. It combines multiple hypotheses of the base learners to form a better hypothesis, thus producing better predictive performance.

Taking advantage of the above two techniques, we propose to predict drug-target interactions from heterogeneous spaces with anchor graph hashing and ensemble learning (AGHEL). First, every drug is encoded as a

vector using its heterogeneous properties including drug chemical structures, drug side-effects, and drug substituents. Analogously, each target is encoded as a vector using its related Gene Ontology (GO). Next, the high-dimensional drug and target vectors are compressed into short vectors (i.e., hashing bits) in Hamming space using AGH. Then we get the compressed vector of arbitrary drug-target pair by appending the compressed target vector to the compressed drug vector. Finally, compressed vectors of drug-target pairs are used as input to train and evaluate the ensemble learning model. We employ random forest (Tomasi 2017) and XGBoost (Chen & Guestrin 2016) as base learners of the ensemble learning model. Our innovations lie in integrating heterogeneous drug and target properties into a unified framework for predictions, in utilizing the advanced anchor graph hashing to reduce vector dimension, and in the employment of ensemble learning to improve the prediction performance. The performance of the proposed method is evaluated on simulative target prediction of 1,094 drugs from DrugBank. Extensive comparison experiments demonstrate that the proposed method can achieve high prediction efficiency while preserving satisfactory accuracy. In fact, it is 99.3 times faster and only 0.001 less in AUC than the best literature method "Pairwise Kernel Method".

6.2 Methods

6.2.1 Prediction framework

The framework of the proposed method is illustrated in Figure 6.1. First, each drug is encoded as a 5,682-bit vector using its chemical substructures (881), side-effects (4,063), and substituents (738) where each bit represents the presence or absence of the feature by +1 or -1. Likewise, every target is encoded as a vector of 4,198 bits using related GO terms. Then, we employ AGH to embed raw drug and target vectors into low-dimensional Hamming Space (12/16/24/32/48/64 bits). The number of output hashing bits is a parameter which can be set as 12/16/24/32/48/64. Following that,

we use all known drug-target interactions as positive samples, and randomly select the same number of negative samples from unobserved interactions. Next, sample vectors formed by appending corresponding target hashing bits to the drug hashing bits are used as input to train and evaluate the ensemble learning model. The ensemble learning model employs random forest and XGBoost as base learners, and produces the final prediction scores by averaging the scores predicted by them. Finally, the prediction results are evaluated through 10-fold cross-validation on both the prediction accuracy and prediction efficiency.

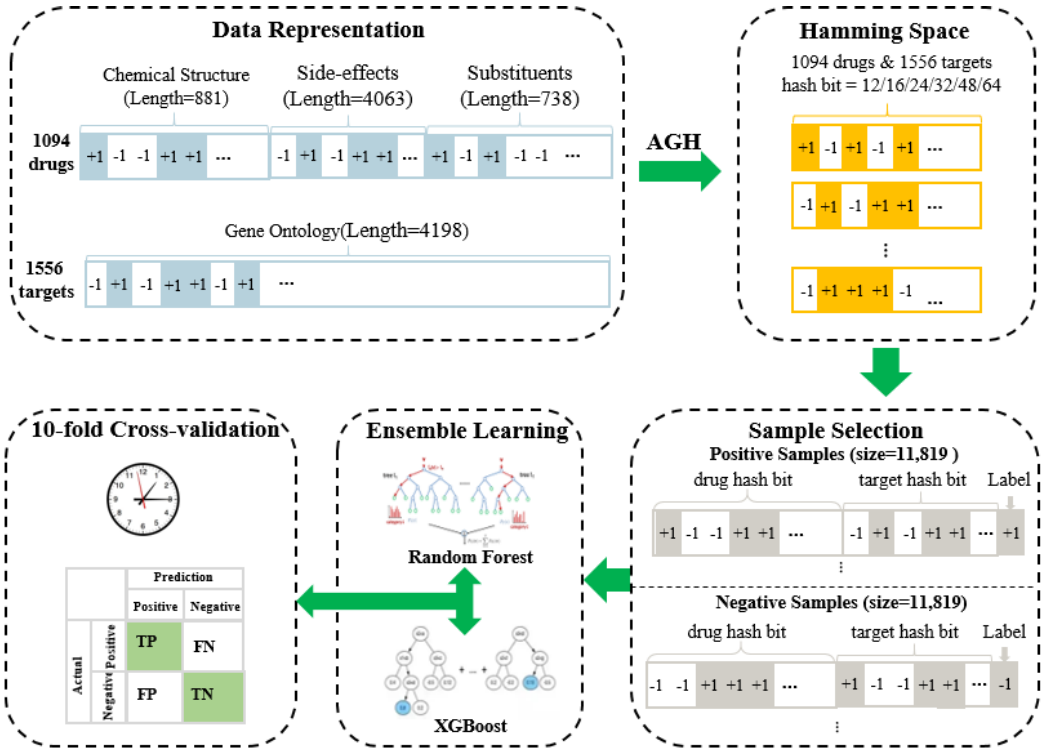


Figure 6.1: **The framework for drug target prediction by anchor graph hashing and ensemble learning.** It consists of five components: data representation, anchor graph hashing compression, sample selection, ensemble learning, and 10-fold cross-validation.

6.2.2 Data representation

We collected 1,094 drugs and 1,556 targets for experiments. Each drug is represented as a 5,682-dimensional binary vector using its chemical structure (881), associated side-effects (4,063), and substituents (738). The drug chemical structure is represented by 881 chemical substructures defined in PubChem (Chen & Wild 2009). We obtained 4,063 associated side-effects from SIDER (Kuhn & Letunic 2015) and 738 substituents from DrugBank (Law & Knox 2013). The elements of the drug vector encode the presence or absence of each feature (i.e., chemical substructure/side-effect/substituent) by +1 or -1. We obtained 4,198 unique GO terms which are associated with the targets from the EMBL-EBI website (Consortium 2004). Analogously, we represented each target as a 4,198-dimensional binary vector where each element encodes the presence or absence of each GO term by +1 or -1.

6.2.3 Anchor graph hashing compression

The goal of AGH is to learn short binary codes such that neighbors in the input space are mapped to neighbors in the Hamming space (Liu & Wang 2011). AGH seeks an r-bit Hamming embedding $Y \in \{1, \text{-}1\}^{n \times r}$ for n points in the input space by minimizing

$$\begin{aligned} \min_{Y} \frac{1}{2} \sum_{i,j=1}^{n} \|Y_i - Y_j\|^2 A_{ij} &= tr(Y^\top L Y) \\ s.t \qquad Y &\in \{1, \text{-}1\}^{n \times r}, 1^\top Y = 0, Y^\top Y = n I_{r \times r} \end{aligned} \tag{6.1}$$

where Y_j is the j^{th} row of Y representing the r-bit code for point x_j, A_{ij} is the similarity between the data pair (x_i, x_j), and L is the graph Laplacian. To solve the above NP-hard problem, AGH uses a small subset of m points called anchors to approximate the data neighborhood structure (Liu & He 2010). Specifically, K-means is performed on all n points to get m ($m << n$) cluster centers (i.e., anchors) $U = \{u_j \in \mathbb{R}^d\}_{j=1}^{m}$ first. Next AGH computes truncated

similarity matrix between all n data points and m anchors as follows:

$$Z_{ij} = \begin{cases} \frac{exp(-D^2(x_i,u_j)/t)}{\sum_{j' \in \langle i \rangle} exp(-D^2(x_i,u_{j'})/t)} & ,\forall j \in \langle i \rangle \\ 0 & \text{otherwise} \end{cases} \tag{6.2}$$

where $\langle i \rangle \in [1:m]$ denotes the indices of s ($s<<m$) nearest anchors of point x_i in U according to a distance function $D()$, and t is the bandwidth parameter. Then the adjacent matrix A can be approximated as $A^{'} = ZV^{-1}Z^{\top}$ where $V = diag(Z^{\top}\mathbf{1}) \in \mathbb{R}^{m \times m}$ (Liu & He 2010). The eigenvectors of $A^{'}$ can be calculated easily by using its low-rank property. In particular, AGH solves the eigenvalue system of a small $m \times m$ matrix $M = V^{-1/2}Z^{\top}ZV^{-1/2}$. This produces r eigenvector-eigenvalue pairs $\{(p_k, \sigma_k)\}_{k=1}^{r}$ where $0 < \sigma_r \leq ... \leq \sigma_1 < 1$. Finally, AGH obtains the desired spectral embedding matrix as

$$Y = \sqrt{n}ZV^{-1/2}P\sum^{-1/2} \tag{6.3}$$

where $P = [p_1, ..., p_r] \in \mathbb{R}^{m \times r}$ and $\sum = diag(\sigma_1, ..., \sigma_r) \in \mathbb{R}^{r \times r}$. Using AGH, the high-dimension input vectors are compressed into low-dimension hashing bits where neighbors in the input space remain neighbors in the output space.

6.2.4 Sample selection

All the 11,819 known drug-target pairs are used as positive samples and assigned the label "+1". An equal number of negative samples are randomly selected from unobserved drug-target pairs and assigned the label "-1". To avoid bias, the negative sample selection process was repeated 5 times and the average results were used for the final evaluation. After we obtained positive samples and negative samples, we represented them as vectors by appending the corresponding target hashing bits to the drug hashing bits. These vectors were subsequently used as input of the ensemble learning model.

6.2.5 Ensemble learning

We chose random forest (Tomasi 2017) and XGBoost (Chen & Guestrin 2016) as the base learners of the ensemble learning model as they perform well on

low-dimension data. We simply took the average value of their prediction scores as the final prediction scores.

6.2.6 10-fold cross-validation

The prediction performance is evaluated via 10-fold cross validation. Since we randomly chose the same number of negative samples as the positive samples (i.e., known drug-target interactions). We repeated the sample selection process and cross-validation 5 times respectively to avoid bias. It means each method would run 25 (5*5) times and the average results were used for the final evaluation. Several metrics, i.e. operating characteristic curve (ROC), the area under ROC curve (AUC), precision, recall, accuracy, F-measure (F1) and the execution time were used to evaluate the prediction performance.

6.3 Results and discussions

6.3.1 Data resources

Drug-target interactions are continually being identified with time. Instead of using old data from the existing literature, we collected the dataset from well-known databases by ourselves. Specifically, we collected drugs from DrugBank (Law & Knox 2013), a comprehensive drug database. We merged targets from DrugBank and DrugCentral (Ursu & Holmes 2017). Table 6.1 and Figure 6.2 show the details of our dataset (Dataset_AGHEL). By comparison, we also list the datasets used in PKM (Ding & Takigawa 2013) (Dataset_PKM) and BLM (Bleakley & Yamanishi 2009) (Dataset_BLM). From table 6.1, we can see that our dataset has much more targets and known interactions.

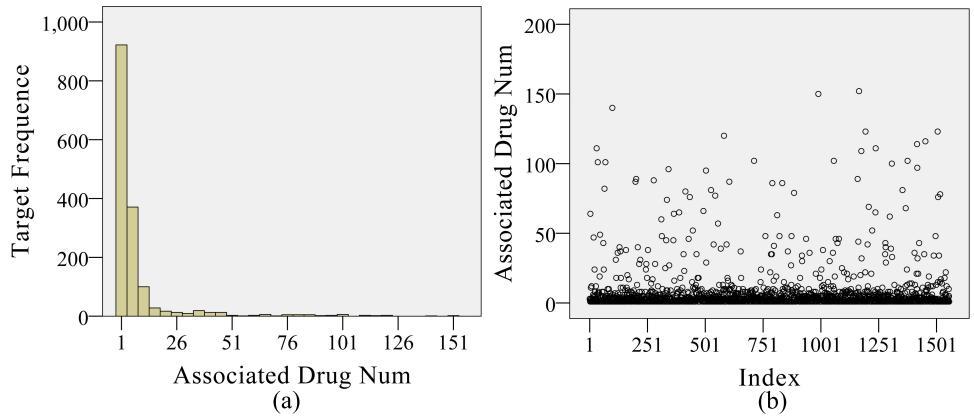


Figure 6.2: **Characteristics of targets and their associated drugs.** The left panel is the histogram of the associated drug number for the targets and the right panel is the index-plot of the number of associated drugs for each target.

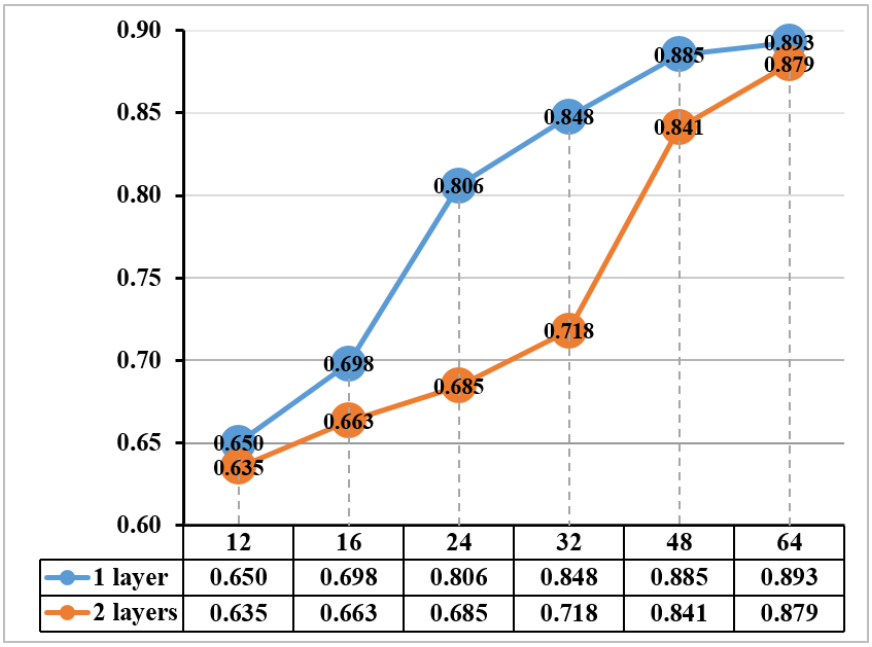
Table 6.1: **Dataset used in this chapter and two datasets used in publications.**

Dataset	Number of drugs	Number of targets	Number of interactions
Dataset_AGHEL	1,094	**1,556**	**11,819**
Dataset_BLM	932	989	5,127
Dataset_PKM	1,205	889	2,782

6.3.2 Parameter optimization

The key parameters of anchor graph hashing are the number of its layer and the output hashing bits (Liu & Wang 2011). Therefore, we first performed experiments to find out the best settings for the hashing layer and output hashing bit. Specifically, we employed k-means to cluster all raw drug (target) vectors to get 100 cluster centers as their anchors first. Then we obtained r-bit ($r \in \{12, 16, 24, 32, 48, 64\}$) hashing bits for drugs (targets) using AGH of different layers (i.e., 1 layer and 2 layers). Finally, the drug hashing bits and the target hashing bits obtained under the same AGH

layer and bit are combined as vectors for predictions and evaluations. The experiment results are illustrated in Figure 6.3. It can be observed that the AUC increases with the number of hashing bits for both 1-layer AGH and 2-layer AGH. Besides, 1-layer AGH outperforms 2-layer AGH on the same number of bits. Based on the above observation, we set the layer of AGH as 1 and the number of output hashing bits as 64 for AGHEL in the following experiment.



	12	16	24	32	48	64
1 layer	0.650	0.698	0.806	0.848	0.885	0.893
2 layers	0.635	0.663	0.685	0.718	0.841	0.879

Figure 6.3: **The AUC of AGHEL with different AGH settings.** The x-axis is the number of output hashing bits and the y-axis is the AUC score.

6.3.3 Comparison with existing methods

To demonstrate the superior performance of AGHEL, we compared it with three different types of prediction methods on both the prediction accuracy and prediction efficiency. BLM and PKM are briefly introduced in Section 2.3. Nearest neighbor (NN) is a commonly-used baseline method. It predicts a test drug to interact with targets known to be targeted by its nearest

drug. Analogously, NN predicts a test target to interact with drugs known to target its nearest target. The above two predictions are combined to produce a definitive prediction. All the experiments were performed on a Linux workstation with dual Intel XEON 3.4GHz processors (14 cores) and 256 GB main memory.

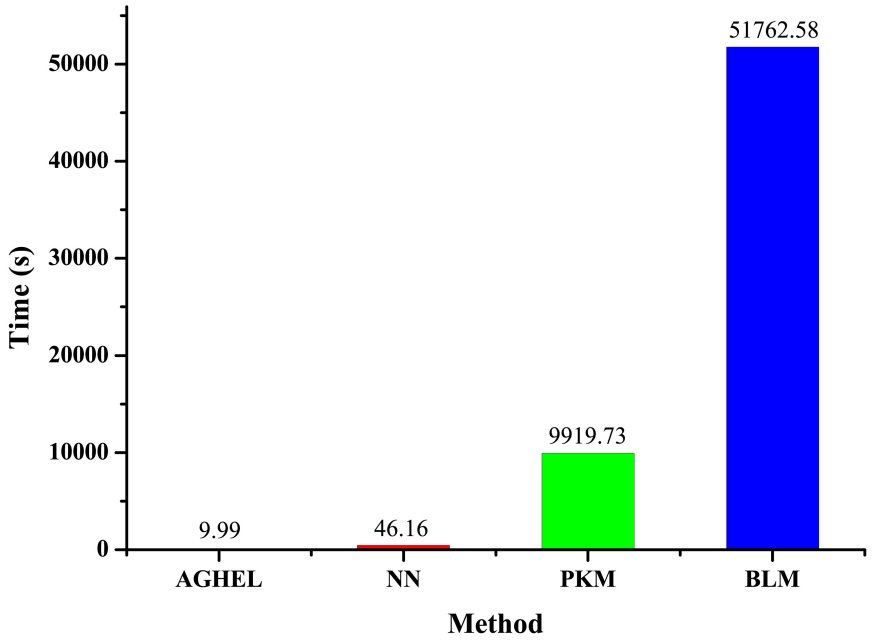


Figure 6.4: **The average execution time (s) of four methods using 10-fold cross-validation.**

Figure 6.4 shows the average execution time of the four methods. We can see that AGHEL is the fastest, completed predictions of 23,638 drug-target interactions in 9.99 s (0.53s for hashing compression and 9.46s for predictions). It is 4.62 times faster than the second fastest method NN, and 5181.44 times faster than the slowest method BLM. Among the three existing comparison methods, NN is the fastest. This is owing to its simple prediction principle i.e., predicting the interacted targets (drugs) for drugs (targets) the same as its nearest drug (target). BLM is the slowest because two classifiers are required to be trained for each interaction. It means 47,276 classifiers have to be trained for predicting the whole 23,638 interactions.

By comparison, PKM is more efficient than BLM, as only one classifier is required to be trained. However, due to the large size of the kernel matrix ($23,638 \times 23,638$), it is still much slower than NN and AGHEL. AGHEL trains one classifier for all predictions as well. However, compared with PKM whose input vector dimension is 23,638, the input vector dimension of AGHEL is largely reduced from 9,880 (5,682 + 4,198) to 128 (64 + 64) by its AGH component. Through this fast process (0.53s), the prediction time of AGHEL decreases largely. This explains why AGHEL costs the least execution time.

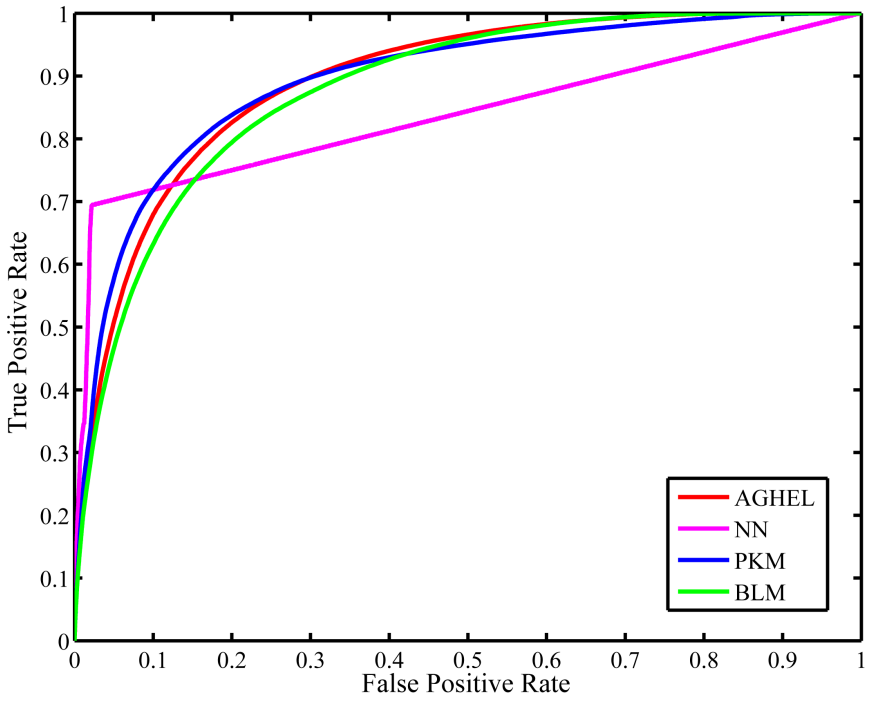


Figure 6.5: **The ROC curves of four methods using 25 runs of 10-fold cross-validation.**

To draw the ROC curves, we merged all prediction scores from the 25 runs for each method and a single ROC curve was drawn. Figure 6.5 illustrates the obtained ROC curves. From it, we can see that PKM (colored blue) achieves the best performance. It has been demonstrated to be the most powerful method in one review of machine learning methods for drug target

Table 6.2: **Average metric scores of four methods evaluated by 25 runs of 10-fold cross-validation.**

Method	AUC	Precision	Recall	Accuracy	F1
AGHEL	0.893	0.797	0.839	0.812	0.817
NN	0.836	0.989	0.119	0.559	0.212
PKM	0.894	0.834	0.803	0.822	0.818
BLM	0.879	0.781	0.824	0.797	0.802

predictions (Ding & Takigawa 2013). The proposed method (colored red) AGHEL achieves the second best performance, slightly below PKM. It may be caused by the information loss when embedding the original drug-target space into Hamming Space.

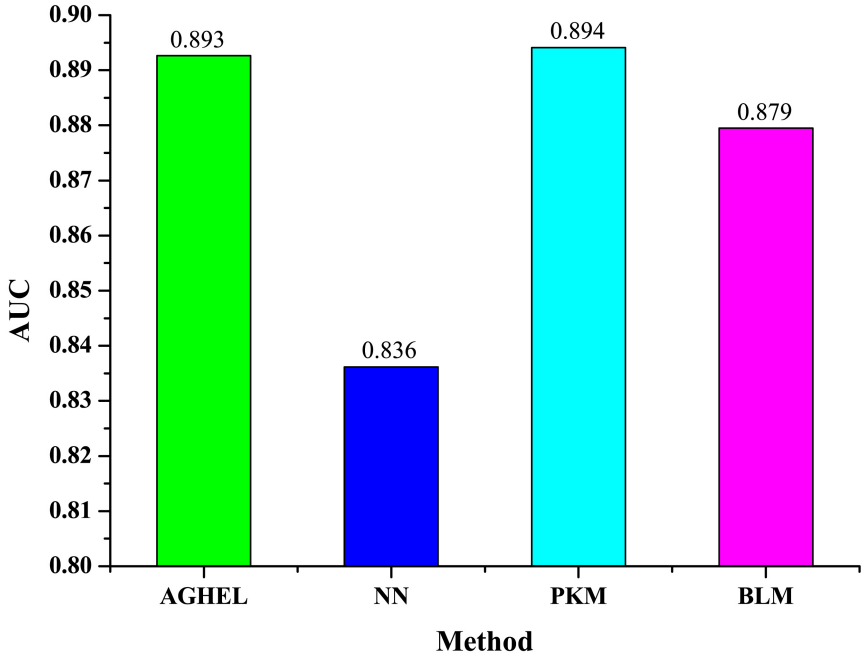


Figure 6.6: **The average AUC of four methods evaluated by 25 runs of 10-fold cross-validation.**

To better visualize the differences between the four methods, we computed the average AUC scores for each method by averaging AUC scores

from its 25 runs of 10-fold cross-validation. The average AUC scores are shown in Figure 6.6. It can be seen that AGHEL (0.893) achieves satisfactory AUC, only 0.001 less than PKM (0.894). Following AGHEL are BLM (0.879) and NN (0.836). Other evaluation metric scores are listed in Table 6.2. It can be observed that NN achieves the highest precision. However, it has much lower recall value than other methods, which leads to the poorest performance on accuracy and F1 measure. AGHEL achieves the highest recall value, and it is close to PKM on F1 measure (AGHEL=0.817, PKM=0.818). The results demonstrate that AGHEL can achieve satisfactory prediction accuracy.

Chapter 7

DDI-PULearn: a novel PU learning method for prediction of drug-drug interaction

7.1 Introduction

As we pointed out in section 2.4, computational methods for drug-drug interaction (DDI) prediction face challenges due to the lack of experimentally verified negative samples. To address this problem, we propose a novel positive-unlabeled learning method named DDI-PULearn for large-scale drug-drug-interaction predictions. DDI-PULearn first generates seeds of reliable negatives via OCSVM (one-class support vector machine) under a high-recall constraint and via the cosine-similarity based KNN (k-nearest neighbors) as well. Then trained with all the labeled positives (i.e., the validated DDIs) and the generated seed negatives, DDI-PULearn employs an iterative SVM to identify a set of entire reliable negatives from the unlabeled samples (i.e., the unobserved DDIs). Following that, DDI-PULearn represents all the labeled positives and the identified negatives as vectors of abundant drug properties by a similarity-based method. Finally, DDI-PULearn transforms these vectors into a lower-dimensional space via

PCA (principal component analysis) and utilizes the compressed vectors as input for binary classifications. The performance of DDI-PULearn is evaluated on simulative prediction for 149,878 possible interactions between 548 drugs, comparing with two baseline methods and five state-of-the-art methods. Related experiment results show that the proposed method for the representation of DDIs characterizes them accurately. DDI-PULearn achieves superior performance owing to the identified reliable negatives, outperforming all other methods significantly. In addition, the predicted novel DDIs suggest that DDI-PULearn is capable to identify novel DDIs.

7.2 Methods

7.2.1 Data resources

Drug properties

We extract drug properties from different data sources. Drug chemical substructures and drug substituents are extracted from DrugBank (Wishart & Feunang 2017), a comprehensive drug database. Drug targets are obtained by fusing drug-target associations from both DrugBank and DrugCentral (Ursu & Holmes 2016). The drug-side-effect associations are downloaded from SIDER (Kuhn & Letunic 2015), a large labeled side-effect database. The drug-indication associations, drug-pathway associations, and drug-gene associations are retrieved from the CTD (comparative toxicogenomics database) (Davis & Grondin 2016).

Drug-drug interactions

We use a recent benchmark dataset (Zhang & Chen 2017) collected from TWOSIDES (Tatonetti & Patrick 2012), a database which contains DDIs mined from FAERS. It contains 548 drugs and 48,584 pairwise drug-drug interactions. The specific drug list and all verified DDIs are available in **Additional file 11**. The source codes and formatted data are included in

Additional file 12.

7.2.2 Proposed methods

The framework of the proposed method is illustrated in Fig. 7.1. It consists of five components listed as follows: reliable negative sample identification, feature vector representation for DDIs, PCA compression, DDI prediction, and performance evaluation. First, reliable negative samples are generated using DDI-PULearn. Then both the labeled positive samples and the reliable negative samples are represented as vectors according to the drug properties, such as chemical substructures, associated side-effects, and indications. Next, the sample vectors are compressed into a lower-dimension space using PCA. Following that, the compressed vectors together with their labels are used as input for DDI prediction. Finally, the prediction performance is evaluated according to the confusion matrix.

Reliable negative sample identification

We propose a novel two-step strategy to generate reliable negative samples. In the first step, we generate reliable negative sample (RNS) seeds from the unlabeled samples using OCSVM and KNN. Then we employ SVM trained with labeled positive samples and RNS seeds to generate reliable negative samples iteratively. Labeled positive samples are validated DDIs and unlabeled samples are unobserved DDIs between every two drugs which are not in labeled positive samples. Fig. 7.2 details the flow for identification of reliable negative samples.

A. RNS seed generation

In the first step, we employ two techniques namely OCSVM and KNN to generate the RNS seeds. For OCSVM, we feed it with all labeled positive samples and optimize its parameters via 5-fold cross-validation. To ensure that the majority of true DDIs are correctly predicted, a high recall (> 0.95) is required for OCSVM. With the optimized parameter settings (nu: 0.05,

gamma: 0.001), OCSVM achieves a recall of 0.951 and generates 1,602 RNS seeds from the 101,294 (C^2_{548}-48,584) unlabeled samples.

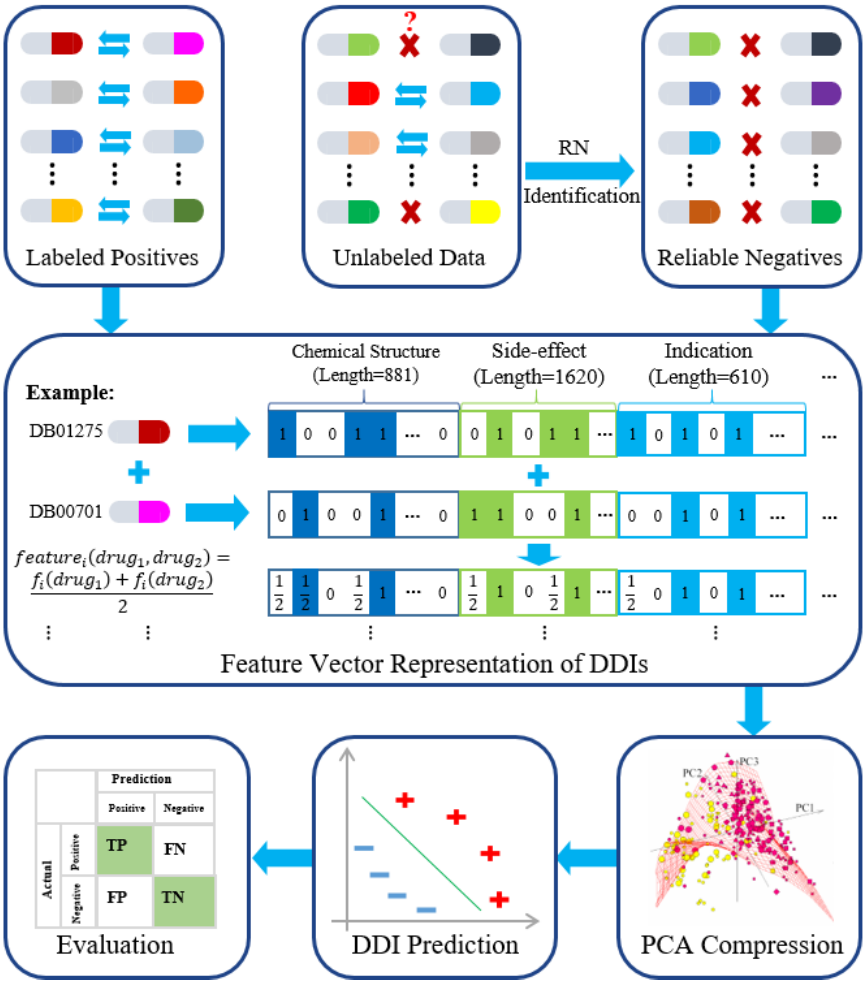


Figure 7.1: **The framework of the proposed method.** It consists of the following five components: reliable negative sample identification, feature vector representation for DDIs, PCA compression, DDI prediction, and performance evaluation. RN: reliable negative samples; PCA: principal component analysis; DDI: drug-drug interaction.

As described in the next subsection, each DDI is represented as a 3,111-dimensional vector. We use the cosine function as the similarity measure for

KNN:

$$\begin{aligned} sim(ddi_i, ddi_j) &= cosine(vector(ddi_i), vector(ddi_j)) \\ &= \frac{\sum_{l=1}^{3,111} [vector_l(ddi_i) * vector_l(ddi_j)]}{\sum_{l=1}^{3,111} vector_l(ddi_i)^2 * \sum_{l=1}^{3,111} vector_l(ddi_j)^2} \end{aligned} \tag{7.1}$$

where $vector(ddi_i)$ and $vector(ddi_j)$ are vectors of the DDI/sample ddi_i and ddi_j respectively. The specific process to generate RNS seeds using KNN is described in Algorithm 7.1. After optimizing, we set k as 5 and the threshold as 4.026. Using the KNN strategy, we obtain 5000 RNS seeds. Merging the RNS seeds generated by OCSVM and KNN, we finally obtain 6,602 RNS seeds (see Table S15 in **Additional file 11**).

Algorithm 7.1 Generating reliable negative sample seeds using KNN.

Input positive samples P, unlabeled samples U, threshold T, the number of nearest neighbors k.

1: $seeds = \{\}$;
2: **for** $u_i \in U$ **do**
3: **for** $p_j \in P$ **do**
4: Computing the similarity $sim(u_i, p_j)$;
5: **end for**
6: Select k nearest neighbors $n_j (j = 1, 2..., k, n_j \in P)$ for u_i;
7: Compute the accumulative similarity $asim(u_i)$ for u_i: $asim(u_i) = \sum_{j=1}^{k} sim(u_i, n_j)$
8: **if** $asim(u_i) < T$ **then**
9: $seeds = seeds \cup u_i$;
10: **end if**
11: **end for**

Output The reliable negative sample seeds $seeds$.

B. Iterative SVM for RNS identification

In the second step, we run SVM trained by labeled positive samples and RNS seeds iteratively to identify all reliable negatives from the remaining unlabeled data. The pseudo-code is shown in Algorithm 7.2. We aim to

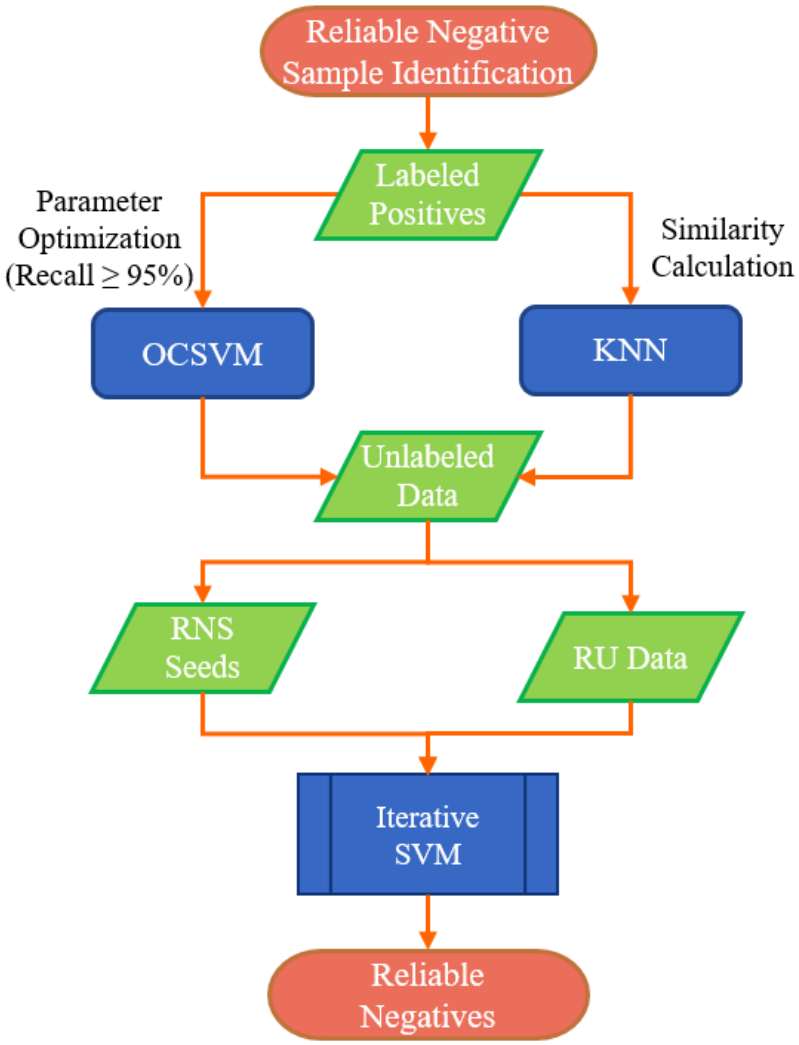


Figure 7.2: **The flow chart for the identification of reliable negative samples.** OCSVM: one-class support vector machine; KNN: k-nearest neighbor; RNS: reliable negative samples; RU: remaining unlabeled.

identify all reliable negative samples from the unlabeled data, thus we use the last SVM classifier at convergence as the best classifier instead of selecting a good classifier from the classifiers built by SVM. Through the iteration, we finally obtained 45,026 reliable negative samples.

Algorithm 7.2 Reliable negative samples extraction with iterative SVM.

Input positive samples P, unlabeled samples U, reliable negative sample seeds $seeds$, reliable negative samples $RNSs$, the remaining unlabeled samples RU (i.e., $U - seeds$).

```
Initiate RNSs as seeds;
Assign label +1 to each sample in P;
Assign label -1 to each sample in RNSs;
while true do
    Train a SVM classifier S using P and RNSs;
    Classify RU using S;
    Let the samples in RU classified as negatives be V;
    if V == {} then
        break;
    else
        RNSs = RNSs ∪ V;
        RU = RU − V;
    end if
end while
```

Output The reliable negative samples $RNSs$.

Feature vector representation for DDIs

We collected a variety of drug properties which may help to improve the prediction, namely drug chemical substructures, drug substituents, drug targets, drug side-effects, drug indications, drug-associated pathways, and drug-associated genes. We investigate which drug property to use for drug representation by feature importance ranking using Random Forrest. The implementation details and experiment results are described in **Additional file 10**. The feature ranking analysis shows that drug properties including drug chemical substructures, drug targets, and drug indications play a leading role in DDI prediction, thus, we decide to employ them for drug representation. Specifically, we represent each drug as a 3,111-dimensional

feature vector using 881 drug chemical substructures, 1,620 side-effects, and 610 indications. The drug chemical substructures correspond to 881 substructures defined in the PubChem database (Kim, Chen, Cheng, Gindulyte, He, He, Li, Shoemaker, Thiessen & Yu 2018). The side-effects and indications are 1,620 unique side-effects in SIDER (Kuhn & Letunic 2015), and 610 unique indications in DrugBank (Wishart & Feunang 2017) respectively. Each bit of the feature vector denotes the absence/presence of the corresponding substructure/side-effect/indication by 0/1. Further, we propose a similarity-based representation for DDIs based on the following formula:

$$vector_k(drug_i, drug_j) = \frac{feature_k(drug_i) + feature_k(drug_j)}{2} \tag{7.2}$$

where $feature_k(drug_i)$ and $feature_k(drug_j)$ are the k-th bit of the feature vectors of drug $drug_i$ and $drug_j$ respectively, $vector_k$ is the k-th bit of vector for the DDI $drug_i$-$drug_j$.

PCA compression

There are 149,878 (C_{548}^2) possible DDIs between the 548 drugs used for experiments. Thus the size of the classification input could be around the order of magnitude of billion ($149,878 * 3,111$). Such high dimensionality inevitably incurs a huge computational cost. To speed up the prediction process, we employ PCA to map the raw vectors of DDIs into lower-dimension space. Specifically, all training DDI vectors are used to fit the PCA first. Then the fitted PCA is used to transform both the training and testing DDI vectors into lower-dimensional vectors. Finally, the compressed vectors are used as input to train and validate the binary classifier.

DDI prediction

We formalize the DDI prediction task as a binary classification problem to predict a DDI is true or not. The inputs for the binary classifiers are the compressed vectors of DDIs and their labels. Specifically, we label labeled

positive samples (i.e., validated DDIs) as +1 and the generated reliable negative samples as -1. Finally, we train and test a binary classifier with the above vectors and labels. Specifically, we employ "Random Forrest" as the binary classifier in this chapter.

Performance evaluation

5-fold CV (cross-validation) is performed to evaluate the prediction performance. To avoid the bias of data split, 5 independent runs of 5-fold CV are implemented and average results are used for final evaluation. Precision, recall, F1-score, and AUC (area under the receiver operating characteristic curve) are used as evaluation metrics.

7.3 Results and discussions

We first report the number of components for PCA. Then we present the prediction performances under different representations of DDIs using multi-source drug property data. Following that, we show the performance improvement brought by reliable negative samples generated by DDI-PULearn via comparing with randomly selected negative samples and all potential negative samples. We also demonstrate the superior prediction performance of DDI-PULearn by comparing with five state-of-the-art methods. Finally, we apply DDI-PULearn to predict unobserved DDIs and verify the results in DrugBank.

7.3.1 Components for PCA

To obtain the best setting for the PCA component number (PCN), we tried the following settings: $PCN \in \{1, 5, 10, 20, 30, 40, 50, 65, 80, 95, 110, 125, 140, 150, 160, 175, 200, 225, 250, 275, 300, 350, 400, 450, 500, 550, 600, 750, 800, 1000, 1250, 1750, 2000\}$. The F1-scores of DDI-PULearn with different PCNs are illustrated in Fig. 7.3. It can be observed that the F1-score

increases with PCN when $PCN \leq 50$. Besides, the F1-score values plateau when the PCN is larger than 50. The same conclusion can be drawn from the AUC results, as shown in Figure S4 in **Additional file 10**. Based on the above observation and considering the computational memory and time cost (computational memory and time increase with PCN), we set PCN as 50 for DDI-PULearn in our experiments.

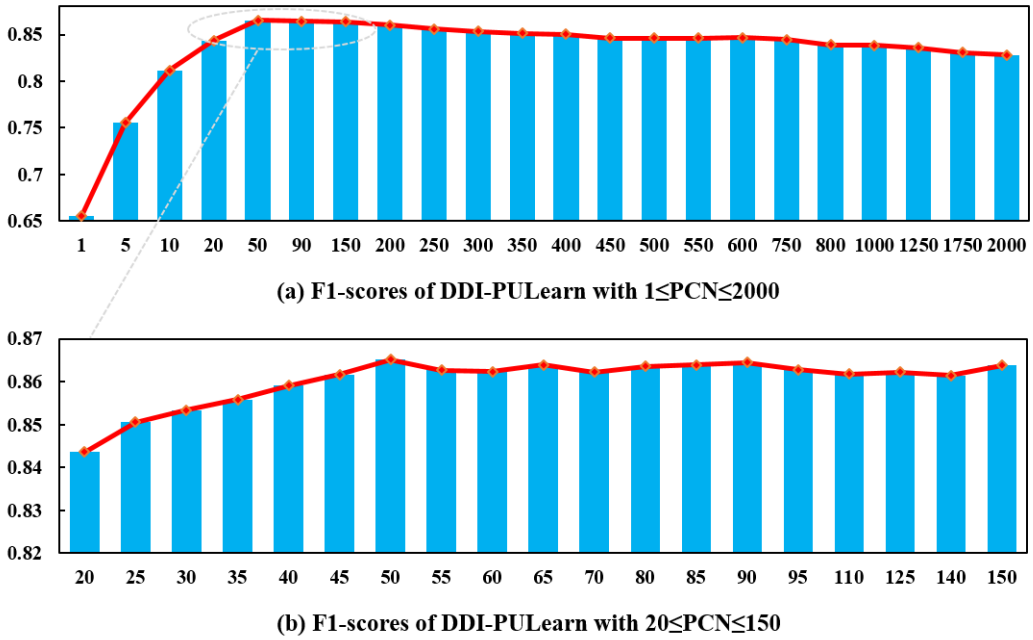


Figure 7.3: ***F*1-scores of DDI-PULearn with different PCNs.** The x-axis is the PCA component number and the y-axis is the F1-score. Panel (a) shows the F1-scores for PCN between 1 and 2,000, and Panel (b) is an amplification of the range [20,150].

7.3.2 Representation of DDIs using multi-source drug property data

As mentioned in the "Feature vector representation for DDI" subsection, we perform the feature ranking analysis to decide which drug property to

use for DDI representation. Here, we conduct more experiments to confirm the analysis results. Specifically, we use the drug chemical substructures, drug targets and drug indications as basic drug properties (BDPs) for representation. Then we test the following 8 combinations of drug features for predictions: (1) BDPs; (2) BDPs + substituents; (3) BDPs + targets; (4) BDPs + pathways; (5) BDPs + substituents + targets; (6) BDPs + substituents + pathways; (7) BDPs + targets + pathways; (8) BDPs + substituents + targets + pathways. Apart from the feature vector representation, other details of the eight combinations are the same with DDI-PULearn. Fig. 7.4 shows the bar charts of the prediction results. It can be observed that all performance evaluation indices (i.e., precision/recall/F1-score) vary very slightly among the above 8 combinations. Employing more drug features for predictions bring redundant information which doesn't improve the prediction performance. It indicates that drug properties including drug substituents, drug targets and drug pathways play a minor role in the DDI predictions while the basic drug properties decide the prediction performance. The results further confirm the conclusion drawn in the previous feature ranking analysis. The detailed evaluation index values of the predictions are listed in Table S10 in **Additional file 11**.

7.3.3 Performance improvement brought by identified reliable negative samples

Existing classification-based models either use all potential negative samples (all-negatives hereafter) or random negative samples (random-negatives hereafter) for predictions (Cheng & Zhao 2014, Cami & Manzi 2013). All-negatives refer to all potential non-DDIs (i.e., unobserved DDIs) which are not in the positive samples. Random-negatives are generated by selecting a random number of negatives from all-negatives. To demonstrate the prediction performance improvement brought by reliable negative samples identified by DDI-PULearn, we compare DDI-PULearn with the above two baseline methods. Specifically, we obtain 101,294 ($C_{548}^2 - 48,584$) negatives

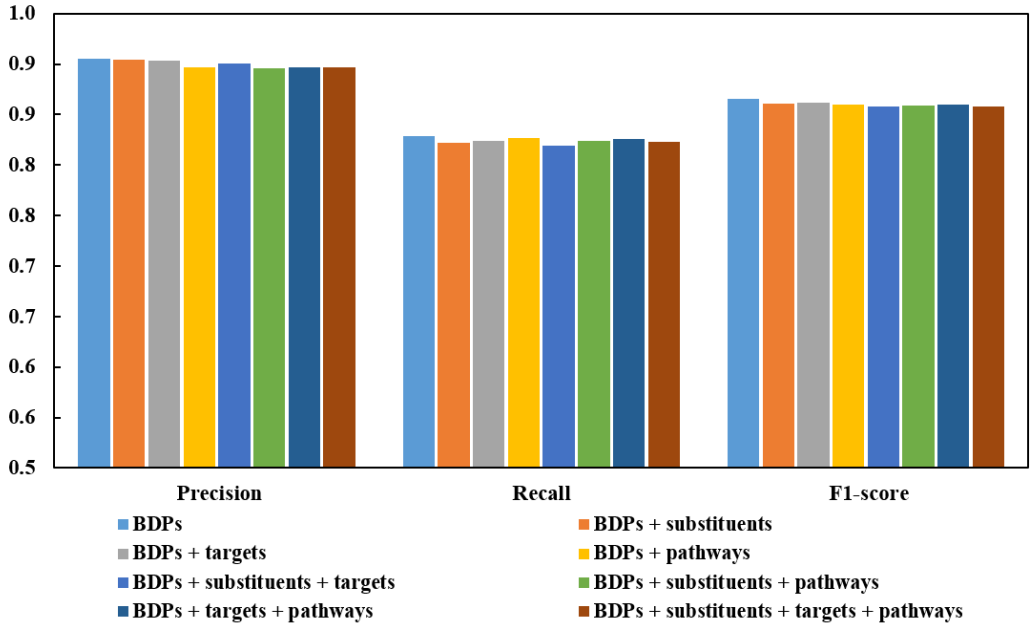


Figure 7.4: **Prediction results using different combinations of drug features.** BDPs refer to the basic drug properties namely drug chemical substructures, drug targets, and drug indications.

for all-negatives. And we randomly select the same number of negatives (i.e., 45,026) with DDI-PULearn as random-negatives. Besides the negative samples, other details of prediction using all-negatives and random-negatives are the same with DDI-PULearn. To avoid bias, random-negatives are repeated 5 times and the average results are used for the final evaluation. Related prediction results are shown Table 7.1. It can be clearly seen that the prediction performances are significantly improved owing to the identified reliable negative samples. For example, the F1-score improvement over random-negatives and all-negatives are 0.147 (20.47%) and 0.315 (57.27%). It suggests that a better decision boundary has been learned with the identified reliable negative samples.

Table 7.1: **Prediction performance comparison with the two baseline methods, namely all-negatives and random-negatives.**

Method	Precision	Recall	F1-score
DDI-PULearn	**0.906**	**0.828**	**0.865**
random-negatives	0.765	0.676	0.718
all-negatives	0.709	0.449	0.550

7.3.4 Comparison with existing state-of-the-art methods

To further confirm the superior performance of DDI-PULearn, we compare it with several state-of-the-art methods reported in a recent study (Zhang & Chen 2017) using the same dataset. Same as (Zhang & Chen 2017), we evaluated DDI-PULearn by 20 runs of 3-fold cross-validation and 5-fold cross-validation under the same condition. The macro-averaging results of the 20 runs are used for final evaluation. The comparison results are listed in Table 7.2 and Table 7.3. Vilar's substructure-based method (Vilar & Harpaz 2012) and Vilar's interaction-fingerprint-based method (Vilar & Uriarte 2013) are two similarity-based methods proposed by Vilar et al.; Zhang's weighted average ensemble method, Zhang's L1 classifier ensemble method and Zhang's L2 classifier ensemble method are three ensemble methods which integrate neighbor recommendation, random walk and matrix perturbation by Zhang et al (Zhang & Chen 2017). The reported performances of the five state-of-the-art methods are extracted from Zhang's study (Zhang & Chen 2017). As shown in Table 7.2 and Table 7.3, DDI-PULearn achieves better performance than other state-of-the-art methods on all metrics. For example, using 5-fold cross-validation, DDI-PULearn outperforms the other five methods by 0.633 (276.6%), 0.415 (92.9%), 0.150 (21.1%), 0.139 (19.3%), 0.143 (19.9%) in F1-score respectively. The results

further confirm the superior performance of the proposed positive-unlabeled learning method DDI-PULearn.

Table 7.2: **Performances of DDI-PULearn and the benchmark methods evaluated by 20 runs of 3-fold cross-validation.**

Method	Precision	Recall	F1-score
Vilar's substructure-based method	0.145	0.535	0.229
Vilar's interaction-fingerprint-based method	0.377	0.553	0.447
Zhang's weighted average ensemble method	0.782	0.703	0.740
Zhang's L1 classifier ensemble method	0.788	0.717	0.751
Zhang's L2 classifier ensemble method	0.784	0.712	0.746
DDI-PULearn	**0.902**	**0.822**	**0.860**

Table 7.3: **Performances of DDI-PULearn and the benchmark methods evaluated by 20 runs of 5-fold cross-validation.**

Method	Precision	Recall	F1-score
Vilar's substructure-based method	0.145	0.535	0.229
Vilar's interaction-fingerprint-based method	0.377	0.553	0.447
Zhang's weighted average ensemble method	0.775	0.659	0.712
Zhang's L1 classifier ensemble method	0.785	0.670	0.723
Zhang's L2 classifier ensemble method	0.783	0.665	0.719
DDI-PULearn	**0.904**	**0.824**	**0.862**

7.3.5 Novel DDIs predicted by DDI-PULearn

We employ DDI-PULearn to predict labels for the 101,294 unobserved DDIs, which are not available in the benchmark dataset. In the prediction, a larger prediction score of a drug pair suggests they have a higher interaction probability. We can obtain a recommendation list of novel DDIs by ranking them in descending order of their prediction scores. Like other data mining results, it is unrealistic to expect all highly ranked DDIs to be of value to domain experts. Therefore, we shortlist the top 25 novel interactions predicted by DDI-PULearn in Table 7.4. We further verify them in the DrugBank database which stores the latest DDI information. We highlight the confirmed DDIs in bold font. From Table 7.4, we can see that a significant ratio of predicted interactions is confirmed in DrugBank (11 out of 25). It indicates that DDI-PULearn does have the capability to predict novel drug-drug interactions.

Table 7.4: **Top 25 novel DDIs predicted by the proposed method DDI-PULearn.** (DDIs which are confirmed in DrugBank are highlighted in bold font.)

Rank	Drug1	Drug1 Name	Drug2	Drug2 Name	Score
1	**DB01137**	**Levofloxacin**	**DB01182**	**Propafenone**	**1**
2	DB00512	Vancomycin	DB00704	Naltrexone	1
3	DB00704	Naltrexone	DB00783	Estradiol	1
4	DB00773	Etoposide	DB00783	Estradiol	1
5	DB01137	Levofloxacin	DB00704	Naltrexone	1
6	DB00635	Prednisone	DB00704	Naltrexone	1
7	**DB00295**	**Morphine**	**DB01037**	**Selegiline**	**0.975**
8	**DB01050**	**Ibuprofen**	**DB00712**	**Flurbiprofen**	**0.975**
9	**DB00218**	**Moxifloxacin**	**DB00328**	**Indomethacin**	**0.975**
10	DB00213	Pantoprazole	DB01189	Desflurane	0.975
11	DB00586	Diclofenac	DB01037	Selegiline	0.975
12	DB00398	Sorafenib	DB00445	Epirubicin	0.975
13	DB00724	Imiquimod	DB00331	Metformin	0.975
14	**DB00203**	**Sildenafil**	**DB00457**	**Prazosin**	**0.975**
15	**DB00999**	**Hydrochlorothiazide**	**DB00880**	**Chlorothiazide**	**0.975**
16	DB00635	Prednisone	DB00324	Fluorometholone	0.975
17	**DB00704**	**Naltrexone**	**DB00327**	**Hydromorphone**	**0.975**
18	**DB00295**	**Morphine**	**DB00704**	**Naltrexone**	**0.975**
19	**DB00813**	**Fentanyl**	**DB01232**	**Saquinavir**	**0.975**
20	**DB00624**	**Testosterone**	**DB00959**	**Methylprednisolone**	**0.95**
21	DB00959	Methylprednisolone	DB00550	Propylthiouracil	0.95
22	**DB01193**	**Acebutolol**	**DB00264**	**Metoprolol**	**0.95**
23	DB00295	Morphine	DB00674	Galantamine	0.95
24	DB01137	Levofloxacin	DB00323	Tolcapone	0.95
25	DB00591	Fluocinolone Acetonide	DB00641	Simvastatin	0.95

Chapter 8

Conclusions and future work

8.1 Conclusions

In this thesis, we mainly address three research problems on drugs namely drug side-effect prediction, drug-target identification, and drug-drug interaction detection. The proposed methods for solving the three problems are detailed in Chapters 3-7 and are presented in my published journal and conference papers (see section **List of Publications**). Specifically, Chapter 3&4 describe two methods to predict side-effects for single drug medication and combined medication of multi-drugs respectively. Chapter 5 presents the method designed to detect adverse drug reactions (i.e. side-effects) from medical forums. The machine learning methods developed for drug-target association prediction, and drug-drug interaction detection are introduced in Chapter 6&7 respectively. The work, innovations, and contributions of this thesis are concluded below.

In Chapter 3, a novel method which learns from reliable negative samples generated by a comprehensive drug similarity framework is introduced to predict side-effects for single drug medication. Comparing with existing related methods, this method has several advantages. Firstly, it integrates drug similarity information from various drug properties including chemical structures, target proteins, substituents, and therapeutic information.

Related experiments demonstrate that the integration characterizes the drug features more accurately. Secondly, it generates reliable negative samples according to inverse drug similarities which improve the prediction performance significantly. In this way, the problem of lacking reliable negative samples is solved and the binary classification of drug-side-effect associations become possible. Extensive evaluations, comparisons, and the case study all confirm the reliability and accuracy of the proposed method.

Chapter 4 focuses on predicting side-effects for combined medication of multi-drugs. A scoring method on a drug-disease-gene tripartite network is developed to prioritize interacting drugs, paving a way to generate credible negative samples for side-effect prediction of combined medication. The scoring method makes use of three types of relationships: drug function relationship, drug therapeutic relationship in diseases, and the relationship between shared target genes and genes of common diseases. Besides, another method is designed to represent drug pairs as feature vectors using its chemical structures, target proteins, substituents, and enriched pathways. The above two methods overcome obstacles of applying machine learning to side-effect prediction of combined medication, namely the lack of proper drug-pair representation and credible negative samples. Comparison experiments using four machine learning algorithms and with two baseline approaches confirm the feasibility of the drug-pair representation method, demonstrating the superior prediction performance of the whole proposed prediction method. The case study reveals that the proposed method is able to predict both known and novel drug-drug-side-effect associations.

Chapter 5 presents an improved method to detect adverse drug reactions from medical forums via constrained information entropy. The web-pages are crawled from two online medical forums namely SteadyHealth.Com and MedHelp.Org using the specially-developed crawler first. Then post contents are extracted from the crawled semi-structured web-pages using a Java library HtmlParser. Following that, entities including drug names and ADRs are identified using multiple lexicons. Finally, the cause-result relationship

between drugs and ADRs are filtered by a keyword dictionary and detected according to their information entropy. Comparing with co-occurrence based methods, the proposed method shows advantages in the following aspects: (1) It captures both high-frequency and low-frequency ADRs simultaneously. (2) It returns drug-related ADRs only, which can largely reduce the false detection rate. (3) Related comparison experiments demonstrate that the proposed method outperforms the state-of-the-art detection methods in terms of the completeness and accuracy.

In Chapter 6, an ensemble learning method to predict drug-target associations with anchor graph hashing is described. The drug-target association prediction problem is turned into a binary classification task where inputs are feature vectors of drug-target pairs and labels are judgments of their associations. A sample (i.e., drug-target pair) is labeled as positive if the drug interacts with the target, otherwise negative. We successfully applied an advanced technique named "anchor graph hashing" to the prediction task which improves the prediction efficiency significantly. Besides, the prediction method achieves satisfactory prediction accuracy via ensemble learning of random forest and XGBoost. Extensive comparison experiments demonstrate that the proposed method has the highest efficiency while achieving comparable prediction accuracy with the best literature method.

The problem on detection of drug-drug interactions is addressed in Chapter 7. The innovations of the work lie in: (1) Design a similarity-based method using abundant drug properties to represent drug-drug interactions. (2) Development of a novel PU learning method which generates reliable negatives with the help of OCSVM under a high-recall constraint and the cosine-similarity based KNN. The proposed methods are validated on simulative prediction for 149,878 possible interactions between 548 drugs, comparing with two baseline methods and five state-of-the-art methods. The experimental results show that the designed method for DDIs representation characterizes them accurately. Learning using the proposed PU learning

method, superior prediction performance has been achieved, outperforming all other methods significantly. 11 out of 25 top-ranked novel DDIs in terms of prediction scores are confirmed in DrugBank. It suggests that the proposed method does have the capability to detect novel DDIs.

8.2 Future work

With the advance of information technology, research on computational drug design and discovery has experienced fast development. However, there are still many problems required to be addressed in this field. In addition, computational methods for different drug discovery stages should be combined and applied to practical applications. Under this background, we will focus on the following research problems in our future work.

- **Identification of potential non-protein drug targets**
  Current drug-target researches mainly focus on predicting targets belong to proteins, such as enzymes, ion channel, G protein-coupled receptors and, nuclear receptors. Non-protein drug targets such as DNA, RNA, and lipid membranes are not researched enough. However, these non-protein drug targets still play an important role in disease development and treatment. For example, microRNAs known as small non-coding RNA molecules capable of suppressing protein synthesis and regulating genes in many organisms (Zheng, Liu, Xu, Li, Luo & Jiang 2013), are related to many disorders (e.g., cancers, metabolic and neurodegenerative diseases). Identification of non-protein drug targets could help to gain knowledge about the cause and treatment of diseases, and boost the new drug discovery as well.

- **Drug repositioning for approved drugs**
  The goal of drug repositioning is to identify new indications of existing drugs (Luo & Wang 2016). Besides identifying novel drug targets, it provides a promising alternative to boost the productivity of the current

drug design process. Since the bioavailability and safety profiles of the approved drugs are usually available, the developing time, financial cost, and failure risks during the drug development process can be reduced significantly. Methods from data mining can be applied to identify novel drug-disease associations for drug repositioning.

- **Design and implement a comprehensive simulation system for drug discovery**

  Existing research methods and tools usually work on a specified task of drug discovery, e.g., drug target identification. To the best of our knowledge, there are no comprehensive simulation systems for drug discovery which are freely available to the public. We want to design and implement an online simulation system for drug discovery. It integrates sub-tasks including drug target identification, drug side-effect prediction, drug-drug interaction detection, and drug repositioning into a comprehensive system. It has the following features: (1) Given a drug, the system predict and visualize its potential targets, side-effects, interacted drugs, and potential treated diseases; (2) Given a target, the system shows drugs which are potential to bind it; (3) Given a side-effect, its associated drugs and drug-drug interactions are visualized. (4) Given a disease, the system evaluates and visualizes drugs which could treat it. (5) The complex network among the above entities (i.e., drugs, targets, side-effects, and diseases) and their interaction probabilities are visualized using knowledge graph.

- **Predicting drug side-effects and drug interactions for drugs from other markets**

  All our previous research focuses on drugs approved by the U.S. FDA. We plan to develop methods to predict side-effects and interactions for drugs from other markets, such as the Chinese market and the Australian market. The drug development process in these markets is similar to the U.S., however, there is not that much public-available

data in these markets. Moreover, the data is not well-organized and the data quality is not that high. Consequently, it is a challenge even to perform the same task for drugs from other markets.

- **Machine learning for drug understanding**
  Successful application of machine learning methods could help us to gain more drug understanding and boost the drug development process. However, there are a few challenges for the machine learning methods: (1) Lacking validated negative samples in many drug-related research problems such as identification of drug-target association and drug-drug interactions; (2) Unbalanced distribution widely exists in drug-related research problems; (3) Heavily rely on proper feature representation captured from the medical domain knowledge. These challenges suggest another promising technical direction: develop or improve machine learning methods and get them adapted to specific drug-related research problems.

Appendix A

Appendix: Methodology foundation

A.1 Applied machine learning algorithms

A.1.1 K-nearest neighbors

k-nearest neighbors algorithm (KNN) is a non-parametric method for classification. The classification procedures are as follows: (1) Calculate the distance between the query-sample and all training samples. (2) Sort the training samples according to their distances and select the k-nearest neighbors. (3) Gather the labels of the k-nearest neighbors. (4) Determine the predicted label and value of the query-sample (e.g., majority label of k-nearest neighbors).

A.1.2 Support vector machine

Given a dataset with l samples x_i ($x_i \in R^n, i \in \{1, 2, ..., l\}$) and their labels y_i ($y_i \in \{1, -1\}$), the SVM solves the following primal optimization problem

(Chang & Lin 2011):

$$\begin{aligned} \min_{\omega,b,\xi} \quad & \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\omega^T\phi(x_i)+b) \geq 1-\xi_i, \\ & \xi_i \geq 0, i = 1,...,l, \end{aligned} \tag{A.1}$$

where $C > 0$ is the regularization parameter and $\phi(x_i)$ maps x_i into a higher-dimensional space. In practice, we solve the following dual problem:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ \text{subject to} \quad & y^T\alpha = 0, \\ 0 \leq \alpha_i \leq C, \quad & i = 1,2,...,l, \end{aligned} \tag{A.2}$$

where Q is a l by l positive semidefinite matrix, $e = [1,...,1]^T$ is a vector of all ones, $Q_{ij} = y_i y_j K(x_i, x_j)$ and $K(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ is the kernel function. After solving problem A.2 and according to primal-dual relationship, the optimal ω satisfies $\omega = \sum_{i=1}^{l} y_i\alpha_i\phi(x_j)$ and the decision function is $sgn(\omega^T\phi(x) + b) = sgn(\sum_{i=1}^{l} y_i\alpha_i K(x_i, x) + b)$. Finally, we store $y_i\alpha_i, b$, support vectors, label names and kernel parameters for further prediction.

A.1.3 Extreme learning machine

ELM is one of the most excellent learning algorithms for single hidden layer feed-forward neural networks (SLFNs). It randomly chooses input weights and analytically determines output weights of SLFNs, thus providing the best generalization performance at extremely fast learning speed (Huang, Huang, Song & You 2015).

Given a training set $\{(x_i, t_i)|x_i \in R^n, t_i \in R^m, i = 1,2,...,N_1\}$, a SLFNs with N2 hidden neurons can be expressed as follow:

$$\sum_{j=1}^{N_2} \beta_j * g(w_j * x_i + b_j) = o_j, i = 1,2,...,N_1 \tag{A.3}$$

where $g(x)$ is the excitation function; β_j is the output weight; b_j is the offset of the ith hidden unit. The learning goal of ELM is to find specific $w_i^{'}, b_i^{'}, \beta_i^{'} (i = 1, 2, ..., N_2)$ such that

$$||H(w_1^{'}, w_{N_2}^{'}, b_1^{'}, b_{N_2}^{'}) * \beta_i^{'} - T|| = min^{w_i, b_i, \beta_i}(||H(w_1, w_{N_2}, b_1, b_{N_2}) * \beta - T||) \tag{A.4}$$

which is equivalent to minimizing the cost function

$$\sum_{j=1}^{N_1}(\sum_{i=1}^{N_2} \beta_j * g(w_j * x_i + b_j - t_j) \tag{A.5}$$

The learning processes of ELM are as follow: (1) The input weight w_j and bias b_j are assigned arbitrarily. (2) The hidden layer output matrix H is calculated. (3) The output weight β is calculated by: $\beta = H^{\dagger}T$, where $\dagger$ is the Moore-Penrose generalized inverse of H.

A.1.4 Radial basis function networks

Radial basis function(RBF) networks usually consist of three layers shown in Fig. A.1: an input layer, a hidden layer, and a linear output layer. The input layer is the signal source, the hidden layer transfers the raw non-linearly separable space into another separable space with RBF activation function, and the output layer is a scalar function of the input vector, $f(X) : R^n \rightarrow R$.

$$y = f(X) = \sum_{i=1}^{n_c} \omega_i \phi(||X - C_i||, \sigma_i) = \sum_{i=1}^{n_c} \omega_i exp[\frac{-(||X - C_i||^2}{2\sigma_i^2}] \tag{A.6}$$

where n_c is the number of hidden nodes, $||.||$ is Euclidean norm, $X, C_i \in R_n$, and w_i is the weight between the ith radial basis function with the output node $(i = 1, 2, ..., n_c)$.

A.1.5 Logistic regression

Logistic regression is mathematically defined as:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m \tag{A.7}$$

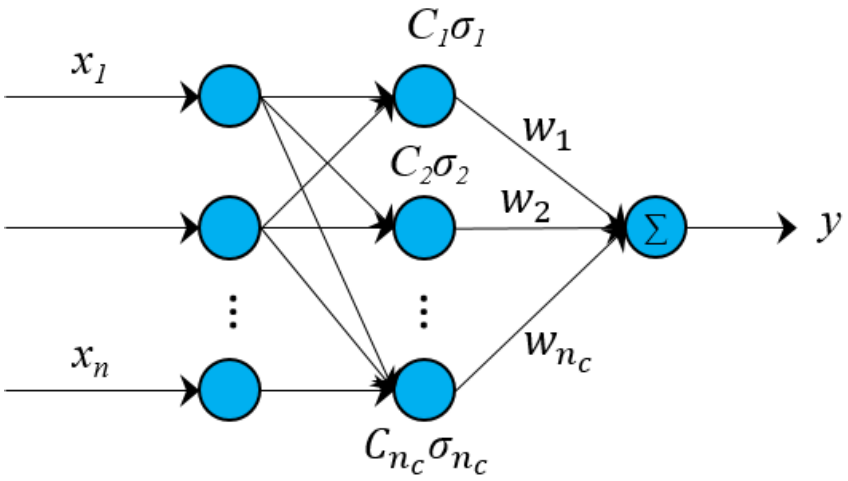


Figure A.1: **The structure of RBF.**

where p is the probability of label $y = 1$, β_i is the regression coefficient, and x_i is the input. Then, we can estimate the value of p as:

$$\hat{p} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_m x_m}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_m x_m}}. \tag{A.8}$$

The regression problem is converted to estimate β_i so that $\hat{p}$ is close to p.

A.1.6 Random forest

Random forest is an ensemble learning method using a multitude of decision trees. First, it generates m sub-training sets using the bagging method. Then a decision tree is built on each sub-training set. Different from conventional decision trees which split the tree nodes according to some indices (e.g., maximum information gain), the decision trees in random forest use the optimal solution of randomly-selected features to split the tree nodes. Finally, the results of all decision trees are averaged/voted to predict the final prediction results.

A.1.7 XGBoost

XGBoost is a scalable end-to-end tree boosting system developed by Chen et al. (Chen & Guestrin 2016). Comparing with other tree boosting methods (e.g., Gradient Boosting Decision Tree), it has the following features: (1) A novel sparsity-aware algorithm is proposed to handle sparse data. (2) A theoretically justified weighted quantile sketch is designed for approximate tree learning. Therefore, XGBoost is efficient and accurate. In Chapter 6, we use XGBoost as a base learner for ensemble learning. We use the python package of XGBoost downloaded from GitHub (https://github.com/dmlc/xgboost). More details of XGBoost can be found in their paper (Chen & Guestrin 2016) and the above GitHub website.

A.2 Cross validation and performance evaluation indices

A.2.1 Cross validation

Cross validation is a widely used strategy to estimate how accurately a predictive model performs. It assesses how the model can be generalized to an independent data set.

Specifically, we employed n-fold cross-validation in our research. The common steps are as follows: (1) Samples in the gold standard set are split into n roughly equal-sized subsets; (2) each subset is used as the test set, and the remaining subsets are taken as the training set in turn to test and train the predictive models; (3) the final performance is evaluated on all results over n-folds.

A.2.2 Performance evaluation indices

The following evaluation indices are used to evaluate the performance of the prediction/classification models:

$$Precision = \frac{TP}{TP + FP} \tag{A.9}$$

$$Recall = \frac{TP}{TP + FN} \tag{A.10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{A.11}$$

$$F1\text{-}score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{A.12}$$

$$Macro\_X = \frac{\sum_{i=1}^{n} X_i}{n} \tag{A.13}$$

where TP and FP are the numbers of correctly and falsely predicted positives respectively; TN and FN are the numbers of correctly and falsely predicted negatives respectively; $X \in$ {Precision, Recall, $F1$-score, Accuracy}. Apart from the above indices, AUC (area under the receiver operating characteristic curve) and AUPR (area under the precision-recall curve) are used as evaluation indices as well. Precision, recall, accuracy, F1-score and their macro values can be calculated using the above formulas. AUC and AUPR can be computed via the ROC curve and PR curve using open-source tools (e.g., perfcurve in Matlab).

Appendix B

Appendix: List of Supplementary files

The Additional file list and the corresponding download links

name	chapter	description	link
Additional file 1	3	supplementary figures for chapter 3	download
Additional file 2	3	supplementary tables for chapter 3	download
Additional file 3	3	codes and data for chapter 3	download
Additional file 4	4	sorted drug-pairs	download
Additional file 5	4	researched drugs and ADRs	download
Additional file 6	4	disease gene associations	download
Additional file 7	4	drug-drug-side-effect associations	download
Additional file 8	4	code and data for chapter 4	download
Additional file 9	4	supplementary figures for chapter 4	download
Additional file 10	7	supplementary figures for chapter 7	download
Additional file 11	7	supplementary tables for chapter 7	download
Additional file 12	7	code and data for chapter 7	download

Appendix C

Appendix: List of Symbols

The following list is neither exhaustive nor exclusive, but may be helpful.

CCA	Canonical correlation analysis
$SCCA$	Sparse canonical correlation analysis
CDK	Chemistry development kit
GO	Gene ontology
ATC	Anatomical therapeutic chemical
$ADRs$	Adverse drug reactions
$SRSs$	Spontaneous reporting systems
$EHRs$	Electronic health records
PBs	Pharmacological databases
$DDIs$	Drug-drug interactions
$DDSAs$	Drug-drug-side-effect associations
$FAERS$	Food and drug administration adverse event reporting system

NSR	Negative sample ratio
PCA	Principal component analysis
PCN	Principal component analysis component number
AUC	Area under the receiver operating characteristic curve
ROC	Receiving operating characteristic curve
CIE	Constrained information entropy
WHO	World health organization
$CADM$	Co-occurrence based adverse drug reaction detection method
ANN	Artificial neural network
BLM	Bipartite local model
AGH	Anchor graph hashing
NN	Nearest neighbor
PKM	Pairwise kernel method
$PUlearning$	Positive and unlabeled learning
NB	Naive Bayesian
$RNSs$	Reliable negative samples
EM	Expectation maximization
SOM	Self organizing map
$DDI-PULearn$	Positive and unlabeled learning for drug-drug interaction
CTD	Comparative toxicogenomics database

CV	Cross validation
$BDPs$	Basic drug properties
KNN	K-nearest neighbors
ELM	Extreme learning machine
SVM	Support vector machine
RBF	Radial basis function
$OCSVM$	one-class support vector machine

Bibliography

Agarwal, P. & Searls, D. B. (2008), 'Literature mining in support of drug discovery', *Briefings in Bioinformatics* **9**(6), 479–492.

Andronis, C., Sharma, A., Virvilis, V., Deftereos, S. & Persidis, A. (2011), 'Literature mining, ontologies and information visualization for drug repurposing', *Briefings in Bioinformatics* **12**(4), 357–368.

Atias, N. & Sharan, R. (2011), 'An algorithmic framework for predicting side effects of drugs', *Journal of Computational Biology* **18**(3), 207–218.

Banda, J. & Callahan, A. (2016), 'Feasibility of prioritizing drug–drug-event associations found in electronic health records', *Drug Safety* **39**(1), 45–57.

Beijnen, J. H. & Schellens, J. H. (2004), 'Drug interactions in oncology', *The Lancet Oncology* **5**(8), 489–496.

Bender, A. & Scheiber, J. (2007), 'Analysis of pharmacology data and the prediction of adverse drug reactions and off-target effects from chemical structure', *ChemMedChem* **2**(6), 861–873.

Benton, A. & Ungar, L. (2011), 'Identifying potential adverse effects using the web: a new approach to medical hypothesis generation', *Journal of Biomedical Informatics* **44**(6), 989–996.

Bleakley, K. & Yamanishi, Y. (2009), 'Supervised prediction of drug–target interactions using bipartite local models', *Bioinformatics* **25**(18), 2397–2403.

Bock, J. R. & Gough, D. A. (2005), 'Virtual screen for ligands of orphan g protein-coupled receptors', *Journal of Chemical Information and Modeling* **45**(5), 1402–1414.

Brouwers, L., Iskar, M., Zeller, G., Van Noort, V. & Bork, P. (2011), 'Network neighbors of drug targets contribute to drug side-effect similarity', *PLoS ONE* **6**(7), e22187.

Bushardt, R. L., Massey, E. B., Simpson, T. W., Ariail, J. C. & Simpson, K. N. (2008), 'Polypharmacy: misleading, but manageable', *Clinical Interventions in Aging* **3**(2), 383.

Cami, A. & Manzi, S. (2013), 'Pharmacointeraction network models predict unknown drug-drug interactions', *PLoS ONE* **8**(4), e61468.

Campillos, M. & Kuhn, M. (2008), 'Drug target identification using side-effect similarity', *Science* **321**(5886), 263–266.

Center, I. K. (2013), 'The data mining process', `http://www.ibm.com/support/knowledgecenter/SSEPGG_9.5.0/com.ibm.im.easy.doc/c_dm_process.html`.

Chang, C.-C. & Lin, C.-J. (2011), 'Libsvm: A library for support vector machines', *ACM Transactions on Intelligent Systems and Technology* **2**(3), 27.

Chen, B. & Wild, D. (2009), 'Pubchem as a source of polypharmacology', *Journal of Chemical Information and Modeling* **49**(9), 2044–2055.

Chen, L., Zeng, W.-M., Cai, Y.-D., Feng, K.-Y. & Chou, K.-C. (2012), 'Predicting anatomical therapeutic chemical (atc) classification of drugs

by integrating chemical-chemical interactions and similarities', *PLoS ONE* **7**(4), e35254.

Chen, T. & Guestrin, C. (2016), Xgboost: A scalable tree boosting system, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 785–794.

Chen, X., Liu, M.-X. & Yan, G.-Y. (2012), 'Drug–target interaction prediction by random walk on the heterogeneous network', *Molecular BioSystems* **8**(7), 1970–1978.

Chen, Y., Zhou, X. S. & Huang, T. S. (2001), One-class svm for learning in image retrieval, *in* '2001 International Conference on Image Processing', Vol. 1, IEEE, pp. 34–37.

Cheng, F., Li, W., Wu, Z., Wang, X., Zhang, C., Li, J., Liu, G. & Tang, Y. (2013), 'Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space', *Journal of Chemical Information and Modeling* **53**(4), 753–762.

Cheng, F. & Zhao, Z. (2014), 'Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties', *Journal of the American Medical Informatics Association* **21**(e2), e278–e286.

Consortium, G. O. (2004), 'The gene ontology (go) database and informatics resource', *Nucleic Acids Research* **32**(suppl 1), D258–D261.

Davis, A. & Grondin, C. (2016), 'The comparative toxicogenomics database: update 2017', *Nucleic Acids Research* **45**(D1), D972–D978.

Ding, H. & Takigawa, I. (2013), 'Similarity-based machine learning methods for predicting drug–target interactions: a brief review', *Briefings in Bioinformatics* **15**(5), 734–747.

Dudley, J. T., Deshpande, T. & Butte, A. J. (2011), 'Exploiting drug–disease relationships for computational drug repositioning', *Briefings in Bioinformatics* **12**(4), 303–311.

Duke, J. D., Han, X., Wang, Z., Subhadarshini, A., Karnik, S. D., Li, X., Hall, S. D., Jin, Y., Callaghan, J. T. & Overhage, M. J. (2012), 'Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions', *PLoS Computational Biology* **8**(8), e1002614.

Edwards, I. R. & Aronson, J. K. (2000), 'Adverse drug reactions: definitions, diagnosis, and management', *The Lancet* **356**(9237), 1255–1259.

Emig, D., Ivliev, A., Pustovalova, O., Lancashire, L., Bureeva, S., Nikolsky, Y. & Bessarabova, M. (2013), 'Drug target prediction and repositioning using an integrated network-based approach', *PLoS ONE* **8**(4), e60618.

Erfani, S. M., Rajasegarar, S., Karunasekera, S. & Leckie, C. (2016), 'High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning', *Pattern Recognition* **58**, 121–134.

European Medicines Agency (2019), 'European adverse drug reactions database'. [Online; accessed 11-May-2019].
URL: *http://www.imi-protect.eu/adverseDrugReactions.shtml*

Faulon, J.-L., Misra, M., Martin, S., Sale, K. & Sapra, R. (2007), 'Genome scale enzyme–metabolite and drug–target interaction predictions using the signature molecular descriptor', *Bioinformatics* **24**(2), 225–233.

Fox, S. (2011), *The social life of health information 2011*, Pew Internet & American Life Project Washington, DC.

Freifeld, C. & Brownstein, J. (2014), 'Digital drug safety surveillance: monitoring pharmaceutical products in twitter', *Drug Safety* **37**(5), 343–350.

Fukuzaki, M. & Seki, M. (2009), Side effect prediction using cooperative pathways, *in* 'IEEE International Conference on Bioinformatics and Biomedicine.', IEEE, pp. 142–147.

Giacomini, K. M., Krauss, R. M., Roden, D. M., Eichelbaum, M., Hayden, M. R. & Nakamura, Y. (2007), 'When good drugs go bad', *Nature* **446**(7139), 975–977.

Ginn, R., Pimpalkhute, P., Nikfarjam, A., Patki, A., O'Connor, K., Sarker, A., Smith, K. & Gonzalez, G. (2014), Mining twitter for adverse drug reaction mentions: a corpus and classification benchmark, *in* 'Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing', Citeseer, pp. 1–8.

Gottlieb, A., Stein, G. Y., Oron, Y., Ruppin, E. & Sharan, R. (2012), 'Indi: a computational framework for inferring drug interactions and their associated recommendations', *Molecular Systems Biology* **8**(1), 592.

Haggarty, S. J., Koeller, K. M., Wong, J. C., Butcher, R. A. & Schreiber, S. L. (2003), 'Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays', *Chemistry & Biology* **10**(5), 383–396.

Halperin, I., Ma, B., Wolfson, H. & Nussinov, R. (2002), 'Principles of docking: an overview of search algorithms and a guide to scoring functions', *Proteins: Structure, Function, and Bioinformatics* **47**(4), 409–443.

Hammann, F. & Gutmann, H. (2010), 'Prediction of adverse drug reactions using decision tree modeling', *Clinical Pharmacology & Therapeutics* **88**(1), 52–59.

Harpaz, R., Chase, H. S. & Friedman, C. (2010), 'Mining multi-item drug adverse effect associations in spontaneous reporting systems', *BMC Bioinformatics* **11**(9), S7.

Harte, J. & Newman, E. A. (2014), 'Maximum information entropy: a foundation for ecological theory', *Trends in Ecology & Evolution* **29**(7), 384–389.

He, L. & Yang, Z. (2013), 'Extracting drug-drug interaction from the biomedical literature using a stacked generalization-based approach', *PLoS ONE* **8**(6), e65814.

Hopkins, A. L. (2009), 'Drug discovery: predicting promiscuity', *Nature* **462**(7270), 167.

Huang, G., Huang, G.-B., Song, S. & You, K. (2015), 'Trends in extreme learning machines: A review', *Neural Networks* **61**, 32–48.

Huang, J., Niu, C., Green, C. D., Yang, L., Mei, H. & Han, J.-D. J. (2013), 'Systematic prediction of pharmacodynamic drug-drug interactions through protein-protein-interaction network', *PLoS Computational Biology* **9**(3), e1002998.

Huang, S.-M., Temple, R., Throckmorton, D. & Lesko, L. (2007), 'Drug interaction studies: study design, data analysis, and implications for dosing and labeling', *Clinical Pharmacology & Therapeutics* **81**(2), 298–304.

Imming, P., Sinning, C. & Meyer, A. (2006), 'Drugs, their targets and the nature and number of drug targets', *Nature Reviews Drug Discovery* **5**(10), 821–834.

Iyer, S. & Harpaz, R. (2013), 'Mining clinical text for signals of adverse drug-drug interactions', *Journal of the American Medical Informatics Association* **21**(2), 353–362.

Jacob, L. & Vert, J.-P. (2008), 'Protein-ligand interaction prediction: an improved chemogenomics approach', *Bioinformatics* **24**(19), 2149–2156.

Jin, H. & Chen, J. (2008), 'Mining unexpected temporal associations: applications in detecting adverse drug reactions', *IEEE Transactions on Information Technology in Biomedicine* **12**(4), 488–500.

John, L. (2010), 'The elements of statistical learning: data mining, inference, and prediction', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**(3), 693–694.

Karimi, S. & Cavedon, L. (2011), Drug side-effects: What do patient forums reveal ?, *in* 'The Second International Workshop on Web Science and Information Exchange in the Medical Web', pp. 10–11.

Karimi, S. & Wang, C. (2015), 'Text and data mining techniques in adverse drug reaction detection', *ACM Computing Surveys* **47**(4), 56.

Kim, M.-H. & Shim, E.-J. (2012), 'A case of allopurinol-induced fixed drug eruption confirmed with a lymphocyte transformation test', *Allergy, Asthma & Immunology Research* **4**(5), 309–310.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B., Thiessen, P. & Yu, B. (2018), 'Pubchem 2019 update: improved access to chemical data', *Nucleic Acids Research* **47**(D1), D1102–D1109.

Kim, S. & Thiessen, P. (2015), 'Pubchem substance and compound databases', *Nucleic Acids Research* **44**(D1), D1202–D1213.

Kuhn, M. & Campillos, M. (2008), 'Large-scale prediction of drug–target relationships', *FEBS Letters* **582**(8), 1283–1290.

Kuhn, M. & Letunic, I. (2015), 'The sider database of drugs and side effects', *Nucleic Acids Research* **44**(D1), D1075–D1079.

Kurgan, L. A. & Musilek, P. (2006), 'A survey of knowledge discovery and data mining process models', *The Knowledge Engineering Review* **21**(01), 1–24.

Lan, C., Chen, Q. & Li, J. (2016), 'Grouping mirnas of similar functions via weighted information content of gene ontology', *BMC Bioinformatics* **17**(19), 159.

Law, V. & Knox, C. (2013), 'Drugbank 4.0: shedding new light on drug metabolism', *Nucleic Acids Research* **42**(D1), D1091–D1097.

Law, V. & Knox, C. (2014), 'Drugbank 4.0: shedding new light on drug metabolism', *Nucleic Acids Research* **42**(D1), D1091–D1097.

Leaman, R. & Wojtulewicz, L. (2010), Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts to health-related social networks, *in* 'Proceedings of the 2010 Workshop on Biomedical Natural Language Processing', Association for Computational Linguistics, pp. 117–125.

Lee, W.-P., Huang, J.-Y., Chang, H.-H., Lee, K.-T. & Lai, C.-T. (2017), 'Predicting drug side effects using data analytics and the integration of multiple data sources', *IEEE Access* **5**, 20449–20462.

Li, K.-L., Huang, H.-K., Tian, S.-F. & Xu, W. (2003), Improving one-class svm for anomaly detection, *in* '2003 International Conference on Machine Learning and Cybernetics', Vol. 5, IEEE, pp. 3077–3081.

Liu, J. & Zhao, S. (2016), 'An ensemble method for extracting adverse drug events from social media', *Artificial Intelligence in Medicine* **70**, 62–76.

Liu, M. & Wu, Y. (2012), 'Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs', *Journal of the American Medical Informatics Association* **19**(e1), e28–e35.

Liu, W. & He, J. (2010), Large graph construction for scalable semi-supervised learning, *in* 'Proceedings of the 27th International Conference on Machine Learning', pp. 679–686.

Liu, W. & Wang, J. (2011), Hashing with graphs, *in* 'Proceedings of the 28th international conference on machine learning', Citeseer, pp. 1–8.

Liu, X. & Chen, H. (2013), Azdrugminer: an information extraction system for mining patient-reported adverse drug events in online patient forums, *in* 'Smart Health', Springer, pp. 134–150.

Loveman, A. B. & Simon, F. A. (1939), 'Fixed eruption and stomatitis due to sulfanilamide', *Archives of Dermatology and Syphilology* **40**(1), 29–34.

Luo, H. & Wang, J. (2016), 'Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm', *Bioinformatics* **32**(17), 2664–2671.

Ma, F., Meng, C., Xiao, H., Li, Q., Gao, J., Su, L. & Zhang, A. (2017), Unsupervised discovery of drug side-effects from heterogeneous data sources, *in* 'Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 967–976.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. & McClosky, D. (2014), 'The stanford corenlp natural language processing toolkit', pp. 55–60.

Martin, Y. C., Kofron, J. L. & Traphagen, L. M. (2002), 'Do structurally similar molecules have similar biological activity?', *Journal of Medicinal Chemistry* **45**(19), 4350–4358.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., Cowley, A. P. & Lopez, R. (2013), 'Analysis tool web services from the embl-ebi', *Nucleic Acids Research* **41**(W1), W597–W600.

MedHelp.org (2019), 'Medhelp.org'. [Online; accessed 11-May-2019].
URL: *http://www.medhelp.org*

MedHelp.org (2019, May 2), 'Medical information, forums and communities'. **URL:** *https://www.medhelp.org*

Mizutani, S., Pauwels, E., Stoven, V., Goto, S. & Yamanishi, Y. (2012), 'Relating drug–protein interaction network with drug side effects', *Bioinformatics* **28**(18), i522–i528.

Muñoz, E., Nováček, V. & Vandenbussche, P.-Y. (2017), 'Using drug similarities for discovery of possible adverse reactions', *American Medical Informatics Association Annual Symposium Proceedings* **2016**, 924.

Muñoz, E. & Nováček, V. (2017), 'Facilitating prediction of adverse drug reactions by using knowledge graphs and multi-label learning models', *Briefings in Bioinformatics* **20**(1), 190–202.

Nagamine, N. & Sakakibara, Y. (2007), 'Statistical prediction of protein–chemical interactions based on chemical structure and mass spectrometry data', *Bioinformatics* **23**(15), 2004–2012.

Nikfarjam, A. & Gonzalez, G. H. (2011), Pattern mining for extraction of mentions of adverse drug reactions from user comments, *in* 'American Medical Informatics Association Annual Symposium Proceedings', Vol. 2011, American Medical Informatics Association, p. 1019.

Niu, S.-Y. & Xin, M.-Y. (2015), 'Dsep: A tool implementing novel method to predict side effects of drugs', *Journal of Computational Biology* **22**(12), 1108–1117.

Palmer, M., Chan, A., Dieckmann, T. & Honek, J. (2013), *Notes to biochemical pharmacology*, John Wiley & Sons.

Park, K., Kim, D., Ha, S. & Lee, D. (2015), 'Predicting pharmacodynamic drug-drug interactions through signaling propagation interference on protein-protein interaction networks', *PLoS ONE* **10**(10), e0140816.

Patki, A., Sarker, A., Pimpalkhute, P., Nikfarjam, A., Ginn, R., O'Connor, K., Smith, K. & Gonzalez, G. (2014), 'Mining adverse drug reaction signals from social media: going beyond extraction', *Proceedings of BioLinkSig* **2014**, 1–8.

Pauwels, E. & Yamanishi, Y. (2011), 'Predicting drug side-effect profiles: a chemical fragment-based approach', *BMC Bioinformatics* **12**(1), 1.

Percha, B., Garten, Y. & Altman, R. B. (2012), Discovery and explanation of drug-drug interactions via text mining, *in* 'Biocomputing', World Scientific, pp. 410–421.

Roughead, E. E. & Semple, S. J. (2009), 'Medication safety in acute care in australia: where are we now? part 1: a review of the extent and causes of medication problems 2002–2008', *Australia and New Zealand Health Policy* **6**(1), 1.

Roujeau, J. C. & Stern, R. S. (1994), 'Severe adverse cutaneous reactions to drugs', *New England Journal of Medicine* **331**(19), 1272–1285.

Scheiber, J., Chen, B., Milik, M., Sukuru, S. C. K., Bender, A., Mikhailov, D., Whitebread, S., Hamon, J., Azzaoui, K. & Urban, L. (2009), 'Gaining insight into off-target mediated effects of drug candidates with a comprehensive systems chemical biology analysis', *Journal of Chemical Information and Modeling* **49**(2), 308–317.

Scheiber, J., Jenkins, J. L., Sukuru, S. C. K., Bender, A., Mikhailov, D., Milik, M., Azzaoui, K., Whitebread, S., Hamon, J. & Urban, L. (2009), 'Mapping adverse drug reactions in chemical space', *Journal of Medicinal Chemistry* **52**(9), 3103–3107.

Sehgal, V. & Gangwaani, O. (1986), 'Hydralazine-induced fixed drug eruption', *International Journal of Dermatology* **25**(6), 394–394.

Shiohara, T. (2009), 'Fixed drug eruption: pathogenesis and diagnostic tests', *Current Opinion in Allergy and Clinical Immunology* **9**(4), 316–321.

Steady Health Comunity (2019), 'Steadyhealth.com'. [Online; accessed 11-May-2019].
URL: *http://www.steadyhealth.com*

Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. (2003), 'The chemistry development kit (cdk): An open-source java library for chemo-and bioinformatics', *Journal of Chemical Information and Computer Sciences* **43**(2), 493–500.

Strandell, J., Bate, A., Lindquist, M., Edwards, I. R. & Swedish, Finnish, I. X.-r. d.-d. i. d. t. S. g. (2008), 'Drug–drug interactions–a preventable patient safety issue?', *British Journal of Clinical Pharmacology* **65**(1), 144–146.

Tan, S. (2005), 'Neighbor-weighted k-nearest neighbor for unbalanced text corpus', *Expert Systems with Applications* **28**(4), 667–671.

Tari, L., Anwar, S., Liang, S., Cai, J. & Baral, C. (2010), 'Discovering drug–drug interactions: a text-mining and reasoning approach based on properties of drug metabolism', *Bioinformatics* **26**(18), i547–i553.

Tatonetti, N. P., Denny, J., Murphy, S., Fernald, G., Krishnan, G., Castro, V., Yue, P., Tsau, P., Kohane, I. & Roden, D. (2011), 'Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels', *Clinical Pharmacology & Therapeutics* **90**(1), 133–142.

Tatonetti, N. P., Fernald, G. H. & Altman, R. B. (2011), 'A novel signal detection algorithm for identifying hidden drug-drug interactions in adverse event reports', *Journal of the American Medical Informatics Association* **19**(1), 79–85.

Tatonetti, N. & Patrick, Y. (2012), 'Data-driven prediction of drug effects and interactions', *Science Translational Medicine* **4**(125), 125ra31–125ra31.

Thakrar, B. T., Grundschober, S. B. & Doessegger, L. (2007), 'Detecting signals of drug–drug interactions in a spontaneous reports database', *British Journal of Clinical Pharmacology* **64**(4), 489–495.

Tomasi, C. (2017), 'Random forest classifiers'.

Triaridis, S., Tsiropoulos, G., Rachovitsas, D., Psillas, G. & Vital, V. (2009), 'Spontaneous haematoma of the pharynx due to a rare drug interaction', *Hippokratia* **13**(3), 175.

Ursu, O. & Holmes, J. (2016), 'Drugcentral: online drug compendium', *Nucleic Acids Research* p. gkw993.

Ursu, O. & Holmes, J. (2017), 'Drugcentral: online drug compendium', *Nucleic Acids Research* **45**(Database issue), D932–D939.

Van, T. & Nabuurs, S. (2011), 'Gaussian interaction profile kernels for predicting drug–target interaction', *Bioinformatics* **27**(21), 3036–3043.

Vidhya, K. & Aghila, G. (2010), 'Text mining process, techniques and tools: an overview', *International Journal of Information Technology and Knowledge Management* **2**(2), 613–622.

Vilar, S. & Harpaz, R. (2012), 'Drug—drug interaction through molecular structure similarity analysis', *Journal of the American Medical Informatics Association* **19**(6), 1066–1074.

Vilar, S. & Uriarte, E. (2013), 'Detection of drug-drug interactions by modeling interaction profile fingerprints', *PLoS ONE* **8**(3), e58321.

Wassermann, A. M., Geppert, H. & Bajorath, J. (2009), 'Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects', *Journal of Chemical Information and Modeling* **49**(10), 2155–2167.

White, R. W., Tatonetti, N. P., Shah, N. H., Altman, R. B. & Horvitz, E. (2013), 'Web-scale pharmacovigilance: listening to signals from the crowd', *Journal of the American Medical Informatics Association* **20**(3), 404–408.

White, R. W., Wang, S., Pant, A., Harpaz, R., Shukla, P., Sun, W., DuMouchel, W. & Horvitz, E. (2016), 'Early identification of adverse drug reactions from search log data', *Journal of Biomedical Informatics* **59**, 42–48.

Whitebread, S., Hamon, J., Bojanic, D. & Urban, L. (2005), 'Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development', *Drug Discovery Today* **10**(21), 1421–1433.

Wiffen, P. (2016), 'Litt's drug eruption and reaction manual 22nd edition'.

Wikipedia (2018, April 1*a*), 'Allopurinol-wikipedia'.
URL: *https://en.wikipedia.org/wiki/Allopurinol*

Wikipedia (2018, April 1*b*), 'Drug eruption-wikipedia'.
URL: *https://en.wikipedia.org/wiki/Drug_eruption*

Wikipedia contributors (2018), 'Drugs.com — Wikipedia, the free encyclopedia'. [Online; accessed 15-May-2018].
URL: *https://en.wikipedia.org/w/index.php?title=Drugs.com&oldid=839918274*

Wishart, D. & Feunang, Y. (2017), 'Drugbank 5.0: a major update to the drugbank database for 2018', *Nucleic Acids Research* **46**(D1), D1074–D1082.

Wu, Z., Wang, Y. & Chen, L. (2013), 'Network-based drug repositioning', *Molecular BioSystems* **9**(6), 1268–1281.

Yamanishi, Y., Kotera, M., Kanehisa, M. & Goto, S. (2010), 'Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework', *Bioinformatics* **26**(12), i246–i254.

Yamanishi, Y., Kotera, M., Moriya, Y., Sawada, R., Kanehisa, M. & Goto, S. (2014), 'Dinies: drug–target interaction network inference engine based on supervised analysis', *Nucleic Acids Research* **42**(W1), W39–W45.

Yamanishi, Y. & Pauwels, E. (2012), 'Drug side-effect prediction based on the integration of chemical and biological spaces', *Journal of Chemical Information and Modeling* **52**(12), 3284–3292.

Yang, C. & Jiang, L. (2012), Detecting signals of adverse drug reactions from health consumer contributed content in social media, *in* 'Proceedings of ACM SIGKDD Workshop on Health Informatics'.

Yang, H. & Yang, C. C. (2016), Discovering drug-drug interactions and associated adverse drug reactions with triad prediction in heterogeneous healthcare networks, *in* '2016 IEEE International Conference on Healthcare Informatics', IEEE, pp. 244–254.

Yeleswarapu, S., Rao, A., Joseph, T., Saipradeep, V. G. & Srinivasan, R. (2014), 'A pipeline to extract drug-adverse event pairs from multiple data sources', *BMC Medical Informatics and Decision Making* **14**(1), 1.

Zhang, L. & Huang, S. (2009), 'Predicting drug–drug interactions: an fda perspective', *The AAPS Journal* **11**(2), 300–306.

Zhang, P., Wang, F., Hu, J. & Sorrentino, R. (2015), 'Label propagation prediction of drug-drug interactions based on clinical side effects', *Scientific Reports* **5**, 12339.

Zhang, W. & Chen, Y. (2016), Drug side effect prediction through linear neighborhoods and multiple data source integration, *in* '2016 IEEE International Conference on Bioinformatics and Biomedicine', IEEE, pp. 427–434.

Zhang, W. & Chen, Y. (2017), 'Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data', *BMC Bioinformatics* **18**(1), 18.

Zhang, W., Liu, F., Luo, L. & Zhang, J. (2015), 'Predicting drug side effects by multi-label learning and ensemble learning', *BMC Bioinformatics* **16**(1), 365.

Zhang, W. & Zou, H. (2016), 'Predicting potential side effects of drugs by recommender methods and ensemble learning', *Neurocomputing* **173**, 979–987.

Zheng, M., Liu, X., Xu, Y., Li, H., Luo, C. & Jiang, H. (2013), 'Computational methods for drug design and discovery: focus on china', *Trends in Pharmacological Sciences* **34**(10), 549–559.

Zheng, Y. & Ghosh, S. (2017), An optimized drug similarity framework for side-effect prediction, *in* 'Computing in Cardiology', IEEE, pp. 1–4.

Zheng, Y., Lan, C., Peng, H. & Li, J. (2016), Using constrained information entropy to detect rare adverse drug reactions from medical forums, *in* '38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society', IEEE, pp. 2460–2463.

Zheng, Y., Peng, H., Zhang, X., Gao, X. & Li, J. (2018), Predicting drug targets from heterogeneous spaces using anchor graph hashing and ensemble learning, *in* 'International Joint Conference on Neural Networks', IEEE, pp. 1–7.

Zheng, Y., Sun, C., Zhu, C., Lan, X., Fu, X. & Han, W. (2015), Lwcs: a large-scale web page classification system based on anchor graph hashing, *in* '2015 6th IEEE International Conference on Software Engineering and Service Science', IEEE, pp. 90–94.

Zhou, X., Peng, Y. & Liu, B. (2010), 'Text mining for traditional chinese medical knowledge discovery: a survey', *Journal of Biomedical Informatics* **43**(4), 650–660.