Briefings in Bioinformatics, 22(2), 2021, 1604–1619

doi: 10.1093/bib/bbz176
Advance Access Publication Date: 10 February 2020
Review Article

Biomedical data and computational models for drug repositioning: a comprehensive review

Huimin Luo, Min Li, Mengyun Yang, Fang-Xiang Wu, Yaohang Li and Jianxin Wang

Corresponding author: Jianxin Wang, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China. Tel.: +86-731-88820212; Fax:+86-731-88877936; E-mail: jxwang@mail.csu.edu.cn

Abstract

Drug repositioning can drastically decrease the cost and duration taken by traditional drug research and development while avoiding the occurrence of unforeseen adverse events. With the rapid advancement of high-throughput technologies and the explosion of various biological data and medical data, computational drug repositioning methods have been appealing and powerful techniques to systematically identify potential drug-target interactions and drug-disease interactions. In this review, we first summarize the available biomedical data and public databases related to drugs, diseases and targets. Then, we discuss existing drug repositioning approaches and group them based on their underlying computational models consisting of classical machine learning, network propagation, matrix factorization and completion, and deep learning based models. We also comprehensively analyze common standard data sets and evaluation metrics used in drug repositioning, and give a brief comparison of various prediction methods on the gold standard data sets. Finally, we conclude our review with a brief discussion on challenges in computational drug repositioning, which includes the problem of reducing the noise and incompleteness of biomedical data, the ensemble of various computation drug repositioning methods, the importance of designing reliable negative samples selection methods, new techniques dealing with the data sparseness problem, the construction of large-scale and comprehensive benchmark data sets and the analysis and explanation of the underlying mechanisms of predicted interactions.

Key words: drug repositioning; drug-target prediction; drug-disease prediction; computational model; data integration; evaluation metric

Huimin Luo is a PhD student in the School of Computer Science and Engineering at Central South University, Hunan, China and with School of Computer and Information Engineering, Henan University, KaiFeng, 475001, China. Her research interests include bioinformatics and computational drug repositioning.
Min Li is a Professor in the School of Computer Science and Engineering at Central South University, Hunan, China. Her research interests include bioinformatics and systems biology.
Mengyun Yang is a PhD student in the School of Computer Science and Engineering at Central South University, Hunan, China. His research interests include bioinformatics and computational drug repositioning.
Fang-Xiang Wu is a Professor in the College of Engineering and the Department of Computer Science at University of Saskatchewan, Saskatoon, Canada. His current research interests include bioinformatics and artificial intelligence.
Yaohang Li is an Associate Professor in the Department of Computer Science at Old Dominion University, Norfolk, USA. His current research interests are in computational biology, Monte Carlo methods, big data analysis and parallel/distributed/grid computing.
Jianxin Wang is a Professor in the School of Computer Science and Engineering at Central South University, Hunan, China. His research interests include computational genomics and proteomics.
Submitted: 23 October 2019; **Received (in revised form):** 7 December 2019

Introduction

Although the investment in the pharmaceutical research and development has increased dramatically over past decades, the number of new drugs approved for the market remains low. Recent advances in genomics, life sciences and technologies have not sped up the drug discovery process. Indeed, ~90% of drug candidates fail during their phase 1 clinical trials and it usually takes billions of dollars and 10–15 years to successfully bring a new drug to the market [1]. Therefore, it is urgent and important to explore useful methods to improve the success rate of the drug research and development. Recently, drug repositioning has become a promising field in the drug discovery and attracted increasing interests from both the pharmaceutical industry and research community [2]. Drug repositioning aims to identify new therapeutic opportunities for existing drugs, which can reduce the time, costs and risk of traditional drug development, shorten the period of drug approval and launch [3, 4]. Successful examples of drug repositioning include sildenafil, thalidomide and retinoic acid. These success stories of drug repositioning further inspired global pharmaceutical industries to explore the potential capacity of existing drug space. In past 10 years, governments, academic researchers and pharmaceutical companies have launched large-scale funding and activities to support drug repositioning-related studies. The National Centre for Advancing Translational Sciences has launched the Discovering New Therapeutic Uses for Existing Molecules project [5]. The Canadian Institutes of Health Research has established funding to support the basic science research for drug repurposing [6].

In order to find new indications for existing drugs, two strategies can be adopted by pharmaceutical industry generally, named activity-based drug repositioning and computational drug repositioning [3, 7]. As an experimental strategy, activity-based drug repositioning refers to testing drugs in assays based on some available comprehensive clinical compound databases. By performing target- or cell-based screens for thousands of medications, the potential indications of drugs can be detected directly. Though structural information of targets or compounds is not needed, the activity-based method is still a time- and labor-consuming process due to the requirement of the entire collection of existing drugs, specialized equipment and screening assays [8]. Besides, the underlying molecular mechanisms are often not clear for many cases, which make it difficult to reposition drugs in a large-scale manner. Unlike the activity-based method, the computational drug repositioning utilizes some online publicly available databases and bioinformatics tools to detect interactions among drugs, targets and diseases. It allows a much faster repositioning process at a reduced cost, and most pharmaceutical companies have already adopted such methods for drug discovery recently [7]. Basically, the rapid development and success achieved in the algorithms for computational drug repositioning can be attributed to the following two aspects. First, large amount of high-throughput data related to drugs at various levels has been rapidly accumulated, such as genomic data, protein structures and phenotypes. The second aspect is the progress made in computer sciences which is the foundation for developing efficient repositioning algorithms [9, 10]. Computational methods are expected to effectively reposition drugs against various targets and diseases.

In general, the traditional computational methods for identifying drug-target interactions mainly include ligand-based approaches and structure-based approaches [11–13]. The ligand-based approaches predict drugs interacting with a target protein by comparing the candidate ligands with the known ones that can bind to the given target protein. Therefore, ligand-based approaches may not perform well when there are no or few known ligands for the target protein. The structure-based approaches use the docking simulation techniques to identify potential drug–target interactions (DTIs) based on the known three-dimensional (3D) structures of targets. This kind of approaches is computationally time-consuming, depends on the reliability of the docking simulation methods and cannot be applied to targets without 3D structure information.

Machine learning techniques are good complementary to the ligand-based and structure-based approaches by exploring the global patterns of drug-target and drug-disease interactions, which have been popularly used to develop effective approaches for drug-target and drug-disease prediction recently. The machine learning-based approaches can utilize and integrate heterogeneous multisource biomedical data on drugs, targets and diseases from studies in drug discovery to systematically identify potential drug-target and drug-disease interactions. Recently, with the growth of biomedical data, various computational drug repositioning methods based on machine learning techniques have been proposed and applied successfully.

Computational drug repositioning methods based on machine learning utilizes publicly available databases and bioinformatics tools to systematically identify interactions among drugs, diseases and genes or proteins. The workflow of computational drug repositioning is shown in Figure 1. First, researchers need to collect available data from various types of publicly available biomedical data sources related to drugs, targets and diseases which could provide effective information for identifying potential drug-target and drug-disease interactions. Next, benefiting from the progress of computer sciences, various models have been developed to identify potential drug-target or drug-disease interactions. For instance, Napolitano *et al.* [14] utilized known drug-related data including gene expression, chemical structure and target information to predict therapeutic classes. These data were used to compute three individual kernels. Then, a joint kernel was defined by integrating these individual kernels and used as a kernel in support vector machine (SVM) classification. Li *et al.* [15] developed a novel similarity-based method to identify new indications of an existing drug through its similar drugs. They assumed that similar drugs had common indications. The similarity between two drugs was calculated based on drug chemical structures and drug target information. Finally, to evaluate and validate the performance of the proposed method, it should be compared with the available state-of-the-art methods under different metrics and prediction tasks on the gold standard data sets and other test data sets.

The knowledge about related biomedical data sources is essential for developing novel drug repositioning tools. It remains an enormous challenge to understand the overall relationships among various biomedical entities due to the heterogeneity and incompleteness of available data [16]. Though many researchers have reviewed some available data resources related with repositioning, a comprehensive description about important data entities or their interactions about current popular data sources is still lacking, which impedes researchers to obtain required biomedical data rapidly and conveniently for developing efficient drug repositioning algorithms. More importantly, although some methods have been proven to be successful in addressing the problem of drug repositioning, there are still outstanding challenges which need to be addressed. First, the known or validated interaction data are

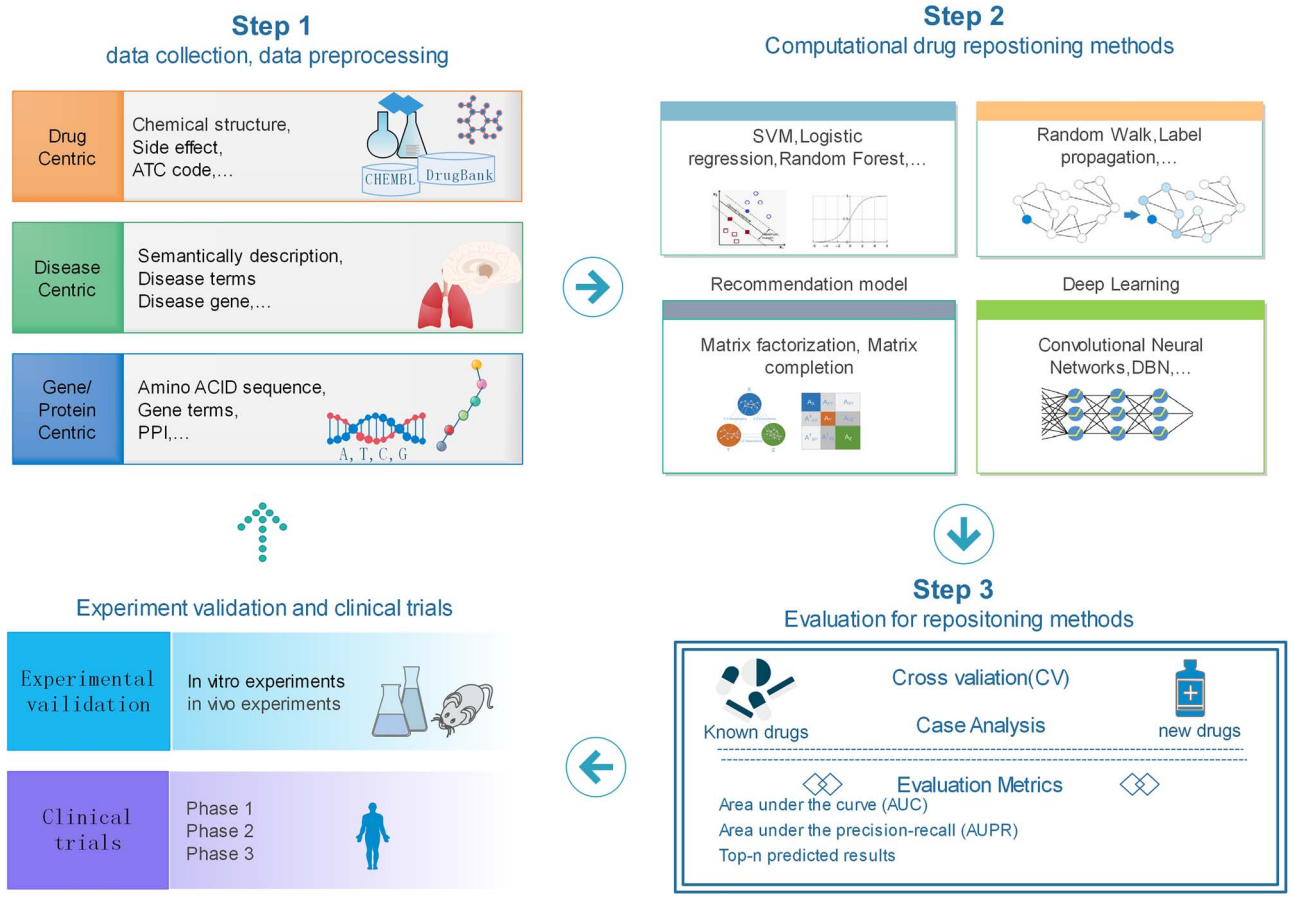

Figure 1. The workflow of computational drug repositioning.

still insufficient. Second, making the unknown interactions as true negatives or ignoring them in the training process can potentially degrade the predictive power of computational model for drug repurposing [17]. Finally, gold standard data sets and evaluation experiments are generally not consistent for analyzing and evaluating drug repositioning methods.

The advances in computational drug repositioning have been reviewed in detail from different aspects in recent years [1, 18–24]. For example, Ding *et al.* [18] provided a review of various similarity-based machine learning methods for DTIs prediction. Zhu *et al.* [19] reviewed existing drug knowledge bases and their applications in various biomedical studies, including drug repositioning. The review by Ezzat *et al.* [22] provided a comprehensive overview and an empirical comparison of computational DTI prediction methods. Compared with previous reviews, this study provided a more comprehensive and integrative analysis of machine learning-based prediction methods and related biomedical data used in drug repositioning. We begin with discussing various biomedical data sources which enable researchers to not only understand and discover new strategies but also validate their results as a part of the study. Second, a comprehensive overview is given on popular drug repositioning approaches based on different machine learning models. Finally, standard data sets and evaluation metrics used in drug repositioning are analyzed comprehensively, and a brief comparison of various methods on the same gold standard data sets is given to evaluate their advantages and limitations.

In the recent review by Zhou *et al.* [25], relevant data repositories and different computational models for DTI identification have been discussed in detail. In this review, we focus on analyzing and classifying existing computational methods for potential drug-target or drug-disease interaction prediction from a machine learning perspective. Although the methods covered in our survey have certain overlaps with [25], we complement it with computational methods based on matrix completion, which have recently demonstrated attractive prediction precision enhancement in drug repositioning. We further discuss the advantages and disadvantages of various prediction models and related issues, such as cold-start, sparsity and noisy data problems. Moreover, our survey extends various types of data related to different biomedical entities, including drugs, diseases, genes and proteins. The common validation strategies, evaluation metrics and benchmark data sets used in drug repositioning studies are also included to guide researchers to efficiently evaluate and verify the predictive ability of their developed methods in future studies.

Available biomedical data

Recent technological advances in the high-throughput biology have generated huge amounts of multi-omic data, such as genomic, proteomic and metabolomic data, and other data from pharmacogenomic, clinical and chemical sources. Most of these data are stored in databases, which are publicly available for further study and analysis. To some extent, they have created unprecedented opportunities for developing efficient drug repositioning methods and tools. For computational drug repositioning, researchers often need to collect or fuse

data from multiple data sources and construct biomedical network containing more hidden information. However, one of major difficulties is how to collect and analyze required biomedical data because they are heterogeneous, and the data generated from different experiments include different types of information such as nucleotide sequences and protein–protein interactions. Moreover, data sets provided by individual groups or research institutes often show inconsistency and disconnection in term of entity annotations and data formats, which cannot be used to link them directly to other databases. As a result, a comprehensive understanding for these available databases can alleviate the issues caused by entity identification and data inconsistencies. Moreover, it can also help improve the accuracy of data, and speed the subsequent data analysis process. The aim of this section is to review various types of data related to different biomedical entities, including drugs, diseases, genes and proteins, which can help extract biomedical information from different sources. A list of those frequently used data sources [26–59] are summarized in Supplementary Table 1, where the brief descriptions and the access links are provided. These popular public data sources can be classified into four groups based on the biomedical entity concerned: drug-centric, disease-centric, gene/protein-centric and integrated databases.

Drug centric data

The past decades have witnessed a steady progress in pharmaceutical industry and academic drug discovery efforts, such as compound synthesis and assays, chemical biology or genomics, which have led to valuable reports about new compounds and their corresponding biological activities. In some popular medicinal chemistry journals, about 20 000–30 000 new compounds are published per year, and the rate has been accelerating in recent years [60]. On this basis, some research organizations collected the available information about drugs or drug-like compounds and their interactions with the human biological system, called drug-centric data, and constructed many valuable databases. As shown in Table 1, some drug databases are freely available and popularly used, which can be utilized to build knowledge-based models for enabling us to conduct analysis and prediction. As one can be, there exist significant differences for these 11 drug-centric databases described in Table 1, which is caused by different biomedical data sources and data types. It is essential for researchers to understand how to characterize and organize drugs or drug-like compounds in each drug-centric database. We have visited these databases, browsed and analyzed their contents step-by-step. Here, we aim to help researchers to understand and take advantage of these drug-centric databases by discussing some popular resources used for drug repositioning. It provides an overview of the main contents in Table 1.

Generally, properties of a drug or a drug-like compound are described or quantified from the following aspects. First, a drug can be identified and quantified based on its physical and chemical properties. For example, some databases including Drugbank and SuperTarget provide molecular structures, which can be used by some computational chemical similarity approaches or toolkits, such as CDK [61] and SIMCOMP [62]. Second, molecular activities and phenotypes are very effective to profile-related drugs. Drugs or drug-like compounds exert their biological activities through binding to the cellular proteins. Some researchers utilize the interactive profile information between drugs and targets or proteins to compute the similarity between drugs [63]. Diseases and side effects can also be used to quantify and characterize drugs or drug-like compounds [64]. Drug–drug interaction may occur when two or more drugs are taken together, and can reflect the similarity between drugs [65].

For drug repositioning purpose, we can choose some popular databases, such as Drugbank, to obtain basic biological data about molecular structures or interactive profiles. Sometimes, researchers want to integrate data from multiple drug-centric databases to obtain a comprehensive view of drugs or drug-like compounds and their interaction information. In this case, several practical problems need to be addressed. First, there often exist many different names and identifiers for a single drug in drug-centric databases. Thus, some mapping mistakes or data missing may occur during the integration process. Therefore, a standardized naming or mapping rule for the data integration is needed in order to obtain better performance and avoid inconsistence of integrated experimental data sets. Fortunately, some researchers have started to focus on this problem and presented initial solutions, such as RxNorm [66]. Second, differences in terms of resource provision modes are also a challenge problem for understanding and integrating drug-centric databases. For example, formats for chemical structures are different, which can be represented by simplified molecular input line entry system (SMILES), MOL, SDF, PDB or other formats. Currently, for this issue, researchers can utilize some open software systems to convert one file format to the other format. In the future, all drug-centric databases are expected to provide a standard, uniform format of molecular structures for data integration. Third, some databases do not provide the downloadable whole data set and researchers can only search related data to construct their data sets. Likewise, differences in the content of drug-centric databases and the inconsistency of the data organizing process lead to the deviation of gold standard data sets used for drug repositioning.

Disease centric data

Disease-centric databases enable researchers to access a wide variety of disease-related data and play a central role in bioinformatics. Over last decades, many disease-related molecular genetics, phenotypes and drugs have been studied while disease-associated databases have been developed to understand the nature of diseases. For instance, Online Mendelian Inheritance in Man (OMIM) database is the main repository of genetic information for disorders.

In previous studies, diseases have been classified according to different criteria, such as pathology, anatomy, prognosis, molecular genetics and drugs [67]. Here, we aim to focus on some most relevant features in popular disease-centric databases to help researchers understand and use these databases. By browsing and analyzing the information in these databases, disease-related data are grouped into four valuable categories in terms of drug repositioning applications, which are shown in Table 2.

The semantic structure is often utilized to compute the similarity between diseases, which is an important factor in exploring molecular mechanisms of diseases. Besides, disease terms usually present some key information about the phenotype of a disease, and some disease-centric databases provide clear definition about human diseases. Researchers generally obtain the structure information or disease terms from disease databases UMLS, DO, Mesh, HPO, etc. In addition, associated genes with specific diseases can be collected from OMIM, DisGeNET, etc. The genetic information can help researchers to understand diseases at molecular level. Comparatively speaking, the drug

Table 1. Public drug-centric databases

Data source	Molecular structure	Molecular activity		Therapeutic & phenotype			Drug–drug interactions
		Target	Pathway	ATC code	Indications	Side effect	
BindingDB [26]	√	√					
CHEMBL [27]	√	√			√		
DrugBank [28]	√	√	√	√	√	√	√
DrugMap [29]		√	√	√			
Drugs.com [30]						√	√
Offsides [31]					√	√	
PROMISCUOUS [32]		√	√	√		√	
PubChem [33]	√			√	√	√	
SIDER [34]					√	√	
STITCH [35]	√						
SuperTarget [36]	√	√	√	√		√	

Table 2. Public disease-centric databases

Data source	Disease phenotypes	Disease terms	Associated genes	Drug information
Disease ontology [37]	√	√		
DISEASE [38]			√	
DiseaseConnect [39]	√		√	√
DisGeNET [40]			√	
eDGAR [41]		√	√	
GAD [42]			√	
HPO [43]	√	√		
ICD [44]	√	√		
MalaCards [45]	√	√	√	√
Mesh [46]	√	√		
OMIM [47]		√	√	
UMLS [48]	√	√		

information associated with diseases is not provided by these databases except for MalaCards and DiseaseConnect.

Generally, these databases are valuable resources for researchers to conduct studies. The quality of disease annotations or ontology structures has great impact on the accuracy of drug repositioning. Moreover, the heterogeneity of terms and the incompleteness of contents generally exist in these sources, which significantly hinder the data sharing and exchanging. DO database plays an important role in the standardization of human disease annotations in biomedical databases. DO database has semantically integrated multiple diseases and medical vocabularies and provided cross-references between these vocabularies. Therefore, many researchers always unify disease terms to DO. We expected the further development of the disease ontology in order to associate genetic and genomic data with human diseases.

Gene/protein centric data

Genes and proteins provide the basis of the rational drug design. Recently, many gene- and protein-centric databases have been successfully developed, which are shown in Table 3. Generally, amino acid sequences, gene terms and protein–protein interactions are often contained in gene-/protein-centric databases. An amino acid sequence refers to a string of amino acids in a particular order constituting the protein molecules. For drug repositioning, UniProtKB and HPRD databases present amino acid sequence information about proteins, which can be utilized to calculate the similarity between targets or represent target features. Likewise, gene terms and protein–protein interactions can also be used to identify potential targets for drugs. Moreover, some researchers construct target networks using protein–protein interactions to find potential drug-target pairs. The 3D structure of a protein also could provide important information for understanding its role in disease.

Table 3. Public gene-/protein-centric databases

Data source	Sequence	Gene terms	PPI	3D Structure
BioGrid [49]		√	√	
Gene ontology [50]		√		
HPRD [51]	√		√	
PDB [52]	√			√
STRING [53]			√	
UniProtKB [54]	√	√	√	√

Integrated data

The studies described above mainly collect related data of one type of the biomedical entity. Nevertheless, single data source may be incomplete or limited. Therefore, it is extremely important to integrate various biomedical data from multiple sources in practice. Some studies show that data integration can help improve the performance of drug repositioning. For example, Wang *et al.* [68] calculated drug similarities on the basis of chemical structures, target proteins and side-effects from multiple

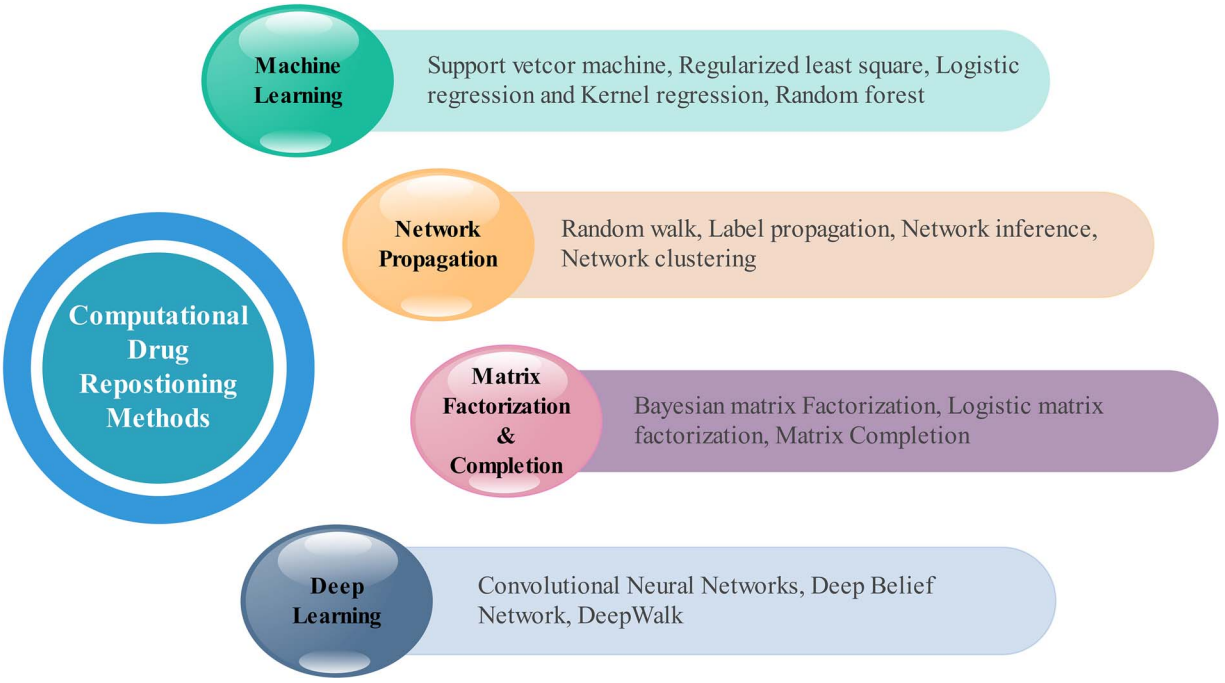

Figure 2. Computational repositioning methods based on various models.

sources, respectively. In most cases, however, researchers need to browse required data from multiple databases and integrate them manually, which slows down the rapid development of drug repositioning approaches. Integrated databases can aid the research of drug repositioning effectively. Some integrated databases containing the information of drugs, diseases, genes, pathways and various interactions are constructed gradually, which are shown in Supplementary Table 1.

In recent studies, most drug repositioning methods are developed by utilizing the similarity or feature data of drugs, diseases and targets. Based on the above data analysis, multiple data which describe drugs, diseases and targets from different aspects and levels can be utilized to measure similarities or construct features. The similarity and feature information of biomedical entities can be further incorporated into different prediction models to identify novel drug-related associations.

Computational drug repositioning methods

Various machine learning technologies have been utilized to develop effective prediction approaches in drug repositioning. According to the machine learning models applied to computational drug repositioning methods, these drug repositioning methods are classified into four main groups: classical machine learning model; network propagation; matrix factorization and completion model; and deep learning, which are shown in Figure 2.

Classical machine learning model-based methods

Recently, the quick development of machine learning techniques provides effective and efficient ways to perform drug repositioning, including DTI and drug–disease interaction prediction. These methods are motivated by the observation that similar drugs tend to target similar proteins and vice versa [69], or similar drugs tend to target similar diseases and vice versa [70]. An intuitive idea is to formulate drug repositioning as a binary classification problem, where the drug-target and drug-disease pairs are treated as instances; the information of drugs, diseases and targets is treated as features. Then, classical classification models can be used, e.g. SVMs, regularized least square (RLS), logistic regression and random forest to develop the drug repositioning methods.

Support vector machine

Bleakley *et al.* [71] proposed a novel supervised inference method, bipartite local model (BLM) which used SVMs as local classifiers, to predict novel DTIs. For a candidate drug-target pair, two independent predictions including predicting the target proteins of the given drug and predicting the drugs targeting the given protein can be combined to give a definitive prediction for each interaction. BLM has achieved better performance on four gold standard data sets involving enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors in human. However, BLM has limitations as it is not able to provide a reasonable prediction for new drug/target candidates without any interactions, which is referred to as the cold-start problem.

Considering that many types of data sources can be integrated efficiently to improve the prediction performance, Wang *et al.* [68] proposed a new method for drug repositioning, PreDR (Predict Drug Repositioning), by integrating molecular structure, molecular activity and phenotype data used to define drug and disease similarity. Specially, drugs and diseases were represented by their similarity profiles, and a kernel function was constructed to calculate the similarity between drug-disease pairs. Then, an SVM with the defined kernel was trained to find novel drug–disease interactions. By utilizing data fusion technology, Wang *et al.* [72] collected drug pharmacological and therapeutic effects, drug chemical structures and proteomic information to measure drug similarity and target similarity, and integrated these similarities by an efficient kernel function within a SVM-based predictor.

Most existing machine learning approaches used experimentally validated interactions as positive samples and negative sampling from unvalidated interactions to train prediction models [73]. However, the randomly generated negative samples may include real positive samples not yet known, which may lead to a biased decision boundary [74]. Therefore, it is important to screen highly reliable negative (RN) samples for supervised learning-based drug repositioning methods. Motivated by such challenge, Liu *et al.* [73] proposed a systematic screening workflow to identify highly RN samples. The screening framework is developed based on the assumption that proteins dissimilar to any known/predicted targets of a given compound are not much likely to be targeted by the compound and vice versa. Extensive experiments were conducted and the results showed that the screened negative samples could effectively improve the accuracy of prediction methods.

Difficulties of the DTI identification include the sparsity of known DTIs and no experimentally verified negative samples. Lan *et al.* [75] presented a DTI prediction method called PUDT. They set known interactions as positive samples and unknown interactions as unlabeled samples. Based on target similarity, the unlabeled samples were grouped into RN samples and likely negative samples using three strategies, including random walk with restarts, KNN and heat kernel diffusion. Then, the final label of unlabeled samples was determined by aggregating the results of three strategies. Finally, a multilevel classifier using weighted SVM was developed to identify potential DTIs.

Peng *et al.* [76] developed a negative sample extraction method to identify positive, negative and ambiguous DTI samples from unknown DTIs based on positive-unlabeled learning. Then, the probabilities of ambiguous samples belonging to the positive and negative classes were computed. By integrating the screened samples, an SVM-based optimization model was constructed to predict new DTIs.

Regularized least square

Based on the Laplacian RLS method, Xia *et al.* [77] proposed a semi-supervised learning approach called NetLapRLS, which integrated information from chemical space, genomic space and drug–protein interaction network space to predict drug–protein interactions. The standard LapRLS was improved by incorporating validated drug-protein interaction information.

Van Laarhoven *et al.* [78] proposed a simple RLS algorithm by incorporating a product of kernels constructed from DTI profiles. By combining the constructed Gaussian interaction profile (GIP) kernels with chemical and genomic information, the RLS-Kron algorithm that combined a kernel on drugs and a kernel on targets using the Kronecker product was introduced and achieved a good performance. Nascimento *et al.* [79] extended KronRLS method by proposing a new prediction algorithm, KronRLS-MKL, which utilized the multiple kernels learning algorithm to automatically select and combine kernels on a bipartite drug-protein prediction problem. A kernel can be seen as a similarity matrix estimated on all pairs of instances. Laarhoven *et al.* [80] integrated a simple weighted nearest neighbor procedure into the RLS-Kron classifier for the prediction of DTIs. Specifically, the weighted nearest neighbor procedure defined a profile for a new drug compound by using interaction profiles and similarity information of all drugs in the training set.

Many kernel-based prediction methods constructed kernel matrix by combining various kernels linearly, which is not appropriate when kernels exists nonlinear relationships. Thus, Hao *et al.* [81] proposed a DTI prediction method RLS-KF by integrating RLS with the nonlinear kernel fusion algorithm. The similarity matrices were combined with Gaussian kernel matrices by adopting kernel fusion technology. Li *et al.* [82] proposed a novel flexible and robust multiple sources learning (FRMSL) framework to integrate distinct sources of biological data to measure drug similarity and disease similarity and solve the prediction problem by utilizing the Kronecker RLS (KronRLS) method.

Logistic regression and kernel regression

Perlman *et al.* [63] incorporated multiple drug-drug and gene-gene similarity measures to develop a similarity-based DTIs inference framework. Based on a given pair of drug-drug and gene-gene similarity measures, association score was calculated for each drug-target pair. Then, a logistic regression classifier that integrated the scores of multiple measures was trained to predict novel DTIs.

Based on the observation that similar drugs are indicated for similar diseases, Gottlieb *et al.* [83] proposed a computational method PREDICT, which constructed multiple drug-drug and disease-disease similarity measures, followed the method in [63] to construct classification features and subsequently learned a logistic regression classifier to predict novel associations between drugs and diseases.

A kernel regression-based method was proposed by Yamanishi *et al.* [84] to predict novel DTIs. First, drugs and targets were embedded into the unified pharmacological space based on the drug-target bipartite graph. Then, the kernel regression model representing the correlation between pharmacological space and chemical/genomic space was learned to infer pharmacological features of new compound/protein. The closeness between compounds and proteins was measured based on their pharmacological features. By investigating relationships among the chemical space, the pharmacological space and the topology of DTI networks, Yamanishi *et al.* [85] found that the pharmacological similarity is more useful than the chemical structure similarity for predicting DTIs. Therefore, they proposed to infer DTIs by applying the learning method proposed in [84] which used the pharmacological space and the genomic space.

Random forest

Olayan *et al.* [86] developed an efficient drug-target prediction computational method called DDR. Multiple similarity measures for drugs and targets were computed by using various drug- and target-related data. Then, these similarity measures were selected and fused to construct the composite similarity. By integrating drug similarities, target similarities and known DTIs, a DTI heterogeneous graph was constructed. DDR derived graph-based features from the constructed DTI heterogeneous graph and trained the random forest model to predict DTIs. Cao *et al.* [87] proposed a DTI prediction method by combining the information from chemical, biological and DTI network. One drug-target pair was represented by three types of features, including drug's chemical information, target's structure information and interaction information. Then, the DTI prediction method was developed using a random forest model with integrated features.

Network propagation-based methods

Besides learning a classifier to predict DTIs and drug–disease interactions, network-based approaches have been widely used

[88]. It aims at organizing relationships among biological and biomedical entities using networks to identify potential novel relationships at the network level [16]. The network-based analysis has become a widely used strategy for computational drug repositioning. In recent years, numerous network-based approaches have been proposed and become a popular tool for drug repositioning.

Random walk

Chen *et al.* [89] assumed that similar drugs interact with similar targets and, thus, proposed a Network-based Random Walk with Restart on the Heterogeneous network (NRWRH) method. Based on the known DTIs, NRWRH constructed a new drug similarity matrix and a new target similarity matrix. The drug chemical similarity matrix and the new drug similarity matrix were combined into one integrated drug integrated similarity matrix, and the target protein sequence similarity matrix and the new target similarity matrix were combined into one integrated target similarity matrix. Then, a drug-target heterogeneous network was constructed by integrating the two integrated similarity networks and the known DTI data. Finally, the random walk algorithm was implemented on the constructed heterogeneous network to predict the interaction probabilities between drugs and targets. Liu *et al.* [74] proposed the two-pass random walks with restart on a drug-disease heterogeneous network, TP-NRWRH, to identify potential drug-disease associations. Firstly, the heterogeneous network consisting of a drug-drug similarity network, a disease-disease similarity network and a known drug-disease association network was constructed. Then, they extended NRWRH by implementing two-pass random walks with restart including drug-centric and disease-centric random walk to determine the interaction likelihood values for unknown drug-disease pairs.

Luo *et al.* [90] exploited known drug-disease associations to improve the drug-drug and disease-disease similarity measures and then constructed drug and disease similarity networks to build a drug-disease heterogeneous network, on which a bi-random walk algorithm is proposed to predict novel potential drug-disease associations. By integrating drug-target and disease-gene data, Luo *et al.* [91] proposed a new drug repositioning method RWHNDR to predict drug–disease interactions. By integrating drug–target and disease–gene interactions, they constructed a drug-target-disease network and extended the basic random walk model to the constructed heterogeneous network.

Label propagation

Considering that the heterogeneous network label propagation could explore global and local features of the network effectively, Yan *et al.* [92] proposed a network-based label propagation method LPMIHN to predict DTIs. For a query drug, LPMIHN performed the label propagation on drug similarity network firstly and then performed the label propagation on target similarity network with mutual interaction information to predict label confidence scores of all targets. Targets with scores exceeding a prespecified threshold are considered to be candidates for the query drug.

Shahreza *et al.* [93] proposed a heterogeneous label propagation algorithm, Heter-LP, to predict potential DTIs by integrating multisource information. First, a heterogeneous network was constructed by integrating drug similarity, disease similarity, target similarity, DTIs, drug–disease interactions and disease–target interactions. Then Heter-LP propagated the label information across the heterogeneous network to identify interactions between drugs and targets.

Network-based inference

Based on the complex network theory, Cheng *et al.* [94] developed three supervised inference methods, namely drug-based similarity inference (DBSI), target-based similarity inference (TBSI) and network-based inference (NBI), to predict novel drug-target associations. Among them, NBI could prioritize candidate targets for a specific drug or candidate drugs for a given target based on the two-phase diffusion on a bipartite drug-target graph. NBI yielded the good predictive performance on four gold standard data sets by using only the known drug-target association information. Cheng *et al.* [95] improved NBI by assigning weighted values to edges or nodes, namely edge-weighted NBI and node-weighted NBI, respectively. The systematic evaluation revealed that the two weighted NBI methods marginally outperformed the original NBI. Alaimo *et al.* [96] presented another NBI method, called domain tuned-hybrid (DT-hybrid), which extended NBI and Hybrid algorithms by adding the domain-based knowledge through a similarity matrix. The drug similarity, target similarity and known drug-target associations were integrated to provide a unified framework for the drug-target prediction. Computational experiment results showed that DT-hybrid clearly outperformed NBI for the DTI prediction by incorporating biological knowledge. Wu *et al.* [97] developed a useful tool, namely the substructure-drug-target NBI (SDTNBI), to find novel targets for old drugs, failed drugs and new chemical entities. By integrating known DTIs, drug-substructure linkages and new chemical entity-substructure linkages, SDTNBI utilized the resource diffusion method to infer new DTIs.

Wang *et al.* [98] proposed a heterogeneous graph-based inference method named HGBI, to predict drug-target associations. A drug-target heterogeneous graph, which incorporated drug similarity, target similarity and known DTIs, was constructed. Then, based on the guilt-by-association principle and an intuitive interpretation of information flow on the heterogeneous network, HGBI iteratively updated edge weights between drug-target pairs by considering all paths connecting them in the graph. The final weights between drugs and targets were obtained when the update procedure converged. By integrating drug, disease and target information simultaneously, Wang *et al.* [99] proposed a novel heterogeneous network model which integrated drug-disease and DTI prediction into a unified computational framework. Martínez *et al.* [100] have proposed a network-based prioritization method, DrugNet, which integrated the information of diseases, drugs and targets to perform drug-disease and disease-drug prioritization simultaneously.

Network clustering and network paths

Based on known drug and disease-related gene and feature information, Wu *et al.* [101] constructed a weighted disease and drug heterogeneous network and used two network clustering methods to find modules which contained candidate drug–disease interactions. Ba-Alawi *et al.* [102] developed a computational drug-target prediction method, DASPind, which used simple paths of particular lengths inferred from a drug-target heterogeneous network. Based on the assumption that a drug and a protein had a higher probability to interact if there are more paths connecting them, DASPind traversed all simple paths between a drug and a target protein and obtained the

aggregated score representing the likelihood that an interaction exists between them.

Matrix factorization- and completion-based methods

In order to develop more efficient and more accurate algorithms, a variety of techniques from many different fields have been incorporated into biomedical domains [111]. In particular, techniques such as matrix factorization and matrix completion have been successfully applied to discover novel drug–target and drug–disease interactions. Compared with the other methods, computational drug repositioning methods based on recommender models did not need negative samples and can integrate more prior information flexibly.

Basic matrix factorization

The matrix factorization methods assume that there are limited factors that determine the drug, target and disease relationship, which can be effectively obtained by matrix factorization. Dai *et al.* [103] proposed a matrix factorization model by integrating drug–gene interactions, disease–gene interactions and gene–gene interactions, to predict novel drug-disease associations. Considering interactions between drugs and diseases have their evidence in gene interaction networks, they integrated genomic space into the proposed model. The genomic space including drug–gene interactions, disease–gene interactions and gene–gene interactions could provide molecular biological information for the identification of novel drug-disease associations. Ezzat et al. [104] proposed two matrix factorization methods based on graph regularization techniques to predict novel DTIs. In addition, they also developed a weighted k-nearest neighbor method WKNKN as preprocess step to estimate the interaction likelihood values for unknown drug-target pairs.

Besides drug chemical and gene sequence information, there are also multiple related information which can be integrated to further improve the performance of drug repositioning. By using chemical structure, ATC code, genomic sequence, gene ontology and protein–protein interaction, Zheng *et al.* [105] defined multiple drug and target similarity matrices and proposed a multiple similarities collaborative matrix factorization (MSCMF) model with additive regularization terms corresponding similarity information. MSCMF projected drugs and targets into low-rank feature spaces which were used to approximate DTIs. Kuang *et al.* [106] have proposed a kernel matrix dimension reduction (KMDR) method to predict novel DTIs using the kernel matrix transformation. KMDR utilized chemical structure, ATC code and amino acid sequence information to define kernels related drug-target pairs. Then, KMDR conducted the eigenvalue decomposition of the defined kernel matrices and used the eigenvectors corresponding to top n largest eigenvalues to construct the link similarity matrix. Based on the link similarity matrix and known drug-target pairs, KMDR identified DTIs and outperformed the RLS classifier and the semi-supervised link prediction classifier.

Bayesian matrix factorization

Gönen [107] proposed a kernelized Bayesian matrix factorization (KBMF2K) method, which is a Bayesian formulation having combined dimensionality reduction, matrix factorization and binary classification to predict interactions for new drugs and new targets. KBMF2K utilized drug chemical structure to calculate drug similarity, and the drug similarity matrix was defined as a drug kernel matrix. Moreover, the protein sequence information is utilized to compute the target similarity, and target similarity matrix was used as a target kernel matrix. KBMF2K projected drugs and targets into a unified subspace to obtain low-dimensional feature representations of drugs and targets and used them to calculate the interaction scores for drug-target pairs. Given one drug and one target, their predicted positive score indicated that they interact with each other. KBMF2K can be applied in predicting candidate targets for new drugs, predicting candidate drugs for new targets and predicting interactions between new drugs and new targets.

To make use of multiple side information sources about objects, Gönen and Kaski [108] proposed a novel probabilistic model KBMF2KMKL, which extended the kernelized matrix factorization by incorporating multiple kernels learning. The drug (or target) kernels with different weight values were combined linearly to construct a combined drug (or target) kernel, and the combined drug kernel and target kernel were projected into a unified low-dimensional space.

Logistic matrix factorization

Liu *et al.* [109] proposed a DTI prediction algorithm, namely neighborhood regularized logistic matrix factorization (NRLMF), which had integrated logistic matrix factorization with neighborhood regularization and neighborhood smoothing. In NRLMF, drugs and targets were projected into a shared low-dimensional space and were represented by latent vectors. Then, the interaction probability for each drug-target pair was modeled by a logistic function using these latent vectors. Moreover, considering the known/observed drug-target pairs had been verified experimentally, NRLMF assigned higher importance levels to known pairs than unknown pairs. The previous kernel-based methods often combined multiple kernels linearly to construct the final kernel matrix. However, it is not reasonable for those kernels which have no clear linear relationship. Hao *et al.* [110] proposed a dual-network integrated logistic matrix factorization algorithm DNILMF to predict novel DTIs. For drugs, the chemical similarity matrix was converted to the drug kernel matrix. Then, the drug kernel matrix was combined with the inferred drug Gaussian kernel matrix by a nonlinear diffusion technique to construct the diffused drug matrix. For targets, the sequence similarity matrix was converted to the target kernel matrix and then combined with the inferred target Gaussian kernel matrix to construct the diffused target matrix. Then, DNILMF used the logistic matrix factorization integrating similarity information to predict interaction probabilities between drugs and targets.

Lim *et al.* [111] presented a computational method named COSINE to improve target prediction for novel chemicals using the collaborative filtering technique. By applying the logistic matrix factorization, COSINE projected chemicals and proteins into the unified low-dimensional latent space to model chemical–protein interactions. COSINE introduced position specific weights which essentially consider the importance of known or validated interactions and imputation of interaction values. Moreover, COSINE implemented a weighted-profile method to solve the cold-start problem.

Matrix completion

The matrix completion methods attempt to fill out the unknown elements in the drug, target and disease association matrices to reveal novel indications. Luo *et al.* [112] proposed a drug repositioning recommendation system (DRRS) to predict

potential indications for drugs, which can efficiently solve the cold-start and sparsity problem in drug repositioning. A drug-disease matrix was constructed by integrating drug- and disease-related data including similarity and interaction information. Based on the assumption that the hidden factors contributing to drug–disease interactions are highly correlated, the corresponding drug-disease matrix is low-rank. Then, a fast singular value thresholding algorithm was utilized to complete the drug-disease matrix with predicted interaction scores for new drug-disease pairs which have no validated interactions.

Yang *et al.* [113] proposed a bounded nuclear norm regularization (BNNR) method to complete the drug-disease matrix under the low-rank assumption. Although BNNR and DRRS were developed based on the same drug-disease matrix, BNNR could deal with noisy data included in similarity information by incorporating a regularization term. Moreover, BNNR could obtain more interpretable predicted values by incorporating additional bounded constraint to restrict predicted matrix entry values within the specific interval. They also developed an overlap matrix completion (OMC) method [114] for drug repositioning. The basic idea of OMC is to combine the prediction results from drug- and disease-side aspects. The OMC method provides an OMC2 algorithm and an OMC3 algorithm for drug-disease bilayer networks and drug-protein-disease tri-layer networks, respectively. The missing entries were filled out using the BNNR model. The experimental results show that OMC not only outperforms many other methods in the cross validation, but also has better computational efficiency.

Moreover, Wang *et al.* [115] proposed a Laplacian graph regularized matrix completion model to predict DTIs. In their work, drug similarities and target similarities were incorporated into the completion model by using a dual Laplacian graph regularization term. They addressed the matrix completion problem using an iterative strategy based on the Augmented Lagrange Multiplier algorithm.

Deep-learning-based methods

Deep learning is an extension of artificial neural network, which employs multiple processing layers to automatically learn the representations of data with multiple levels of the abstraction, and has been successfully applied to many fields, such as computer vision, speech recognition, natural language processing, chemoinformatics and bioinformatics [116–119].

Considering deep learning methods can extract topological structural features of vertices in the network [121], and the extracted features can be further used to compute topological similarities of two vertices. Zong *et al.* [120] proposed a similarity-based drug-target prediction method based on a deep learning algorithm DeepWalk [121]. First, they constructed a heterogeneous network, in which the drugs, targets and diseases as the vertices and associations among them as edges. Then, DeepWalk was used to obtain low-dimensional vector representations of vertices based on the topology of the constructed heterogeneous network, and similarity between two vertices was calculated using their representations. Finally, two similarity-based inference methods DBSI and TBSI [94] were utilized to predict novel DTIs using drug-drug similarities and target-target similarities.

Based on transcriptome data of drugs and targets, Xie *et al.* [122] modeled the DTI prediction as a binary classification task and designed a deep neural network (DNN) model to predict potential interactions. Wen *et al.* [119] developed a deep-learning-based algorithmic framework named DeepDTIs, which applied a deep-belief network to accurately predict new DTIs. The features of drugs and target were represented by the molecular chemical substructures and protein sequence information, respectively. The descriptors of drug-target pairs were constructed by concatenating the features of their drugs and targets. Then, these descriptors were used as input raw data of DeepDTIs. DeepDTIs learned representations of drug-target pairs from raw input descriptors through the unsupervised training process, and then build a classification model based on positive drug-target pairs and negative pairs randomly selected from unlabeled pairs.

Öztürk *et al.* [123] proposed a deep-learning-based model DeepDTA, which used the sequence information of targets and SMILES representations of drugs to predict DTI binding affinities. First, DeepDTA learned representations of drugs and targets from SMILES and protein sequences with convolutional neural networks. Then, these learned representations were used as inputs to three fully connected layers to predict binding affinities.

By integrating drug-related information, Zeng *et al.* [124] constructed multiple drug networks and learned low-dimensional features of drugs by a multimodal deep auto encoder. Then, the learned features and interactions with diseases are encoded and decoded collectively via a variant auto encoder to identify potential drug–disease interactions.

Experiment and Comparison

Gold standard data sets

In drug repositioning, two standard gold data sets have been used mostly to evaluate the performance of prediction methods. One is the drug-target data set from [84], which contained a set of known DTIs, drug similarities and protein target similarities. In this gold standard data set, known DTIs were collected from multiple public databases including KEGG BRITE, BRENDA, SuperTarget and DrugBank; drug-drug similarities were calculated by SIMCOMP based on their chemical structures; target-target similarities were defined as amino acid sequences similarities which were calculated using a normalized version of Smith-Waterman alignment scores. Chemical structure and protein sequence data were taken from the public database KEGG. This gold standard data set was divided into four groups according to the protein target types including enzyme, GPCR, ion channel and nuclear receptor. Each group contains the corresponding targets, drugs and interactions of drugs and targets.

Another gold standard data set is the drug-disease data set from [83], which contained 1933 associations between 593 drugs taken from DrugBank and 313 diseases listed in the OMIM database. In addition, the drug-disease data set contained five drug-drug similarity measures and two disease-disease similarity measures. The five drug-drug similarity measures are implemented based on drug chemical structure, drug side effect and drug target information. The pairwise disease-disease similarity is measured based on the semantic similarity of disease phenotypes. In the following evaluation experiments, this gold standard drug-disease data set is corresponding to Fdata set used in some studies. Moreover, the other two drug-disease data sets, including Cdata set and DNdata set, are collected in [90, 100]. Cdata set contains 663 drugs, 409 diseases and 2352 interacting drug-disease pairs. DNdata set contains 1490 drugs, 4516 diseases and 1008 interacting drug-disease pairs.

Evaluation metrics

For most of the aforementioned computational methods, the prediction methods is generally evaluated based on the *n*-fold or leave-one-out cross validation. The *n*-fold CV experiments are usually conducted in two different manners: (1) interaction prediction (CV-pairs): all known or verified interactions are divided into *n* folds randomly. In each round, $(n-1)/n$ of interactions are used as the training set, and the remaining $1/n$ of interactions are used as the test set; (2) new drug prediction (CV-drugs): all drugs are divided into *n* folds. In each round, $(n-1)/n$ of drugs are used as the training set, and the remaining $1/n$ of drugs are used as the test set. For the leave-one-out CV, the interaction prediction is done by selecting one interaction as the test set, and the remaining of interactions are used as the training set; the new drug prediction is done by selecting one drug as the test set, and the remaining of drugs are used as the training set. The drug-target or drug-disease pairs without known interactions are considered as candidate pairs.

A receiver operating characteristic (ROC) curve plots the true-positive rate versus the false-positive rate at various thresholds. A precision-recall (PR) curve plots the precision versus the recall at various thresholds. In the cross validation experiments, the drug repositioning methods predict the interaction probabilities of all candidate drug-target or drug-disease pairs. True positive is the number of correctly predicted known interacting pairs, and false positive is the number of incorrectly predicted noninteracting pairs. After performing the cross validation, the area under the ROC curve (AUC) and the area under the PR curve (AUPR) are calculated to evaluate the performance of different methods. For the two common used evaluation metrics, AUPR has been reported that it can provide more informative assessment than AUC for highly imbalanced data sets [125].

Experiment results

The performance of various prediction methods on the drug-target and drug-disease gold standard data sets was compared via the cross validation experiments. For the task of the drug-target prediction, the experiment results in terms of AUC and AUPR values obtained from the corresponding studies [92, 109, 110, 126] are shown in Tables 4 and 5, where the best experiment result in each row is indicated with bold fonts and the second best result is underlined. It can be observed in Table 4 that NRLMF have shown better predicting performance. The success of NRLMF can be contributed to the utilization of the neighborhood regularized logistic factorization, which assigned higher importance to known/positive DTIs, and considered influences from the nearest neighbors of drugs and targets via the neighborhood regularization. DNILMF employed the nonlinear diffusion technique to construct diffused drug similarity matrix and target similarity matrix. Then, DNILMF predicted the interaction probability scores by considering the diffused similarity and DTI information and achieved the best prediction results.

Moreover, it can be observed in Table 5 that the label propagation with the mutual interaction information derived from heterogeneous networks (LPMIHN) tends to dominate others, and MINProp achieves suboptimal results. The experiment results showed that the strategy of integrating the similarity information with topological information of the interactions used by LPMIHN method can help improve the prediction performance. In addition, MINProp is also a label propagation-based prediction method, which is different from LPMIHN in the label propagating pathways.

For drug-disease prediction tasks, the experiment results in terms of AUC and precision results are obtained from our previous study [113] and shown in Table 6. The *de novo* test in [113] is conducted by selecting drugs with only one known interactions to evaluate the capability of prediction method in identifying novel indications for new drugs with no verified interactions. One can find that BNNR and DRRS, which are designed based on matrix completion models, have achieved better prediction performance than other methods on all data sets in both 10-fold CV and *de novo* tests. This also indicates that BNNR and DRRS provide effective solutions to the cold-start problem which is prediction for new drugs in drug repositioning.

Challenges and Future Work

Drug repositioning is an important and promising methodology for drug discovery and development. Compared with traditional drug discovery, drug repositioning can shorten the time, save the cost and reduce the probability of failure as it starts from compounds candidates with well-known safety and pharmacokinetic profiles [4, 127]. Computational drug repositioning including drug–target and drug–disease interaction identification is implemented based on large-scale biological data and various models. In this review, biological and medical data, public databases, computational prediction methods, evaluation metrics and common gold standard data sets involved in drug repositioning are summarized and discussed.

Drug repositioning is a complicated process, in which the preciseness of biological information and the accuracy of computational models can affect the performance of prediction methods. We reviewed public databases and recent applications relevant to drug repositioning from the data perspective. It can be found that various types of biomedical data, such as chemical structure, bioactivity profile, side effect, therapeutic effect, gene expression, drug binding site, drug–drug interaction, ontology and semantic data have been utilized in recent studies. However, these biomedical data tends to be uncertain in practice due to high data noise, incompleteness and inaccuracy. By making full use of various biological and medical data from different heterogeneous data sources, the computational predictive models could realize more accurate identification of drug–target and drug–disease interactions. Generally, one type of data reflects only a single or a few aspects of the biomedical entity. Future efforts should be more thoroughly directed toward integrating the huge and heterogeneous amount of available data (chemical, biological, structural, clinical) into a unified workflow, which is obviously a challenging task. For data integration, the drug repositioning methods should consider different contributions of various biomedical data, and design rational weights assignment mechanisms for improvement.

By utilizing these data, many prediction approaches based on various computational models have been applied to identify drug–target or drug–disease interactions. We grouped these prediction methods based on their adopted computational models. Among these methods, matrix factorization and matrix completion based ones have shown better prediction performance according to the recently reported studies, and the integration of multisource information could further improve the prediction accuracy of computational drug repositioning methods. Generally, each computational method has its strength, applicability, drawbacks and limitations. For example, some methods can predict candidates for approved drugs having known interactions but not for new drugs which have no any known interactions, and some cannot integrate multiple types of biological

Table 4. Comparison on drug-target gold standard data sets under 10-fold cross validation

	10-fold cross validation						
	NetLapRLS	BLM-NII	WNN-GIP	KBMF2K	CMF	NRLMF	DNILMF
Data set	AUC value | AUPR value						
	CV_pairs						
NR	0.850 | 0.465	0.905 | 0.659	0.901 | 0.589	0.877 | 0.534	0.864 | 0.584	0.950 | 0.728	**0.955** | **0.751**
GPCR	0.915 | 0.616	0.950 | 0.524	0.944 | 0.520	0.926 | 0.578	0.940 | 0.745	0.969 | 0.749	**0.975** | **0.812**
IC	0.969 | 0.837	0.981 | 0.821	0.959 | 0.717	0.961 | 0.771	0.981 | 0.923	0.989 | 0.906	**0.990** | **0.938**
Enzyme	0.972 | 0.789	0.978 | 0.752	0.964 | 0.706	0.905 | 0.654	0.969 | 0.877	0.987 | 0.892	**0.989** | **0.922**
	CV_drugs						
NR	0.789 | 0.417	0.799 | 0.438	0.890 | 0.504	0.844 | 0.477	0.818 | 0.488	0.900 | 0.545	**0.956** | **0.776**
GPCR	0.817 | 0.229	0.838 | 0.315	0.891 | 0.295	0.839 | 0.366	0.857 | 0.365	0.895 | 0.364	**0.967** | **0.781**
IC	0.757 | 0.200	0.796 | 0.302	0.797 | 0.258	0.799 | 0.308	0.743 | 0.286	0.813 | 0.344	**0.961** | **0.822**
Enzyme	0.786 | 0.123	0.813 | 0.253	0.882 | 0.278	0.713 | 0.263	0.829 | 0.229	0.871 | 0.358	**0.964** | **0.796**

Table 5. Comparison on drug-target gold standard data sets under leave-one-out cross validation

	Leave-one-out cross validation					
	Weighted profile	RLS-Kron	BLM-NII	NRWRH	MINProp	LPMIHN
Data set	AUC | AUPR value					
NR	0.749 | 0.171	0.922 | 0.684	0.981 | 0.866	0.867 | 0.663	0.973 | 0.938	**0.996** | **0.970**
GPCR	0.765 | 0.109	0.954 | 0.790	0.984 | 0.865	0.945 | 0.674	0.987 | 0.937	**0.999** | **0.973**
IC	0.819 | 0.172	0.984 | 0.943	0.990 | 0.950	0.971 | 0.591	0.984 | 0.841	**0.999** | **0.961**
Enzyme	0.864 | 0.063	0.978 | 0.915	0.988 | 0.929	0.953 | 0.634	0.990 | 0.849	**0.999** | **0.929**

Table 6. Comparison on drug-disease gold standard data sets in terms of AUC values and precision results under 10-fold cross validation and *De novo* test

	10-fold cross validation				
	BNNR	DRRS	MBiRW	HGBI	DrugNet
Data set	AUC value | precision				
Fdata set	**0.932** | **0.440**	0.930 | 0.375	0.917 | 0.304	0.829 | 0.130	0.868 | 0.192
Cdata set	**0.948** | **0.471**	0.947 | 0.403	0.933 | 0.351	0.858 | 0.168	0.903 | 0.239
DNdata set	0.955 | **0.347**	0.934 | 0.346	0.956 | 0.321	0.921 0.204	0.950 | 0.150
	De novo test				
Data set	AUC value | precision				
Fdata set	**0.830** | 0.252	0.824 | **0.269**	0.818 | 0.234	0.746 | 0.099	0.782 | 0.135
Cdata set	0.812 | **0.254**	**0.819** | 0.243	0.804 | 0.232	0.732 | 0.107	0.785 | 0.147
DNdata set	0.956 | **0.418**	0.946 | 0.385	**0.970** | 0.242	0.928 | 0.320	0.969 | 0.242

information in drug repositioning. Therefore, by analyzing their characteristics, various computational methods can be assembled efficiently. Then, their applicability domain could be further extended, and their prediction accuracy could be improved.

The number of verified or labeled drug–target and drug–disease interactions is much less than that of unlabeled interactions in practice, which brings about the data sparseness problem in drug repositioning. Therefore, some prior information of biomedical entities could be utilized to preprocess the interaction data and complement missing values to relieve the sparseness problem. Some drug repositioning methods need positive and negative samples to train prediction models. However, the negative samples are often scarce or missing in practice; most studies solve this problem by selecting negative samples randomly from unlabeled data, which may affect the performance of prediction models. Therefore, rational negative sample selection methods should be designed to generate RN samples used in the training set. In order to fairly compare and evaluate various computational drug repositioning models, the benchmark data sets containing comprehensive biomedical data should be constructed by collecting useful information from multiple data sources. The analysis and explanation of the underlying mechanisms of predicted novel interactions are a significant problem in drug repositioning. Many studies performed case studies to analyze the predicted interactions by searching evidences from current public databases and literatures. However, such analysis

needs expert knowledge and is very time-consuming. Therefore, more research efforts should be conducted to solve the problem of verifying novel interactions predicted by drug repositioning methods.

Key Points

- Multiple types of biomedical data can be utilized and integrated to develop computational drug repositioning methods and validate their prediction results.
- This article analyzed and classified existing computational methods for potential drug–target or drug–disease interaction prediction from a machine learning perspective.
- Comprehensive benchmark data sets could be constructed by collecting biological and medical information from multiple data sources.
- Rational negative sample selection methods can generate reliable negative samples to train prediction models.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. U1909208, 61802113 and 61828205), Hunan Provincial Science and technology Program(No. 2018wk4001), and 111Project(No. B18059).

Conflict of Interest

None declared.

References

1. Li J, Zheng S, Chen B, *et al*. A survey of current trends in computational drug repositioning. *Brief Bioinform* 2015;**17**:2–12.
2. Hurle M, Yang L, Xie Q, *et al*. Computational drug repositioning: from data to therapeutics. *Clin Pharmacol Ther* 2013;**93**:335–41.
3. Kim T-W. Drug repositioning approaches for the discovery of new therapeutics for Alzheimer's disease. *Neurotherapeutics* 2015;**12**:132–42.
4. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. *Nat Rev Drug Discov* 2004;**3**:673.
5. Vanhaelen Q, Mamoshina P, Aliper AM, *et al*. Design of efficient computational workflows for in silico drug repurposing. *Drug Discov Today* 2017;**22**:210–22.
6. Hernandez JJ, Pryszlak M, Smith L, *et al*. Giving drugs a second chance: overcoming regulatory and financial hurdles in repurposing approved drugs as cancer therapeutics. *Front Oncol* 2017;**7**:273.
7. Shim JS, Liu JO. Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int J Biol Sci* 2014;**10**:654.
8. Turanli B, Grøtli M, Boren J, *et al*. Drug repositioning for effective prostate cancer treatment. *Front Physiol* 2018;**9**.
9. Zou J, Zheng M-W, Li G, *et al*. Advanced systems biology methods in drug discovery and translational biomedicine. *Biomed Res Int* 2013;**2013**:742835.
10. Lavecchia A, Cerchia C. In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov Today* 2016;**21**:288–98.
11. Keiser MJ, Roth BL, Armbruster BN, *et al*. Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* 2007;**25**:197.
12. Acharya C, Coop A, E Polli J, *et al*. Recent advances in ligand-based drug design: relevance and utility of the conformationally sampled pharmacophore approach. *Curr Comput Aided Drug Des* 2011;**7**:10–22.
13. Cheng AC, Coleman RG, Smyth KT, *et al*. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**:71.
14. Napolitano F, Zhao Y, Moreira VM, *et al*. Drug repositioning: a machine-learning approach through data integration. *J Chem* 2013;**5**:30.
15. Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity. In: *2012 IEEE International Conference on Bioinformatics and Biomedicine*. *IEEE*, 2012, 1–4.
16. March-Vila E, Pinzi L, Sturm N, *et al*. On the integration of in silico drug design methods for drug repurposing. *Front Pharmacol* 2017;**8**:298.
17. Yella J, Yaddanapudi S, Wang Y, *et al*. Changing trends in computational drug repositioning. *Pharm* 2018;**11**:57.
18. Ding H, Takigawa I, Mamitsuka H, *et al*. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief Bioinform* 2013;**15**:734–47.
19. Zhu Y, Elemento O, Pathak J, *et al*. Drug knowledge bases and their applications in biomedical informatics research. *Brief Bioinform* 2019;**20**:1308–21.
20. Chen X, Yan CC, Zhang X, *et al*. Drug–target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016;**17**:696–712.
21. Hao M, Bryant SH, Wang Y. Open-source chemogenomic data-driven algorithms for predicting drug–target interactions. *Brief Bioinform* 2019;**20**:1465–74.
22. Ezzat A, Wu M, Li X-L, *et al*. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief Bioinform* 2019;**20**:1337–57.
23. Lotfi Shahreza M, Ghadiri N, Mousavi SR, *et al*. A review of network-based approaches to drug repositioning. *Brief Bioinform* 2017;**19**:878–92.
24. Rifaioglu AS, Atas H, Martin MJ, *et al*. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* 2019;**20**:1878–912.
25. Zhou L, Li Z, Yang J, *et al*. Revealing drug-target interactions with computational models and algorithms. *Molecules* 2019;**24**:1714.
26. Liu T, Lin Y, Wen X, *et al*. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res* 2006;**35**:D198–201.
27. Gaulton A, Bellis LJ, Bento AP, *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res* 2011;**40**:D1100–7.
28. Law V, Knox C, Djoumbou Y, *et al*. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* 2013;**42**:D1091–7.
29. Fu C, Jin G, Gao J, *et al*. DrugMap Central: an on-line query and visualization tool to facilitate drug repositioning studies. *Bioinformatics* 2013;**29**:1834–6.

30. Drugs.com. http://www.drugs.com/.
31. Tatonetti NP, Patrick PY, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**:125ra131–1.
32. Von Eichborn J, Murgueitio MS, Dunkel M, *et al*. PROMISCUOUS: a database for network-based drug-repositioning. *Nucleic Acids Res* 2010;**39**:D1060–6.
33. Kim S, Thiessen PA, Bolton EE, *et al*. PubChem substance and compound databases. *Nucleic Acids Res* 2015;**44**:D1202–13.
34. Kuhn M, Letunic I, Jensen LJ, *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2015;**44**:D1075–9.
35. Kuhn M, von Mering C, Campillos M, *et al*. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2007;**36**:D684–8.
36. Günther S, Kuhn M, Dunkel M, *et al*. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2007;**36**:D919–22.
37. Kibbe WA, Arze C, Felix V, *et al*. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2014;**43**:D1071–8.
38. Pletscher-Frankild S, Pallejà A, Tsafou K, *et al*. DISEASES: text mining and data integration of disease–gene associations. *Methods* 2015;**74**:83–9.
39. Liu C-C, Tseng Y-T, Li W, *et al*. DiseaseConnect: a comprehensive web server for mechanism-based disease–disease connections. *Nucleic Acids Res* 2014;**42**:W137–46.
40. Piñero J, Bravo À, Queralt-Rosinach N, *et al*. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res* 2017;**45**:D833–D839.
41. Babbi G, Martelli PL, Profiti G, *et al*. eDGAR: a database of Disease-Gene Associations with annotated Relationships among genes. *BMC Genomics* 2017;**18**:554.
42. Becker KG, Barnes KC, Bright TJ, *et al*. The genetic association database. *Nat Genet* 2004;**36**:431.
43. Köhler S, Vasilevsky NA, Engelstad M, *et al*. The human phenotype ontology in 2017. *Nucleic Acids Res* 2016;**45**:D865–76.
44. World Health Organization. ICD-11 for Mortality and Morbidity Statistics (ICD-11 MMS) 2018 version. Available at: https://icd.who.int/browse11/l-m/en.
45. Rappaport N, Twik M, Plaschkes I, *et al*. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Res* 2016;**45**:D877–87.
46. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc* 2000;**88**:265.
47. Hamosh A, Scott AF, Amberger JS, *et al*. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 2005;**33**:D514–7.
48. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–70.
49. Stark C, Breitkreutz B-J, Reguly T, *et al*. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 2006;**34**:D535–9.
50. Consortium GO. The gene ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;**32**:D258–61.
51. Keshava Prasad T, Goel R, Kandasamy K, *et al*. Human protein reference database—2009 update. *Nucleic Acids Res* 2008;**37**:D767–72.
52. Rose PW, Prlić A, Altunkaya A, *et al*. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res* 2016;gkw1000.
53. Mering C, Huynen M, Jaeggi D, *et al*. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**:258–61.
54. Apweiler R, Bairoch A, Wu CH, *et al*. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:D115–9.
55. Davis AP, Grondin CJ, Johnson RJ, *et al*. The comparative toxicogenomics database: update 2017. *Nucleic Acids Res* 2016;**45**:D972–8.
56. Kanehisa M, Furumichi M, Tanabe M, *et al*. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* 2016;**45**:D353–61.
57. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenomics knowledge base. *Methods Mol Biol* 2013;311–20.
58. Deng Z, Tu W, Deng Z, *et al*. PhID: An open-access integrated pharmacology interactions database for drugs, targets, diseases, genes, side-effects, and pathways. *J Chem Inf Model* 2017;**57**:2395–400.
59. Li YH, Yu CY, Li XX, *et al*. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2017;**46**:D1121–7.
60. Nicola G, Liu T, Gilson MK. Public domain databases for medicinal chemistry. *J Med Chem* 2012;**55**:6987–7002.
61. Steinbeck C, Han Y, Kuhn S, *et al*. The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics. *J Chem Inf Comput Sci* 2003;**43**:493–500.
62. Hattori M, Tanaka N, Kanehisa M, *et al*. SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. *Nucleic Acids Res* 2010;**38**:W652–6.
63. Perlman L, Gottlieb A, Atias N, *et al*. Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol* 2011;**18**:133–45.
64. Hodos RA, Kidd BA, Khader S, *et al*. Computational approaches to drug repurposing and pharmacology. *Wiley Interdiscip Rev Syst Biol Med* 2016;**8**:186.
65. Vilar S, Uriarte E, Santana L, *et al*. Similarity-based modeling in large-scale prediction of drug-drug interactions. *Nat Protoc* 2014;**9**:2147.
66. Zhou L, Plasek JM, Mahoney LM, *et al*. Mapping partners master drug dictionary to RxNorm using an NLP-based approach. *J Biomed Inform* 2012;**45**:626–33.
67. Rodriguez-Esteban R. A drug-centric view of drug development: how drugs spread from disease to disease. *PLoS Comput Biol* 2016;**12**:e1004852.
68. Wang Y, Chen S, Deng N, *et al*. Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. *PLoS One* 2013;**8**:e78518.
69. Schuffenhauer A, Floersheim P, Acklin P, *et al*. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci* 2003;**43**:391–405.
70. Chiang AP, Butte AJ. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clinical Pharmacology & Therapeutics* 2009;**86**:507–10.
71. Bleakley K, Yamanishi Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 2009;**25**:2397–403.
72. Wang Y-C, Zhang C-H, Deng N-Y, *et al*. Kernel-based data fusion improves the drug–protein interaction prediction. *Comput Biol Chem* 2011;**35**:353–62.

73. Liu H, Sun J, Guan J, *et al*. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics* 2015;**31**:i221–9.
74. Liu H, Song Y, Guan J, *et al*. Inferring new indications for approved drugs via random walk on drug-disease heterogeneous networks. *BMC Bioinformatics* 2016;**17**:539.
75. Lan W, Wang J, Li M, *et al*. Predicting drug–target interaction using positive-unlabeled learning. *Neurocomputing* 2016;**206**:50–7.
76. Peng L, Zhu W, Liao B, *et al*. Screening drug-target interactions with positive-unlabeled learning. *Sci Rep* 2017;**7**:8087.
77. Xia Z, Wu L-Y, Zhou X, *et al*. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. In: *BMC Systems Biology BioMed Central*, 2010, S6.
78. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* 2011;**27**:3036–43.
79. Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics* 2016;**17**:46.
80. Van Laarhoven T, Marchiori E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One* 2013;**8**:e66952.
81. Hao M, Wang Y, Bryant SH. Improved prediction of drug-target interactions using regularized least squares integrating with kernel fusion technique. *Anal Chim Acta* 2016;**909**:41–50.
82. Chen H, Li J. A flexible and robust multi-source learning algorithm for drug repositioning. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, ACM*, 2017, 510–5.
83. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**.
84. Yamanishi Y, Araki M, Gutteridge A, *et al*. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**: i232–40.
85. Yamanishi Y, Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**:i246–54.
86. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics* 2017;**34**:1164–73.
87. Cao DS, Zhang LX, Tan GS, *et al*. Computational prediction of drug-target interactions using chemical, biological, and network features. *Mol Inform* 2014;**33**:669–81.
88. Huang Y-F, Yeh H-Y, Soo V-W. Inferring drug-disease associations from integration of chemical, genomic and phenotype data using network propagation. *BMC Med Genet* 2013;**6**:S4.
89. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. *Mol BioSyst* 2012;**8**:1970–8.
90. Luo H, Wang J, Li M, *et al*. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 2016;**32**:2664–71.
91. Luo H, Wang J, Li M, *et al*. Computational drug repositioning with random walk on a heterogeneous network. *IEEE/ACM Trans Comput Biol Bioinform*.
92. Yan X-Y, Zhang S-W, Zhang S-Y. Prediction of drug–target interaction by label propagation with mutual interaction information derived from heterogeneous network. *Mol BioSyst* 2016;**12**:520–31.
93. Shahreza ML, Ghadiri N, Mousavi SR, *et al*. Heter-LP: a heterogeneous label propagation algorithm and its application in drug repositioning. *J Biomed Inform* 2017;**68**: 167–83.
94. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**:e1002503.
95. Cheng F, Zhou Y, Li W, *et al*. Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One* 2012;**7**:e41064.
96. Alaimo S, Pulvirenti A, Giugno R, *et al*. Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 2013;**29**:2004–8.
97. Wu Z, Cheng F, Li J, *et al*. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning. *Brief Bioinform* 2016;**18**:333–47.
98. Wang W, Yang S, Li J. *Drug target predictions based on heterogeneous graph inference*. *Biocomputing 2013 World Scientific*, 2013, 53–64.
99. Wang W, Yang S, Zhang X, *et al*. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics* 2014;**30**:2923–30.
100. Martínez V, Navarro C, Cano C, *et al*. DrugNet: network-based drug–disease prioritization by integrating heterogeneous data. *Artif Intell Med* 2015;**63**:41–9.
101. Wu C, Gudivada RC, Aronow BJ, *et al*. Computational drug repositioning through heterogeneous network clustering. *BMC Syst Biol* 2013;**7**:S6.
102. Ba-Alawi W, Soufan O, Essack M, *et al*. DASPfind: new efficient method to predict drug–target interactions. *J Chem* 2016;**8**:15.
103. Dai W, Liu X, Gao Y, *et al*. Matrix factorization-based prediction of novel drug indications by integrating genomic space. *Comput Math Methods Med* 2015;**2015**:275045.
104. Ezzat A, Zhao P, Wu M, *et al*. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE ACM T Comput Bi* 2017;**14**:646–56.
105. Zheng X, Ding H, Mamitsuka H, *et al*. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, 1025–33.
106. Kuang Q, Li Y, Wu Y, *et al*. A kernel matrix dimension reduction method for predicting drug-target interaction. *Chemom Intell Lab Syst* 2017;**162**:104–10.
107. Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;**28**:2304–10.
108. Gönen M, Khan S, Kaski S. Kernelized Bayesian matrix factorization. In: *International Conference on Machine Learning*, 2013, 864–72.
109. Liu Y, Wu M, Miao C, *et al*. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol* 2016;**12**:e1004760.
110. Hao M, Bryant SH, Wang Y. Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci Rep* 2017;**7**:40376.
111. Lim H, Gray P, Xie L, *et al*. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci Rep* 2016;**6**: 38860.

112. Luo H, Li M, Wang S, *et al*. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. *Bioinformatics* 2018;**34**:1904–12.
113. Yang M, Luo H, Li Y, *et al*. Drug repositioning based on bounded nuclear norm regularization. *Bioinformatics* 2019;**35**:i455–63.
114. Yang M, Luo H, Li Y, *et al*. Overlapped matrix completion for predicting drug-associated indications. *PLoS Comput Biol* 2019; doi:10.1371/journal.pcbi.1007541.
115. Wang M, Tang C, Chen J. Drug-target interaction prediction via dual laplacian graph regularized matrix completion. *Biomed Res Int* 2018;**2018**:1425608.
116. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;**521**:436.
117. Singaravel S, Suykens J, Geyer P. Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction. *Adv Eng Inform* 2018;**38**:81–90.
118. Xu Y, Dai Z, Chen F, *et al*. Deep learning for drug-induced liver injury. *J Chem Inf Model* 2015;**55**:2085–93.
119. Wen M, Zhang Z, Niu S, *et al*. Deep-learning-based drug–target interaction prediction. *J Proteome Res* 2017;**16**:1401–9.
120. Zong N, Kim H, Ngo V, *et al*. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. *Bioinformatics* 2017;**33**:2337–44.
121. Perozzi B, Al-Rfou R. Skiena S. Deepwalk: Online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, 701–10.
122. Xie L, Zhang Z, He S, *et al*. Drug-target interaction prediction with a deep-learning-based model. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. *IEEE*, 2017, 469–76.
123. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 2018;**34**:i821–9.
124. Zeng X, Zhu S, Liu X, *et al*. deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 2019; doi:10.1093/bioinformatics/btz418.
125. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**:e0118432.
126. Mei J-P, Kwoh C-K, Yang P, *et al*. Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics* 2012;**29**:238–45.
127. Nagaraj A, Wang Q, Joseph P, *et al*. Using a novel computational drug-repositioning approach (DrugPredict) to rapidly identify potent drug candidates for cancer treatment. *Oncogene* 2018;**37**:403.