Machine learning approaches and databases for prediction of drug–target interaction: a survey paper

Maryam Bagherian†, Elyas Sabeti†, Kai Wang, Maureen A. Sartor, Zaneta Nikolovska-Coleska and Kayvan Najarian

Abstract

The task of predicting the interactions between drugs and targets plays a key role in the process of drug discovery. There is a need to develop novel and efficient prediction approaches in order to avoid costly and laborious yet not-always-deterministic experiments to determine drug–target interactions (DTIs) by experiments alone. These approaches should be capable of identifying the potential DTIs in a timely manner. In this article, we describe the data required for the task of DTI prediction followed by a comprehensive catalog consisting of machine learning methods and databases, which have been proposed and utilized to predict DTIs. The advantages and disadvantages of each set of methods are also briefly discussed. Lastly, the challenges one may face in prediction of DTI using machine learning approaches are highlighted and we conclude by shedding some lights on important future research directions.

Key words: Machine learning; drug–target interaction prediction; DTI software; DTI database

Introduction

In recent years, pharmaceutical scientists have been highly focused on novel drug development strategies that rely on knowledge about existing drugs [1–5]. Indeed, the difficulty of the drug discovery task lies in the rarity of existing drug–gene interactions [6], and a major risk is in unexpected/unintended interaction of drugs with off-target proteins, i.e. side effects [7–9]. While most of these side effects are undesired and harmful, occasionally they lead to interesting therapeutic discoveries. For instance, minoxidil was primarily developed to treat ulcers, and Sildenafil (Viagra) was developed to treat angina; however, they are currently used for treatment of hair loss and erectile dysfunction, respectively. As such, novel drug development strategies are currently the principle focus of many pharmacologists. It has been reported that several terms such as drug repositioning, drug repurposing, drug repurposing, drug redirecting, drug rediscovery and drug delivery have been used in the literature to describe these novel drug development strategies [3]. While various definitions have been used for these terms [3], drug repurposing usually refers to the studies that reinvestigate existing drugs that failed approval for new therapeutic indications [10], while drug repurposing suggests the application of already approved drugs and compounds to treat a different disease [11, 12].

Maryam Bagherian is a postdoctoral research fellow at Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor. Her Ph.D. degree is in applied mathematics and her research includes mathematical physics and mathematical biology. Elyas Sabeti is a postdoctoral research fellow at the Michigan Institute for Data Science, University of Michigan, Ann Arbor. He conducts research on data science methodology and its application in healthcare research. Maureen Sartor is an associate professor at Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor. Zaneta Nikolovska-Coleska is an associate professor at the Department of Pathology, University of Michigan, Ann Arbor. Kayvan Najarian is a Professor at Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor. His research focuses on signal/image processing and machine learning methods for medical applications.

Submitted: 4 September 2019; Received (in revised form): 1 November 2019

© The Author(s) 2020. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
A major step in the drug discovery process is to identify interactions between drugs and targets (e.g., genes), which can be reliably performed by in vitro experiments. In order to reduce temporal and monetary costs, in silico approaches are gaining more attention [2]. As such, instead of an exhausting in vitro search, virtual screening is initially performed and possible candidates are then experimentally verified [2]. Generally, there are two principle approaches for in silico prediction of drug–target interaction (DTI, also referred to as compound–protein interactions): docking simulations and machine learning methods [2].

In docking simulations, the 3D structure of drug molecules and targets are considered and potential binding sites are identified. While biologically well accepted, the docking simulation process is time-consuming [2]. Additionally, this process cannot be applied if the 3D structure of the protein is unknown [13]. For instance, for a class of proteins called G-protein-coupled receptors (GPCR), very few structures have been crystallized (orphan GPCR) [14, 15], so docking simulations cannot be applied.

To tackle this issue, chemogenomics was introduced as a way to aim at mining the entire chemical space for interaction with the biological space (also referred to as genomic space), instead of considering each protein target independently from other proteins [14, 16, 17].

The aim of chemogenomics research is to relate this chemical space of possible compounds with the genomic space in order to identify potentially useful compounds such as imaging probes and drug leads [13]. Chemogenomics approaches are usually categorized as ligand based, target based and target–ligand [14, 17], all of which are based on similarities between molecules proteins and targets. In fact, this salient similarity-based point of view of chemogenomics allowed the machine learning approaches to be suitable for prediction of DTIs. In machine learning methods [18], knowledge about drugs, targets and already confirmed DTIs are translated into features that are used to train a predictive model, which in turn is used to predict interactions between new drugs and/or new targets. The main assumption of these studies is that if drug d is interacting with protein p, then (i) drug compounds similar to d are likely to interact with protein p, (ii) proteins similar to p are likely to interact with drug d and (iii) drug compounds similar to d are likely to interact with proteins similar to p. The similarities between drug compounds and protein sequences are usually measured by kernels specifically designed for this purpose [19]. In practice, based on the availability of knowledge about interacting drug compounds and target proteins, the DTI prediction problem can be categorized into four classes: (i) known drug versus known target, (ii) known drug versus new target candidate, (iii) new drug candidate versus known target and (iv) new drug candidate versus new target candidate. While the ultimate goal of the machine learning methods is interaction prediction for new drug and target candidates, most of the methods in the literature are limited to the 1st three classes.

In this paper, the state of the art methods, which used machine learning methods for prediction of DTIs, are reviewed. The following studies were excluded:

- studies that do not use machine learning methods for prediction or (e.g. [20–25]).
- studies that focus on bioactivity (quantitative structure–activity relationship (SAR), proteochemometric) relationships (e.g. [26–32]).
- studies that rely on 3D structures of targets (e.g. [33–36]).
- studies that consider only the genomic space or chemical space (e.g. [4, 37–52]).
- studies that focus on gene expression for drug response (e.g. [53–58]).
- studies that only use side effect similarities or only predict side effects (e.g. [59–63]).
- studies that use disease–gene associations (e.g. [64–67]).
- studies that focus on drug–drug interactions or protein–protein interactions (PP) (e.g. [68–72])
- studies that use biomedical documents from which information is extracted by text mining techniques (e.g. [73]).

It is worth mentioning that the machine learning methods used in DTI prediction can be thought of as a broader problem of ‘link predictions’ in complex networks [74]. A section is dedicated to summarize the databases used in these studies as well. An overview of the paper is illustrated in Figure 1.
Previous review papers

There have been few reviews on DTI prediction with various emphases [79–83]; however, none of these studies had a machine learning focus. For previous reviews on machine learning methods for DTI prediction, please see [84–94]. In particular, [84] is a brief review of similarity-based machine learning methods used for DTI prediction. As reported in this work, similarity-based approaches have four advantages: (i) they do not need feature extraction and feature selection, (ii) similarity measure kernels for both drugs and genes have been fully studied before, (iii) they can be easily incorporated with kernel-based learning methods such as support vector machine (SVM), (iv) they can be used to connect chemical space and the genomic space. In [85], the focus of the review is on the methods that use both drug chemical structure and target protein sequence to predict DTIs. Mousavian et al. [90] reviewed machine learning-based methods from supervised and semi-supervised perspectives. Chen et al. [91] reviewed the well-known databases, web servers and computational models used for DTI prediction. In this paper, computational approaches are divided into network-based methods and machine learning-based methods. Ezzat et al. [92] provided an ‘empirical’ overview on chemogenomic DTI prediction methods and the databases used. In their work, the chemogenomic methodologies are separated into five models: neighborhood models, bipartite local models, network diffusion models, matrix factorization models and feature-based classification models. Chen et al. [87] reviewed the machine learning methods and databases that used chemogenomic approaches of DTI prediction. As such, based on the way negative samples are handled, chemogenomic approaches are divided into two categories: (i) supervised learning methods such as similarity-based and feature-based methods, (ii) semi-supervised learning methods. Kurgan et al. [88] wrote one of the most comprehensive surveys of DTI predictions before April 2018. Sachdev et al. [93] reviewed feature-based chemogenomic approaches (excluding similarity-based chemogenomic approaches) used for DTI prediction. In this survey, feature-based methods are categorized as: (i) SVM-based methods, (ii) ensemble-based methods (methods that employ decision tree or random forest) and (iii) miscellaneous techniques (neither SVM-based nor ensemble-based). Sercinoğlu et al. [94] reviewed all the available databases for drug repurposing.

Machine learning methods used in DTI prediction

Although all the DTI prediction frameworks that uses machine learning are summarized in this manuscript, recent methods that use matrix factorization algorithms have outperformed other methods in terms of efficiency. These methods take advantage of the recommender system approaches [75, 76], while using both chemical and genomic information is optimal for the DTI prediction problem. This problem is very similar to the famous Netflix challenge [77].

Machine learning methods used in DTI prediction date back to an early work in pharmacological DTI prediction [78]. While the focus of their work was not specifically ‘drug discovery’, they aimed at finding a ranked list of molecule ligands that bind with each orphan GPCR where due to lack of crystalized 3D structures, docking simulation could not be used [15]. Here, the machine learning approaches have been categorized into six groups (Figure 2). In the coming section, a description of each category along with a list of methods for each is provided. Moreover, advantages and disadvantages of each group of methods are briefly discussed.

Similarity/distance-based methods

The most popular group of methods used for DTI prediction incorporate drug–drug and target–target similarity measures through similarity or distance functions that are utilized to perform the prediction. These methods have been proposed and employed by several authors, mainly [13, 95–109].

Generally, the methods consist of a similarity score scheme for either drug–drug, target–target or drug–target associations based on a known pair of drug–drug and target–target similarity measures. Similarly, the similarity measure could be obtained by a distance function that defines how similar (or here ‘close’) a new drug is with respect to the known pairs. There are several ways to define the ‘nearness’ through a distance function for nearest neighbor (NN) algorithms [96, 102] among which the Euclidean distance is well known. For instance, authors in [102] employed the following definition for the NN algorithm; assuming two vector spaces (aka sample spaces) $V_1$ and $V_2$, with the same dimension, the distance (nearness) of the two samples is denoted by $D(V_1, V_2)$.
In addition to the above, the similarity/distance function could be also defined based on the pharmacological similarity of drugs and genomic similarity of protein sequences as well as the topological properties of a multipartite network of the existing drugs and protein targets [9, 110]. To this end, authors in [95] defined five drug–drug similarity measures as chemical based, ligand based, expression based, side effect based and annotation based. The main disadvantage of this group of methods lies in the fact that only a small number of drugs and their interactions are known while there exists copious unlabeled data among the datasets (see Section 3). Even though some efforts have attempted to deal with the lack of labeled data [5, 106, 107, 111, 112], the challenge has not yet been overcome. A comprehensive list of the methods proposed based on similarity/distance is provided in Table 1.

Deep learning methods

Deep learning is becoming more and more popular given its great performance in many areas, such as speech recognition, image recognition and natural language processing. Applying deep learning methods to drug discovery has been consistently increasing in recent years [113, 114].

Deep learning approaches appear to overcome certain limitations by reducing the loss of feature information in predicting DTIs. One of the drawbacks in using deep learning methods lays in the fact that there is not always sufficient information available in order to perform deep learning methods. Recently, in order to deal with high dimensional and oftentimes noisy data in DTI predictions in general and in drug repurposing in particular, authors in [115–117] proposed and developed deep learning algorithms in the DTIs machine learning approaches.

Most of the deep learning-based DTI prediction methods consist of two major steps: generating feature vectors and then applying deep learning to known DTIs. Usually, three types of properties (i.e. biological, topological and physico-chemical information) of drugs and/or targets can be used for generating feature vectors/matrix for deep learning based DTI methods. In recently published works [116–122], methods such as deep belief neural networks [118, 119], convolutional neural networks [120, 122] and multiple layer perceptrons [121, 122] were used to establish DTI prediction programs.

In [117], instead of using a bipartite network to represent the DTI, a Tripartite Linked Network [117], derived from the existing linked open datasets in the biomedical domain [125] were used for new DTI predictions. One advantage of methods employing deep learning over the state-of-the-art feature extraction methods and SVM classifiers is the ability to mine the hidden interactions between drugs and targets.

Although all of the aforementioned deep learning methods show good performance, there is room for improvement in several aspects. First, creating robust negative datasets for supervised deep learning method is a challenging task. Most previously published deep learning based DTI prediction programs are supervised machine learning methods, so how to establish an unbiased negative DTI dataset for model fitting and testing is a key step. In addition, DTI prediction is to discover new DTIs. How to select real no-interaction drug–target pairs

Table 1. Similarity/distance-based methods

| Abbreviations | Algorithms | Description |
|---------------|------------|-------------|
| SITAR         | Similarity-based Inference of drug-TARGets | A prediction scheme that integrates multiple drug–drug and gene–gene similarity measures to facilitate the prediction task using logistic regression [95]. |
| SRP           | Similarity-Rank-based Predictor | A lazy supervised non-parametric model using quantitative index to measure the tendency of interacting similar drugs and similar targets to predict DTIs. [97]. |
| ECKNN/HLM     | K-Nearest Neighbor Regression with Error Correction or Hubness-aware Local Models | A KNN method with an error correction method (hubness-aware regression technique) in order to alleviate the detrimental effect of bad hubs [98, 99] (with substantially different labels from those instances [100]). |
| NP, WP        | Nearest Profile & Weighted Profile | Given a test drug candidate, it finds a known drug sharing the highest similarity with the test drug, and predict the test drug to interact with target known to interact with the nearest drug [13, 101, 102]. |
| MDTI          | MultiviewDTI | A clustering algorithm, based on spectral clustering, integrating drug data and target data from both structural and chemical views and known DTIs [103]. |
| STC           | Super-Target Clustering | A clustering of similar targets by introducing the concept of super target to handle the missing interactions. [104]. |
| LPLNI, LPLNI-II | Label Propagation method with Linear Neighborhood Information | A framework in which first drug–drug linear neighborhood similarity is calculated, then the manifold of drugs are taken as similarities and finally unobserved DTIs are predicted using drug–drug similarities, interaction profiles and label propagation [105]. |
| WNN-GIP, RLS-WNN | Weighted Nearest Neighbors-GIP | A weighted NN algorithm directly incorporated into the GIP method, for constructing an interaction score profile for a new drug compound using information about known compounds [106]. |
| BLM           | Bipartite Local Models | In a bipartite graph model, predicts presence or absence of edges between drug and target using local models trained on known drugs and targets [98, 101, 107, 108]. |
| BLM-NII       | BLM with Neighbor-based Interaction-profile Inferring | An inferring integrated into the BLM method to handle the new candidate problem of pure BLM [107]. |
| WBRDTI        | Weighted Bayesian Ranking method | An improvement of BRDTI method by incorporating interaction weights for unknown DTs calculated based on known neighboring DTs [109]. |

where

\[
D(V_1, V_2) = 1 - \frac{V_1 \cdot V_2}{||V_1|| \cdot ||V_2||}
\]

where \((\cdot \cdot)\) denote the inner product and the Euclidean norm, respectively. One could easily verify that \(D\) is indeed a distance function satisfying the definition of the distance.
is a tricky task. Second, with more and more different types of drug/target data available, how to incorporate heterogeneous data into high-dimensional features from drug and/or target for deep learning methods is also a challenge. Last but not least, deep learning methods that show great performance on the testing dataset do not mean they also can achieve great performance in real drug discovery. More details about applying deep learning in drug discovery can be found in [126]. In Table 2, a brief list of deep learning-based methods mentioned in this paper is provided.

**Feature-based methods**

The vast majority of machine learning methods performing DTI prediction fall into this category. It is a broad range of methods including SVM, tree-based methods and other kernel-based methods. Any pairs of drugs and targets would be represented in terms of feature vectors with certain length, often with binary labels that classify the pair vectors into two classes with positive and negative interaction. In other words, assuming feature space $F$ where

$$F = \{ f : d \oplus t \mid d = [d_1, d_2, \ldots, d_n] \& t = [t_1, t_2, \ldots, t_m] \},$$

where $d$ and $t$ denote the target and drug feature vectors of length $n$ and $m$, respectively. Once the feature space is defined, assorted machine learning methods can be established to perform the DTI prediction task [5, 6, 9, 13, 14, 78, 89, 102, 106, 112, 127-178]. The lack of 3D structures of membrane proteins prevents extracting the main features, which otherwise would have yielded to better prediction performances. Tables 3 and 4 provides a broad list of feature-based methods along with a short description and the papers in which those methods were proposed and employed.

**Matrix factorization methods**

The matrix factorization methods have been shown to outperform other groups of machine learning methods in the

![Figure 3. Matrix factorization method.](image)

The primarily goal in DTI prediction is to decompose matrix $X_{n \times m}$ into two matrices, $Y_{n \times k}$ and $Z_{k \times m}$, where $X \simeq YZ^T$ with $k < n, m$ (Figure 3). Here $Z^T$ denotes the transposed matrix of $Z$. This will factorize matrix $X_{n \times m}$ into two matrices with lower orders (i.e. rank reduction), which make it easier to perform the matrix completion techniques in order to handle the missing data.

In contrast to most machine learning methods used for DTI prediction that need (2D) drug structural similarities, certain matrix factorization methods do not rely on chemical similarity or drug similarities and instead utilize collaborative filtering algorithms, among which one could name probabilistic matrix factorization (PMF) [179]. Some other methods are inspired by the idea of low-rank embedding (LRE) [180, 181] with the goal of finding a low-rank representation $R$ of the dataset $X$ by an optimization problem and then fixing $R$ and minimizing the reconstruction error in the embedded space in a way that the pointwise linear reconstruction (local structure of original samples) is preserved.

In this group of methods, it is assumed that the drugs and targets are lying in the same distance space such that the distance among drugs and targets can be used to measure the

---

Table 2. Deep learning methods

| Abbreviations | Algorithms | Description |
|---------------|------------|-------------|
| DeepDTIs      | Deep Learning in predicting DTIs | A deep-learning approach utilizing DBN [123] to abstract raw input vectors and predict new DTIs between FDA approved drugs and targets [118]. |
| DeepWalk      | Stacked Autoencoder Deep Neural Network | A deep learning similarity-based DTI prediction method based on the topology of multipartite network of the existing drugs and targets [117]. |
| AutoDNP       | Deep learning with convolution-DTI | A deep learning method capturing local residue patterns of proteins participating in DTIs[122]. |
| LASSO-DNN     | Least absolute shrinkage and selection operator-Deep Neural Network | A deep learning method based on features extracted from the LASSO regression models fitted using the protein-specific and drug-specific features respectively [121]. |
| DeepDTA       | Deep DT Binding Affinity Prediction | A deep learning-based model using only character representations (raw sequence information) for both drugs and targets simply [120]. |
| DeepNP        | Deep Neural Representation | An interpretable end-to-end deep learning architecture to predict DTIs from low level representations [119]. |
| DeepTrans     | Deep Transcriptome data | A framework for DTI prediction based on transcriptome data in the L1000 database gathered from drug perturbation and gene knockout trials [124]. |
In Section 4, some constraints. Although this group of methods has been shown to be more reliable than the others, rapid growth in the quantity and variety of data related to a certain drug and/or a target far exceeds the capacity of matrix-based data representations and many current analysis algorithms. A solution to this issue has been proposed in Section 4. In Table 5, the matrix factorization methods and the paper(s) in which they are proposed, developed, and employed are listed.

**Network-based methods**

The network-based methods refer to those that utilize graph-based techniques in order to perform the task of DTI prediction (Figure 4). Among the methods is network-based inference (NBI). In a bipartite graph model, predicts the presence or absence of edges between drug and target based on graph-based similarity to known drugs and targets in a unified Euclidean space of chemical and genomic space called pharmacological space [13, 142, 143].

### Table 3. Feature-based methods: part I

| Abbreviations | Algorithms                        | Description                                                                 |
|---------------|-----------------------------------|-----------------------------------------------------------------------------|
| SVM, K SVM, MH-SVM | Support Vector Machine          | A support vector machine constructs a hyperplane or set of hyperplanes, which can be used for prediction of presence or absence of interaction between drugs and targets [14, 78, 127–141]. |
| BGL/KRM       | Bipartite Graph Learning or Kernel Regression-based Method | In a bipartite graph model, predicts the presence or absence of edges between drug and target based on graph-based similarity to known drugs and targets in a unified Euclidean space of chemical and genomic space called pharmacological space [13, 142, 143]. |
| NetLapRLS     | RLS with kernels derived from known DTIs | The improved version of LapRLS by incorporating a new kernel established from the known DTI network [6]. |
| PKR           | Pairwise Kernel Regression        | A regression model similar to KRM without requirement of any unified chemical and genomic space [9]. |
| RF, DDR       | Random Forest                     | A robust model against the overfitting problem of traditional statistical methods that performs more efficiently for large-scale databases [5, 131, 144, 145] (using [137, 146–150]). |
| iDTI-ESBoost  | Positive-Unlabeled learning for DT prediction | A prediction model for identification of DTIs using evolutionary and structural features [151]. |
| PUDT          | RLS with Gaussian interaction profile kernel | A framework treating unknown DTI as unlabeled samples and using weighted SVM predictor [152]. |
| GIP           | Regularized Least Square, also RLS-Kron, RLS-avg, LapRLS, KRLS, RLS-KF, KronRLS-MKL | An RLS algorithm that incorporates the topology of known DTI network as source information through GIP kernel [5, 153]. |
| RLS           | SimBoost, SimBoostQuant          | A semi-supervised framework that incorporates known DTIs and unknown DTIs in a general-purpose learner [6, 106, 153–158]. |

### Table 4. Feature-based methods: part II

| Abbreviations | Algorithms                        | Description                                                                 |
|---------------|-----------------------------------|-----------------------------------------------------------------------------|
| RFDT          | Rotation Forest-based DTI prediction | A computational model based on the assumptions that the protein sequences are encoded as Position Specific Scoring Matrix (PSSM) [160] descriptor and the drug molecules are encoded as fingerprint feature vector [161]. |
| DrugRPE       | ChemoGenomics-Based Virtual Screening DASPfind | A random projection ensemble approach for based on the REPTree algorithm [162] and using random projection [102, 162, 163, 163–165]. |
| CGBVS         | Structure-Activity Relationship method | A kernel-based state-of-the-art method using virtual screening (VS) [89] and pairwise kernel method (PKM) [14] [166]. |
| SAR           | Discriminative Vector Machine     | A computational DTI prediction method relying on the topological structure of the heterogeneous graph interaction model [167]. |
| EnsL          | Ensemble Learning (with dimensionality reduction, or class imbalance-aware) | A framework predicts DTI based on average voting of its base classifiers: Decision Tree (EnsemDT) [171–173] (based on Singular Value Decomposition (SVD), Partial Least Squares (PLS) [174] and Laplacian Eigenmaps (LapEig) [175]), Kernel Ridge Regression (EnsemKRR), Random Forest (EnsemRF) [112], stacked (EnsemSTACK) [176], DrugE-Rank [177]. |
| BE-DTI        | Bagging-based Ensemble method     | A bagging-based ensemble framework that involves dimensionality reduction and active learning [178]. |
Table 5. Matrix factorization methods

| Abbreviations | Algorithms | Description |
|---------------|------------|-------------|
| MSCMF         | Multiple Similarities one-Class Matrix Factorization | An approach to approximate the input DTI matrix by two low-rank matrices, which share the same feature space and are generated by the weighted similarity matrices of drugs and those of targets, respectively [182] using [183-186]. |
| NRLMF         | Neighborhood Regularized Logistic Matrix Factorization | A mode that integrates logistic matrix factorization with neighborhood regularization for DTI prediction [187]. |
| PMF           | Probabilistic Matrix Factorization | A collaborative filtering method that decomposes the DT bipartite connectivity matrix as a product of two matrices of latent variables that will be used for prediction, irrespective of the drug or target similarities [179]. |
| DLGRMC        | Dual Laplacian Graph Regularized Matrix Completion | An optimization framework for low-rank approximation of interaction matrix based on matrix completion in which drug similarity and target similarity are used as dual Laplacian graph regularization term [188]. |
| GRMF-WGRMF    | Graph Regularized Matrix Factorization and Weighted GRMF | Two manifold learners for extracting low-dimensional non-linear manifolds of DTI bipartite graph [189]. |
| Pseudo-SMR    | Pseudo Substitution Matrix Representation | An extension to SAR classification problem [137], employing a python package called scikit-learn for machine learning to implement Extremely Randomized Tree (ER-Tree) introduced in [190, 191]. |
| BRDTI         | Bayesian Ranking method | A method based on Bayesian Personalized Ranking matrix factorization (BPR) that incorporates target bias and content alignment for drug and target similarities [2, 99, 192]. |
| LRE, SLRE, MLRE | Low Rank Embedding | An algorithm of finding a low-rank representation (by optimization problem) and fixing and minimizing the reconstruction error in the embedded space in a way that the pointwise linear reconstruction (local structure of original samples) is preserved [181]. LRE for an arbitrary view (structure or chemical) is called SLRE and for multiview is called MLRE [180]. |
| VB-MK-LMF     | Variational Bayesian Multiple Kernel Logistic Matrix Factorization | A method integrating multiple kernel learning, weighted observations, graph Laplacian regularization and explicit modeling of probabilities of binary DTIs [193]. |
| KBMF, KBMF2K  | Kernelized Bayesian Matrix Factorization | A method for factorizing the interaction score matrix in terms of kernel matrices (similarity matrices), which can be used as DTI predictors for new drugs and protein [194]. |

Figure 4. Drug-target interaction heterogeneous network.

inference methods and uses only DT bipartite network topology similarity [195].

Moreover, in certain methods three networks of protein–protein similarity, drug–drug similarity and known DTIs are integrated into a heterogeneous network and assumed similar drugs often target similar proteins [196, 203]. A two-layer undirected graphical representation of the network could also be adopted in order to train to predict direct DTIs (usually caused by protein–ligand binding), indirect DTIs and drug mode of actions (binding interaction, activation interaction and inhibition interaction) in addition to performing the DTI prediction task. A pertinent example is proposed in [204] using Restricted Boltzmann Machine (RBM) [123]. A list of network-based methods with a short description for each method is provided in Table 6.

Hybrid methods

Hybrid methods refer to all the approaches in which any combination of the feature-based, matrix factorization, deep learning and network-based methods are exploited. This can extend the capability of a prediction algorithm by integrating different sets of information. The hybrid methods in general serve two purposes; they address the problems of unknown interaction in DTIs as well as taking the most advantage of machine learning methods, simultaneously. For instance, authors in [177] proposed a method integrating feature-based and similarity-based machine learning approaches [205, 206]. The hybrid methods performed superior to other state-of-the-art methods by optimizing the feature extraction process by extracting the complex hidden features of drugs and targets [134, 144, 172, 173, 182, 197, 201, 207, 208]. Integrating two machine learning methods in DTI prediction often has a leverage in terms of results as they fully exploit the potential of two methods, simultaneously. However, one should be able to deal with the high complexity (either computational or operational) caused by integrating two groups of methods. A short description of such methods are listed in Table 7.
Table 6. Network-based methods

| Abbreviations | Algorithms | Description |
|---------------|------------|-------------|
| NBI           | Network-Based Inference | A method based on DT bipartite network topology similarity [195]. |
| NRWRH         | Network-based Random Walk with Restart on the Heterogeneous network | A method based on the framework of RWR to infer potential DTIs on a bipartite graph network [196]. |
| NetCBP        | Network-Consistency-based Prediction Method | A semi-supervised inference method, utilizing both labeled and unlabeled data [111]. |
| DTINet        | inter/intra-network RWR or Co-rank NormMulInf | A computational network integration pipeline for DTI prediction [197]. |
| IN-RWR        | Beta-distribution-rescored NRLMF | An improved NRLMF algorithm that rescores the score of NRLMF as the expected value of the $\beta$-distribution, which is determined based on interaction information and NRLMF score [201]. |
| RWR           | Random Walk with Restart | A method that requires a matrix inversion and provides a good relevance score between two nodes in a weighted graph of DTIs [202]. |

Table 7. Hybrid methods

| Abbreviations | Algorithms | Description |
|---------------|------------|-------------|
| DT-Hybrid     | Domain tuned-hybrid | An extended NBI technique that incorporates domain-based knowledge such as drug similarities and target similarities [209] (also look [195, 210, 210, 211, 211?]) for extension of the capability of recommender systems. |
| KMDR          | Kernel Matrix Dimension Reduction | A framework for construction of link similarity matrix from kernel matrix and feature transformation for DTI prediction [208]. |
| MGRNNM, DGRMC | Multi Graph Regularized Nuclear Norm Minimization | A computational method that adds multiple drug–graph and target–graph Laplacian regularization terms to the standard matrix completion framework to predict DTIs [212, 213]. |
| WLNM          | Weisfeiler-Lehman Neural Machine | An algorithm for extraction of the adjacency matrix that represents the interactions between potential drugs and targets [214]. |
| PDTPS         | Predicting Drug Targets with Protein Sequence $L_1$-regularized Classifier | A framework based on Relevance Vector Machine that integrates Bi-gram probabilities, PSSM and PCA [215]. |
| RBM           | Restricted Boltzmann Machine | A two-layer undirected graphical model to represent a multidimensional DTI network and encode different types of DTIs [204]. |
| LRF-DTI       | Lasso-based Random Forest method | A method of DTI prediction based on Lasso dimensionality reduction and random forest predictor [144]. |
| COSINE        | COld Start INtegrations | A statistical dual-regularized, one-class collaborative filtering method [217] framework and a corresponding computational method for multi-target virtual screening using one-class collaborative filtering technique that can employ either logistic or linear factorization [218]. |
| DMF           | Deep Matrix Factorization | A deep learning approach in the context of recommendation systems to extract the non-linearity of latent variables [219] (DMF was originally introduced in [220] as a deep learning method in the context of recommendation systems to extract the non-linearity of latent variables). |
| CoDe-DTI      | COllaborative DEep learning-based DTI predictor | A method using both PMF and a denoising autoencoder [221]. |

Software and packages

Sakakibara et al. [222] developed a web service called Comprehensive Predictor of Interactions between Chemical compounds And Target proteins based on their previous works [127, 129] that uses SVM as the DTI predictor. This server seems to be no longer available.

Cao et al. [223] developed a Python package called PyDPI based on Random Forest [150] that integrates chemoinformatics, bioinformatics, proteochemometrics and chemogenomics for DTI prediction. The proposed framework involves the selection of molecular features and uses predefined dictionaries for classification. This package can be used to construct web-based servers and provides an interface for databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG), PubChem, Drugbank and Uniprot. The same group in the same year [224] also developed a web-based server called PreDPI-KI (which seems to be no longer available) based on a random forest predictor that takes binding affinities of DT pairs into account in order to better predict interactions.
Xiao et al. [225] established a web server called iGPCR-drug, which is accessible at iGPCR-drug. Moreover, they developed a sequence-based classifier also called iGPCR-drug. In the predictor, the drug compound is formulated by a 2D fingerprint via a 256D vector, GPCRs by the pseudo amino acid composition [226] generated with the gray model theory and the prediction engine is operated by the fuzzy K-nearest neighbor (KNN) classification method [227]. The authors validated their method with the jackknife test [228].

Yamanishi et al. [229] designed a web server called DINIES (DTI network inference engine based on supervised analysis) for predicting DTI using various types of biological data such as chemical structures, protein domain and drug side effects (note that studies that primarily focused on side effect are excluded in this paper [59–62]) and three supervised algorithms (BGL [13, 143], BLM [101] and pairwise kernel regression [9]). This is due to the work by Scheiber et al. [230] that enables the calculation of correlation between any drug compound and pharmacological effects in chemical space. While the training can be performed using KEGG DRUG database, the principle advantage of their web server is the flexibility of the input data, as long as it’s represented a similarity matrix or gene/drug profile.

Seal et al. [231] developed a standalone R and Shiny package called Netpredictor based on Random Walk with Restart (NRWRVH) [196, 202] and NBI [195, 209] to predict any missing links between drugs, proteins and drug–proteins in any unipartite or bipartite. The main advantage of this package is the friendly user interface that is provided by package installation.

Hao et al. [232] review, compare and reimplemented five state-of-the-art methods (BLM [101], KronRLS-MKL [158], DT-Hybrid [209], the proposed method by Shi et al. [104] and DNILMF [233]) and published the source codes in R.

### Databases used in DTI-pPrediction

To support the above methods, many drug-related databases have been established. These databases contain different types of drug-related information and are critical resources for DTI predictions in silico. In this paper, we review all popular used databases associated with this topic. Based on the content of these databases, we classify them into four categories, DTI databases, drug-centered or target centered databases, drug–target binding affinity databases and supporting databases.

### DTI databases

DTI databases are established for collecting DTIs and other related information. In this paper, we list 11 databases in this category. Within these databases, some are not directly proposed as ‘DTI’ databases, but the data contained can be used for DTI research. For example, KEGG is an extensive database that covers many types of biological data from genes/proteins to biological pathways and human diseases. In KEGG [234], two subdatabases, KEGGDRUG [235] and KEGG-BRITE [236] contain data that can be used for DTI predictions. ChEMBL [237–239] is also not specifically a ‘drug-target’ database and it was established based on collecting bioactive compounds. However, combined with targets and other related biological information, this database can also be used in drug-target repositioning and repurposing. Similar to ChEMBL [237–239], IntAct [240] is a database that contains molecular interactions and can be used for drug research. LINCS is different from the aforementioned two databases. This data portal contains biochemistry data that aims to understand changes in gene expression and cellular processes that are caused by different perturbing agents. Many of the perturbing agents used in LINCS are drugs, so this is also a great data source for DTI research. Other databases included in this group are SuperTarget [241], Guide to PHARMACOLOGY (GtoPdb) [246], DrugBank [242–246], Therapeutic Targets Database (TTD) [247], STITCH [248–252], ChemProt 3.0 [253] and DGIdb 3.0 [254]. The general information for these databases is summarized in Table 8.

### ChEMBL

The data stored in the ChEMBL database [237–239] were manually extracted from published literatures. This database was published by European Molecular Biology Laboratory (EMBL)-European Bioinformatics Institute in 2002. Since the latest update in 2018, this database contains more than 1.9 million chemical compounds. Within these compounds, over 10 thousand drugs and more than 12 thousand targets are included in ChEMBL.

### ChemProt 3.0

ChemProt [253, 255, 256] was proposed as a disease chemical biology database that integrated data from multiple chemical–protein annotation databases and disease-associated PPI. The first release of ChemProt was in 2011, which collected data from eight public databases, i.e. ChEMBL [238], BindingDB [257], PDSP Ki database [258], DrugBank [244], PharmGKB [259], PubChem bioassay [260], CTD [261] and STITCH [248] and two commercial
databases, WOMBAT and WOMBAT-FK [262]. The second update of ChemProt was in 2012 integrated therapeutic effects and adverse drug reactions into the 2.0 version. The latest update (version 3.0) was released in 2015. The third version updated the disease chemical biology data. In addition, several computational methods, such as network biology based enrichment analysis, were also incorporated.

**DGIdb 3.0**

The first release (in 2013) of DGIdb integrated 13 data sources that cover information in disease-related human genes, drugs, drug interactions and potential druggability [263, 264]. The latest update of DGIdb was in 2017 and in total 30 data sources are included in the 3.0 version [254]. Six new data sources were added and nine of the previous data sources were updated.

**DrugBank**

DrugBank [242–246] is one of the most popular databases and has been widely used as a drug reference resource. This database was first released in 2006. As a database both in bioinformatics and cheminformatics, DrugBank contains detailed drug data with comprehensive drug target information. The DTI relationships in DrugBank were originally collected from textbooks, published articles and other electronic databases. All data can be freely downloaded from DrugBank.

**GtoPdb**

This database was established by the International Union of Basic and Clinical Pharmacology/British Pharmacological Society. The GtoPdb [240] contains the ligand–activity–target relationships data that were collected from pharmacological and medicine chemistry literature.

**IntAct**

IntAct [265] is an open source database of molecular interactions populated by data from literature and other data sources. In total, 11 molecular interaction databases (including IntAct) were incorporated into IntAct including AgBase [266–269], MINT [270–273], UniProt [274] [41], I2D [275], MBINFO, MatrixDB [276], Molecular Connections, InnateDN [277], IMEx [278] and GOA.

**KEGG**

KEGG is a comprehensive database that provides many types of knowledge about genes and genomes [234, 235]. The whole database can be summarized in four major categories. The first one is systems information, contains three databases: KEGG PATHWAY, KEGG BRITE, and KEGG MODULE. The second category contain genomic information. In this group, four databases are included: KEGG ORTHOLOGY, KEGG GENOME, KEGG GENES and KEGG SSDb. The third category holds the chemical information. Five databases are in this category: KEGG COMPOUND, KEGG GLYCAN, KEGG REACTION, KEGG RCLASS and KEGG ENZYME. The last category is health information that covers four databases: KEGG DISEASE, KEGGDRUG, KEGG DGROUP and KEGG ENVIRON. The KEGG DGROUP database contains information regarding drug interaction networks including DTIs, drug metabolism and indirect interactions with enzymes and target genes.

**LINCS**

The LINCS program aims to establish a network-based landscape to describe how different perturbing agents influence cellular processes. In total, there are 398 datasets collected in the LINCS database including fluorescence imaging, ELISA and ATAC-seq data, etc. The majority datasets (177 datasets) in LINCS are KINOMEscan kinase-small molecule binding assays. This assay is used to measure binding interactions between test compounds.

**PROMISCUOUS**

PROMISCUOUS was established in 2011 and proposed as a database for network-based drug repositioning. This database contains three different types of data: drugs, proteins and side effects. The protein data are extracted from UniProt and incorporated with the 3D structure information from Protein Data Bank (PDB). Drugs and side effects are extracted and incorporated from SuperDrug and SIDER, respectively. In addition to DTIs and drug side effects linkages, PROMISCUOUS also includes data on drug–drug similarities and PPI.

**STITCH**

STITCH [248–252] is a database that stores information for interactions between proteins and small molecules. The interaction data are collected from predicted results, other databases (e.g. PubChem [279]), and literature. The first release of STITCH was in 2008.

**SuperTarget**

SuperTarget [241] is a database that covers DTI information with drug metabolism, pathways and Gene Ontology (GO) terms. Medical indications and adverse drug effects are also included in this database. The DTIs information in this database were extracted starting with text mining from 15 million public literature listed in PubMed. Also, potential drug–target relations were also extracted from Medline. Furthermore, the relationships of DTIs from other databases (i.e. DrugBank [244], KEGG [234], PDB [280], SuperLigands [281] and TTD [282]) were also used to obtain any missed DTIs that were not included from the previous two strategies.

**Drug-centered or target-centered databases**

In this category, six databases are included. They are BRENDA [283], PubChem [279], SuperDRUG2 [284], DrugCentral [285, 286], PDID [287], Pharos [288] and ECOdrug [289].

Among these databases, SuperDRUG2 and DrugCentral are proposed as ‘drug-centered’ databases. Since PubChem is a database established on collecting millions of chemical compounds, in this paper, we also list this one as a ‘drug-centered’ database. PDID and Pharos are classified as ‘target-centered’ databases. We also included BRENDA as a ‘target database’. The huge amount of enzymes and related ligands stored in BRENDA can be used as targets in DTI research. In addition, we also list ECOdrug here as a target-centered database. Different from the aforementioned ones, this database contains target information in non-human model species. Relative information can be found in Table 9.

**BRENDA**

BRENDA [283, 290] is a comprehensive enzyme database that was first established in 1987. This database contains `84 000
enzymes and their corresponding enzyme–ligand related information. All data collected in this database was manually evaluated and extracted from 140,000 literature references based on the Enzyme Commission (EC) classification system of the International Union of Biochemistry and Molecular Biology. All compounds related to enzyme catalyzed reactions are labeled as 'ligands' in BRENDA, such as substrates, products, activators, inhibitors and cofactors. In total, about 205,000 enzyme ligands were collected and stored in the associated ligand database. Users can search the ligand database through the search box on the home page. BRENDA also provides download functionality for users to download all BRENDA data.

DrugCentral

DrugCentral is a comprehensive database that focuses on drug collection [285, 286]. This database was released in 2016 and contains approved active pharmaceutical ingredients (drugs) from FDA and other regulatory agencies. For each drug, structure information, bioactivity and regulatory records, as well as pharmacologic actions and indications were incorporated. In this database, all drugs are simply classified into three categories, small molecule active ingredients, biological active ingredients and others.

ECOdrug

In drug discovery research, non-human model species are important in that they are used for drug testing. ECOdrug [289] is a database that contains DTI data for 640 eukaryotic species. The data stored in ECOdrug can help researchers investigate the conservation of human drug targets across species. The drug information and drug targets are from previous research [291] and DrugBank [244].

PDIID

PDIID [287] was released in 2014 and covers all known protein–drug interactions and predicted protein–drug interactions for the entire structural human proteome. The known interactions were extracted from DrugBank [244], BindingDB [257] and PDB [280]. The predictions were made by using three different softwares (i.e. ILbind [292], SMAP [45] and eFindSite [293, 294]).

Pharos

Pharos [288] is a platform that was established for presenting the data in the Target Central Resource Database (TCRD). TCRD is a comprehensive database that was initially developed for discovering new druggable proteins. The data stored in TCRD came from many different sources. It includes biomedical literature, expression data, disease and phenotype association data, bioactivity data, DTI data and databases from Harmonizome [295].

Table 9. Drug-centered or Target-centered databases

| Database     | Latest updates | Type              | No. of targets | No. of drugs/Compounds | Predicted DTIs |
|--------------|----------------|-------------------|----------------|------------------------|---------------|
| BRENDA       | Jan 2019       | Target centered   | >84,000        | >205,000               | ✓             |
| DrugCentral  | Apr, 2019      | Drug centered     | -              | 4543                   | ✓             |
| ECOdrug      | Oct 2017       | Target centered   | -              | -                      | ✓             |
| PDID         | Apr 2015       | Target centered   | 3746           | 51                     | ✓             |
| Pharos       | Nov 2018       | Target centered   | 20,244         | 130,166                | ✓             |
| PubChem      | Mar 2019       | Drug centered     | 79,622         | 96,157,016             | ✓             |
| SuperDRUG2   | Mar 2018       | Drug centered     | 4456           | 4605                   | ✓             |

PubChem

PubChem [279, 296] stores the information of chemical substances and corresponding biological actives. This database consists of three sub-databases: Substance, Compound and BioAssay. Substance is the primary repository to store chemical information provided from individual data contributors. The Compound database contains the unique chemical structures extracted from the Substance database. All biological related data of these chemical substance data are saved in the BioAssay database.

SuperDRUG2

SuperDRUG2 [284] is proposed as a one-stop data source that offers all crucial features of approved and marketed drugs. The drug items in SuperDRUG2 are classified into two categories: small molecules and biological/other drugs. Several public resources like US FDA, CFDA and EMA, etc. were used for drug collections. Drug target information in SuperDRUG2 was extracted from DrugBank [244], TTD [247] and ChEMBL [238]. Besides these drugs and targets information, SuperDRUG2 also provides 2D and 3D structure information of small molecule drugs, drug side effects, drug–drug interactions and drug pharmacokinetic parameters.

Binding affinity databases

In this category, BindingDB [257, 297–299], PDBBind [300] and PDSP Ki [301] are included. All of them contain the data on chemical-protein binding affinities. BindingDB is mainly focused on collection of binding affinity data between drugs (drug-like molecules) and target proteins. PDBbind is established based on binding affinity measurements of biomolecular complexes from PDB. PDSP Ki is similar to BindingDB, which also contains a large number of binding affinity data on DTIs. Table 10 shows the relative information of these three databases.

BindingDB

BindingDB [257, 297–299] is a repository that contains experimental protein–small molecule interaction information. All of these data were extracted from scientific literature and US patents. In addition, other databases (e.g. ChEMBL [238], PubChem [296], etc.) are also linked with BindingDB.

PDBbind

PDBbind [300] was first released in 2004 and the purpose of this database is to bridge the gap between protein structural information and energetic properties. The data stored in PDBbind were classified by the biomolecular complex data from PDB. Then, the binding affinity data were collected from the
Table 10. Binding affinity databases

| Database   | Latest updates | No. of targets | No. of drugs/compounds | No. of DTI | No. of TTI |
|------------|----------------|----------------|------------------------|------------|-----------|
| BindingDB  | May 2019       | 7269           | 733198                 | 1651120    | -         |
| PDBBind    | Jan 2018       | -              | -                      | 16276      | 3312      |
| PDSP Ki    | 2019           | -              | -                      | -          | -         |

Figure 5. Coupled matrix–matrix versus coupled tensor–matrix.

associated literature on PDB. PDBbind has regular updates with the growth of PDB database.

**PDSP Ki**

PDSP Ki [301] is a public database that stored binding affinities data of drugs/chemical compounds for four different types of proteins, i.e. receptors, neurotransmitter transporters, ion channels and enzymes. This database was developed and maintained by University of North Carolina at Chapel Hill. Search function for both drugs and targets are provided.

**DTI database challenges and future work**

The challenges in making reliable predictions of DTI can be classified into two main categories: the challenges concerning the databases and those concerning computations. Oftentimes, one may overcome the computational difficulties using different prediction methods depending on the nature of the problem. However, major challenges arise due to the source of the databases. Here, we provide some challenges of the first type, also discussed by authors in [88, 92], followed by some suggestions on how to deal with the challenges in future work.

**Database challenges and future work**

Almost all the methods used in DTI prediction, particularly similarity-based methods, heavily rely on assertions concerning similar drugs and similar targets, the type of database used for the prediction plays a significant role. In terms of databases, lacking a uniform definition of drugs and targets as well as a consistent way of calling and identifying compounds and biomolecules, overlapping with at least one other source in the pool, adopting different identifiers to represent drug and targets are among the main challenges [88, 92]. Additionally, incorporating heterogenous data in a database is another challenge to be pointed out. Not all the drugs and targets included in a database have 3D structures and GO/PPI sequences, respectively, which makes similarity scores. As a consequence, the resulting data could vary even if the same literature is used.

Future predictions should rely on more comprehensive internal databases, which would require a significant effort to map and curate data across the sources that utilize different ways to define, name and identify the drugs and targets. From the data perspective, there is an issue of datasets being of a binary nature; i.e. given an interaction matrix $X_{n \times m}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$, one may define

$$x_{ij} = \begin{cases} 
1 & \text{if drug } d_i \text{ and target } t_j \text{ interact} \\
0 & \text{in the absence of any known interaction.}
\end{cases}$$

This causes a significant problem. Some of the 0’s in $X_{n \times m}$ may be interactions that are yet undiscovered, which may throw off the training process for the different classifiers. Another point is that in reality DT pairs have binding affinities that vary over a spectrum (interactions are not binary on/off). One suggestion to overcome this challenge is to utilize datasets with continuous values representing DT binding affinities. This have been previously proposed by authors in [5, 131, 153, 302, 303]. Our suggestion is to replace each $x_{ij}$ with continuous-valued parameters. Based on the probability of interaction, one may define $x_{ij} = \mu$ where $\mu \in [0, 1]$. 0, as it should, indicates no interaction while 1 denotes complete interaction. Any number within $(0, 1)$ represents the probability that drug $d_i$ and target $t_j$ interact.
### Table 11. The summary of all algorithms and databases

| Study                          | Algorithm       | Database                                                                 |
|--------------------------------|-----------------|-------------------------------------------------------------------------|
| Bock et al. [78]               | SVM             | PDSP Ki, Swiss-Prot (UniProt), Ligand.Info, ExPaSy                      |
| Faulon et al. [130]            | SVM             | PTC, KEGG, DrugBank                                                    |
| Nagamine et al. [129]          | SVM             | DrugBank, UniProt, PubChem, PDSP Ki, GLIDA                              |
| Nagamine et al. [127]          | SVM             | DrugBank, UniProt, NIST05, CE-MS                                       |
| Wassermann et al. [128]        | SVM             | MEROPS, CutDB, SCOP, MDDR, PDB, BindingDB                               |
| Jacob et al. [14]              | SVM             | KEGG BRITE                                                             |
| Cao et al. [137]               | SVM             | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Liu et al. [138]               | SVM             | DrugBank, Matador, STITCH, PubChem, SIDER                              |
| Mousavian et al. [136]         | SVM             | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Shen et al. [135]              | SVM             | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Ding et al. [134]              | SVM             | KEGG BRITE, BRENDA, SuperTarget, DrugBank, CheMBL, Matador             |
| Yamanishi et al. [143]         | BGL             | KEGG DRUG, KEGG LIGAND, KEGG GENES, KEGG BRITE, BRENDA, SuperTarget,   |
|                               |                 | DrugBank, JAPIC                                                        |
| Yamanishi et al. [13]          | BGL or KRM, NN  | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Bleakley et al. [101]          | BLM, KRM, NN    | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| He et al. [102]                | NN              | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank     |
| Xia et al. [6]                 | LaRLS           | KEGG LIGAND, KEGG GENES                                                 |
| Van Laarhoven et al. [153]     | GIP, RLS        | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank     |
| Perlman et al. [95]            | SITAR           | KEGG DRUG, DrugBank, DCD8, SuperTarget, Reactome, CTD                  |
| Takarabe et al. [9]            | PKR             | AERS, SIDER, JAPIC, KEGG DRUG, KEGG GENES                               |
| Gonen [194]                    | KBMF            | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank     |
| Cheng et al. [195]             | NBI, TBSI, DBSI | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Chen et al. [196]              | NWKRH           | KEGG LIGAND, KEGG BRITE, BRENDA, SuperTarget, DrugBank                 |
| Mei et al. [107]               | BLM-NII         | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank     |
| Yu et al. [131]                | SVM, RF         | DrugBank                                                                |
| Tai et al. [216]               | $L_1$-regularized | KEGG BRITE, BRENDA, SuperTarget, DrugBank                          |
| Wang et al. [204]              | RBM             | MATADOR, STITCH                                                        |
| Zheng et al. [182]             | MSCMF           | KEGG BRITE, BRENDA, SuperTarget, DrugBank                               |
| Van Laarhoven et al. [106]     | WNN-GIP         | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank     |
| Cobanoglu et al. [179]         | PMF             | DrugBank                                                                |
| Alaimo et al. [209]            | DT-Hybrid       | KEGG BRITE, BRENDA, SuperTarget, DrugBank ([195])                      |
| Chen et al. [111]              | NetCBP          | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Tai et al. [139]               | MH-SVM          | STITCH, PubChem, UniProt, PFAM                                          |
| Pahikkala et al. [5]           | RF, RLS         | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Niu et al. [112]               | EnsemRF         | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Bharadwaja [156]              | KRLS            | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Kuang et al. [155]             | RLS-Kron        | DrugBank, KEGG LIGAND, UniProt                                         |
| Peng et al. [199]              | NormMultInf     | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Zhang [176]                    | EnsemSTACK      | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Seal et al. [202]              | RWR             | DrugBank, CheMBL                                                        |
| Shi et al. [104]               | Super-Target    | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Shi et al. [97]                | SRP             | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Lan et al. [152]               | PUDT            | KEGG BRITE, BRENDA, SuperTarget, DrugBank, KEGG LIGAND                 |
| Liu et al. [187]               | NRLMf           | KEGG BRITE, BRENDA, SuperTarget, DrugBank, CheMBL, KEGG LIGAND         |
| Elzatt et al. [171]            | EnsemDT         | DrugBank                                                                |
| Ba-alawi et al. [167]          | DASPfind        | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Yuan et al. [177]              | DrugE-Rank      | DrugBank                                                                |
| Hao et al. [157]               | RLS-KF          | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Nascimento et al. [158]        | KronRLS-MKL     | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Lim et al. [216]               | COSINE          | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Buza et al. [98]               | ECKNN, HLM      | KEGG BRITE, BRENDA, SuperTarget, DrugBank, Kinase, KEGG GENES          |
| Peska et al. [2]               | BRA, BRDTI      | KEGG BRITE, BRENDA, SuperTarget, DrugBank, Kinase                      |
| Meng et al. [215]              | PDTPS           | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Zhang et al. [105]             | LPLNI           | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Ezzat et al. [172, 173]        | EnsemDT,        | DrugBank ([171]), KEGG BRITE, BRENDA, SuperTarget, DrugBank            |
|                               | EnsemKRR        | EnsemKRR                                                                |
| Ezzat et al. [189]             | GRMF            | KEGG BRITE, BRENDA, SuperTarget, DrugBank                              |
| Kuang et al. [208]             | KMDR            | DrugBank, KEGG LIGAND, UniProt                                         |

(Continued)
The trend of using such continuous-valued datasets may eventually catch on as it is more useful and more meaningful, in the sense that it represents the reality better than the binary datasets that have been used in the majority of previous work in DTI prediction. The main challenge, however, lies in the fact that to date, there is a large number of small molecule compounds that have not yet been used as drugs and for the majority of them, their interaction profiles with proteins are still unknown.

Future work on DTI predictions could be categorized in two main approaches. Modifications and suggestions toward the databases in general seem inescapable. On the one hand, the databases should be combined together to collect the most complete set of known drug–protein interactions. On the other hand, the sources should regularly be updated and disseminated, which results in improvements and completeness. A larger number of source databases should be integrated to derive the internal database.

**DTI prediction method challenges and future work**

Future research should focus on methods that combine multiple similarities. The ensemble-based models that combine multiple types of similarities are likely to provide more accurate results than the methods that use one similarity. For instance, repurposed drugs have been identified via retrospective clinical analysis (e.g. reviewing side effects), pharmacological analysis or simply serendipity. Given the surprisingly successful early examples (repurposing minoxidil from hypertension to hair loss, sildenafil from angina to erectile dysfunction and thalidomide from morning sickness to multiple myeloma), research is now focusing on how best to adopt a more comprehensive, systematic approach. In addition, a great amount of work is invested to identify molecular drivers of disease development, progression and treatment resistance, providing many candidate targets for drugs across the spectrum of human disease. However, a majority of these molecular drivers have no known drug to target them. Thus, a comprehensive, improved methodology for predicting DTIs would have great benefit. Due to challenges listed in Section 4.1, current knowledge of which cellular molecules are targeted by a drug is scarce and is derived from various, sometimes complementary sources.

As per the formulation of the problem, appropriate representation of datasets seems crucial for gaining insight and effectiveness in DTI predictions. In Big Data applications it is common that data is sparse (mostly zeros) and partially missing. Missing data imputation, especially in the context of sparse, noisy data, is therefore a central problem. To infer the missing entries from the known ones, reasonable assumptions should be made based

---

**Table 11. Continued**

| Study                  | Algorithm       | Database                                                                 |
|------------------------|-----------------|--------------------------------------------------------------------------|
| Olayan et al. [145]    | RF (DDR)        | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Zhang et al. [103]     | MultiviewDTI    | DrugBank                                                                 |
| Li et al. [180]        | LRE             | DrugBank, KEGG                                                           |
| Wen et al. [118]       | DeepDTIs        | DrugBank                                                                 |
| Luo et al. [197]       | DTINet          | DrugBank, HPRD                                                           |
| Zong et al. [117]      | DeepWalk        | DrugBank                                                                 |
| He et al. [159]        | SimBoost,       | Kinome Datasets in [307, 308]                                            |
|                        | SimBoostQuant   |                                                                          |
| Li et al. [169]        | DVM             | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Zhang et al. [164]     | DrugRPE         | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank ([102]) |
| Rayhan et al. [151]    | iDTI-ESBoost    | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Hao et al. [233]       | DNILMF          | KEGG BRITE, BRENDA, SuperTarget, DrugBank, KEGG GENES, KEGG DRUG, KEGG   |
|                        |                 | COMPOUND                                                                |
| Ohue et al. [166]      | CGBVS           | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Wang et al. [188]      | DLGRMC          | KEGG BRITE, BRENDA, SuperTarget, DrugBank, KEGG LIGAND, KEGG GENES       |
| Sharma et al. [178]    | BE-DTI’         | DrugBank, KEGG                                                           |
| Shi et al. [109]       | WBRDTI          | KEGG BRITE, BRENDA, SuperTarget, DrugBank, Kinase                        |
| Huang et al. [198]     | IN-RWR          | DrugBank, DGI, TTD                                                       |
|                        | Co-rank         |                                                                          |
| Shi et al. [144]       | LRF-DTI         | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Kadiyala [214]         | WLNMI           | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank ([106]) |
| Manoochehri et al. [219]| DMP            | KEGG BRITE, KEGG LIGAND, KEGG GENES, BRENDA, SuperTarget, DrugBank ([106]) |
| Mongia et al. [212, 213]| MGRNNM,        | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
|                        | DGRMC           |                                                                          |
| Wan et al. [207]       | NeoDTI          | DrugBank, HPRD ([197])                                                  |
| Wang et al. [116]      | AutoDNP         | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Huang et al. [191]     | Pseudo-SMR      | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Wang et al. [161]      | RFDT            | KEGG BRITE, BRENDA, SuperTarget, DrugBank                                |
| Ban et al. [201]       | NRLMF’          | KEGG BRITE, BRENDA, SuperTarget, DrugBank, KEGG LIGAND, KEGG GENES      |
| Bolgar et al. [193]    | VB-MK-LMF       | KEGG DRUG, KEGG BRITE, BRENDA, SuperTarget, DrugBank                     |
| Lee et al. [222]       | DeepConv-DTI    | DrugBank 4.0 [243], KEGG, International Union of Basic and Clinical Pharmacology (IUPHAR) [309] |
|                        |                 |                                                                          |
| You et al. [121]       | LASSO-DNN       | Drugbank                                                                 |
| Özgür et al. [120]     | DeepDTA         | Kinase [308], KIBA [310]                                                |
| Gao et al. [119]       | DeepNP          | BindingDB [257]                                                         |
| Xie et al. [124]       | DeepTrans       | DrugBank                                                                 |
on commonly observed challenges in the structure of data. Considering matrix factorization methods in predicting DTIs, a common situation is a matrix with missing entries (such as the famous Netflix problem.) Under the assumption that the completed matrix has low rank, the low-rank matrix completion problem is NP hard and highly non-convex \[ \text{NP hard and highly non-convex} \] [304], but there are various algorithms that work under certain assumptions of the data. One approach to low rank matrix completion is to use the nuclear norm as a convex relaxation of the matrix rank, and use semidefinite programming to find a completion that minimizes the nuclear norm (see \[ \text{see} \] [305, 306]). Although the low-rank matrix completion problem does not depend on any metric, most approaches utilize some kind of metric (such as the nuclear norm, the Euclidean metric or an \( \ell_2 \)-norm). Such approaches may perform well in completion of certain matrix types but do not cover all types of matrices. Moreover, the structure of the data may be more complicated than a matrix with dimension \( d = 2 \). To this end, it is our belief that coupled matrices and tensors are very powerful tools to visualize DT data while maintaining the structural information. For \( d \geq 3 \), such a dataset is a tensor (a multidimensional array) of order \( d \). Tensors are ubiquitous in Big Data. The importance of using tensors in Big Data is illustrated by the fact that they preserve the structure of the data and allow more effective data analysis by incorporating the structure throughout the process. An illustration of coupled matrix–matrix versus coupled tensor–matrix completion is shown in Figure 5.

Summary of materials and methodologies

Table 11 summarizes all the methods we reviewed in this paper along with the databases.

Key Points

- **Machine learning**: To our best knowledge, this manuscript is the first which provides a comprehensive list of all the machine learning methods that have been proposed, developed and employed to carry out the task of DTI prediction. A classification of these methods along with advantages and disadvantages of each class of method have been provided.

- **DTI software and packages**: A list and a short description of all the key software used in DTI predictions is provided. This could help future research, based on their approach to the problem, by helping researchers decide which software and packages suit their problem the best.

- **DTI databases**: One of the main challenges in the prediction of DTIs is the fact that not all the interactions between drugs and targets are known. In fact, the number of unknown interactions far exceeds the number of known interactions. As a partial solution, a comprehensive list of all databases along with the most recent update dates and the focus are provided.

References

1. Raju TN. The nobel chronicles. The Lancet 2000;355:1022.
2. Peska L, Buza K, Koller J. Drug-target interaction prediction: a bayesian ranking approach. Comput Methods Programs Biomed 2017;152:15–21.
3. Langedijk J, Mantel-Teeuwisse AK, Slijkerman DS, et al. Drug repositioning and repurposing: terminology and definitions in literature. Drug Discov Today 2015;20(8):1027–34.
4. Keiser MJ, Setola V, Irwin JJ, et al. Predicting new molecular targets for known drugs. Nature 2009;462(7270):175.
5. Pahikkala T, Airola A, Pietilä S, et al. Toward more realistic drug-target interaction predictions. Brief Bioinform 2014; 16(2):325–37.
6. Xia Z, Wu L-Y, Zhou X, et al. Semi-supervised drug–protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol 2010;4:S6. BioMed Central.
7. Blagg J. Structure–activity relationships for in vitro and in vivo toxicity. Annu Rep Med Chem 2006;41:353–68.
8. Whitebread S, Hamon J, Bojanic D, et al. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. Drug Discov Today 2005; 10(21):1421–33.
9. Takarabe M, Koteria M, Nishimura Y, et al. Drug target prediction using adverse event report systems: a pharmacogenomic approach. Bioinformatics 2012;28(18):i611–8.
10. Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Brief Bioinform 2011;12(4):303–11.
11. Swamidass SJ. Mining small-molecule screens to repurpose drugs. Brief Bioinform 2011;12(4):327–35.
12. Moriaud F, Richard SB, Adcock SA, et al. Identify drug repurposing candidates by mining the protein data bank. Brief Bioinform 2011;12(4):336–40.
13. 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(13):i232–40.
14. Jacob L, Vert J-P. Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 2008; 24(19):2149–56.
15. Ballesteros J, Palczewski K. G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. Curr Opin Drug Discov Devel 2001;4(5):561.
16. Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. Br J Pharmacol 2007;152(1):5–7.
17. Rognand D. Chemogenomic approaches to rational drug design. Br J Pharmacol 2007;152(1):38–52.
18. Nath A, Kumari P, Chaube R. Prediction of human drug targets and their interactions using machine learning methods: current and future perspectives. In: Computational Drug Discovery and Design. Springer, NY, USA. 2018, 21–30.
19. Schölkopf B, Tsuda K, Vert J-P. Kernel Methods in Computational Biology. Cambridge, MA: MIT Press, 2004.
20. Yildirim MA, Goh K-I, Cusick ME, et al. Drug-target network. Nat Biotechnol 2007;25(10):i119–26.
21. Iorio F, Bosotti R, Scarcheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci 2010;107(33):14621–6.
23. Chen B, Ding Y, Wild DJ. Assessing drug target association using semantic linked data. PLoS Comput Biol 2012;8(7):e1002574.

24. Wu Z, Cheng F, Li J, et al. SDTDB: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. Brief Bioinform 2016;18(2):335–47.

25. Bansal A, Srivastava PA, Singh TR. An integrative approach to develop computational pipeline for drug-target interaction network analysis. Sci Rep 2018;8(1):10238.

26. Swann SL, Brown SP, Muchmore SW, et al. A unified, probabilistic framework for structure-and ligand-based virtual screening. J Med Chem 2011;54(5):1223–32.

27. Cheng T, Li Q, Wang Y, Bryant SH. Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. J Chem Inf Model 2011;51(9):2440–8.

28. Cheng F, Li W, Wu Z, et al. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. J Chem Inf Model 2013;53(4):753–62.

29. van Westen GJ, Wegner JK, IJzerman AP, et al. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. MedChemComm 2011;2(1):16–30.

30. Paricharak S, Cortés-Ciriano I, Ijzerman AP, et al. Proteochemometric modelling coupled to in silico target prediction: an integrated approach for the simultaneous prediction of polypharmacology and binding affinity/potency of small molecules. J Chem 2015;7(1):15.

31. Yang M, Simm J, Lam CC, et al. Linking drug target and pathway activation for effective therapy using multi-task learning. Sci Rep 2018;8:8322.

32. Fu G, Ding Y, Seal A, et al. Predicting drug target interactions using meta-path-based semantic network analysis. BMC Bioinformatics 2016;17(1):160.

33. González-Diaz H, Prado-Prado F, García-Mera X, et al. Mind-best: web server for drugs and target discovery; design, synthesis, and assay of MAO-B inhibitors and theoretical-experimental study of GPDH protein from Trichomonas gallinae. J Proteome Res 2011;10(4):1698–718.

34. Xie L, Evangelidis T, Xie L, et al. Drug discovery using chemical systems biology: weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS Comput Biol 2011;7(4):e1002037.

35. Li H, Gao Z, Kang L, et al. Tarfishock: a web server for identifying drug targets with docking approach. Nucleic Acids Res 2006;34(suppl_2):W219–24.

36. Yang L, Wang K, Chen J, et al. Exploring off-targets and off-systems for adverse drug reactions via chemical-protein interactome–clozapine-induced agranulocytosis as a case study. PLoS Comput Biol 2011;7(3):e1002016.

37. Hansen NT, Brunak S, Altman R. Generating genome-scale candidate gene lists for pharmacogenomics. Clin Pharmacol Ther 2009;86(2):183–9.

38. Keiser MJ, Roth BL, Armbruster BN, et al. Relating protein pharmacology by ligand chemistry. Nat Biotechnol 2007;25(2):197.

39. Butina D, Segall MD, Frankcombe K. Predicting adme properties in silico: methods and models. Drug Discov Today 2002;7(11):S83–8.

40. Byvatov E, Fechner U, Sadowski J, et al. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. J Chem Inf Model 2003;43(6):1882–9.

41. Li YY, An J, Jones SJ. A computational approach to finding novel targets for existing drugs. PLoS Comput Biol 2011;7(9):e1002139.

42. Hopkins AL. Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 2008;4(11):682.

43. Li YY, Jones SJ. Drug repositioning for personalized medicine. Genome Med 2012;4(3):27.

44. Kinnings SL, Liu N, Buchmeier M, et al. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol 2009;5(7):e1000423.

45. Xie L, Bourne PE. Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. Proc Natl Acad Sci 2008;105(14):5441–6.
60. Lounkine E, Keiser MJ, Whitebread S, et al. Large-scale prediction and testing of drug activity on side-effect targets. Nature 2012;486(7403):361.
61. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. BMC Bioinformatics 2011;12(1):169.
62. Atias N, Sharan R. An algorithmic framework for predicting side effects of drugs. J Comput Biol 2011;18(3):207–18.
63. Ding Y, Tang J, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing 2019;325:211–24.
64. Chiang AP, Butte AJ. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 2009;86(5):507–10.
65. Yang K, Bai H, Ouyang Q, et al. Finding multiple target optimal intervention in disease-related molecular network. Mol Syst Biol 2008;4(1):228.
66. Li J, Zhu X, Chen JY. Building disease-specific drug-protein connectivity maps from molecular interaction networks and PubMed abstracts. PLoS Comput Biol 2009;5(7):e1000450.
67. Emig D, Ivliev A, Pustovalova O, et al. Drug target prediction and repositioning using an integrated network-based approach. PLoS One 2013;8(4):e60618.
68. Tatonetti NP, Denny J, Murphy S, et al. Detecting drug interactions from adverse-event reports: interaction between paroxetine and pravastatin increases blood glucose levels. Clin Pharmacol Ther 2011;90(1):133–42.
69. Tatonetti NP, Fernald GH, Altman RB. A novel signal detection algorithm for identifying hidden drug–drug interactions in adverse event reports. J Am Med Inform Assoc 19(1):79–85, 2011.
70. Zeng J, Li D, Wu Y, et al. An empirical study of features fusion techniques for protein–protein interaction prediction. Curr Bioinfo 2016;11(4):1–12.
71. Wei L, Xing P, Zeng J, et al. Improved prediction of protein–protein interactions using novel negative samples, features, and an ensemble classifier. Artif Intell Med 2017;83:67–74.
72. Kim S, Jin D, Lee H. Predicting drug–target interactions using drug–drug interactions. PloS One 2013;8(1):e80129.
73. Zhu S, Okuno Y, Tsujimoto G, et al. A probabilistic model for mining implicit ‘chemical compound–gene’ relations from literature. Bioinformatics 2005;21(suppl_2):ii245–51.
74. Li L, Zhou T. Link prediction in complex networks: a survey. Physica A 2011;390(6):1150–70.
75. Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 2005;6(6):734–49.
76. Su X, Khoshgoftaar TM. A survey of collaborative filtering techniques. Adv Artif Intell 2009;2009.
77. Nguyen J, Zhu M. Content-boosted matrix factorization techniques for recommender systems. Stat Anal Data Min 2013;6(4):286–301.
78. Bock JR, Gough DA. Virtual screen for ligands of orphan g protein-coupled receptors. J Chem Inf Model 2005;45(5):1402–14.
79. Kuhn M, Campillos M, González P, et al. Large-scale prediction of drug–target relationships. FEBS Lett 2008;582(8):1283–90.
80. Iskar M, Zeller G, Zhao X-M, et al. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. Curr Opin Biotechnol 2012;23(4):609–16.
81. Koutsoukas A, Simma B, Kirchmair J, et al. From in silico target prediction to multi-target drug design: current databases, methods and applications. J Proteomics 2011;74(12):2554–74.
82. Dai Y-F, Zhao X-M. A survey on the computational approaches to identify drug targets in the postgenomic era. Biomed Res Int 2015;2015.
83. Cichonska A, Rousu J, Aittokallio T. Identification of drug candidates and repurposing opportunities through compound–target interaction networks. Expert Opin Drug Discovery 2015;10(12):1333–45.
84. 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(5):734–47.
85. Yamanishi Y. Chemogenomic approaches to infer drug–target interaction networks. In: Data Mining for Systems Biology. Springer, Totowa, NJ, 2013, 97–113.
86. Zhang W, Lin W, Zhang D, et al. Recent advances in the machine learning-based drug–target interaction prediction. Curr Drug Metab 2019;20(3):194–202.
87. Chen R, Liu X, Jin S, et al. Machine learning for drug-target interaction prediction. Molecules 2018;23(9):2208.
88. Wang C, Kurgan L. Survey of similarity-based prediction of drug–protein interactions. Curr Med Chem 2019;26(1). https://doi.org/10.2174/0929867326666190808154841
89. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. Drug Discov Today 2015;20(3):318–31.
90. Mousavian Z, Masoudi-Nejad A. Drug–target interaction prediction via chemogenomic space: learning-based methods. Expert Opin Drug Metab Toxicol 2014;10(9):1273–87.
91. Chen X, Yan CC, Zhang X, et al. Drug–target interaction prediction: databases, web servers and computational models. Brief Bioinform 2015;17(4):696–712.
92. Ezzat A, Wu M, Li X-L, et al. Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 2018;9.
93. Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform, 2019, 93:103159.
94. Serçinogló O, Sarica PO. In silico databases and tools for drug repurposing. In: In Silico Drug Design. Elsevier, London, United Kingdom, 2019, 703–42.
95. Perlman L, Gottlieb A, Atias N, et al. Combining drug and gene similarity measures for drug-target elucidation. J Comput Biol 2011;18(2):133–45.
96. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;28(8):1226–38.
97. Shi J-Y, Yiu S-M. SRP: a concise non-parametric similarity-rank-based model for predicting drug-target interactions. In: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, NY, USA, 2015, 1636–41.
98. Buza K, Peška L. Drug–target interaction prediction with bipartite local models and hubness-aware regression. Neurocomputing 2017;260:284–93.
99. Buza K. Drug–target interaction prediction with hubness-aware machine learning. In: 2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE, NY, USA, 2016, 437–40.
100. Buza K, Nanopoulos A, Nagy G. Nearest neighbor regression in the presence of bad hubs. Knowl-Based Syst 2015;86:250–60.
142. Yamanishi Y. Supervised bipartite graph inference. In: Advances in Neural Information Processing Systems, NIPS, Vancouver, BC, CA, 2009, 1841–8.
143. Yamanishi Y, Kotera M, Kancheva M, et al. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 2010;26(12):i246–54.
144. Shi H, Liu S, Chen J, et al. Predicting drug–target interactions using lasso with random forest based on evolutionary information and chemical structure. Genomics 2018; 111(6):1839–1852.
145. 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(7):1164–73.
146. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292(2): 195–202.
147. O’Boyle NM, Banck M, James CA, et al. Open babel: an open chemical toolbox. J Chem 2011;3(1):33.
148. Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Series B Stat Methodol 2011;73(3):273–82.
149. Chawla NV, Bowyer KW, Hall LO, et al. Smote: synthetic minority over-sampling technique. J Artif Intell Res 2002; 16:321–57.
150. Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
151. Rayhan F, Ahmed S, Shatabda S, et al. iDTI-ESBoost: identification of drug target interaction using evolutionary and structural features with boosting. Sci Rep 2017;7(1):17731.
152. Lan W, Wang J, Li M, et al. Predicting drug–target interaction using positive-unlabeled learning. Neurocomputing 2016;206:50–7.
153. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 2011;27(21):3036–43.
154. Belkin M, Niyogi P, Sindhwani V. Manifold regularization: a similarity based learning method for labeled examples. J Mach Learn Res 2006;7:2399–434.
155. Kuang Q, Xu X, Li R, et al. An eigenvalue transformation technique for predicting drug–target interaction. Sci Rep 2015;5:13867.
156. Belkin M, Niyogi P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in Neural Information Processing Systems, NIPS, Vancouver, BC, CA, 2002, 585–91.
157. Deprem E, Gunay A, Eren A, et al. Improved prediction of drug–target interaction using kernel fusion technique. J Theor Biol 2013;320:41–6.
158. Wang L, You Z-H, Chen X, et al. RFDT: a rotation forest-based predictor for predicting drug–target interactions using drug structure and protein sequence information. Curr Protein Pept Sci 2018;19(5):445–54.
181. Liu R, Hao R, Su Z. Mixture of manifolds clustering via low rank embedding. J Inform Comput Sci 2011;8:725–37.

182. 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, Chicago, Illinois, USA, 2013, 1025–33.

183. Ding CH, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. IEEE Trans Pattern Anal Mach Intell 2010;32(1):45–55.

184. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. In: Linear Algebra. Springer, Berlin, Heidelberg, 1971, 134–51.

185. Ye J. Generalized low rank approximations of matrices. Mach Learn 2005;61(1–3):167–91.

186. Mnih A, Salakhutdinov RR. Probabilistic matrix factorization. In: Advances in Neural Information Processing Systems, NIPS, Vancouver, BC, CA, 2008, 1257–64.

187. Liu Y, Wu M, Xiao C, et al. Neighborhood regularized logistic matrix factorization for drug–target interaction prediction. PLoS Comput Biol 2016;12(2):e1004760.

188. Wang M, Tang C, Chen J. Drug–target interaction prediction via dual laplacian graph regularized matrix completion. Biomed Res Int 2018;2018.

189. Ezzat A, Zhao P, Wu M, et al. Drug–target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinform 2017;14(3):646–56.

190. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.

191. Huang Y-A, You Z-H, Chen X. A systematic prediction of drug–target interactions using molecular fingerprints and protein sequences. Curr Protein Pept Sci 2018;19(5):468–78.

192. Rendle S, Freudenthaler C, Gantner Z, et al. BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence. AUAI Press, McGill, Canada, 2009, 452–61.

193. Bolgä R, Antal P. VB-MK-LMF: fusion of drugs, targets and interactions using variational Bayesian multiple kernel logistic matrix factorization. BMC Bioinformatics 2017;18(1):440.

194. Gönen M. Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics 2012;28(18):2304–10.

195. 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(5):e1002503.

196. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. Mol Biosyst 2012;8(7):1970–8.

197. Luo Y, Zhao X, Zhou J, et al. A network integration approach for drug–target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 2017;8(1):573.

198. Huang Y, Zhu L, Tan H, et al. Predicting drug-target on heterogeneous network with co-rank. In: International Conference on Computer Engineering and Networks. Springer, Cham, Switzerland, 2018, 571–81.

199. Peng L, Liao B, Zhu W, et al. Predicting drug–target interactions with multi-information fusion. IEEE J Biomed Health Inform 2015;21(2):561–72.

200. Wright J, Ganesh A, Rao S, et al. Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In: Advances in Neural Information Processing Systems, NIPS, Vancouver, BC, CA. 2009, 2080–8.
221. Yasuo N, Nakashima Y, Sekijima M. CoDe-DTI: Collaborative deep learning-based drug–target interaction prediction. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, NY, USA, 2018, 792–7.

222. Sakakibara Y, Hachiya T, Uchida M, et al. COPICAT: a software system for predicting interactions between proteins and chemical compounds. Bioinformatics 2012; 28(5):745–6.

223. Xiao X, Min J-L, Wang P, Chou KC. Igpcr-drug: a web server on supervised analysis. Nucleic Acids Res 2014; 42(W1): W39–45.

224. Hao M, Bryant SH, Wang Y. Predicting drug–target interactions by dual-network integrated logistic matrix factorization. Nature Neus, 2017; 4036.

225. Kanehisa M, Goto S, Hattori M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 2016; 45(D1):D353–61.

226. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014; 42(D1): D1083–90.

227. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2013; 41(D1):D1091–7.

228. Gaulton A, Hersey A, Nowotka M, et al. The ChEMBL bioactivity database: an update. Nucleic Acids Res 2014; 42(D1): D1083–90.

229. Keller JM, Gray MR, Givens JA. A fuzzy K-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 1985; 4(4): 580–5.

230. Scheiber J, Jenkins JL, Sukuru SCK, et al. Mapping adverse drug reactions in chemical space. J Med Chem 2009; 52(9):3103–7.

231. Seal A, Wild DJ. Netpredictor: R and Shiny package for predicting interaction between gpcrs and drugs in cellular networking. PLoS One 2013; 8(8):e72234.

232. Hao M, Bryant SH, Wang Y. Drug-target interaction network inference engine based on supervised analysis. Nucleic Acids Res 2014; 42(W1): W39–45.

233. Kuhn M, von Mering C, Campillos M, et al. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res 2015; 44(D1): D380–4.

234. Cotto KC, Wagner AH, Feng Y-Y, et al. DGIldb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Res 2017; 46(D1):D1068–73.

235. Kim Kjærulff S, Wich L, Kringelum J, et al. ChemProt-2.0: a global chemical biology diseases mapping. Database 2016; 2016.

236. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2011; 40(D1): D1100–7.

237. Bento AP, Gaulton A, Hersey A, et al. The chembl bioactivity database: an update. Nucleic Acids Res 2014; 42(D1): D1083–90.

238. Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2011; 40(D1): D1100–7.

239. Gunther S, Kuhn M, Dunkel M, et al. Supertarget and matador: resources for exploring drug-target relationships. Nucleic Acids Res 2007; 36(suppl_1):D919–22.

240. Knox C, Law V, Jewison T, et al. Drugbank 3.0: a comprehensive resource for ‘omics’ research on drugs. Nucleic Acids Res 2010; 39(suppl_1):D1035–41.

241. Law V, Knox C, Djoumbou Y, et al. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2013; 42(D1): D1091–7.

242. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res 2017; 46(D1): D1074–82.

243. Wishart DS, Knox C, Guo AC, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006; 34(suppl_1):D668–72.

244. Wishart DS, Knox C, Guo AC, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2007; 36(suppl_1): D901–6.

245. Chen X, Ji ZL, Chen YZ. TTDB: therapeutic target database. Nucleic Acids Res 2002; 30(1):412–5.

246. Kuhn M, Szklarczyk D, Franceschini A, et al. STITCH 2: an interaction network database for small molecules and proteins. Nucleic Acids Res 2009; 38(suppl_1):D552–6.

247. Kuhn M, Szklarczyk D, Franceschini A, et al. STITCH 3: zooming in on protein–chemical interactions. Nucleic Acids Res 2011; 40(D1): D876–80.

248. Kuhn M, Szklarczyk D, Pletscher-Frankild S, et al. STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res 2013; 42(D1): D401–7.

249. Kringelum J, Kjaerulff SK, Brunak S, et al. ChemProt-2.0: a global chemical biology diseases mapping. Database 2016; 2016.

250. Cotto KC, Wagner AH, Feng Y-Y, et al. DGIldb 3.0: a redesign and expansion of the drug–gene interaction database. Nucleic Acids Res 2017; 46(D1):D1068–73.

251. Kim Kjærulff S, Wich L, Kringelum J, et al. ChemProt-2.0: visual navigation in a disease chemical biology database. Nucleic Acids Res 2012; 41(D1): D464–9.

252. Taboureau O, Nielsen SK, Audouze K, et al. ChemProt: a disease chemical biology database. Nucleic Acids Res 2010; 38(suppl_1): D367–72.

253. Gilton MK, Liu T, Baitaluk M, et al. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 2015; 44(D1): D1045–53.

254. Roth BL, Lopez E, Beischel S, et al. Screening the receptorome to discover the molecular targets for plant-derived psychoactive compounds: a novel approach for CNS drug discovery. Pharmacol Ther 2004; 102(2): 99–110.

255. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res 2002; 30(1):163–5.

256. Tatusova T. Genomic databases and resources at the National center for biotechnology information. Data Mining Techniques for the Life Sciences. Springer, New York, USA, 2010, 17–44.

257. Davis AP, Murphy CG, Saraceni-Richards CA, et al. Comparative toxicogenomics database: a knowledgebase and
discovery tool for chemical–gene–disease networks. Nucleic Acids Res 2008;37(suppl_1):D786–92.

262. Olah M, Rad R, Ostopovici L, et al. WOMBAT and WOMBAT-PK: bioactivity databases for lead and drug discovery. In: Chemical Biology: From Small Molecules to Systems Biology and Drug Design 2007;1:760–86.

263. Wagner AH, Coffman AC, Ainscough BJ, et al. DGIdb 2.0: mining clinically relevant drug–gene interactions. Nucleic Acids Res 2015;44(D1):D1036–44.

264. Griffith M, Griffith OL, Coffman AC, et al. DGIdb: mining the druggable genome. Nat Methods 2013;10(12): 1209.

265. Orchard S, Ammari M, Aranda B, et al. The mintact project—intact as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res 2013;42(D1): D358–63.

266. Pillai I, Chouvarine P, Tudor CO, et al. Developing a bio-curated workflow for agbase, a non-model organism database. Database 2012;2012.

267. McCarthy FM, Bridges SM, Wang N, et al. AgBase: a unified resource for functional analysis in agriculture. Nucleic Acids Res 2006;35(suppl_1):D599–603.

268. McCarthy FM, Gresham CR, Buza TJ, et al. AgBase: supporting functional modeling in agricultural organisms. Nucleic Acids Res 2010;39(suppl_1): D497–506.

269. McCarthy FM, Wang N, Magee GB, et al. AgBase: a functional genomics resource for agriculture. BMC Genomics 2006;7(1):229.

270. Licata L, Briganti L, Peluso D, et al. MINT, the molecular interaction database: 2012 update. Nucleic Acids Res 2011;40(D1):D857–61.

271. Ceol A, Chatr-Aryamontri A, Licata L, et al. MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 2009;38(suppl_1):D532–9.

272. Chatr-Aryamontri A, Ceol A, Palazzi LM, et al. MINT: the molecular interaction database. Nucleic Acids Res 2006;35(suppl_1):D572–4.

273. Zanzoni A, Montecchi-Palazzi L, Quondam M, et al. MINT: a molecular interaction database. FEBS Lett 2002;513(1): 135–40.

274. Dimmer EC, Huntley RP, Alam-Faruque Y, et al. The UniProt-GO annotation database in 2011. Nucleic Acids Res 2011;40(D1):D565–70.

275. Kotlyar M, Pastrello C, Sheahan N, et al. Integrated interactions database: tissue-specific view of the human and model organism interactomes. Nucleic Acids Res 2015;44(D1):D356–41.

276. Launay G, Salza R, Multedo D, et al. MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res 2014;43(D1):D321–7.

277. Breuer K, Foroushani AK, Laird MR, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res 2012; 41(D1):D1228–33.

278. Orchard S, Kerrien S, Abbani S, et al. Protein interaction data curation: the international molecular exchange (iMEx) consortium. Nat Methods 2012;9(4):345.

279. Kim S, Thiessen PA, Bolton EE, et al. PubChem substance and compound databases. Nucleic Acids Res 2015;44(D1):D1202–12.

280. Deshpande N, Address KJ, Bluhm WF, et al. The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res 2005;33(suppl_1):D233–7.

281. Michalsky E, Dunkel M, Goede A, et al. SuperLigands—a database of ligand structures derived from the protein data bank. BMC Bioinformatics 2005;6(1):122.

282. Li YH, Yu CY, Li XQ, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. Nucleic Acids Res 2017;46(D1):D1121–7.

283. Jeske L, Placzek S, Schomburg I, et al. Brenda in 2019: a European ELIXIR core data resource. Nucleic Acids Res 2018;47(D1):D542–9.

284. Siramshetty VB, Eckert OA, Gohlke B-O, et al. SuperDUG2: a one stop resource for approved/marketed drugs. Nucleic Acids Res 2017;46(D1):D1137–43.

285. Ursu O, Holmes J, Bologa CG, et al. DrugCentral 2018: an update. Nucleic Acids Res 2018;47(D1):D963–70.

286. Ursu O, Holmes J, Knockel J, et al. DrugCentral: online drug compendium. Nucleic Acids Res 2016;44(1):kw993.

287. Wang C, Hu G, Wang K, et al. PDID: database of molecular-level putative protein–drug interactions in the structural human proteome. Bioinformatics 2015;32(4):579–86.

288. Nguyen D-T, Mathias S, Bologa C, et al. Pharos: collating protein information to shed light on the druggable genome. Nucleic Acids Res 2016;45(D1):D995–D1002.

289. Verbruggen B, Gunnarsson L, Kristiansson E, et al. ECO-drug: a database connecting drugs and conservation of their targets across species. Nucleic Acids Res 2017;46(D1): D930–6.

290. Schomburg I, Chang A, Ebeling C, et al. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res 2004;32(suppl_1):D431–3.

291. Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16(1):19.

292. Hu G, Gao J, Wang K, et al. Finding protein targets for small biologically relevant ligands across fold space using inverse ligand binding predictions. Structure 2012;20(11): 1815–22.

293. Feinstein WP, Brylinski M. eFindSite: enhanced fingerprint-based virtual screening against predicted ligand binding sites in protein models. Mol Inform 2014;33(2):135–50.

294. Brylinski M, Feinstein WP. eFindSite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. J Comput Aided Mol Des 2013;27(6):551–67.

295. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database 2016;2016.

296. Capecci A, Awale M, Probst D, et al. PubChem and CHEMBL beyond Lipinski. Mol inform 2019;38(5):1900016.

297. 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(suppl_1): D198–201.

298. Chen X, Liu M, Gilson MK. BindingDB: a web-accessible molecular recognition database. Comb Chem High Throughput Screen 2001;4(8):719–25.

299. Nicola G, Liu T, Hwang L, Gilson M. BindingDB: a protein–ligand database for drug discovery. Biophys J 2012;102(3):61a.

300. Liu Z, Li Y, Han L, et al. PDB-wide collection of binding data: current status of the pdbbind database. Bioinformatics 2014;31(3):405–12.
301. Roth BL, Lopez E, Patel S, et al. The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? Neuroscientist 2000;6(4):252–62.
302. Pahikkala T, Waegeman W, Airola A, et al. Conditional ranking on relational data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Heidelberg, Germany, 2010, 499–514.
303. Pahikkala T, Airola A, Stock M, et al. Efficient regularized least-squares algorithms for conditional ranking on relational data. Mach Learn 2013;93(2–3):321–56.
304. Friedland S, Lim L-H. Nuclear norm of higher-order tensors. Math Comput 2018;87(311):1255–81.
305. Fazel, M. and Hindi, H. and Boyd, S., Rank minimization and applications in system theory. Proceedings of the 2004 American control conference, IEEE, vol. 4, pp. 3273–3278, 2004.
306. Candès EJ, Benjamin R. Exact matrix completion via convex optimization. Found Comput Math 9(6):717–72.
307. Metz JT, Johnson EF, Soni NB, et al. Navigating the kinome. Nat Chem Biol 2011;7(4):200.
308. Davis MI, Hunt JP, Herrgard S, et al. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 2011; 29(11):1046.
309. Southan C, Sharman JL, Benson HE, et al. The IUPHAR/BPS guide to PHARMACOLOGY in 2016: towards curated quantitative interactions between 1300 protein targets and 6000 ligands. Nucleic Acids Res 2015;44(D1): D1054–68.
310. Tang J, Szwajda A, Shakyawar S, et al. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model 2014;54(3): 735–43.