Machine Learning for Drug-Target Interaction Prediction

Ruolan Chen [1] , **Xiangrong Liu** [1] , **Shuting Jin** [1] **, Jiawei Lin** [1] **and Juan Liu** [2,*]

[1] Department of Computer Science, School of Information Science and Technology, Xiamen University, Xiamen 361005, China; chenruolan@stu.xmu.edu.cn (R.C.); xrliu@xmu.edu.cn (X.L.); stjin.xmu@gmail.com (S.J.); 23020161153321@stu.xmu.edu.cn (J.L.)

[2] Department of Instrumental and Electrical Engineering, School of Aerospace Engineering, Xiamen University, Xiamen 361005, China

* Correspondence: cecyliu@xmu.edu.cn

Abstract: Identifying drug-target interactions will greatly narrow down the scope of search of candidate medications, and thus can serve as the vital first step in drug discovery. Considering that in vitro experiments are extremely costly and time-consuming, high efficiency computational prediction methods could serve as promising strategies for drug-target interaction (DTI) prediction. In this review, our goal is to focus on machine learning approaches and provide a comprehensive overview. First, we summarize a brief list of databases frequently used in drug discovery. Next, we adopt a hierarchical classification scheme and introduce several representative methods of each category, especially the recent state-of-the-art methods. In addition, we compare the advantages and limitations of methods in each category. Lastly, we discuss the remaining challenges and future outlook of machine learning in DTI prediction. This article may provide a reference and tutorial insights on machine learning-based DTI prediction for future researchers.

1. Introduction

Most drugs demonstrate efficacy via the in-vivo interactions with their target molecules such as enzymes, ion channels, nuclear receptors and G protein-coupled receptors (GPCRs). Therefore, identifying drug-target interactions (DTIs) has become a vital precondition in cognate areas including poly-pharmacology, drug repositioning, drug discovery, side-effect prediction and drug resistance [1]. The experimentation and confirmation of drug-target pairs have been great hindrances to many drug researches. On top of that biochemical experiments for undiscovered drug-target interactions involve significantly costly, time-consuming and challenging work. For instance, it takes around 1.8 billion dollars for each new molecular entity (NME) [2] as well as an average time span of 9 to 12 years for the approval of a new drug application (NDA) [3].

Besides the known interactions already stored in various databases, there exist countless unpaired small molecule compounds that could potentially be discovered and developed into new medications. Only a small number of drug-target pairs have been experimentally validated in the current data set. In fact, although there are more than 90 million compounds described in the PubChem database, a large proportion of interactions still remain to be discovered [4]. Furthermore, the number of truly innovative drugs approved by regulatory agencies has decreased in recent years, despite the progress in biotechnology. For instance, it is reported that US Food and Drug Administration (FDA) only approves approximately 20 novel drugs every year with high investment costs [5]. These large time, money and resource costs, both human and material, have motivated researchers to constantly develop

innovative technology for the exploitation of new drugs. Interaction prediction helps to screen new drugs candidates effectively and efficiently.

Identifying new targets for existing or abandoned drugs, namely drug repositioning, is another important part in drug discovery. The "multi-target, multi-drug" in place of "one target, one drug" model has been widely accepted as our understanding of pharmacology deepens [1]. The important fact is that drugs typically target multiple proteins rather than only one. The anticancer drugs sunitinib (Sutent) and imatinib (Gleevec) are both concrete evidence. What's more, drugs may interact with other proteins in addition to the primary therapeutic targets, namely off-target effects. Off-target effects are typically considered harmful side effects. However, in some cases, they may be beneficial since they could lead to unexpected therapeutic effects and provide a new perspective on the molecular mechanisms of drug side effects. The purpose of drug repositioning is the detection for new clinical uses for existing drugs. An obvious benefit of drug repositioning is that existing drugs have already been strictly verified for their safety and bioavailability. Omitting some previously completed steps can greatly speed up the drug development process. Governments, academic institutions and non-trading organizations around the world have made more effort into drug repositioning recently which will effectively facilitate the repositioning research [6].

For all the reasons mentioned above, detecting drug-target interactions is fundamental to both new drug discovery and old drug repositioning. The known drug-target interactions based on wet-lab experiments are limited to a very small number. The huge gap between known and unknown drug-target pairs has prompted interest in DTI prediction. Traditional prediction strategies in vitro have faced the limitations of time and monetary costs, while recently developed computational or in silico methods can more efficiently predict potential interaction candidates. Computational methods have achieved favorable performance in many related bioinformatics fields, such as disease-related miRNA prediction [7–9], disease genes prediction [10], protein-protein interaction prediction [11] and protein subcellular location prediction [12]. They greatly narrow the broad scope of research of experimental DTI validation. Therefore, there is a continuous and urgent demand for the development of computational techniques on DTI predictions.

Currently, the ligand-based, docking simulation, and chemogenomic approaches are the three main classes of computational methods for predicting DTIs. Ligand-based methods [13] like Quantitative Structure Activity Relationship (QSAR) utilize the idea that similar molecules usually bind to similar proteins. Specifically, these methods predict interactions by comparing a new ligand to known proteins ligands. However, ligand-based methods perform poorly when the number of known ligands is insufficient.

As for docking simulation methods [14], the three-dimensional (3D) structures of proteins are required for simulation hence becoming inapplicable when there are numerous proteins with unavailable 3D structures. Moreover they cannot be applied to membrane proteins like ion channel and G-Protein Coupled Receptors (GPCRs) whose structures are too complex to obtain. Docking simulations usually take significant time and thus it can be especially inefficient.

To address the difficulties of traditional methods, chemogenomic approaches [15] have recently been performed successfully in drug discovery and repositioning on a large scale. There are four main types of target frequently involved in DTI prediction, namely protein, disease, gene and side effect. For the purpose of drug-target pair prediction, these methods integrate both the chemical space of compounds and the genomic space of target proteins into a unified space: pharmacological space. Hence, chemogenomic approaches can make full use of abundant biological data that is favorable for prediction. In such a DTI prediction problem, the major challenge is the scarcity of known drug-protein interactions and unverified negative drug-target interaction samples. These chemogenomic approaches can be classified into different categories, such as machine learning-based methods, graph-based methods and network-based methods [16]. Among all the chemogenomic approaches, machine learning-based methods have gained the most attention for their reliable prediction results. Most of these methods generally utilize the chemical and biological features of drugs and targets, and adopt

various machine learning techniques to predict interactions between drugs and targets. Figure 1 is a branch diagram of recent computational methods for DTI prediction.
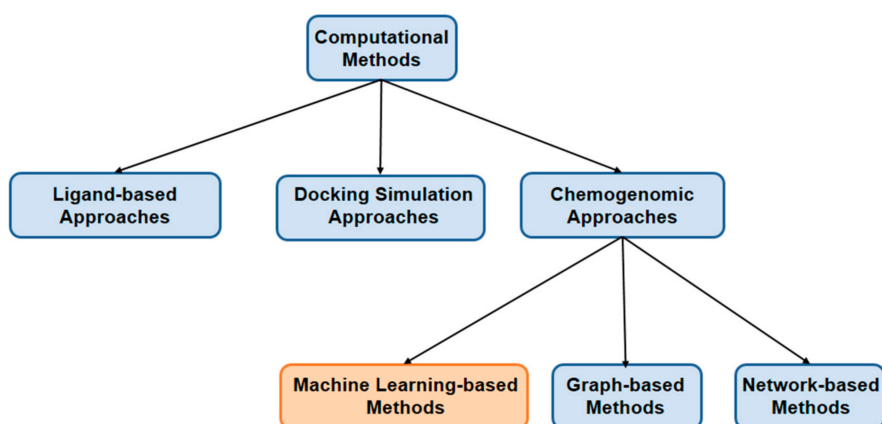

Figure 1. Branch diagram of recent computational methods for DTI prediction.

In this review, we focus on machine learning methods applied to DTI prediction. To be specific, we aim to provide a comprehensive overview on a subclass of chemogenomic approaches exploiting machine learning frameworks. Compared with those ligand-based methods that also apply machine learning strategies, the methods discussed in this review can be applicable to target proteins with insufficient known ligands. Firstly, we summarize a brief list of databases frequently used in drug discovery. Next, we adopt a hierarchical classification scheme. In particular, we classify the machine learning methods into two major categories i.e., supervised and semi-supervised methods, and provide more subclasses. We attempt to introduce several representative methods of each category, respectively. Furthermore, we present the advantages and disadvantages for methods of each category. Finally, we will discuss the challenges and further outlook for current machine learning methods in DTI prediction domain from our point of view.

1. Supervised Learning Methods Both positive labels and negative labels are required in the training set. Then these labeled samples are used to train the learning models for subsequent DTI prediction.

   - Similarity-based methods The similarities among drugs or among targets are calculated via various similarity measurement strategies. Similarity matrices can be utilized in various types of kernel functions:

     (i) The nearest neighbor methods: The nearest neighbor methods make predictions based on the information of the nearest neighbors.

     (ii) Bipartite local models: Two local models are firstly trained for drugs and targets respectively. The final prediction result for each drug-target pair is computed based on the operation of the two independent prediction scores.

     (iii) Matrix factorization methods: Drug-target interaction matrix is factorized into two latent feature matrices that when multiplied together can approximate the original matrix.

   - Feature vector-based methods The training data is represented as feature vectors. Then some machine learning models, like Random Forest, can be utilized for prediction based on these vectors.

2. Semi-Supervised Learning Methods Semi-supervised learning methods make predictions only based on a small amount of labeled data and a large amount of unlabeled data. To our best

knowledge, there are already some excellent reviews on chemogenomic approaches for DTI prediction [6,15–19]. Compared to previous works, we focus on the special topic of machine learning methods used in DTI prediction. Besides, we utilize a hierarchical classification scheme and summarize several latest prediction methods such as [20–23] which are hardly mentioned in any previous review. In particular, review [17] is written only from a narrow viewpoint, namely similarity-based approaches, which are a subclass of machine learning methods. Surveys [6,15,18,19] all provide a more general and comprehensive overview of chemogenomic approaches rather than emphasizing machine learning. In recent years, machine learning has made breakthroughs and attracted a lot of public attention. Discussing state-of-the-art DTI prediction strategies from this special perspective can demonstrate more methodology details. Although review [16] also focuses on learning-based methods, its emphasis is only on supervised learning. In comparison, we provide more detailed sub-classes and introduce newly developed methods after review [16] was published. The rest of this article is organized as follows: The "Databases" section describes current available data sources for DTI prediction research. The "Methods" section briefly introduces several representative machine learning methods via a hierarchical classification scheme. Then we discuss advantages and limitations of methods in each category as well as remaining challenges. Finally, the "Conclusions and Outlook" section makes a future perspective for machine leaning in DTI prediction.

2. Databases

Data mining and utilization based on the existing bioinformatics databases is a significant methodology for drug discovery. With the development of molecular biology, abundant information about drugs and targets has accumulated. Thus, it is necessary to establish databases for managing and maintaining the data. There exist a number of different professional databases involving potential cellular targets for various families of chemical compounds up to now. A large portion of them are publicly available. Moreover, the data size is increasing owing to the contributions of researchers from around the world. As more information about drugs and targets is collected, there are more opportunities for drug discovery research. To a certain degree, these databases have promoted the development of latest methodologies for drug discovery. In Table 1, we list frequently used databases, their web servers and brief descriptions. Table 2 shows the statistics of the number of compounds, targets and compound-target interactions in these databases. Note that not all databases provide complete information in their databases and published papers.

Some of these databases are being updated frequently, such as DrugBank, KEGG, and STITCH and so on, while the data in other databases has remained almost the same for several years, such as SuperPred which was last updated in April 2014. It is, however, encouraging that more new databases and easy-to-use web servers have been recently established. On one hand, the existing databases provide plentiful data sources of drug space and target space. It is time for the researchers to make efforts to integrate more different types of heterogeneous data. On the other hand, current databases do not involve any non-interaction information. This common drawback has limited the prediction result of supervised learning methods. Thus it would be meaningful to make public both interactions and non-interactions between drugs and targets in the future.

Table 1. Databases supporting drug discovery methods.

Database and URL	Brief Descriptions
KEGG [29] http://www.genome.jp/kegg	An encyclopedia of genes and genomes for both functional interpretation and practical application of genomic information.
BRENDA [30] http://www.brenda-enzymes.org/	The main enzyme and enzyme-ligand information system.
PubChem [31] https://pubchem.ncbi.nlm.nih.gov/	A database for information on chemical substances and their biological activities involving three inter-linked databases, i.e., Substance, Compound and BioAssay.
TTD [32] http://bidd.nus.edu.sg/group/ttd/ttd.asp	Therapeutic Target Database providing comprehensive information about the drug resistance mutations, gene expressions and target combinations data.
DrugBank [33] http://www.drugbank.ca	Consisting of two parts information involving detailed drug data (i.e., chemical, pharmacological and pharmaceutical) and drug target information (i.e., sequence, structure, and pathway) respectively.
SuperTarget [34] http://bioinf-apache.charite.de/supertarget	A database integrating drug-related information with more than 330,000 compound-target protein relations.
ChEMBL [35] https://www.ebi.ac.uk/chembldb	Data resource for molecule structures and molecule-protein interactions collected from the primary published literature on a regular basis.
STITCH [36] http://stitch.embl.de/	Repository of known and predicted chemical-protein interactions.
MATADOR [37] http://matador.embl.de/	A database of protein-chemical interactions including as many direct and indirect interactions as possible.
BindingDB [38] http://www.bindingdb.org/bind	A public database of protein-ligand binding affinities.
TDR targets [39] http://tdrtargets.org/	A chemogenomics resource for neglected tropical diseases.
SIDER [40] http://sideeffects.embl.de/	Serving information on marketed medicines and their recorded adverse drug reactions.
ChemBank [41] http://chembank.broad.harvard.edu/	Collections of available data derived from small molecules and small-molecule screens and resources for studying their properties.
DCDB [42] http://www.cls.zju.edu.cn/dcdb/	The Drug Combination Database for collecting and organizing known examples of drug combinations.
CancerDR [43] http://crdd.osdd.net/raghava/cancerdr/	Cancer Drug Resistance Database of 148 anticancer drugs and their effectiveness against around 1000 cancer cell lines.
ASDCD [44] http://asdcd.amss.ac.cn/	The first Antifungal Synergistic Drug Combination Database including published synergistic antifungal drug combinations, targets, indications, and other pertinent data.
SuperPred [45] http://prediction.charite.de/	Resource of compound-target interactions.

Table 2. The statistics of the number of compounds, targets and compound-target interactions in the databases covered in the review.

Databases	The Number of Compounds	The Number of Targets	The Number of Compound-Target Interactions
KEGG	18,380	26,885,475	
BRENDA		7341	
PubChem	96,479,316	68,868	
TTD	34,019	3101	
DrugBank	11,682	26,889	131,724
SuperTarget	195,770	6219	332,828
ChEMBL	2,275,906	12,091	
STITCH	500,000	9,600,000	1,600,000,000
MATADOR	775		
BindingDB	652,068	7082	1,454,892
TDR targets	2,000,000	5300	
SIDER	5868	1430	139,756
ChemBank	1,700,000		
DCDB	904	805	
CancerDR	148	116	
ASDCD	105	1225	210
SuperPred	341,000	1800	665,000

3. Methods

In the era of big data, machine learning methods are designed to generate predictive models based on some underlying algorithm and a given big data set. For biological and biomedical research, machine learning plays a pivotal role in filtering large amounts of data into patterns [24–27]. The general machine learning workflow in DTI prediction can be divided into three steps. First, preprocessing the input data of the drug and the target; second, training the underlying model based on a set of learning rules; third, utilizing the predictive model to make predictions for a test data set.

From our research, study [28] is the first work that applies machine learning to protein-chemical interaction prediction. This work establishes a SVM analysis framework of amino acid sequence data, chemical structure data and mass spectrometry data. This pioneering study has inspired subsequent studies. Machine learning for drug discovery has become a field of long-standing and growing interest since then.

For simplicity, we classify machine learning methods for drug-target interaction prediction into two major categories, i.e., supervised learning and semi-supervised methods. Specifically, the supervised learning methods can be further classified into two sub-classes including similarity-based methods and feature-based methods.

3.1. Supervised Learning Methods

Supervised learning methods are applied to train the learning model and identify patterns when labels are available. For the DIT prediction problem, known drug-target interactions are labeled as positive samples and the rest are labeled as negative ones. Next, these labels are used to train the model for subsequent interaction predicting. In fact, those drug-target pairs without explicit interaction information may correspond to unknown or missing interactions rather than

non-interactions. In general results of non-interactions between drugs and targets are not published. Methods of this category regard all the unknown drug-target interactions as non-interaction despite inaccuracy. In the section, we will review the supervised methods proposed so far in two categories, i.e., similarity-based methods and feature-based methods.

3.1.1. Similarity-Based Methods

A key underlying assumption of similarity-based machine learning methods is the "guilt-by-association" assumption, that is, similar drugs tend to share similar targets and vice versa. In this kind of approach, the similarity among drugs or among targets is computed by various similarity measures. The constructed similarity matrices define several types of kernel functions.

● The Nearest Neighbor Methods

The nearest neighbor methods generally adopt relatively simple similarity functions. Researchers often integrate these methods with some other approaches to help predict new drugs or targets, such as models in paper [46,47]. In the early stage, study [48] proposed two exploratory approaches, namely the nearest profile method (NN) and the weighted profile method. The nearest profile method follows the key concept that similar drugs or targets tend to be close in the network. This method was used in [49] as the baseline. In contrast, the weighted profile method utilizes the similarities of all the other drugs and targets and then adopts a weighted average. However, these methods show poor performance in the case when targets bound to similar drug share low sequence similarity or vice versa.

In the studies [23,50] by Zhang et al., methods that make drug-drug pair predictions based on neighbors were developed. These studies further extended the classic neighbor recommender method to the integrated neighborhood-based method (INBM). In simple terms, neighbor recommender method generally uses the weighted average information of neighbors for prediction. INBM is an ensemble model that integrates several neighborhood-based models for a robust prediction. For each drug-drug pair, three commonly used formulas, namely Jaccard similarity, Cosine similarity and Pearson correlation similarity, are used to calculate similarity score.

Another novel methodology in this category is Similarity-Rank-based predictor (SRP) [51]. Two indices, i.e., tendency index and inverse tendency index, are computed to construct a SRP. To be specific, the former represents the likelihood that each drug–target pair tends to interact, while the latter measures the tendency that each drug–target pair does not interact. The calculation formulas involve both similarity and similarity rank. Then an interaction likelihood score is computed as the likelihood ratio of the two indices. This method can generate two interaction likelihood scores, one from the drug side and the other from the target side. The final prediction score is the average of the two scores. The clear advantage of SRP is that it is a lazy and non-parametric model without the requirements of an optimization solver, prior statistical knowledge as well as tunable parameters.

In recent years, other new similarity-based methods have been proposed one after another, such as rule-based inference. Due to the limitation of the previous topology-based methods, a similarity-based deep learning method [52] merges the similarity measure with two rule-based inference methods. In other words, drug-based similarity inference (DBSI) and target-based similarity inference (TBSI) [48,53] are adopted to discover the drug-target interactions with the similarities. Though it is flexible to assemble any kernel functions, the method cannot predict new drugs or targets.

Note that most of similarity measures only utilize some important drug-related or disease-related properties to perform drug-disease prediction and ignore the known drug-disease interaction information [54]. Some researchers have proposed new similarity measures. Luo et al. [54] have designed a comprehensive similarity measure. In order to improve traditional similarity measures for drug-disease prediction, the comprehensive similarity measure has integrated drug or disease feature information with known drug–disease interactions. The similarity measure can be broken down into three steps. In the first step, drug similarity and disease similarity are calculated based on drug-related properties or disease-related properties respectively. In the second step, these similarity values are

adjusted by a logistic function based on the analysis and evaluation results. In the last step, a weighted drug network can be established for the drug similarity. The edge weight represents the number of common diseases between corresponding drugs. Then a cluster method, ClusterONE, is applied to identify potential drug clusters. Similarity between drugs belonging to the same cluster is enhanced and thus comprehensive drug similarity is obtained. Disease similarity can be improved in the same way as for drugs.

- Bipartite Local Models

Bipartite local models (BLMs) firstly generate two independent prediction for drugs and targets respectively. The final prediction result is then obtained by aggregating the two prediction scores.

The concept of BLM was first introduced in the pioneering work by Bleakley and Yamanishi [49]. This method can transform the drug-target interaction prediction problem into a binary classification problem. More specifically, a local model is trained for drugs based on chemical similarity. Another one is trained for proteins based on sequence structure. Therefore, two SVM classifiers can generate two independent prediction results from the drug or target side respectively. Final prediction result for each drug-target pair is computed based on the average of these two independent prediction scores.

Analogously, another method [55] developed a regularized least square classifier introducing two algorithms, called RLS-avg and RLS-kron. In particular, Regularized Least Squares (RLS-avg) utilizes kernel ridge regression to perform prediction. While in RLS-kron, all pairs of drugs and targets are combined into one to make Kronecker product, bringing the runtime down greatly.

Considering the limitation of the BLM-based methods above of predicting new drug or target without any known interactions available, Mei et al. [46] extended existing BLM by adding a preprocessing to infer training data from neighbors' interaction profiles. The method is called Bipartite Local Models with Neighbor-based Interaction Profile Inferring (BLM-NII). BLM-NII involves RLS-avg algorithm and is proven to be effective in new candidate problem.

- Matrix Factorization Methods

Matrix factorization methods are typically used in recommendation systems to find potential user-item interactions. The DTI prediction can be regarded as a matrix completion problem that aims to look for missing interactions. Therefore, drug-target interaction matrix can be factorized into two other matrices that when multiplied together can approximate the original matrix.

Kernelized Bayesian Matrix Factorization with Twin Kernels (KBMF2K) [56] is the original method that introduced matrix factorization to DTI prediction. Following some previous approaches, KBMF2K defines two kernel matrices only based on chemical similarity between drug compounds and genomic similarity between target proteins. It combines Bayesian probabilistic formulation, matrix factorization and binary classification for prediction problem.

Another study adopting probabilistic formulations is Probabilistic Matrix Factorization (PMF) [57]. PMF is distinguished greatly from KBMF2K by its independence of drug or target similarity matrices. Furthermore, the study presented the active learning (AL) strategy along with probabilistic matrix factorization.

Zheng et al. [58] proposed an extension of weighted low-rank approximation from one-class collaborative filtering (CMF), namely Multiple Similarities Collaborative Matrix Factorization (MSCMF). MSCMF integrates multiple similarity matrices, including chemical structure similarity, genomic sequence similarity, ATC similarity, GO similarity and PPI network similarity. Weights over the matrices are estimated to select similarities automatically. This strategy improves predictive performance in the experiment. Drugs and targets are projected into low-rank matrices. Then weights over similarity matrices are estimated using an alternating least squares algorithm. However, regardless of its performance, under this data integration strategy, a large amount of information may be lost, thus leading to sub-optimal solution.

The method developed by Ezzat et al. [59], employed two matrix factorization methods (i.e., GRMF and WGRMF). It was revealed in previous work [60] that data usually lies on or nears to the low-dimensional and non-linear manifold. Therefore, GRMF and WGRMF perform manifold learning implicitly by means of graph regularization. In addition, a preprocessing step (WKNKN) was applied to new drug or target prediction by transforming all the 0's in the original drug-target matrix into interaction likelihood values. This important step distinguishes this method from other work that regards all the 0's of given drug-target matrix as non-interaction roughly, and thus enhances the prediction results.

3.1.2. Feature Vector-Based Methods

Generally, similarity-based prediction algorithms do not take heterogeneous types and interactions defined in semantic networks into consideration. In addition, it may be difficult to add the long indirect connections between two nodes. Therefore, feature vector-based methods have been utilized for DTI prediction. The input of feature vector-based methods is drug-target pairs represented by fixed-length feature vectors. The feature vectors are encoded by various properties of drugs and targets.

In the systematic approach [61], chemical descriptors are calculated using DRAGON program (http://www.talete.mi.it/index.htm). Finally, each drug is represented as a set of 1080 descriptors, including constitutional descriptors, topological descriptors, 2D autocorrelations, eigenvalue-based indices and so on. Likewise, each protein is represented by a set of structural and physicochemical descriptors via PROFEAT WEBSEVER (http://jing.cz3.nus.edu.sg/cgi-bin/prof/prof.cgi). The descriptors involve Amino acid composition descriptors, Dipeptide composition descriptors, and Autocorrelation descriptors and so on. Then each protein sequence with changeable length can be transformed into a standard feature vector of 1080 dimensions. Hence, a set of 2160-dimensional feature vectors for each drug-target pair can be constructed. Subsequent prediction step performs Random Forest (RF) algorithm which introduces random training set (bootstrap) and random input vectors into the trees. The comprehensive framework shows its robustness against the over fitting problem and performs more efficiently for a large-scale data set in experiments.

In order to integrate diverse information from heterogeneous data sources, a method named DTINet was proposed by Luo et al. [20]. Through DTINet, a low dimensional feature vector that accurately explains the topological properties of each node in the heterogeneous network is first learned. In the further step, DTINet applies inductive matrix completion to best project drug space onto protein space.

Due to the fact that DTINet separates features and may result in loss of the optimal solution, Wan et al. [21] created a new framework called neural integration of neighbor information for DTI prediction (NeoDTI). The inspiration of NeoDTI came from convolution neural networks (CNNs). It integrates the neighbor information in heterogeneous network. After extracting the complex hidden features vectors of drugs and targets, NeoDTI automatically learns topology-preserving representations to achieve superior prediction performance.

The pioneering effort in [62] introduced a two-layer undirected graphical model, namely restricted Boltzmann machine (RBM), into a large-scale drug-target interaction prediction. There are no intra-layer connections in these layers. What's more, RBM model is trained via a practical learning algorithm, i.e., Contrastive Divergence (CD). Where the method significantly outperforms other existing approaches is in that it can predict different types of DTIs on a multidimensional network. In other words, the method can identify binary DTIs as well as their corresponding types of interactions, including relationships and drug modes of action.

In the paper published by Fu and cooperators [63], a state-of-the-art machine learning model was constructed based on meta-path-based topological features. Two measures of topological features are calculated, including the number of path instances between nodes and a normalization process to it. Given features, a Random Forest algorithm is used as supervised classification. Furthermore, intrinsic

feature ranking algorithm embedded in Random Forest selects the important topological features for better prediction. This framework has shown precise predictability.

3.2. Semi-Supervised Learning Methods

Considering the negative sample selection has a great influence on the accuracy of DTI prediction results, some researchers have proposed semi-supervised methods to address the problem. These methods use only a small amount of labeled data and a large amount of unlabeled data. Semi-supervised methods typically use the labeled data to infer labels for unlabeled data. On the other hand, the unlabeled data can also help provide insights into the structure of training set.

Having no use of negative samples, study [64] first employed a manifold Laplacian regularized least square (LapRLS) based on the BLM concept. Furthermore, an extension of the standard LapRLS, namely NetLapRLS, was proposed. NetLapRLS integrates information from chemical space, genomic space and drug-protein interaction for a new kernel. These semi-supervised methods have achieved encouraging results than using the labeled data alone. However, it is time-consuming when implementing them on a large scale.

Another method is designed for both semi-supervised and unsupervised settings. Ma et al. [22] presented a new framework to learn accurate and interpretable similarity measures when labels are scarce. This framework constructs a set of Graph Auto-Encoder (GAE)-based models and integrates multi-view drug similarities. Besides, an attentive mechanism is used for view selection and better interpretability.

3.3. Discussion

Each machine learning model possesses its unique advantages as well as disadvantages. Note that just as the popular concept in computer science, namely "no free lunch theorem" [65], machine learning methods are context-specific. Therefore, in this review we can only evaluate the advantages and disadvantages of each method category based on DTI prediction context.

A number of supervised models have been already proven feasible for DTI prediction. However, most supervised methods simply regard all the unlabeled drug-target pairs as negative samples and thus generate inaccurate predictive results. What's more, each similarity-based method has its limitation when extending to large a data set because of high complexity of similarity matrices computation.

Consider the three sub-classes of similarity-based methods respectively. Although the nearest neighbor methods generally apply relatively simple similarity functions, most of them construct neighborhoods only based on first-order similarity and do not involve the transitivity of similarity [66]. A key advantage of bipartite local models is that they process much fewer drug-target pairs, and thus they have much lower complexity than global models. Nevertheless, bipartite local models cannot handle the scenario that both drugs and targets are not involved in the training set unless combined with other methods. According to the experiment result in [19], matrix factorization methods generally have more superior performance than other methods including the nearest neighbor models and bipartite local models.

A small number of known drug-target interactions results in an imbalanced dataset. As an effective solution for imbalanced datasets, semi-supervised learning uses only a small amount of labeled data with a large amount of unlabeled data and generates more reliable prediction than supervised one.

In addition to the aforementioned single machine learning methods, we also have introduced several ensemble methods [61,63]. A better and robust prediction generally results from the biases trade-off of each single method. Generally, ensemble methods can combine different learning models. For more ensemble methods applied to drug-target interaction prediction task, please refer to [67–69].

Generally, machine learning has achieved favorable performance in DTI prediction. Nonetheless, a number of challenges still remain. Above all, recently, some researchers have emphasized that

predictive models based on machine learning are usually established and evaluated with overly simplified settings. Prediction results under such experiment settings may be over optimistic and deviate from the real case. Particularly, most of machine learning methods simply regard drug-target interaction as an on-off relationship and ignore other vital factors like molecule concentrations and quantitative affinities. Pahikkala et al. [24] have pointed out four factors having significant impact on prediction results, including problem formulation, evaluation data set, evaluation procedure and experimental setting. Considering the binding affinities and dose-dependence of drug-target pairs, the DTI prediction problem should be formulated as a regression or rank prediction problem rather than a standard binary classification problem. The second challenge is the imbalanced dataset problem. Due to the small number of known drug-target pairs, the current dataset is imbalanced. Some models like decision trees and SVMs, have a great bias for recognizing the majority class and thus result in poor performance [16]. Thirdly, most machine learning models possess "poor interpretability" properties. In other words, it is difficult to understand the underlying drug mechanism of action from a biological perspective. Note that in most case, it is easier to explain relatively simple models. This case is consistent with one of the "rules of thumb" [70], that is "simple is often better". Nonetheless, for most current state-of-the-art approaches achieving high DTI prediction accuracy, such as deep learning methods, it is difficult to interpret them from a pharmacology perspective. Last but not least, there are still no uniform evaluation metrics special for DTI prediction. Previous studies have adopted some common evaluation metrics in bioinformatics [71], such as sensitivity, specificity, Area Under the Precision-Recall (AUPR) curve and Area under the ROC curve (AUC). The fact is that if the sensitivity increases, the specificity decreases. Considering the limitation of using sensitivity or specificity alone, AUPR and AUC may be better choices in evaluation tasks. In the currently accessible datasets, the number of unknown samples is much more than the known ones, and thus false positives should be weighed more. AUPR can reduce the impact of false positive data on evaluation results as possible [72], and AUC is insensitive to imbalance dataset [73]. Thus both AUPR and AUC are generally adequate metrics for evaluating the performance of machine learning-based methods.

4. Conclusions and Outlook

DTIs contribute to the selection of potential drugs and thus effectively reduce the scope of research for biochemical experiments. Besides, they can provide deep insights into the side effects and the mechanism(s) of action of drugs. Hence, DTI prediction is a vital prerequisite for drug discovery. In fact, a number of public available databases have been established and promoted the development of innovatory DTI prediction strategies.

In this review, we focus on machine learning-based methods integrating chemical space and genomic space. We summarize the databases and machine learning methods frequently used in DTI prediction. In particular, we focus on several state-of-the-art predictive models appearing in recent years. We adopt a hierarchical classification scheme. We classify machine learning methods into two major categories: supervised and semi-supervised methods, and provide more subclasses.

Machine learning will be promising in DTI prediction for the next several years. However, there is still much room for improvement. Hence, we conclude with some advice as a reference for the future researchers.

Firstly, ensemble approaches combine multiple independent classifiers into one model and typically achieve a better prediction results. Next, semi-supervised learning is a powerful tool for addressing the imbalanced dataset problem. However, only a small number of semi-supervised learning methods have been proposed recently. Hence, the research on semi-supervised learning methods needs more attention. Furthermore, note the fact that drug-target pairs involve binding affinities and dose-dependence. It is more practical and meaningful to study new regression methods for DTI prediction problem. The using of quantitative bioactivity data will lead to a more accurate and reliable predictive result. Finally, with the development of high throughput biotechnology, the available

data has been growing quickly recently. It is time for further machine learning technology to take full advantage of more different types of heterogeneous data.

5. Key Points

1. Identifying drug-target interactions is the vital first step in drug discovery research.
2. A number of existing professional databases serve known data resources for DTI prediction and thus promote the drug discovery.
3. Machine learning-base methods are generally effective and reliable for DTI prediction.
4. Different machine learning methods have their merits and demerits. Hence, it is essential to choose appropriate methods or assemble models for special prediction tasks.
5. A more effective prediction model can be established by integrating more heterogeneous data sources of drugs and targets.
6. In reality, DTI prediction is a regression problem with quantitative bioactivity data.

Author Contributions: Conceptualization, R.C.; Writing-Original Draft Preparation, R.C.; Writing-Review & Editing, R.C., X.L., S.J. and J.L. (Jiawei Lin); Funding Acquisition, X.L.; Supervision, J.L. (Juan Liu).

References

1. Masoudi-Nejad, A.; Mousavian, Z.; Bozorgmehr, J.H. Drug-target and disease networks: Polypharmacology in the post-genomic era. *In Silico Pharmacol.* **2013**, *1*, 17. [CrossRef] [PubMed]
2. Paul, S.M.; Mytelka, D.S.; Dunwiddie, C.T.; Persinger, C.C.; Munos, B.H.; Lindborg, S.R.; Schacht, A.L. How to improve R&D productivity: The pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214. [CrossRef] [PubMed]
3. Dickson, M.; Gagnon, J.P. Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.* **2004**, *3*, 417–429. [CrossRef] [PubMed]
4. Wang, Y.; Bryant, S.H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B.A.; Thiessen, P.A.; He, S.; Zhang, J. Pubchem bioassay: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D955–D963. [CrossRef] [PubMed]
5. Chen, H.; Zhang, Z. A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS ONE* **2013**, *8*, e62975. [CrossRef] [PubMed]
6. Li, J.; Zheng, S.; Chen, B.; Butte, A.J.; Swamidass, S.J.; Lu, Z. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **2016**, *17*, 2–12. [CrossRef] [PubMed]
7. Zeng, X.; Liu, L.; Lu, L.; Zou, Q. Prediction of potential disease-associated micrornas using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [CrossRef] [PubMed]
8. Zhang, X.; Zou, Q.; Rodríguez-Patón, A.; Zeng, X. Meta-path methods for prioritizing candidate disease mirnas. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**. [CrossRef]
9. Hua, S.; Yun, W.; Zhiqiang, Z.; Zou, Q. A discussion of micrornas in cancers. *Curr. Bioinform.* **2014**, *9*, 453–462. [CrossRef]
10. Zeng, X.; Liao, Y.; Liu, Y.; Zou, Q. Prediction and validation of disease genes using hetesim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 687–695. [CrossRef] [PubMed]
11. Zeng, J.; Li, D.; Wu, Y.; Zou, Q.; Liu, X. An empirical study of features fusion techniques for protein-protein interaction prediction. *Curr. Bioinform.* **2016**, *11*, 4–12. [CrossRef]
12. Wang, Z.; Zou, Q.; Jiang, Y.; Ju, Y.; Zeng, X. Review of protein subcellular localization prediction. *Curr. Bioinform.* **2014**, *9*, 331–342. [CrossRef]

13. Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206. [CrossRef] [PubMed]

14. Arola, L.; Fernandez-Larrea, J.; Blay, M.; Salvado, M.J.; Blade, C.; Ardevol, A.; Vaque, M.; Pujadas, G. Protein-ligand docking: A review of recent advances and future perspectives. *Curr. Pharm. Anal.* **2008**, *4*, 1–19. [CrossRef]

15. Yamanishi, Y. Chemogenomic approaches to infer drug–target interaction networks. In *Data Mining for Systems Biology: Methods and Protocols*; Mamitsuka, H., DeLisi, C., Kanehisa, M., Eds.; Humana Press: Totowa, NJ, USA, 2013; Volume 939, pp. 97–113. ISBN 978-1-62703-107-3.

16. Mousavian, Z.; Masoudi-Nejad, A. Drug-target interaction prediction via chemogenomic space: Learning-based methods. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 1273–1287. [CrossRef] [PubMed]

17. Ding, H.; Takigawa, I.; Mamitsuka, H.; Zhu, S. Similarity-based machine learning methods for predicting drug-target interactions: A brief review. *Brief. Bioinform.* **2014**, *15*, 734–747. [CrossRef] [PubMed]

18. Chen, X.; Yan, C.C.; Zhang, X.; Zhang, X.; Dai, F.; Yin, J.; Zhang, Y. Drug-target interaction prediction: Databases, web servers and computational models. *Brief. Bioinform.* **2016**, *17*, 696–712. [CrossRef] [PubMed]

19. Ezzat, A.; Wu, M.; Li, X.L.; Kwoh, C.K. Computational prediction of drug-target interactions using chemogenomic approaches: An empirical survey. *Brief. Bioinform.* **2018**. [CrossRef] [PubMed]

20. Luo, Y.; Zhao, X.; Zhou, J.; Yang, J.; Zhang, Y.; Kuang, W.; Peng, J.; Chen, L.; Zeng, J. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.* **2017**, *8*, 573. [CrossRef] [PubMed]

21. Wan, F.; Hong, L.; Xiao, A.; Jiang, T.; Zeng, J. Neodti: Neural integration of neighbor information from a heterogeneous network for discovering new drug-target interactions. *Bioinformatics* **2018**. [CrossRef]

22. Ma, T.; Xiao, C.; Zhou, J.; Wang, F. Drug similarity integration through attentive multi-view graph auto-encoders. *arXiv*, **2018**; arXiv:1804.10850.

23. Zhang, W.; Chen, Y.; Liu, F.; Luo, F.; Tian, G.; Li, X. Predicting potential drug-drug interactions by integrating chemical, biological, phenotypic and network data. *BMC Bioinform.* **2017**, *18*, 18. [CrossRef] [PubMed]

24. Pahikkala, T.; Airola, A.; Pietila, S.; Shakyawar, S.; Szwajda, A.; Tang, J.; Aittokallio, T. Toward more realistic drug-target interaction predictions. *Brief. Bioinform.* **2015**, *16*, 325–337. [CrossRef] [PubMed]

25. Zeng, X.; Zhang, X.; Zou, Q. Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks. *Brief. Bioinform.* **2016**, *17*, 193–203. [CrossRef] [PubMed]

26. Zou, Q.; Ju, Y.; Li, D. Protein folds prediction with hierarchical structured SVM. *Curr. Proteom.* **2016**, *13*, 79–85. [CrossRef]

27. Wang, X.; Zeng, X.; Ju, Y.; Jiang, Y.; Zhang, Z.; Chen, W. A classification method for microarrays based on diversity. *Curr. Bioinform.* **2016**, *11*, 590–597. [CrossRef]

28. Nagamine, N.; Sakakibara, Y. Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* **2007**, *23*, 2004–2012. [CrossRef] [PubMed]

29. Kanehisa, M.; Furumichi, M.; Mao, T.; Sato, Y.; Morishima, K. Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361. [CrossRef] [PubMed]

30. Placzek, S.; Schomburg, I.; Chang, A.; Jeske, L.; Ulbrich, M.; Tillack, J.; Schomburg, D. Brenda in 2017: New perspectives and new tools in brenda. *Nucleic Acids Res.* **2017**, *45*, D380–D388. [CrossRef] [PubMed]

31. Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A. Pubchem substance and compound databases. *Nucleic Acids Res.* **2016**, *44*, D1202–D1213. [CrossRef] [PubMed]

32. Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S.Y.; Zhang, P.; Li, Y.H.; Yang, S.Y.; Wei, Y.Q.; Tao, L. Therapeutic target database update 2014: A resource for targeted therapeutics. *Nucleic Acids Res.* **2014**, *42*, D1118–D1123. [CrossRef] [PubMed]

33. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z. Drugbank 5.0: A major update to the drugbank database for 2018. *Nucleic Acids Res.* **2017**, *46*, D1074–D1082. [CrossRef] [PubMed]

34. Hecker, N.; Ahmed, J.; Von, E.J.; Dunkel, M.; Macha, K.; Eckert, A.; Gilson, M.K.; Bourne, P.E.; Preissner, R. Supertarget goes quantitative: Update on drug-target interactions. *Nucleic Acids Res.* **2012**, *40*, D1113–D1117. [CrossRef] [PubMed]

35. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; Mcglinchey, S.; Michalovich, D.; Allazikani, B. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef] [PubMed]

36. Szklarczyk, D.; Santos, A.; Von, M.C.; Jensen, L.J.; Bork, P.; Kuhn, M. STITCH 5: Augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **2016**, *44*, D380–D384. [CrossRef] [PubMed]

37. Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J. Supertarget and matador: Resources for exploring drug-target relationships. *Nucleic Acids Res.* **2008**, *36*, D919–D922. [CrossRef] [PubMed]

38. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R.N.; Gilson, M.K. Bindingdb: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201. [CrossRef] [PubMed]

39. Magariños, M.P.; Carmona, S.J.; Crowther, G.J.; Ralph, S.A.; Roos, D.S.; Shanmugam, D.; Voorhis, W.C.V.; Agüero, F. TDR targets: A chemogenomics resource for neglected diseases. *Nucleic Acids Res.* **2012**, *40*, D1118–D1127. [CrossRef] [PubMed]

40. Kuhn, M.; Campillos, M.; Letunic, I.; Jensen, L.J.; Bork, P. A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.* **2010**, *6*, 343–348. [CrossRef] [PubMed]

41. Seiler, K.P.; George, G.A.; Happ, M.P.; Bodycombe, N.E.; Carrinski, H.A.; Norton, S.; Brudz, S.; Sullivan, J.P.; Muhlich, J.; Serrano, M. Chembank: A small-molecule screening and cheminformatics resource database. *Nucleic Acids Res.* **2008**, *36*, D351–D359. [CrossRef] [PubMed]

42. Liu, Y.; Wei, Q.; Yu, G.; Gai, W.; Li, Y.; Chen, X. DCDB 2.0: A major update of the drug combination database. *Database* **2014**, *2014*. [CrossRef] [PubMed]

43. Kumar, R.; Chaudhary, K.; Gupta, S.; Singh, H.; Kumar, S.; Gautam, A.; Kapoor, P.; Raghava, G.P.S. CancerDR: Cancer drug resistance database. *Sci. Rep.* **2013**, *3*, 1445. [CrossRef] [PubMed]

44. Chen, X.; Ren, B.; Chen, M.; Liu, M.X.; Ren, W.; Wang, Q.X.; Zhang, L.X.; Yan, G.Y. ASDCD: Antifungal synergistic drug combination database. *PLoS ONE* **2014**, *9*, e86499. [CrossRef] [PubMed]

45. Nickel, J.; Gohlke, B.O.; Erehman, J.; Banerjee, P.; Rong, W.W.; Goede, A.; Dunkel, M.; Preissner, R. SuperPred: Update on drug classification and target prediction. *Nucleic Acids Res.* **2014**, *42*, W26–W31. [CrossRef] [PubMed]

46. Mei, J.P.; Kwoh, C.K.; Yang, P.; Li, X.L.; Zheng, J. Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics* **2013**, *29*, 238–245. [CrossRef] [PubMed]

47. Van Laarhoven, T.; Marchiori, E. Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS ONE* **2013**, *8*, e66952. [CrossRef] [PubMed]

48. Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232–i240. [CrossRef] [PubMed]

49. Bleakley, K.; Yamanishi, Y. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* **2009**, *25*, 2397–2403. [CrossRef] [PubMed]

50. Zhang, W.; Zou, H.; Luo, L.; Liu, Q.; Wu, W.; Xiao, W. Predicting potential side effects of drugs by recommender methods and ensemble learning. *Neurocomputing* **2016**, *173*, 979–987. [CrossRef]

51. Shi, J.Y.; Yiu, S.M. SRP: A concise non-parametric similarity-rank-based model for predicting drug-target interactions. In Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Washington, DC, USA, 9–12 November 2015; IEEE: New York, NY, USA, 2015; pp. 1636–1641.

52. Zong, N.; Kim, H.; Ngo, V.; Harismendy, O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics* **2017**, *33*, 2337–2344. [CrossRef] [PubMed]

53. Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503. [CrossRef] [PubMed]

54. Luo, H.; Wang, J.; Li, M.; Luo, J.; Peng, X.; Wu, F.X.; Pan, Y. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics* **2016**, *32*, 2664–2671. [CrossRef] [PubMed]

55. Van Laarhoven, T.; Nabuurs, S.B.; Marchiori, E. Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics* **2011**, *27*, 3036–3043. [CrossRef] [PubMed]

56. Gönen, M. Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics* **2012**, *28*, 2304–2310. [CrossRef] [PubMed]

57. Cobanoglu, M.C.; Liu, C.; Hu, F.; Oltvai, Z.N.; Bahar, I. Predicting drug–target interactions using probabilistic matrix factorization. *J. Chem. Inf. Model.* **2013**, *53*, 3399–3409. [CrossRef] [PubMed]

58. Zheng, X.; Ding, H.; Mamitsuka, H.; Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL, USA, 11–14 August 2013; ACM: New York, NY, USA, 2013; pp. 1025–1033.

59. Ezzat, A.; Zhao, P.; Wu, M.; Li, X.L.; Kwoh, C.K. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2016**, *14*, 646–656. [CrossRef] [PubMed]

60. Tenenbaum, J.B.; Silva, V.D.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]

61. Yu, H.; Chen, J.; Xu, X.; Li, Y.; Zhao, H.; Fang, Y.; Li, X.; Zhou, W.; Wang, W.; Wang, Y. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. *PLoS ONE* **2012**, *7*, e37608. [CrossRef] [PubMed]

62. Wang, Y.; Zeng, J. Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics* **2013**, *29*, i126–i134. [CrossRef] [PubMed]

63. Fu, G.; Ding, Y.; Seal, A.; Chen, B.; Sun, Y.; Bolton, E. Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinform.* **2016**, *17*, 160. [CrossRef] [PubMed]

64. Xia, Z.; Wu, L.Y.; Zhou, X.; Wong, S.T. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.* **2010**, *4*, S6. [CrossRef] [PubMed]

65. Wolpert, D.H.; Macready, W.G. No free lunch theorems for optimization. *IEEE Trans Evol. Comput.* **1997**, *1*, 67–82. [CrossRef]

66. Zhang, P.; Wang, F.; Hu, J.; Sorrentino, R. Label propagation prediction of drug-drug interactions based on clinical side effects. *Sci. Rep.* **2015**, *5*, 12339. [CrossRef] [PubMed]

67. Ezzat, A.; Wu, M.; Li, X.L.; Kwoh, C.K. Drug-target interaction prediction using ensemble learning and dimensionality reduction. *Methods* **2017**, *129*, 81–88. [CrossRef] [PubMed]

68. Ezzat, A.; Wu, M.; Li, X.L.; Kwoh, C.K. Drug-target interaction prediction via class imbalance-aware ensemble learning. *BMC Bioinform.* **2016**, *17*, 267–276. [CrossRef] [PubMed]

69. Zhang, R. An ensemble learning approach for improving drug–target interactions prediction. In Proceedings of the 4th International Conference on Computer Engineering and Networks, Shanghai, China, 19–20 July 2015; Wong, W.E., Ed.; Springer International Publishing: Cham, Switzerland, 2015; pp. 433–442.

70. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-generation machine learning for biological networks. *Cell* **2018**, *173*, 1581–1592. [CrossRef] [PubMed]

71. Zeng, X.; Lin, W.; Guo, M.; Zou, Q. A comprehensive overview and evaluation of circular rna detection tools. *PLoS Comput. Biol.* **2017**, *13*, e1005420. [CrossRef] [PubMed]

72. Davis, J.; Goadrich, M. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (ICML '06), Pittsburgh, PA, USA, 25–29 June 2006; ACM Press: New York, NY, USA, 2006; pp. 233–240.

73. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]