REVIEW ARTICLE

Chemogenomic Approaches for Revealing Drug Target Interactions in Drug Discovery

Harshita Bhargava[1,*], Amita Sharma[1] and Prashanth Suravajhala[2,3,4]

[1]Department of Computer Science & IT, IIS (Deemed to be University), Jaipur, India; [2]Bioclues.org, Kukatpally, Hyderabad, 500072, India; [3]Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, India; [4]Amrita School of Biotechnology Amrita University, Amritapuri, Kerala 690525, India

ARTICLE HISTORY

Received: August 24, 2020
Revised: May 30, 2021
Accepted: June 25, 2021

DOI:
10.2174/1389202922666210920125800

Abstract: The drug discovery process has been a crucial and cost-intensive process. This cost is not only monetary but also involves risks, time, and labour that are incurred while introducing a drug in the market. In order to reduce this cost and the risks associated with the drugs that may result in severe side effects, the in silico methods have gained popularity in recent years. These methods have had a significant impact on not only drug discovery but also the related areas such as drug repositioning, drug-target interaction prediction, drug side effect prediction, personalised medicine, *etc*. Amongst these research areas predicting interactions between drugs and targets forms the basis for drug discovery. The availability of big data in the form of bioinformatics, genetic databases, along with computational methods, have further supported data-driven decision-making. The results obtained through these methods may be further validated using *in vitro* or *in vivo* experiments. This validation step can further justify the predictions resulting from *in silico* approaches, further increasing the accuracy of the overall result in subsequent stages. A variety of approaches are used in predicting drug-target interactions, including ligand-based, molecular docking based and chemogenomic-based approaches. This paper discusses the chemogenomic methods, considering drug target interaction as a classification problem on whether or not an interaction between a particular drug and target would serve as a basis for understanding drug discovery/drug repositioning. We present the advantages and disadvantages associated with their application.

Keywords: Chemogenomic approaches, drug target, drug discovery, *in silico* methods, chemogenomic methods, personalised medicine.

1. INTRODUCTION

Drug discovery and drug development is basically a time-taking activity that involve several stages such as target identification, target validation, lead compound identification, lead compound optimisation, preclinical trials, and clinical evaluation followed by approval and post-marketing stages [1]. The earlier stages in drug development may start screening from a large set of compounds which is then filtered at each stage resulting in approval of a single drug. Hence drugs failing early in the initial stages of development can decrease the cost considerably. Thus the time taken from the initial to final stage involving a high amount of investment at each stage, gets wasted if the drug is withdrawn from the market or if it fails for approval. The study [2] conducted using data from 50 pharmaceutical companies over a specific time period shows that the clinical success rate of approval was only 19% as compared to the expected success rate. As the drug discovery process starts with target identification hence, this needs to be accurate and reliable. With the substantial increase in open source databases and the relevant datasets, the computational *in silico* approaches are used as opposed to the conventional wet-lab experiments in predicting targets for drugs or *vice versa.* The targets range from RNA, DNA, proteins, biological pathways, disease-associated microRNAs, lncRNAs, biomarkers, crucial nodes of biological networks to molecular functions [3-5]. The study [6] revealed that the majority of the drug targets are proteins; hence in this paper, we have only included studies that consider proteins as targets.

The prediction of drug-target interactions reduces the drug/target search space, which indirectly reduces the incurred cost, time, and labour in the drug discovery pipeline. This paper includes five sections; section 1 discusses the basic stages of the drug discovery process; section 2 discusses the problem of drug-target interaction prediction considering proteins as targets; section 3 lists the different *in silico* approaches used for predicting drug-target interactions while discussing the related advantages and disadvantages associated with each approach (Table **1**); section 4 discusses the different open source databases used for predicting drug-target interactions; section 5 concludes this paper.

1.1. Drug Discovery Process

Drug development follows the drug discovery process where the latter includes target identification; target validation; lead compound identification; lead compound optimisation, while the former includes preclinical trials; clinical trials followed by approval of the drug [7] (Fig. **1**).

*Address correspondence to this author at the Department of Computer Science & IT, IIS (Deemed to be University), Jaipur, India; Tel: 8875021461; E-mail: harshita.bhargava@iissunivac.in

Initially, a disease is selected for which drug needs to be developed while studying the conditions such as the number of people affected across the globe, probable revenue to be generated, funding available, *etc*. The next step is to conduct thorough research about the disease. A conventional approach is to study the physiological effects using animal models or patient models. With the growth of next-generation sequencing (NGS), a comparative analysis between gene sequences of healthy and unhealthy tissues may help in gaining insights into the disease pathways. This helps to identify the genes that may be targeted by drugs to produce the intended effect even though the current studies show that genomics, proteomics, and gene association studies play a key role in identifying targets [8]. Generally, many targets are involved in disease, but not all of the targets have an equal role in the disease pathway. Once the target is identified, it needs to be validated to determine the operational role of the target with respect to the selected disease. Though the true validation can be done only through clinical trials in the early phases, it is carried out using assay development, small interfering RNA (miRNA), animal models, *etc*. [9].

The drug targets are either known targets characterised by their functions and interactions with specific drugs, or these are potential targets whose functions are not known and no interactions have been reported with the drugs [10], with the latter class is often taken up for completely new drug research or for those proteins whose function is not ascertained, also known as "hypothetical proteins" [11,12]. As the next step in the drug discovery process, the lead compound needs to be identified with the use of conventional high throughput screening (HTS) techniques that filter the

Table 1. Overall advantages and disadvantages of each category of methods from the Chemogenomic class.

Chemogenomic Category	Advantages	Disadvantages
NBI series methods	The network-based methodsdo not require three-dimensional structures of the targets as in the case of molecular docking-based methods, nor do they require negative samples, which is a basic requirement in the case of supervised learning approaches.	These methods suffer from the cold start problem of drugs ie. are unable to predict targets for new drugs, and are biased in prediction towards high degree drug nodes. This issue has also been highlighted while using NBI as the recommendation technique. These methods do not consider the side information in the form of drug and target features while predicting targets for drugs.
Similarity inference methods	Since these methods are based on "wisdom of crowd" principle, hence interpretability is one of the key advantages in order to justify the predictions.	Drugs(targets) having similar structure, fingerprint, or side effect (sequence)_may bind to different targets(drugs). Hence such similarity principles may not produce serendipic results. Secondl, none of these methods consider the continuous binding affinity scores that are more indicative than the binary values of the interaction matrix.
Random walk based methods	Were able to address the problem of cold start with respect to drugs The transitive relationships could be traversed in the sparse DTI network to find the vicinity of the query drug/target with all other targets/drugs.	These methods do not consider the continuous binding affinity scores between the drugs and targets. These methods are computationally intensive and hence may take time for convergence.
Local community paradigm(LCP) methods	These methods depend only on the topology of the bipartite network and do not require the similarity information of drug/target nodes.	These methods cannot address the cold start problem with respect to drugs or targets. These methods do not consider the continuous binding affinity scores between the drugs and targets.
Feature based methods	The benefit of such methods is that they can handle new drugs and targets without considering any similar information of chemical drugs and target sequences. Since the features can always be extracted for both drugs and proteins hence even in the case of new drugs/new targets, the machine learning model can predict the interactions by studying the dependence on features.	The feature selection is a difficult and crucial task as the interaction may be a function of only a subset of drug and target features. Secondly, in the case of supervised classification, based techniques class imbalance remains an issue.
Bipartite local models	These supervised models do not require negative samples as in the case of a general supervised machine learning approach.	The computational cost is too high in terms of training a classifier/regressor for each drug target pair in question from both drug and target sides, respectively.
Matrix factorization methods	They do not require negative samples, as in the case of supervised learning machine learning approaches.	The matrix factorization-based techniques are good at modeling linear relationships, but in the case of non-linear relationships, as in the case of drug target entities, neural networks can be a good choice.
Deep learning methods	The manual feature extraction can be surpassed with the use of deep learning models, which is a labor-intensive task in the case of feature-based machine learning models.	The reliability of the automatically learned feature representations is one of the issues which may not match the manually extracted drug or target features based on chemical structure/sequence information, respectively. Secondly, the interpretability of the deep learning classifier/regressor is low, and it is difficult to justify the model results. Thirdly when we use images as input, then the quality of such data is a matter of concern for identifying the underlying relationships.

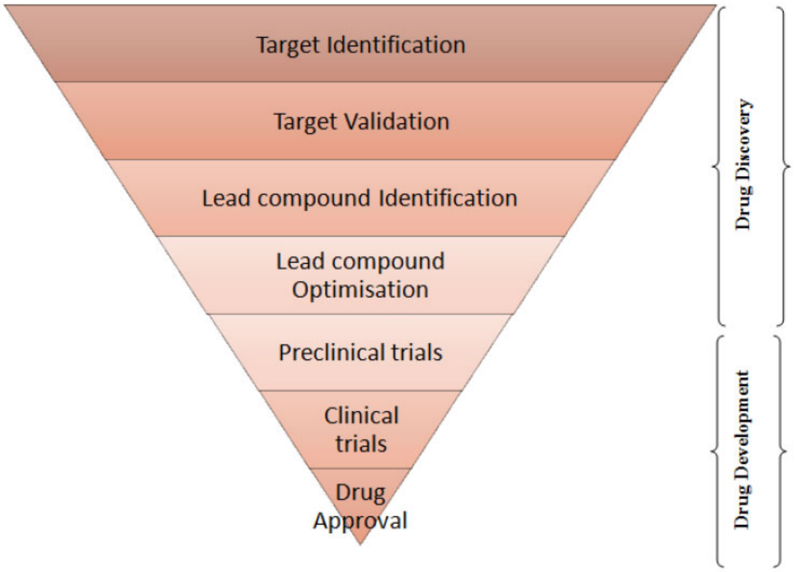

Fig. (1). Drug discovery and drug development process. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

leads from the available compound libraries for a specific target. As an alternative to HTS, many other computational methods are being employed in order to reduce the time taken by traditional approaches. For example, virtual screening is one of the computational techniques to identify the active compounds that are likely to bind to a specific target that may substitute the costly HTS technique. After the lead compounds are identified, they are optimized through iterative modifications to improve the ADME characteristics of the drug. In the next step, preclinical studies are carried out to check the safety and effectiveness of the drug before the drug is tested on humans in clinical studies. These studies deal with the pharmacodynamic and pharmacokinetic properties of the drug with minimum toxicity and side effects. Pharmacodynamics gives the relationship between the drug dosage and its intended biological effect while addressing the potency issues and the side effects. On the other hand, pharmacodynamics identifies the effect of drugs on the body in contrast to pharmacokinetics which identifies the effect of the body on the drug [13]. Since the drug dosage is in turn related to absorption, distribution, metabolism, excretion(ADME) properties of a drug and hence, it plays an important role in studying the biological effect of the drug at this stage. This is usually done using *in vitro, in vivo* models, but now *in silico* approaches are also being used to support preclinical studies/trials.

Clinical trials are taken up to further validate the results from prior preclinical trials, which involve the testing on humans through four different phases [14]. Phase I trials include testing approximately 10-100 healthy individuals with a minimum dose of the drug to study pharmacodynamics as well as pharmacokinetics. Phase II trials include a wider group of infected individuals from 50-500, examining the safety and effectiveness of the drug. The dose is increased successively to determine the best dose and the relevant adverse effects on individuals. Phase III trials are conducted on a larger group of patients at different geographical locations ranging from hundreds to thousands to confirm the efficacy of the drug while uncovering the rare side effects. The various statistical tests performed on data from multiple groups of individuals further confirm the frequency and the best drug dosage. Phase IV trials are conducted on the large diversified real world population to monitor the effectiveness of the drug and any rare adverse effects. Phase IV trials are done only for the drugs that were approved in earlier phases. After clinical trials, the post-marketing research studies are generally carried out to discover new indications for existing ones or abandoned/market withdrawn drugs [15].

2. DRUG TARGET INTERACTION PREDICTION

The drugs inhibit or activate the specific targets to produce the intended therapeutic effect. RNA, DNA, or proteins form the targets, but the majority of targets are proteins [15]. The rapid increase in bioinformatics data is majorly due to the use of open-source genetic databases, NGS HTS techniques [16]. The stored data in these open source databases can be harnessed to support drug discovery and development processes in the form of data-based predictions. These predictions may include drug-target associations, side effects of drugs, protein-protein associations, drug-drug associations, gene-disease associations, *etc*. The resultant predictions can support not only drug discovery but also the complex process of drug repositioning. Drug repositioning is the reuse of existing/abandoned drugs for the disease other than the one for which it was intentionally developed. Detecting novel targets for existing drugs through drug target interac-

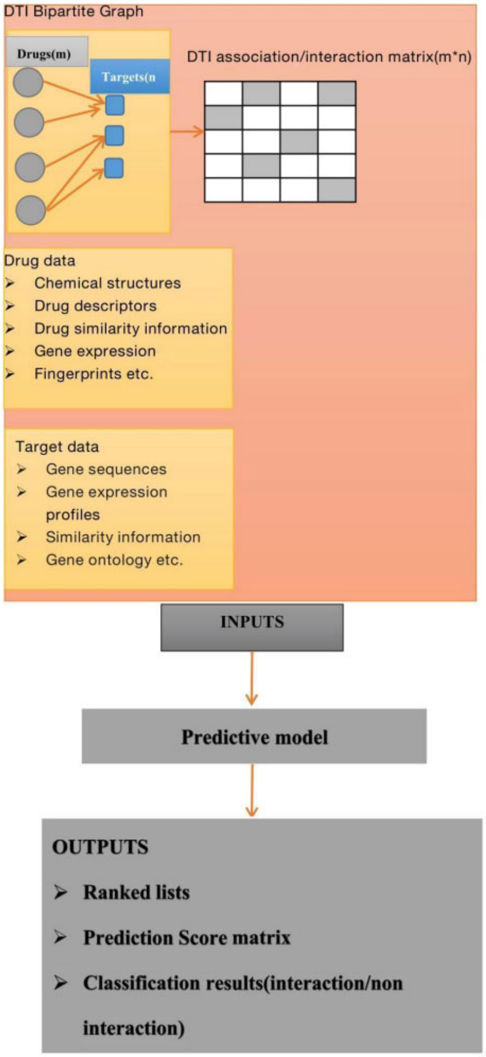

Fig. (2). A basic framework for DTI prediction. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

tion prediction can act as a stepping stone for drug repurposing. The task of predicting drug-target interactions (DTI) also assists polypharmacology, wherein a drug may have multiple associated targets [17]. As indicated in the previous section, the traditional drug discovery process takes a single target identified for a particular disease while testing specific drugs. However, the polypharmacology concept stresses that drugs have multiple targets. Amongst these multiple targets, some off-targets are either involved in side effects/toxic effects, while some may have unintended therapeutic effects, which may be further taken up for repositioning of existing drugs. A basic framework for DTI prediction consists of inputs in the form of drug data, target data, and the drug target association/interaction matrix, which can be in binary(1 or 0) or continuous form. The prediction model takes these inputs and produces outputs in the form of ranked lists, prediction score matrix, or classification results as interactions or non-interactions as depicted in Fig. (**2**)

3. *IN SILICO* METHODS FOR DRUG-TARGET INTERACTION PREDICTION

The drug-target interactions are quantified through experimental approaches, which specify the binding affinity between the drugs and the targets. These affinities are measured using the “inhibition constant (K_i)”, “dissociation constant (K_d)”, “half-maximal inhibitory concentration (IC_{50})” or “half-maximal effective concentration (EC_{50})” values between drugs and target, wherein a high numerical value indicates a lowe binding affinity. One can treat the problem of drug-target/drug-protein interaction prediction as either classification, regression, or a link prediction problem. The biological databases record the associations/interactions resulting from experiments in a continuous/binary format which can be further modelled using an interaction/association matrix. The substantial cost and time reductions are a major advantages of using *in silico* methods compared to the routine wet-lab experiments for predicting drug-target interactions. These methods typically include ligand-based, molecular docking based and chemogenomics-based methods [18]. Ligand-based methods work by matching the query ligand with the known ligands of the drug target. The matching principle includes similarity on the basis of physicochemical properties or structural similarity, indicating that similar ligands map to similar drug targets [19]. The downside of this method is that it generates poor results for the targets having only a few known ligands [20]. In addition, there are many ligands that may differ in structure or physicochemical properties but still have a common drug target. A similar argument applies to ligands that have a similar structure or physicochemical properties need not bind to a common target.

Molecular docking is a target-based method involving the use of three-dimensional protein structures extracted using nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, or cryo-electron microscopy (cryo-EM). The basic disadvantage of this method is that the three-dimensional structure is either not known for all proteins or is too difficult to be derived particularly for membrane proteins [21]. This is further complicated if the role of hypothetical proteins (HPs) is considered with respect to the drug in question. The available number of three-dimensional structures of proteins in Protein Data Bank(PDB) compared to the actual number of known proteins as targets in the human body further conforms to the stated disadvantage of the molecular docking-based approach. The chemogenomic-based methods use both the information related to drug and target space at the same time for inferring the drug target/drug-protein interactions(DTIs/DPIs). The chemogenomic methods can be further categorised as machine learning-based and network-based methods. The machine learning-based method can be further classified as feature-based and similarity-based methods. In this review, we classify them differently as feature-based methods, non-feature-based methods. If explainability is the evaluation parameter of DTI prediction results, then the similarity-based inference from the network-based category and feature-based methods from the machine learning category rank higher than all other methods (Fig. **3**) [22].

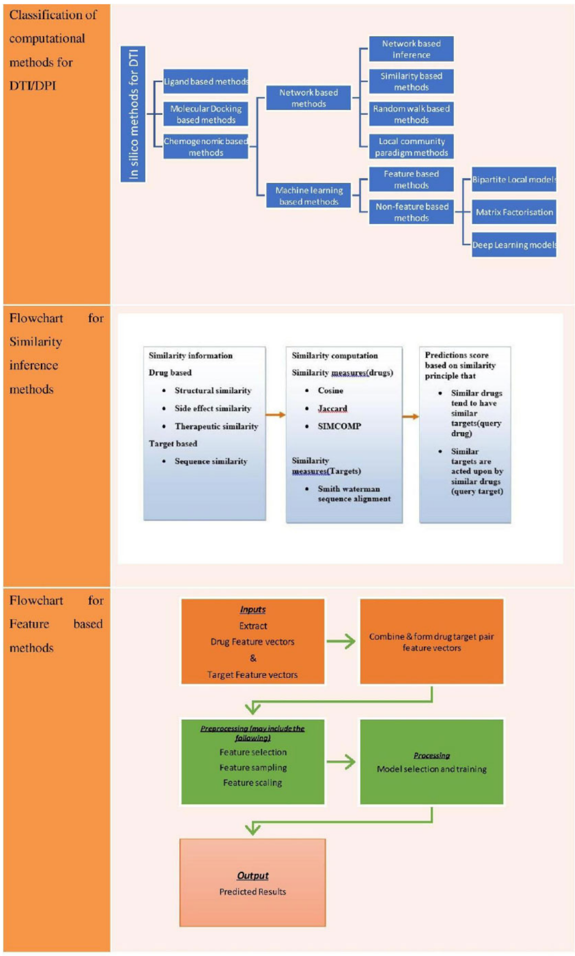
Classification of computational methods for DTI/DPI
In silico methods for DTI
Ligand based methods
Molecular Docking based methods
Chemogenomic based methods
Network based methods
Machine learning based methods
Network based inference
Similarity based methods
Random walk based methods
Local community paradigm methods
Feature based methods
Non-feature based methods
Bipartite Local models
Matrix Factorisation
Deep Learning models
Flowchart for Similarity inference methods
Similarity information
Drug based
Structural similarity
Side effect similarity
Therapeutic similarity
Target based
Sequence similarity
Similarity computation
Similarity measures(drugs)
Cosine
Jaccard
SIMCOMP
Similarity measures(Targets)
Smith waterman sequence alignment
Predictions score based on similarity principle that
Similar drugs tend to have similar targets(query drug)
Similar targets are acted upon by similar drugs (query target)
Flowchart for Feature based methods
Inputs
Extract
Drug Feature vectors
&
Target Feature vectors
Combine & form drug target pair feature vectors
Preprocessing (may include the following)
Feature selection
Feature sampling
Feature scaling
Processing
Model selection and training
Output
Predicted Results

Fig. (3). Computational methods for DTI/DPI. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

3.1. Network-based Methods

The network-based methods utilize an underlying network for predicting drug-target interactions. These methods can be further classified as network-based inference (NBI) methods, similarity-based inference methods, random walk-based methods, and local community paradigm methods [23].

The NBI methods originally developed for recommender systems were used for predicting the missing/unknown subsequent ratings of users for the items. A known binary interaction matrix is used to represent the interaction/non-interaction between drugs and targets. An entry of 1 indicates an interaction, while 0 indicates a non-interaction/unknown/undetected interaction in this known interaction matrix. With the basic NBI method, the interaction matrix can be modeled as a bipartite graph. The drugs and targets serve as nodes, while the known interactions serve as edges. This method generally utilises a concept of two-step resource diffusion process.The diffusion process initiates once from target to drug and then from drug to target, resulting in a scored list of targets for a given drug [24]. "EWNBI" and "NWNBI" [25] were proposed as the edge-weighted and node-weighted versions of the NBI method. In EWNBI (Edge weighted NBI), edges were weighted using the affinity score of the respective drug and target;. In contrast, NWNBI(Node weighted NBI) used the degree of nodes as node-weights followed by the resource diffusion process as suggested by the regular NBI method. If these methods are compared in terms of accuracy, then NBI ranked higher than EWNBI while NWNBI achieved comparable performance after optimising the associated parameter. Another method, "SDTNBI" (substructure-drug-target NBI), was proposed in order to predict targets even for new compounds [26]. The differentiating feature was the use of a tripartite network as opposed to a bipartite network to represent the interactions/associations with the drugs, their substructures, and targets as nodes. It was able to solve the cold start problem with respect to drugs, but the cold start problem with respect to targets remained unaddressed. Another parameterised version of SDTNBI termed "bSDTNBI"(balanced substructure drug target NBI) was proposed to realize higher performance [27]. In this sequence, another NBI based method, DT-hybrid, was proposed, which used the protein-protein and drug-drug similarity information along with the interaction information [28]. This method also suffered from the cold start problem but ranked higher than the NBI & Hybrid methods.

Similaritybased inference methods mimic the underlying concepts from collaborative filtering algorithms used in developing recommendation systems. They utilize the similarity information defined in terms of chemical structural similarity, protein sequence similarity [25], therapeutic similarity, or drug side effect similarity [29] to derive the predictions. In drug-based similarity inference (DBSI), the drugs having similar structures are considered to have similar kinds of probable targets. In the same way, target-based similarity inference (TBSI) was based on the concept that targets having similar sequences are likely to be acted upon by similar kinds of drugs. A similar argument was used for drug side effect similarity inference (DSESI) and drug therapeutic similarity inference (DTSI) to predict targets for drugs.

The random walk-based methods not only detect the direct linkages between the nodes of a graph but also the indirect/transitive linkages, which are inherently found in sparse graphs [30]. A random walk-based method NRWRH was proposed for a heterogeneous network while considering a network of drug-drug similarity, target-target similarity, and drug-target interactions [31]. The method was improvised with parameter optimization while experimenting with different similarity fingerprints, including ECFP, 2D pharmacophore fingerprints, *etc*. The local community paradigm (L-CP) methods used the Cannistraci variations extended to bipartite networks: "cannistraci–alanis–ravasi (CAR)", "cannistraci jaccard (CJC)", "cannistraci preferential attachment (CPA)", "cannistraci–adamic–adar (CAA)", and "cannistraci resource allocation (CRA)" [32]. These link prediction methods rely only on the topology information of the known DTI network of drugs and targets.

3.2. Machine Learning-based Methods

These methods can be classified as feature-based and non-feature-based methods. The feature-based methods employ a set of features from each of the drug and target spaces, respectively, while non-feature-based methods include bipartite local models, matrix factorisation models, and deep learning models. As per the review [33], machine learning has been the most researched category with respect to DTI, indicated by the increase in publications.

3.2.1. Feature-based Methods

The feature-based methods consider the features of the drugs and targets by forming the drug target pair feature vectors. These drug-target pair feature vectors are fed as input to a machine learning model that is trained with the labelled data to predict the DTI interactions. In general, the problem of drug-target interaction is modelled as a classification problem. The supervised learning classification models include SVM [34] and RVM [35]. The supervised classification approach requires not only the positive samples but also negative samples in order to generate unbiased results. On the basis of this intuition, the authors [36] used the random forest as a classification model while utilizing molecular docking to validate the proposed model. The negative dataset was constructed using BindingDB and BioLip databases by considering the measured bioactivity data above a threshold and removing the redundant entries. Similarly, the positive dataset was constructed using DrugBank and Yaminishi data, and the redundant entries were removed. Finally, each instance from the negative dataset was randomly added to either of the constructed positive datasets until all instances are exhausted.

The ensemble methods also have recently been proved to be more accurate as compared to a single classifier [37]. They implemented the homogeneous ensemble of classifiers while proposing the use of heterogeneous ensembles as future work for predicting DTI. Apart from random forest or rotation forest used as ensemble methods, boosting algorithms were used specifically for handling the imbalanced nature of available datasets. The minority class of the DTI dataset included the known positive samples of drug-target interactions, while the negative or unknown samples formed the majority class. The authors [38] proposed clustering-based under-sampling technique with boosting (CUSBoost) and compared the same with the previous methods such as random under-sampling combined with boosting (RusBoost) [39], synthetic minority oversampling technique combined with boosting (SMOTEBoost) [40] and ensemble-based learning without sampling the given data. The basic drawback of this method was that since k-means clustering was used with the majority subset. Another method, *viz*. cumulative feature subspacing with boosting (CFSBoost), was proposed, which used a subset of features to train the weak learners and an ensemble of classifiers obtained as the final prediction model [41]. The feature set was formed using the drug features in each subset while incrementally adding protein-related features in each subset. The dimensionality reduction and selection of best features from drug space and target space are too closely related. However, DTI prediction issues occur differently in both spaces. The dimensionality issue has also been studied by using SVD [42], thereby reducing the underlying complexity while building machine learning models. The feature evaluation followed by feature selection was used in [43] to select the best features for DTI prediction problems.

3.2.2. Non-Feature Based Methods

3.2.2.1. Bipartite Local Models

Bipartite local models (BLM) use a graph of drug and target nodes with an edge indicating the known interaction between these nodes. BLM works on the premise that it predicts targets for a given drug (drug side) and then for a given target (target side) before finally producing the score by aggregating the two predictions [44].To establish if drug d interacts with target t, all targets engaging with the given drug are labelled as +1, while others are labelled as -1, with the exception of target t. The SVM classifier is then trained with labelled data of targets with the sequence similarity information as kernel to find the prediction score between drug d and target t. Similarly, from the target side, all drugs interacting with the given target t are labelled as +1 while others are as -1, excluding drug d. Again the SVM classifier is trained with labelled data of drugs with the chemical similarity information as kernel to find the prediction score between drug d and target t. The final prediction is obtained by using a maximum or an aggregate function. The main drawback of this method is the prediction for new drugs or new targets without any prior interactions. This problem is addressed by introducing BLM-NII(BLM with neighbour-based interaction profile inferring), which considers not only the local interaction profile of drug (or target) but also the similarity information of drugs(or targets) with the new drug [45]. This was a prime advantage in the case of new drugs or targets with no prior interactions. The local models, *i.e*., Classifiers or regressors used can be experimented with further to improve the predictions.

3.2.2.2. Matrix Factorization

The implementation of matrix factorization techniques in recommendation systems motivated the application of these techniques to DTI prediction problems. They consider the interaction matrix of drugs and targets and are approximated by the product of low-rank matrices for predicting the blank or missing entries. The variants include probabilistic matrix factorization(PMF) [46], kernelized bayesian matrix factorization with twin kernels (KBMF2K) [47], multiple similarities collaborative matrix factorization (MSCMF)[48], graph regularised matrix factorization(GRMF) and weighted graph regularised matrix factorization (WGRMF) [49], neighbourhood regularised logistic matrix factorization (NRLMF) [50], and triple matrix factorization(TMF) [51]. These methods provide a mathematical approximation of the original interaction matrix. The distinguishing feature of MSCMF among all other matrix factorization-based methods is that it considers multiple similarity matrices of drugs and targets simultaneously while approximating these matrices along with the original interaction matrix.

3.2.2.3. Deep Learning Models

Deep learning approaches have found applications in big data analytics ranging from pattern recognition, natural language processing, speech recognition to recommendation systems. The feature selection is an essential and important preprocessing step in feature-based methods to know which feature contributes the most in predicting DTIs. Hence deep learning models eliminate this step by learning these representations from the raw data input. A deep neural network-based model was proposed with input compound protein pairs to study the nonlinear relationship between drugs and targets [52] with an appropriate set of hyperparameters, and the model was evaluated on both balanced and imbalanced datasets. DeepDTI [53] used the drug fingerprint information and protein sequence composition descriptor to form concatenated drug-target pair feature vectors. These features were then fed into a deep belief network to classify the input drug-target pair. DeepCPI [54] was proposed to address the high dimensionality issue of feature vectors used in the DeepDTI method by learning the low dimensional feature representations as feature embeddings. DeepDTA[55] and DeepConv-DTI [56] were proposed for predicting the affinity scores between the drug-target pair. The advantage of DeepConv-DTI over DeepDTA or DeepDTI was the ability to learn the local residue patterns in protein sequences which are the key entities contributing to interactions. Collaborative deep learning-based DTI predictor (CoDe-DTI) was proposed to address the cold start problem that exists in the case of new drugs with no prior interaction information [57]. This method was based on deep collaborative learning, which itself uses a combination of probabilistic matrix factorization and denoising autoencoder. The latent features are learned from the autoencoder and then used as input to matrix factorization. The cold start problem of a new drug is well addressed using this method while considering the common substructures interacting with similar targets. The drawback of this method is that it did not address the cold start problem of new targets. Autoencoders have been used as a dimensionality reduction technique and using the same characteristic, a stacked version of autoencoders was used to infer the hidden features [58]. The inferred features were then fed to the rotation forest classifier to determine an interaction or noninteraction. Another deep learning solution to feature extraction was provided using a Least Absolute Shrinkage and Selection Operator (LASSO) based models [59]. These models were used for the drug features and protein features separately in order to receive the most significant features. The retrieved features were concatenated as drug target pairs and then used as input to the deep neural network model. The distinguishing property of this model was that instead of using raw sequences and structures as done in most deep learning models, it used a selective feature set of drugs and proteins, respectively. DEEPScreen [60] used a collection of deep convolutional networks trained separately for each target protein. The system was experimented with differently sized 100 by 100 pixel, 200 by 200 pixel or 400 by 400 pixel two-dimensional images of drugs. Each target protein out of a total of 704 proteins has been considered, has at least 100 interacting drugs, and the performance for each target protein with respect to different CNN architectures was evaluated.

Some methods belong to the hybrid category, which can combine any of the above methods from different categories. DTiGEMS+ [61] was one of these methods that utilised machine learning, graph mining, and graph embedding by considering a heterogeneous graph derived from drug-drug similarity,protein-protein similarity, and drug-target interaction graph. The key idea was to integrate multiple similarity information from various sources while performing a forward similarity selection and keeping only the reliable ones. Two heterogeneous weighted graphs were used, one extracted using integrated similarity information of drug-drug and protein-protein pairs and the other built using cosine similarity of drug-drug and protein-protein feature embeddings. Graph mining was further used to extract path-based score features for each graph in order to be fed into a machine learning classifier.

Table 2. A gist of databases for DTI

Database	Description	URL
DrugBank [62]	It is an online database that combines the information about drugs (including approved, experimental (phase I/II/III) & biotech drugs), targets (including DNA, RNA, proteins, and other macromolecules) along with their mechanisms and interactions. The latest release DrugBank 5.0, has proven to be a comprehensive resource for researchers, pharmacists, the pharmaceutical industry, and educators.	https://www.drugbank.ca/
Pubchem [63]	It is an online cheminformatic database providing programmatic access to its data using its built API. It also includes data related to substances, compounds, targets, bioassays and pathways.	https://pubchem.ncbi.nlm.nih.gov/
BindingDB [64]	It is an online database containing binding affinity scores between small molecule drugs and protein targets. The regression-based datasets can be sourced from this database with affinity scores in terms of Ki, Kd, IC50, or EC50.	https://www.bindingdb.org/bind/index.jsp
SuperTarget [65]	It is a web-based online repository for analysing 195770 drugs,6219 targets, and 332828 drug-target interactions in the form of binary as well as continuous binding affinity data. It contains information derived from DrugBank, BindingDB, and SuperCyp databases.	http://bioinformatics.charite.de/supertarget/
ChEMBL [66]	It is a chemical database having drug-like like properties with bioactivity data in terms of Ki, Kd, IC50, EC50 against the targets collected from literature, patents, *etc*.	https://www.ebi.ac.uk/chembl/
SIDER [67]	A side effect resource containing the reported side effects of drugs or adverse drug reactions with respect to marketed medicines. It also provides the ATC code based classification along with the respective frequency of the side effect for each drug.	http://sideeffects.embl.de/
MATADOR [68]	Manually Annotated Targets and Drugs Online Resource(MATADOR) captures both the direct and indirect interactions between chemicals and proteins either using text mining or manual collection.	http://matador.embl.de
STITCH [69]	Search Tool For Interaction of Chemicals (STITCH)is a database that integrates chemical protein interaction information from several databases, texts, and other experiments. The chemical protein interactions can also be visualised as a network with labelled edges indicating the type of action.	http://stitch.embl.de/
ZINC [70]	One of the largest ligand databases containing more than 230 million purchasable compounds in docking specific 3D formats. It provides an easy-to-use interface for querying over 750 million purchasable compounds.	http://zinc.docking.org
BioLip [71]	It is a weekly updated database for studying protein-ligand binding interactions for available protein structures in the Protein data bank database. It can be used to derive continuous datasets with binding affinity data between proteins and ligands.	https://zhanglab.ccmb.med.umich.edu/BioLiP/

4. OPEN SOURCE DATABASES FOR DRUG-TARGET INTERACTION PREDICTION

The availability of public databases has paved the way to design better predictors for DTI. The major concern lies in the frequency of updation of these source databases, which directly influences the accuracy of prediction. These databases may contain overlapping information that has been extracted from patents, research papers, *etc*., or has been collected from a subset of these databases. The datasets used are extracted from these databases depending upon the modelling of the DTI problem, either as classification or regression. Hence the datasets can be binary or continuous for designing the respective predictors. The following table lists the available databases used for DTI (Table **2**).

CONCLUSION

This paper presents the basic drug discovery and drug development process, followed by an overview of the chemogenomic approaches used for drug-target interaction prediction. We have also discussed the associated advantages and disadvantages of each method while considering drug-target interaction prediction as a classification problem. This review clearly indicates that no single approach can be a feasible solution, but a new method can be designed that achieves the accuracy of the deep learning models and, at the same time, is also capable of explaining the classified interactions. The non-linear relationship between drugs and targets can be further justified with the proper selection of features. Hence feature selection methods may be redesigned to get significant features that play a key role in these non-linear relationships. We firmly believe that in the near future, the imbalance of datasets would be addressed while developing machine learning models which tend to generate biased results.

AUTHORS' CONTRIBUTIONS

HB ideated the concept and wrote the first draft with AS contributing to Figures. PS wrote abstract and conclusions and proofread the manuscript.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] *Pharmaceutical Medicine and Translational Clinical Research*; Vohora, D.; Singh, G., Eds.; Academic Press, **2018**.

[2] DiMasi, J.A.; Feldman, L.; Seckler, A.; Wilson, A. Trends in risks associated with new drug development: success rates for investigational drugs. *Clin. Pharmacol. Ther.,* **2010**, *87*(3), 272-277. http://dx.doi.org/10.1038/clpt.2009.295 PMID: 20130567

[3] Agamah, F.E.; Mazandu, G.K.; Hassan, R.; Bope, C.D.; Thomford, N.E.; Ghansah, A.; Chimusa, E.R. Computational/*in silico* methods in drug target and lead prediction. *Brief. Bioinform.,* **2019**, *21*(5), 1663-1675. http://dx.doi.org/10.1093/bib/bbz103 PMID: 31711157

[4] Chen, X.; Guan, N.N.; Sun, Y.Z.; Li, J.Q.; Qu, J. MicroRNA-small molecule association identification: from experimental results to computational models. *Brief. Bioinform.,* **2018**, *21*(1), 47-61. http://dx.doi.org/10.1093/bib/bby098 PMID: 30325405

[5] Wang, C.C.; Zhao, Y.; Chen, X. Drug-pathway association prediction: from experimental results to computational models. *Brief. Bioinform.,* **2020**. http://dx.doi.org/10.1093/bib/bbaa061 PMID: 32393976

[6] Santos, R.; Ursu, O.; Gaulton, A.; Bento, A.P.; Donadi, R.S.; Bologa, C.G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T.I.; Overington, J.P. A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.,* **2017**, *16*(1), 19-34. http://dx.doi.org/10.1038/nrd.2016.230 PMID: 27910877

[7] Patwardhan, B.; Vaidya, A.D. *Natural products drug discovery: accelerating the clinical candidate development using reverse pharmacology approaches.,* **2010**.

[8] Renaud, J.P. The evolving role of structural biology in drug discovery., **2020**, p. 1-122.

[9] Jain, K.K. RNAi and siRNA in target validation. *Drug Discov. Today,* **2004**, *9*(7), 307-309. http://dx.doi.org/10.1016/S1359-6446(04)03050-8 PMID: 15037229

[10] Umashankar, V.G.; Gurunathan, S. Drug discovery: An appraisal. *Int. J. Pharm. Pharm. Sci.,* **2015**, *7*, 59-66.

[11] Suravajhala, P.; Burri, H.V.; Heiskanen, A. Combining aptamers and *in silico* interaction studies to decipher the function of hypothetical proteins. *Eur. Chem. Bull.,* **2014**, *3*(8), 809-810.

[12] Ijaq, J.; Malik, G.; Kumar, A.; Das, P.S.; Meena, N.; Bethi, N.; Sundararajan, V.S.; Suravajhala, P. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics,* **2019**, *20*(1), 14. http://dx.doi.org/10.1186/s12859-018-2554-y PMID: 30621574

[13] Honek, J. Preclinical research in drug development. *Med. Writing,* **2017**, *26*, 5-8.

[14] Friedman, L.M.; Furberg, C.; DeMets, D.L.; Reboussin, D.M.; Granger, C.B. *Fundamentals of clinical trials*; Springer: New York, **2010**. http://dx.doi.org/10.1007/978-1-4419-1586-3

[15] 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*(4), 696-712. http://dx.doi.org/10.1093/bib/bbv066 PMID: 26283676

[16] Pareek, C.S.; Smoczynski, R.; Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.,* **2011**, *52*(4), 413-435. http://dx.doi.org/10.1007/s13353-011-0057-x PMID: 21698376

[17] Reddy, A.S.; Zhang, S. Polypharmacology: drug discovery for the future. *Expert Rev. Clin. Pharmacol.,* **2013**, *6*(1), 41-47. http://dx.doi.org/10.1586/ecp.12.74 PMID: 23272792

[18] Moumbock, A.F.A.; Li, J.; Mishra, P.; Gao, M.; Günther, S. Current computational methods for predicting protein interactions of natural products. *Comput. Struct. Biotechnol. J.,* **2019**, *17*, 1367-1376. http://dx.doi.org/10.1016/j.csbj.2019.08.008 PMID: 31762960

[19] Bender, A.; Glen, R.C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem.,* **2004**, *2*(22), 3204-3218. http://dx.doi.org/10.1039/b409813g PMID: 15534697

[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*(1), 573.

http://dx.doi.org/10.1038/s41467-017-00680-8 PMID: 28924171

[21] Opella, S.J. Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. *Annu. Rev. Anal. Chem. (Palo Alto, Calif.),* **2013**, *6*, 305-328. http://dx.doi.org/10.1146/annurev-anchem-062012-092631 PMID: 23577669

[22] Chen, R.; Liu, X.; Jin, S.; Lin, J.; Liu, J. Machine learning for drug-target interaction prediction. *Molecules,* **2018**, *23*(9), 2208. http://dx.doi.org/10.3390/molecules23092208 PMID: 30200333

[23] Wu, Z.; Li, W.; Liu, G.; Tang, Y. Network-based methods for prediction of drug-target interactions. *Front. Pharmacol.,* **2018**, *9*, 1134. http://dx.doi.org/10.3389/fphar.2018.01134 PMID: 30356768

[24] 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*(5), e1002503. http://dx.doi.org/10.1371/journal.pcbi.1002503 PMID: 22589709

[25] Cheng, F.; Zhou, Y.; Li, W.; Liu, G.; Tang, Y. Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One,* **2012**, *7*(7), e41064. http://dx.doi.org/10.1371/journal.pone.0041064 PMID: 22815915

[26] Wu, Z.; Cheng, F.; Li, J.; Li, W.; Liu, G.; Tang, Y. SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. *Brief. Bioinform.,* **2017**, *18*(2), 333-347. PMID: 26944082

[27] Wu, Z.; Lu, W.; Wu, D.; Luo, A.; Bian, H.; Li, J.; Li, W.; Liu, G.; Huang, J.; Cheng, F.; Tang, Y. *In silico* prediction of chemical mechanism of action *via* an improved network-based inference method. *Br. J. Pharmacol.,* **2016**, *173*(23), 3372-3385. http://dx.doi.org/10.1111/bph.13629 PMID: 27646592

[28] Alaimo, S.; Giugno, R.; Pulvirenti, A. *Recommendation Techniques for Drug–Target Interaction Prediction and Drug Repositioning.Data Mining Techniques for the Life Sciences. Methods in Molecular Biology; Carugo, O*; Eisenhaber, F., Ed.; Humana Press: New York, NY, **2016**, Vol. 1415, pp. 441-462. http://dx.doi.org/10.1007/978-1-4939-3572-7_23

[29] Cheng, F.; Li, W.; Wu, Z.; Wang, X.; Zhang, C.; Li, J.; Liu, G.; Tang, Y. Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J. Chem. Inf. Model.,* **2013**, *53*(4), 753-762. http://dx.doi.org/10.1021/ci400010x PMID: 23527559

[30] Chen, X.; Liu, M.X.; Yan, G.Y. Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.,* **2012**, *8*(7), 1970-1978. http://dx.doi.org/10.1039/c2mb00002d PMID: 22538619

[31] Seal, A.; Ahn, Y.Y.; Wild, D.J. Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. *J. Cheminform.,* **2015**, *7*(1), 40. http://dx.doi.org/10.1186/s13321-015-0089-z PMID: 26300984

[32] Durán, C.; Daminelli, S.; Thomas, J.M.; Haupt, V.J.; Schroeder, M.; Cannistraci, C.V. Pioneering topological methods for network-based drug-target prediction by exploiting a brain-network self-organization theory. *Brief. Bioinform.,* **2018**, *19*(6), 1183-1202. http://dx.doi.org/10.1093/bib/bbx041 PMID: 28453640

[33] Zhang, W.; Lin, W.; Zhang, D.; Wang, S.; Shi, J.; Niu, Y. Recent advances in the machine learning-based drug-target interaction prediction. *Curr. Drug Metab.,* **2019**, *20*(3), 194-202. http://dx.doi.org/10.2174/1389200219666180821094047 PMID: 30129407

[34] 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*(5), e37608. http://dx.doi.org/10.1371/journal.pone.0037608 PMID: 22666371

[35] Meng, F.R.; You, Z.H.; Chen, X.; Zhou, Y.; An, J.Y. Prediction of drug-target interaction networks from the integration of protein sequences and drug chemical structures. *Molecules,* **2017**, *22*(7), 1119. http://dx.doi.org/10.3390/molecules22071119 PMID: 28678206

[36] Ezzat, A.; Wu, M.; Li, X.; Kwoh, C.K. Computational prediction of drug-target interactions *via* Ensemble Learning. *Methods Mol. Biol.,* **2019**, *1903*, 239-254. http://dx.doi.org/10.1007/978-1-4939-8955-3_14 PMID: 30547446

[37] Coelho, E.D.; Arrais, J.P.; Oliveira, J.L. Computational discovery of putative leads for drug repositioning through drug-target interaction prediction. *PLOS Comput. Biol.,* **2016**, *12*(11), e1005219. http://dx.doi.org/10.1371/journal.pcbi.1005219 PMID: 27893735

[38] Rayhan, F.; Ahmed, S.; Mahbub, A.; Jani, R.; Shatabda, S.; Farid, D.M. Cusboost: cluster-based under-sampling with boosting for imbalanced classification. **2017**, p. 1-5.

[39] Seiffert, C.; Khoshgoftaar, T.M.; Van Hulse, J.; Napolitano, A. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.,* **2009**, *40*(1), 185-197. http://dx.doi.org/10.1109/TSMCA.2009.2029559

[40] Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.,* **2002**, *16*, 321-357. http://dx.doi.org/10.1613/jair.953

[41] Rayhan, F.; Ahmed, S.; Md Farid, D.; Dehzangi, A.; Shatabda, S. CFSBoost: Cumulative feature subspace boosting for drug-target interaction prediction. *J. Theor. Biol.,* **2019**, *464*, 1-8. http://dx.doi.org/10.1016/j.jtbi.2018.12.024 PMID: 30578798

[42] 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. http://dx.doi.org/10.1016/j.ymeth.2017.05.016 PMID: 28549952

[43] Chen, L.; Huang, T.; Shi, X.H.; Cai, Y.D.; Chou, K.C. Analysis of protein pathway networks using hybrid properties. *Molecules,* **2010**, *15*(11), 8177-8192. http://dx.doi.org/10.3390/molecules15118177 PMID: 21076385

[44] Bleakley, K.; Yamanishi, Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics,* **2009**, *25*(18), 2397-2403. http://dx.doi.org/10.1093/bioinformatics/btp433 PMID: 19605421

[45] 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*(2), 238-245. http://dx.doi.org/10.1093/bioinformatics/bts670 PMID: 23162055

[46] 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*(12), 3399-3409. http://dx.doi.org/10.1021/ci400219z PMID: 24289468

[47] Gönen, M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics,* **2012**, *28*(18), 2304-2310. http://dx.doi.org/10.1093/bioinformatics/bts360 PMID: 22730431

[48] Zheng, X.; Ding, H.; Mamitsuka, H.; Zhu, S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining,* **2013**, pp. 1025-1033.

[49] 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. Bioinformatics,* **2017**, *14*(3), 646-656. http://dx.doi.org/10.1109/TCBB.2016.2530062 PMID: 26890921

[50] Liu, Y.; Wu, M.; Miao, C.; Zhao, P.; Li, X.L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLOS Comput. Biol.,* **2016**, *12*(2), e1004760. http://dx.doi.org/10.1371/journal.pcbi.1004760 PMID: 26872142

[51] Shi, J.Y.; Zhang, A.Q.; Zhang, S.W.; Mao, K.T.; Yiu, S.M. A unified solution for different scenarios of predicting drug-target interactions *via* triple matrix factorization. *BMC Syst. Biol.,* **2018**, *12*(9), 136. http://dx.doi.org/10.1186/s12918-018-0663-x PMID: 30598094

[52] Tian, K.; Shao, M.; Wang, Y.; Guan, J.; Zhou, S. Boosting compound-protein interaction prediction by deep learning. *Methods,* **2016**, *110*, 64-72. http://dx.doi.org/10.1016/j.ymeth.2016.06.024 PMID: 27378654

[53] Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-learning-based drug–target interaction prediction. *J. Pro-*

teome Res., **2017**, *16*(4), 1401-1409.
http://dx.doi.org/10.1021/acs.jproteome.6b00618 PMID: 28264154

[54] Wan, F.; Zhu, Y.; Hu, H.; Dai, A.; Cai, X.; Chen, L.; Gong, H.; Xia, T.; Yang, D.; Wang, M.W.; Zeng, J. DeepCPI: a deep learning-based framework for large-scale *in silico* drug screening. *Genomics Proteomics Bioinformatics,* **2019**, *17*(5), 478-495.
http://dx.doi.org/10.1016/j.gpb.2019.04.003 PMID: 32035227

[55] Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics,* **2018**, *34*(17), i821-i829.
http://dx.doi.org/10.1093/bioinformatics/bty593 PMID: 30423097

[56] Lee, I.; Keum, J.; Nam, H. DeepConv-DTI: Prediction of drug-target interactions *via* deep learning with convolution on protein sequences. *PLOS Comput. Biol.,* **2019**, *15*(6), e1007129.
http://dx.doi.org/10.1371/journal.pcbi.1007129 PMID: 31199797

[57] Yasuo, N.; Nakashima, Y.; Sekijima, M. CoDe-DTI: Collaborative Deep Learning-based Drug-Target Interaction Prediction. *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),* **2018**, pp. 792-797.

[58] Wang, L.; You, ZH.; Chen, X.; Xia, SX.; Liu, F.; Yan, X.; Zhou, Y.; Song, KJ. A computational-based method for predicting drug–-target interactions by using stacked autoencoder deep neural network. *J. Comput. Biol.,* **2018**, Mar 1;*25*(3), 361-73.

[59] You, J.; McLeod, R.D.; Hu, P. Predicting drug-target interaction network using deep learning model. *Comput. Biol. Chem.,* **2019**, *80*, 90-101.
http://dx.doi.org/10.1016/j.compbiolchem.2019.03.016 PMID: 30939415

[60] Rifaioglu, A.S.; Atalay, V.; Martin, M.J.; Cetin-Atalay, R.; Dogan, T. DEEPScreen: High performance drug-target interaction prediction with convolutional neural networks using 2-DS tructural compound representations. *J. Chem. Inf. Model.,* **2019**, *59*(10), 4438-4449.
PMID: 31518132

[61] Thafar, M.A.; Olayan, R.S.; Ashoor, H.; Albaradei, S.; Bajic, V.B.; Gao, X.; Gojobori, T.; Essack, M. DTiGEMS+: drug-target interaction prediction using graph embedding, graph mining, and similarity-based techniques. *J. Cheminform.,* **2020**, *12*(1), 44.
http://dx.doi.org/10.1186/s13321-020-00447-2 PMID: 33431036

[62] 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.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.,* **2018**, *46*(D1), D1074-D1082.
http://dx.doi.org/10.1093/nar/gkx1037 PMID: 29126136

[63] Kim, S.; Thiessen, P.A.; Bolton, E.E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B.A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S.H. PubChem substance and compound databases. *Nucleic Acids Res.,* **2016**, *44*(D1), D1202-D1213.
http://dx.doi.org/10.1093/nar/gkv951 PMID: 26400175

[64] 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*(Database issue)(Suppl. 1), D198-D201.
http://dx.doi.org/10.1093/nar/gkl999 PMID: 17145705

[65] Hecker, N.; Ahmed, J.; von Eichborn, 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*(Database issue), D1113-D1117.
http://dx.doi.org/10.1093/nar/gkr912 PMID: 22067455

[66] Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M.P.; Overington, J.P.; Papadatos, G.; Smit, I.; Leach, A.R. The ChEMBL database in 2017. *Nucleic Acids Res.,* **2017**, *45*(D1), D945-D954.
http://dx.doi.org/10.1093/nar/gkw1074 PMID: 27899562

[67] Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.,* **2016**, *44*(D1), D1075-D1079.
http://dx.doi.org/10.1093/nar/gkv1075 PMID: 26481350

[68] Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; Schneider, R.; Skoblo, R.; Russell, R.B.; Bourne, P.E.; Bork, P.; Preissner, R. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.,* **2008**, *36*(Database issue)(Suppl. 1), D919-D922.
PMID: 17942422

[69] Kuhn, M.; von Mering, C.; Campillos, M.; Jensen, L.J.; Bork, P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.,* **2008**, *36*(Database issue)(Suppl. 1), D684-D688.
PMID: 18084021

[70] Sterling, T.; Irwin, J.J. ZINC 15–ligand discovery for everyone. *J. Chem. Inf. Model.,* **2015**, *55*(11), 2324-2337.
http://dx.doi.org/10.1021/acs.jcim.5b00559 PMID: 26479676

[71] Yang, J.; Roy, A.; Zhang, Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.,* **2013**, *41*(Database issue), D1096-D1103.
PMID: 23087378