Computational and Structural Biotechnology Journal 20 (2022) 2831–2838

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini review

A brief review of protein–ligand interaction prediction

Lingling Zhao[a,1], Yan Zhu[a,1], Junjie Wang[b,1], Naifeng Wen[c], Chunyu Wang[a,*], Liang Cheng[d,e,*]

[a] *Faculty of Computing, Harbin Institute of Technology, Harbin, China*
[b] *Department of Medical Informatics, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing, China*
[c] *School of Mechanical and Electrical Engineering, Dalian Minzu University, Dalian, China*
[d] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, China*
[e] *NHC and CAMS Key Laboratory of Molecular Probe and Targeted Theranostics, Harbin Medical University, Harbin, China*

ARTICLE INFO

Article history:
Received 14 March 2022
Received in revised form 30 May 2022
Accepted 1 June 2022
Available online 3 June 2022

Keywords:
Protein–ligand interactions
Drug-target binding affinity
Drug discovery
Machine learning

ABSTRACT

The task of identifying protein–ligand interactions (PLIs) plays a prominent role in the field of drug discovery. However, it is infeasible to identify potential PLIs via costly and laborious *in vitro* experiments. There is a need to develop PLI computational prediction approaches to speed up the drug discovery process. In this review, we summarize a brief introduction to various computation-based PLIs. We discuss these approaches, in particular, machine learning-based methods, with illustrations of different emphases based on mainstream trends. Moreover, we analyzed three research dynamics that can be further explored in future studies.

Contents

1. Introduction

Drug discovery is a time-consuming and labor-intensive process that includes the selection, design, and optimization of molecules based on disease-specific target proteins [1]. The task of predicting the interactions between compounds and proteins is the core and foundation of drug discovery, which consists of drug-target interaction (DTI) [2], drug-target binding affinity (DTA) [3], drug-target interaction sites and drug bioactivity on proteins [4,5]. Protein-ligand interaction (PLI), also known as compound-protein interaction (CPI), is most reliably determined by *in vitro* experiments or biochips; however, this is extremely costly in the first screening of a compound, which requires a prohibitively enormous search space [6,7]. To narrow the search space,

[*] Corresponding authors.
E-mail addresses: chunyu@hit.edu.cn (C. Wang), liangcheng@hrbmu.edu.cn (L. Cheng).
[1] These authors contributed equally.

https://doi.org/10.1016/j.csbj.2022.06.004

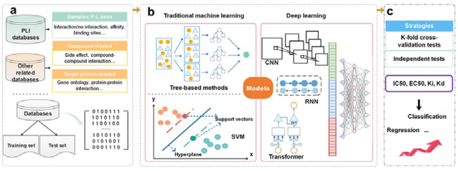

Fig. 1. Workflow of ML methods used in PLI prediction, including (a) benchmark data collection and preprocessing; (b) framework building and model training; and (c) model evaluation.

there is an urgent need to develop more efficient computational approaches.

The increasing publication of large-scale PLI datasets enables the development of traditional machine learning (ML) and deep learning (DL) methods for the prediction of PLIs. The workflow for predicting PLIs using ML methods is shown in Fig. 1. First, the compound–protein pairs and corresponding labels are retrieved from PLI databases and other related databases. In each compound–protein pair, the compound and protein are represented by the feature vectors/matrix from different types of properties (i.e., biological, topological and physicochemical information). Next, the generated feature vectors/matrix and corresponding labels are fed into the ML-based methods for training. After the training stage, the trained model can be tested by different evaluation mechanisms.

2. Research status of PLI prediction

The traditional determination of PLIs using humidity tests involves *in vitro* experiments, biochips, and other classic methods. Due to exorbitant costs, a computational PLI prediction study field has emerged. Researchers have put much effort into this field and have produced excellent results. Currently, there are four types of computation-based PLI methods: ligand-based methods [8], structural methods [9], network-based methods [10] and feature based methods [11].

2.1. Ligand based methods

Ligand based methods have been developed to predict potential PLIs under the hypothesis that ligands with chemical similarity also have similar biological activities and they tend to bind to similar protein targets [12]. Therefore, these methods compare candidate molecules with known protein ligands and predict the interactions based on the similarities between them. These methods do not rely on any knowledge about the target protein, but meanwhile performs poorly for targets with an insufficient number of known ligands.

2.2. Structural methods

Structural methods use the three-dimensional structure of proteins and ligands and molecular docking to simulate the interaction between proteins and ligands and finally utilize the scoring function to evaluate the conformation [13,14]. Structural methods can be divided into three categories by the type of scoring function: the classic scoring function method [15], machine learning scoring function method [16,17] and deep learning scoring function method [18,19]. The core of structural methods is to accurately model the three-dimensional structure of proteins and compounds. Although structural methods can obtain better prediction performance, they often take a certain amount of computing time. In addition, they fail to predict interactions with unknown structures of proteins or compounds. Therefore, it is difficult to screen compound–protein pairs on a large scale, which seriously limits the application scope of this kind of approach.

2.3. Network based methods

Network based methods predict the PLI based on various biological networks and graph theory. A number of computational methods model the relationship between compounds and proteins as a bipartite network [20]. Moreover, PLI-related biological networks, such as protein–protein interactions, drug–drug interactions and drug–disease interactions, have been integrated into a heterogeneous network [9]. The potential interaction information is learned from heterogeneous data from diverse sources to boost the accuracy of DTI prediction tasks [21–23]. However, those prediction approaches are shallow-learning methods that cannot fully extract deep and complex associations between compounds and proteins.

2.4. Feature based methods

Feature based methods are widely used in drug-target interaction prediction studies [24]. These methods predict PLI in a machine learning framework. Feature vectors of drug-target pairs are obtained from their properties or by learning from raw data, and then fed into various classifiers or regressors [25]. Researchers have conducted a plenty of research from many perspectives and their studies are introduced in detail in the following sections. In addition, since both ligand-based and target-based aspects are considered in feature based methods, they can be assigned to the so-called "chemogenomics" approaches [26].

3. Machine learning in PLI prediction

Existing models typically employ the simplified molecular-input line entry system (SMILES) [27,28], molecular structure [29], protein sequences [30], secondary structure of protein [31], gene ontology [32], and other descriptors of predefined molecules and proteins as input features. Then, these inputs were trained by a variety of network frameworks, such as convolutional neural networks (CNNs) [33], recurrent neural networks (RNNs) [34], graph neural networks (GNNs) [35], and Transformer network structures and their variants, to realize the prediction of PLI-related tasks, such as DTI, DTA, and activity [36]. Fig. 1 illustrates a flowchart describing the three generic steps used by these computational approaches for predicting PLIs. Table 1 and Table 2 summarize the typical methods to predict PLIs based on ML in recent years in terms of the input protein/compound features, protein/compound feature extractors, final computational methods, and website. Studies regard the DTI prediction task as a binary classification problem corresponding to the articles in Table 1. These methods, which yield 1 if the DTI is active and 0 otherwise, are concerned about the existence of a DTI. However, other researchers doubt that using classification methods to address the DTI prediction problem loses valuable information about the strength of the interaction between proteins and ligands. The studies in Table 2 considered the PLI problem as a regression task to predict the binding affinity score. It can also be seen in Table 2 that methods [37–39] solve both tasks. The binding affinity, which can be determined by experimental methods, is defined as the strength of the binding interaction between a protein and a ligand.

From Tables 1 and 2, it can be seen that the traditional ML methods are gradually being phased out and replaced by DL technologies, particularly the utilization of diverse neural networks and learning mechanisms. In the following section, we summarized

Table 1
PLI prediction methods as classification tasks based on the ML framework in recent years[a].

Tool[b]	Date	Input protein features	Input compound features	Protein feature extractor	Compound feature extractor	Methods
DeepDTIs [69]	03/2017	Protein sequence composition descriptors	Extended connectivity fingerprints	–	–	DBN
DDR [70]	01/2018	Similarity measures	Similarity measures	–	–	RF
CPI-GNN [19]	07/2018	N-gram amino acids	Molecular graphs	CNN	GNN	Softmax classifier
DeepConv-DTI [18]	06/2019	Local residue patterns	PubChem fingerprints	Convolution and global max-pooling layers	Fully connected layer	Fully connected layer
DTI-CDF [71]	12/2019	Similarity-based features	Similarity-based features	–	–	Cascade deep forest
DEEPScreen [72]	01/2020	–	2-D compound images	–	Convolutional and pooling layers	Fully connected layers
TransformerCPI [54]	05/2020	Amino acid sequence	CNN	Graph structure	GCNs	Transformer with self-attention mechanism
DTI-CNN [73]	08/2020	Similarity matrix	Similarity matrix	Random walk with restart	Random walk with restart	Fully connected layer
MolTrans [52]	10/2020	Substructure embedding	Substructure embedding	Transformer encoder	Transformer encoder	Linear layer
BridgeDPI [35]	02/2021	K-mer/sequence features	Fingerprint/sequence features	Perceptron layers	Perceptron layers	GNN and a full connected layer
CSConv2d [74]	04/2021	–	2-D structural representations	–	A channel and spatial attention mechanism	Fully connected layer
GADTI [75]	04/2021	Similarity data	Similarity data	Heterogeneous network	Heterogeneous network	Graph autoencoder
LGDTI [76]	04/2021	K-mer	Molecular fingerprint	Graph convolutional network and DeepWalk	Graph convolutional network and DeepWalk	RF
PretrainDPI [77]	05/2021	Pretrained models	Molecular graph	CNN	GraphNet	Fully connected layers
X-DPI [51]	06/2021	Structure and sequence features	Atomic features	TAPE embedding	Mol2vec embedding	Transformer decoder
MultiDTI [78]	07/2021	N-gram embedding	N-gram embedding	Deep downsampling residual module	Deep downsampling residual module	Multilayer perceptron
HyperAttentionDTI [79]	10/2021	Amino acid sequences	SMILES strings	CNN and attention mechanism	CNN and attention mechanism	Fully connected layer
DTIHNC [80]	02/2022	Protein-protein interactions, protein-disease associations	Drug-drug interactions, drug-disease associations, drug-side-effects associations	Denoising autoencoder	Denoising autoencoder	CNN module
HIDTI [81]	03/2022	Protein sequences, protein–protein similarities, protein–protein interactions, protein-disease interactions	SMILES strings, drug-drug interactions, drug-side effect associations, drug-disease associations	A residual block	A residual block	Fully connected layers
HGDTI [82]	04/2022	Node features encoding (interactions, similarities, associations)	Node features encoding (interactions, similarities, associations)	BiLSTM	BiLSTM	Fully connected layers

Note: "-" in the table indicates that there is no such information in the corresponding article.

[a] Abbreviations: DBN – deep belief network; RF – random forest; CNN – convolutional neural network; GNN – graph neural network; GCNs – graph convolutional networks; TAPE – tasks assessing protein embeddings; SMILES – simplified molecular-input line-entry system; BiLSTM – bidirectional long short-term memory;

[b] URL addresses for the listed tools: DeepDTIs – https://github.com/Bjoux2/DeepDTIs; DDR – https://bitbucket.org/RSO24/ddr; CPI-GNN – https://github.com/masashit-subaki; DeepConv-DTI – https://github.com/GIST-CSBL/DeepConv-DTI; DTI-CDF – https://github.com//a96123155/DTI-CDF; DEEPscreen – https://github.com/cansyl/DEEPscreen; transformerCPI – https://github.com/lifanchen-simm/transformerCPI; DTI-CNN – https://github.com/MedicineBiology-AI/DTI-CNN; MolTrans – https://github.com/kexinhuang12345/moltrans; BridgeDPI – https://github.com/DeepAAI/BridgeDPI; CSConv2d – https://doi.org/10.4121/uuid:547e8014-d662-4852–9840-c1e-f065d03ef; GADTI – https://github.com/shulijiuba/GADTI; PretrainDPI – https://github.com/QHwan/PretrainDPI; MultiDTI – https://github.com/Deshan-Zhou/MultiDTI; HyperAttentionDTI – https://github.com/zhaoqichang/HpyerAttentionDTI; DTIHNC – https://github.com/ningq669/DTIHNC; HIDTI – https://github.com/DMCB-GIST/HIDTI; HGDTI – https://bioinfo.jcu.edu.cn/hgdti.

several research trends of the machine learning based PLI prediction from the relevant literature in recent decades.

Input protein and compound features Many previous studies have applied manually operated descriptors such as similarity and molecular fingerprints, as well as other composition information, to drive PLI predictions [40–43]. Sequence descriptors, which include SMILES strings and amino acid sequences, are commonly used by encoding sequences in numerical matrices via one-hot or word embedding (such as Prot2Vec and Mol2Vec) [38,44,45]. The sequence representation only considers the primary structure information and limits the learning capability. To more effectively represent compounds and proteins, graph-based features have also

Table 2
PLI prediction methods as regression tasks based on the ML framework in recent years[a].

Tool[b]	Date	Input protein features	Input compound features	Protein feature extractor	Compound feature extractor	Methods
SimBoost [26]	04/2017	Target similarity	Drug similarity	–	–	Gradient boosting tree model
ACNN [83]	2017	Atomic coordinates	Atomic coordinates	Atomic convolution layer	Atomic convolution layer	Atomic fully connected layer
DeepDTA [84]	09/2018	Label encoding	Label encoding	CNN blocks	CNN blocks	Fully connected layer
DeepAffinity [46]	02/2019	Structural property sequence representation	Structural property sequence representation	Seq2seq autoencoders	Seq2seq autoencoders	Unified RNN-CNN
WideDTA [85]	02/2019	Textual information	Textual information	CNN blocks	CNN blocks	Fully connected layers
GraphDTA [86]	06/2019	One-hot encoding	Molecular graph	Convolutional layers	4 graph neural network variants	Fully connected layers
RFScore [17]	08/2019	36 intermolecular features	36 intermolecular features	–	–	Random forest
AttentionDTA [36]	11/2019	Label encoding	Label encoding	CNN block	CNN block	Attention block- fully connected layers
Taba [87]	01/2020	The average distance between pairs of atoms	The average distance between pairs of atoms	–	–	Machine-learning model
GAT_GCN [88]	04/2020	Peptide frequency	Graph structure	CNN	GCN	Fully connected layers
SAnDReS [89]	05/2020	Docking scores	Docking scores	–	–	Machine-learning model
DeepCDA [90]	05/2020	N-gram embedding	SMILES sequence	CNN-LSTM-Two-sided attention mechanism	CNN-LSTM-Two-sided attention mechanism	Fully connected layers
DGraphDTA [91]	06/2020	Protein graph	Molecular graph	GNN	GNN	Fully connected layers
JoVA [92]	08/2020	Multiple unimodal representations	Multiple unimodal representations	Joint view attention module	Joint view attention module	Prediction model
Fusion [93]	11/2020	Atomic representation	Atomic representation	CNNs	SG-GCNs	Fully connected layers
DeepGS [44]	2020	Symbolic sequences	Molecular structure	Prot2Vec-CNN-BiGRU blocks	Smi2Vec-CNN-BiGRU blocks	Fully connected layer
DeepDTAF [94]	01/2021	Sequence, structural property information	SMILES string	Dilated/traditional convolution layers	Dilated convolution layers	Fully connected layers
GanDTI [37] (classification and regression)	03/2021	Protein sequences	Molecule fingerprints-adjacency matrix	Attention module	Residual graph neural network	MLP
Multi-PLI [38] (classification and regression)	04/2021	One-hot vectors	One-hot vectors	CNN blocks	CNN blocks	Fully connected layers
ML-DTI [95]	04/2021	Protein sequences	SMILES string	CNN block (mutual learning)	CNN block (mutual learning)	Linear transformation layers
DEELIG [47]	06/2021	Atomic level-structural information-sequences	Physical properties-fingerprints	CNN	Fully connected layers	Fully connected layers
GEFA [55]	07/2021	Sequence embedding features	Graph representation	GCN	GCN	Linear layers
SAG-DTA [96]	08/2021	Label encoding	Molecular graph	CNN	Graph convolutional layer-SAGPooling layer	Fully connected layers
Tanoori et al. [97]	08/2021	SW sequence similarity	CS similarity	–	–	GBM
EmbedDTI [56]	11/2021	Amino acids	Structural information	CNN	Attention-GCNs	Fully connected layers
DeepPLA [45]	12/2021	Protein sequences (ProSE)	SMILES strings (Mol2Vec)	Head CNN modules-ResNet-based CNN module	Head CNN modules-ResNet-based CNN module	BiLSTM module-MLP module
DeepGLSTM [98]	01/2022	Amino acids	Adjacency representation	BiLSTM	GCN	Fully connected layers

Table 2 (*continued*)

Tool[b]	Date	Input protein features	Input compound features	Protein feature extractor	Compound feature extractor	Methods
MGraphDTA [99]	01/ 2022	Integers	Graph structure	Multiscale convolutional neural network	GNN	MLP
FusionDTA [100]	01/ 2022	word embeddings	SMILES strings	BiLSTM	BiLSTM	Multi-head linear attention blocks/Fully connected layer
HoTS [39] (classification and regression)	02/ 2022	Protein sequences	Morgan/circular fingerprints	Transformer blocks	Transformer blocks	Fully connected layers
ELECTRA-DTA [101]	03/ 2022	Protein sequences	SMILES string	Squeeze-and-excitation convolutional neural network blocks	Squeeze-and-excitation convolutional neural network blocks	Fully connected layers

Note: “-” in the table indicates that there is no such information in the corresponding article.

[a] Abbreviations: CNN – convolutional neural network; GNN – graph neural network; GCNs – graph convolutional networks; LSTM – long short-term memory; SG-CNNs – spatial graph neural networks; BiGRU – bidirectional gate recurrent unit; MLP – multilayer perceptron; GCN – graph convolutional network; SW – Smith-Waterman; CS – chemical structure; GBM – gradient boosting machine; BiLSTM – bidirectional long short-term memory;

[b] URL addresses for the listed tools: SimBoost – https://zenodo.org/record/164436; ACNN – https://github.com/deepchem/deepchem; DeepDTA – https://github.com/hkmztrk/DeepDTA; DeepAffinity – https://github.com/Shen-Lab/DeepAffinity; GraphDTA – https://github.com/thinng/GraphDTA; Taba – https://github.com/azevedolab/taba; SAnDReS – https://github.com/azevedolab/sandres; DeepCDA – https://github.com/LBBSoft/DeepCDA; Fusion – https://github.com/llnl/fast; DeepGS – https://github.com/jacklin18/DeepGS; DeepDTAF – https://github.com/KailiWang1/DeepDTAF; GanDTI – https://github.com/shuyu-wang/GanDTI; Multi-PLI – https://github.com/Deshan-Zhou/Multi-PLI; ML-DTI – https://github.com/guaguabujianle/ML-DTI.git; DEELIG – https://github.com/asadahmedtech/DEELIG; GEFA – https://github.com/ngminhtri0394/GEFA; EmbedDTI – https://github.com/Aurorayuan/EmbedDTI; DeepPLA – https://github.com/David-BominWei/DeepPLA; DeepGLSTM – https://github.com/MLlab4CS/DeepGLSTM.git; MGraphDTA – https://github.com/guaguabujianle/MGraphDTA; FusionDTA – https://github.com/yuanweining/FusionDTA; HoTS – https:// github. com/ GIST- CSBL/ HoTS; ELECTRA-DTA – https://github.com/IILab-Resource/ELECTRA-DTA.

been widely employed. In the graph representing PLIs, the protein is modeled as a graph structure where nodes are residues and the edge information is provided by the contact map [46]. Researchers are also working on leveraging 3D structural information. For instance, a complex is cropped into a cubic box [47]. With advancements in protein structure prediction and the intuitiveness of 3D information, structural information will have significant research value in predicting PLIs and will be a promising study topic in the future.

Protein and compound feature extractors Some works adopted the same structure to handle the representation of proteins and compounds, while others created separate feature extractors for the two inputs [48]. These extractors include CNN-based models, RNN-based models, attention mechanism-based models, and GNN-based models. CNN has the benefit of being able to catch crucial local patterns in the whole space. However, there are certain drawbacks. Protein residues that are not adjacent can be quite close in structure. CNN has failed to obtain this long-distance dependence. RNN-based modules, such as the long short-term memory (LSTM) network, are suitable for learning long-term dependency from compound and protein sequence inputs, compensating for the CNN disadvantage [49]. However, due to the difficulty of encoding long-range dependencies, the training of RNN becomes problematic when the sequence is long. Furthermore, to overcome the difficulty in the interpretation of black-box-like neural networks, researchers have solved this problem with attention mechanism-based models. The attention mechanism can be effectively visualized by mapping regions with high weight to the known 3D protein–compound complex structures, thus indicating the biological significance of the model [50]. However, its operation, as in the case of Transformer with attention mechanism, requires a large amount of computer memory. However, Transformer has released a series of new and updated versions that offer broad prospects for predicting PLI tasks [51,52]. The GNN is a kind of neural network dedicated to extracting graph structure information [53]. GNN-based models, such as the graph convolutional neural network (GCN) and Graph Isomorphism Network (GIN), Graph Attention Networks (GAT), are commonly applied in computer-aided drug design [54–57].

4. Challenges of machine learning in PLIs

ML methods have attracted increasing attention in the fields of bioinformatics and chemical informatics [58,59]. However, the complexity of proteins, compounds and their interactions make ML-based PLI prediction challenging for the following reasons:

(i) In the field of ML, feature engineering is used in traditional ML frameworks to select related features for downstream tasks [60]. DL methods try to avoid complicated feature engineering and learn abstract representation automatically [61]. Since the rise of large-scale data and improvements in computing power, DL techniques have enabled unprecedented breakthroughs in many areas, including image processing, natural language processing and bioinformatics [62]. PLIs involve complex physical, chemical, and biological processes. The combination of compounds and proteins is the consequence of various processes that are highly concentrated. Therefore, proteins and molecules are far more sophisticated than images, language, and other items.

(ii) PLI prediction is mainly modeled as a supervised classification or regression problem in the ML-based method [63]. Supervised learning requires large-scale high-quality labeled datasets. In the case of an insufficient quantity of labeled PLI datasets, research works apply unsupervised learning, semi-supervised learning, or self-supervised learning to predict the PLIs [64–66]. In particular, unsupervised pretrained models on large text corpora have shown remarkable performance on various natural language processing tasks. Consequently, some unsupervised pretrained models for embedding the amino acid sequence and SMILES have been proposed in recent years [67]. Unfortunately, due to the relatively immature understanding of the interaction mechanism between proteins and compounds, there remains a lack of specific unsupervised DL models for the PLI task.

(iii) In addition to unlabeled data, existing ML methods also do not take full advantage of knowledge about proteins and compounds. The related knowledge can be expressed in various forms. Protein-related knowledge includes primary structure, secondary structure, tertiary structure, functional annotation, motif, and various physical and chemical attributes. Compound-related knowledge includes molecular structure, functional groups and molecular properties. Which type of knowledge is connected to PLIs and how to select, represent, and incorporate knowledge into data-driven ML models are progressive theoretical questions.

5. Discussion and analysis

The increase in high-quality and large-scale PLI datasets has enabled the development of traditional ML or DL methods for the prediction of PLIs. Compared with traditional ML methods, DL methods have shown significant advantages, such as feature generation automation and the ability to capture complex nonlinear relationships. It is also worth noting that there is still much room for improvement in prediction accuracy, robustness, generalization, and interpretability.

First, the performance of existing DL methods for PLIs is still poor due to the complexity of the PLI problem itself and the limited data available. Several DL-based models also fail to make good use of large-scale unlabeled data. In addition, the selection of input representation is a vital part of PLI prediction [68]. Most of the existing DL methods train deep neural networks directly on low-level representations, such as amino acid sequences and SMILESs. The primary structure input may affect the model generalizability in predicting the novel PLI. Researchers should pay more attention to improving the generalizability of models in future studies. Furthermore, the lack of interpretability of DL-based methods limits their practical applications, as the potential factors influencing the prediction results are unknown. Some methods use attention mechanisms to capture interaction sites, but they are still unable to explain the mechanisms behind the PLI. Researchers should attempt to design an interpretable DL model to predict PLIs.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (grant no. 62171164, 62102191, 61872114 & 62131004).

References

[1] Lavecchia A, Di Giovanni C. Virtual Screening Strategies in Drug Discovery: A Critical Review. Curr Med Chem 2013;20:2839–60.

[2] Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics 2009;25:2397–403.

[3] Ma W, Yang L, He L. Overview of the detection methods for equilibrium dissociation constant KD of drug-receptor interaction. J Pharm Anal 2018;8:147–52.

[4] Wang P, Zhuo X, Chu W, Tang X. Exenatide-loaded microsphere/thermosensitive hydrogel long-acting delivery system with high drug bioactivity. Int J Pharm 2017;528:62–75.

[5] Zhou HY, Gao M, Skolnick J. Comprehensive prediction of drug-protein interactions and side effects for the human proteome. Sci Rep-Uk 2015;5.

[6] Kim S, Chen J, Cheng TJ, Gindulyte A, He J, He SQ, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47:D1102–9.

[7] Gilson MK, Liu TQ, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 2016;44:D1045–53.

[8] Ahneman DT, Estrada JG, Lin S, Dreher SD, Doyle AG. Predicting reaction performance in C-N cross-coupling using machine learning. Science 2018;360:186–90.

[9] Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. Bioinformatics 2017;33:2337–44.

[10] Ding YJ, Tang JJ, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. Neurocomputing 2019;325:211–24.

[11] You JY, McLeod RD, Hu PZ. Predicting drug-target interaction network using deep learning model. Comput Biol Chem 2019;80:90–101.

[12] Balakin KV, Tkachenko SE, Lang SA, Okun I, Ivashchenko AA, Savchuk NP. Property-based design of GPCR-targeted library. J Chem Inf Comp Sci 2002;42:1332–42.

[13] Filgueira de Azevedo W, Jr., dos Santos GC, dos Santos DM, Olivieri JR, Canduri F, Silva RG, et al. Docking and small angle X-ray scattering studies of purine nucleoside phosphorylase. Biochem Biophys Res Commun 2003;309:923–8.

[14] Levin NMB, Pintro VO, de Avila MB, de Mattos BB, De Azevedo WF, Jr. Understanding the Structural Basis for Inhibition of Cyclin-Dependent Kinases. New Pieces in the Molecular Puzzle. Curr Drug Targets 2017;18:1104–11.

[15] Wang L, You ZH, Chen X, Xia SX, Liu F, Yan X, et al. A Computational-Based Method for Predicting Drug-Target Interactions by Using Stacked Autoencoder Deep Neural Network. J Comput Biol 2018;25:361–73.

[16] Xie LW, He S, Song XY, Bo XC, Zhang ZN. Deep learning-based transcriptome data classification for drug-target interaction prediction. Bmc. Genomics 2018;19.

[17] Wojcikowski M, Siedlecki P, Ballester PJ. Building Machine-Learning Scoring Functions for Structure-Based Prediction of Intermolecular Binding Affinity. Methods Mol Biol 2019;2053:1–12.

[18] Lee I, Keum J, DeepConv-DTI Nam H. Prediction of drug-target interactions via deep learning with convolution on protein sequences. Plos Comput Biol 2019;15.

[19] Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics 2019;35:309–18.

[20] Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol 2007;25:1119–26.

[21] Chen X, Liu MX, Yan GY. Drug-target interaction prediction by random walk on the heterogeneous network. Mol Biosyst 2012;8:1970–8.

[22] Hu BF, Wang H, Wang LT, Yuan WH. Adverse Drug Reaction Predictions Using Stacking Deep Heterogeneous Information Network Embedding Approach. Molecules 2018;23.

[23] Zeng XX, Zhu SY, Lu WQ, Liu ZH, Huang J, Zhou YD, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci 2020;11:1775–97.

[24] Sachdev K, Gupta MK. A comprehensive review of feature based methods for drug target interaction prediction. J Biomed Inform 2019;93:103159.

[25] Bijral RK, Singh I, Manhas J, Sharma V. Exploring Artificial Intelligence in Drug Discovery. A Comprehensive Review Arch Comput Method E 2021.

[26] He T, Heidemeyer M, Ban FQ, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. J Cheminformatics 2017;9.

[27] Toropov AA, Toropova AP, Mukhamedzhanova DV, Gutman I. Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). Indian J Chem A 2005;44:1545–52.

[28] Weininger D. Smiles, a Chemical Language and Information-System .1. Introduction To Methodology and Encoding Rules. J Chem Inf Comp Sci. 1988;28:31-6.

[29] Zhou WW, Wang X, Chang J, Cheng CL, Miao CG. The molecular structure and biological functions of RNA methylation, with special emphasis on the roles of RNA methylation in autoimmune diseases. Crit Rev Cl Lab Sci 2021.

[30] Kumar S, Nei M, Dudley J, Tamura K. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings Bioinf 2008;9:299–306.

[31] Chen Z, Zhao P, Li FY, Marquez-Lago TT, Leier A, Revote J, et al. iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. Briefings Bioinf 2020;21:1047–57.

[32] Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S, et al. AmiGO: online access to ontology and annotation data. Bioinformatics 2009;25:288–9.
[33] Sun YA, Xue B, Zhang MJ, Yen GG. Evolving Deep Convolutional Neural Networks for Image Classification. Ieee T Evolut Comput 2020;24:394–407.
[34] Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR. ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. Future Gener Comp Sy 2021;115:279–94.
[35] Wu Y, Gao M, Zeng M, Chen F, Zhang J. BridgeDPI: A Novel Graph Neural Network for Predicting Drug-Protein Interactions. 2021.
[36] Zhao QC, Xiao F, Yang MY, Li YH, Wang JX. AttentionDTA: prediction of drug-target binding affinity using attention model. Ieee Int C Bioinform 2019:64–9.
[37] Wang SY, Shan P, Zhao YL, GanDTI Zuo L. A multi-task neural network for drug-target interaction prediction. Comput Biol Chem 2021;92.
[38] Hu F, Jiang J, Wang D, Zhu M, Yin P. Multi-PLI: interpretable multi-task deep learning model for unifying protein-ligand interaction datasets. J Cheminform 2021;13:30.
[39] Lee I, Nam H. Sequence-based prediction of protein binding regions and drug-target interactions. J Cheminform 2022;14:5.
[40] Faulon JL, Misra M, Martin S, Sale K, Sapra R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. Bioinformatics 2008;24:225–33.
[41] Steffen A, Kogej T, Tyrchan C, Engkvist O. Comparison of Molecular Fingerprint Methods on the Basis of Biological Profile Data. J Chem Inf Model 2009;49:338–47.
[42] Ding H, Takigawa I, Mamitsuka H, Zhu SF. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Briefings Bioinf 2014;15:734–47.
[43] O'Boyle NM, Sayle RA. Comparing structural fingerprints using a literature-based similarity benchmark. J Cheminformatics 2016;8.
[44] Lin X, Zhao KQ, Xiao T, Quan Z, Wang ZJ, Yu PS. DeepGS: Deep Representation Learning of Graphs and Sequences for Drug-Target Binding Affinity Prediction. Front Artif Intel Ap 2020;325:1301–8.
[45] Wei B, Gong X. DeepPLA: a novel deep learning-based model for protein-ligand binding affinity prediction. 2021.
[46] Karimi M, Wu D, Wang Z, Shen Y. DeepAffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. Bioinformatics 2019;35:3329–38.
[47] Ahmed A, Mam B, Sowdhamini R. DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. Bioinform Biol Insig. 2021;15.
[48] Dhakal A, McKay C, Tanner JJ, Cheng J. Artificial intelligence in the prediction of protein-ligand interactions: recent advances and future directions. Brief Bioinform 2022;23.
[49] Ghimire S, Deo RC, Raj N, Mi JC. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. Appl. Energ 2019;253.
[50] Niu ZY, Zhong GQ, Yu H. A review on the attention mechanism of deep learning. Neurocomputing 2021;452:48–62.
[51] Wang P, Zheng S, Jiang Y, Li C, Liu J, Wen C, et al. X-DPI: A structure-aware multi-modal deep learning model for drug-protein interactions prediction. 2021.
[52] Huang KX, Xiao C, Glass LM, Sun JM. MolTrans: Molecular Interaction Transformer for drug-target interaction prediction. Bioinformatics 2021;37:830–6.
[53] Asif NA, Sarker Y, Chakrabortty RK, Ryan MJ, Ahamed MH, Saha DK, et al. Graph Neural Network: A Comprehensive Review on Non-Euclidean Space. IEEE Access 2021;9:60588–606.
[54] Chen LF, Tan XQ, Wang DY, Zhong FS, Liu XH, Yang TB, et al. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Bioinformatics 2020;36:4406–14.
[55] Nguyen TM, Nguyen T, Le TM, Tran T. GEFA: early fusion approach in drug-target affinity prediction. IEEE/ACM Trans Comput Biol Bioinf 2021;PP:1-..
[56] Jin Y, Lu JR, Shi RH, EmbedDTI Yang Y. Enhancing the Molecular Representations via Sequence Embedding and Graph Convolutional Network for the Prediction of Drug-Target Interaction. Biomolecules 2021;11.
[57] Wekesa JS, Meng J, Luan YS. A deep learning model for plant lncRNA-protein interaction prediction with graph attention. Mol Genet Genomics 2020;295:1091–102.
[58] Riera-Fernandez P, Munteanu CR, Dorado J, Martin-Romalde R, Duardo-Sanchez A, Gonzalez-Diaz H. From Chemical Graphs in Computer-Aided Drug Design to General Markov-Galvez Indices of Drug-Target, Proteome, Drug-Parasitic Disease, Technological, and Social-Legal Networks. Curr Comput-Aid Drug 2011;7:315–37.
[59] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotec 2015;13:8–17.
[60] Zhang M, Li FY, Marquez-Lago TT, Leier A, Fan C, Kwoh CK, et al. MULTiPly: a novel multi-layer predictor for discovering general and specific types of promoters. Bioinformatics 2019;35:2957–65.
[61] Zhu Y, Li F, Xiang D, Akutsu T, Song J, Jia C. Computational identification of eukaryotic promoters based on cascaded deep capsule neural networks. Brief Bioinform 2021;22.
[62] Otter DW, Medina JR, Kalita JK. A Survey of the Usages of Deep Learning for Natural Language Processing. Ieee T Neur Net Lear 2021;32:604–24.
[63] Sinha K, Sun CC, Kamari R, Bettermann K. Current status and future prospects of pathophysiology-based neuroprotective drugs for the treatment of vascular dementia. Drug Discov Today 2020;25:793–9.
[64] Sajadi SZ, Chahooki MAZ, Gharaghani S, Abbasi K. AutoDTI plus plus : deep unsupervised learning for DTI prediction by autoencoders. Bmc. Bioinformatics 2021;22.
[65] Yu WM, Cheng XA, Li ZB, Jiang ZR. Predicting Drug-Target Interactions Based on an Improved Semi-Supervised Learning Approach. Drug Develop Res 2011;72:219–24.
[66] Wang XQ, Yang YN, Li KL, Li WT, Li F, Peng SL. BioERP: biomedical heterogeneous network-based self-supervised representation learning approach for entity relationship predictions. Bioinformatics 2021;37:4793–800.
[67] Yu LH, Su YS, Liu YS, Zeng XX. Review of unsupervised pretraining strategies for molecules representation. Brief Funct Genomics 2021;20:323–32.
[68] Zhu Y, Jia CZ, Li FY, Song JN. Inspector: a lysine succinylation predictor based on edited nearest-neighbor undersampling and adaptive synthetic oversampling. Anal Biochem 2020;593.
[69] Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, et al. Deep-Learning-Based Drug-Target Interaction Prediction. J Proteome Res 2017;16:1401–9.
[70] Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug-target interactions using graph mining and machine learning approaches. Bioinformatics 2018;34:3779.
[71] Chu YY, Kaushik AC, Wang XG, Wang W, Zhang YF, Shan XQ, et al. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Briefings Bioinf 2021;22:451–62.
[72] Rifaioglu AS, Nalbat E, Atalay V, Martin MJ, Cetin-Atalay R, Dogan T. DEEPScreen: high performance drug-target interaction prediction with convolutional neural networks using 2-D structural compound representations. Chem Sci 2020;11:2531–57.
[73] Peng JJ, Li JY, Shang XQ. A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network. Bmc. Bioinformatics 2020;21.
[74] Wang X, Liu D, Zhu J, Rodriguez-Paton A, Song T. CSConv2d: A 2-D Structural Convolution Neural Network with a Channel and Spatial Attention Mechanism for Protein-Ligand Binding Affinity Prediction. Biomolecules 2021;11.
[75] Liu Z, Chen Q, Lan W, Pan H, Hao X, Pan S. GADTI: Graph Autoencoder Approach for DTI Prediction From Heterogeneous Network. Front Genet 2021;12:650821.
[76] Zhao BW, You ZH, Hu L, Guo ZH, Wang L, Chen ZH, et al. A Novel Method to Predict Drug-Target Interactions Based on Large-Scale Graph Representation Learning. Cancers (Basel) 2021;13.
[77] Kim Q, Ko JH, Kim S, Park N, Jhe W. Bayesian neural network with pretrained protein embedding enhances prediction accuracy of drug-protein interaction. Bioinformatics 2021.
[78] Zhou DS, Xu ZJ, Li WT, Xie XL, Peng SL. MultiDTI: drug-target interaction prediction based on multi-modal representation learning to bridge the gap between new chemical entities and known heterogeneous network. Bioinformatics 2021;37:4485–92.
[79] Zhao QC, Zhao HC, Zheng K, Wang JX. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 2022;38:655–62.
[80] Jiang L, Sun JH, Wang Y, Ning Q, Luo N, Yin MH. Identifying drug-target interactions via heterogeneous graph attention networks combined with cross-modal similarities. Briefings Bioinf 2022.
[81] Soh J, Park S, Lee H. HIDTI: integration of heterogeneous information to predict drug-target interactions. Sci Rep 2022;12:3793.
[82] Yu LY, Qiu WR, Lin WZ, Cheng X, Xiao X, Dai JX. HGDTI: predicting drug-target interaction by using information aggregation based on heterogeneous graph neural network. Bmc. Bioinformatics 2022;23.
[83] Gomes J, Ramsundar B, Feinberg EN, Pande VS. Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. 2017.
[84] Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 2018;34:i821–9.
[85] Öztürk H, Ozkirimli E, Özgür A. WideDTA: prediction of drug-target binding affinity. 2019.
[86] Nguyen T, Le H, Quinn TP, Nguyen T, Le TD, Venkatesh S. GraphDTA: predicting drug-target binding affinity with graph neural networks. Bioinformatics 2021;37:1140–7.
[87] da Silva AD, Bitencourt-Ferreira G, de Azevedo WF, Jr. Taba: A Tool to Analyze the Binding Affinity. J Comput Chem 2020;41:69–73.
[88] Wang XF, Liu YF, Lu F, Li HF, Gao P, Wei DQ. Dipeptide Frequency of Word Frequency and Graph Convolutional Networks for DTA Prediction. Front Bioeng. Biotech 2020;8.
[89] Bitencourt-Ferreira G, Rizzotto C, de Azevedo Junior WF. Machine Learning-Based Scoring Functions, Development and Applications with SAnDReS. Curr Med Chem 2021;28:1746–56.
[90] Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. DeepCDA: deep cross-domain compound-protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics 2020;36:4633–42.
[91] Jiang MJ, Li Z, Zhang SG, Wang S, Wang XF, Yuan Q, et al. Drug-target affinity prediction using graph neural network and contact maps. Rsc Adv 2020;10:20701–12.

[92] Agyemang B, Wu WP, Kpiebaareh MY, Lei ZH, Nanor E, Chen L. Multi-view self-attention for interpretable drug-target interaction prediction. J Biomed Inform 2020;110.

[93] Jones D, Kim H, Zhang X, Zemla A, Stevenson G, Bennett WFD, et al. Improved Protein-Ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. J Chem Inf Model 2021;61:1583–92.

[94] Wang KL, Zhou RY, Li YH, Li M. DeepDTAF: a deep learning method to predict protein-ligand binding affinity. Briefings Bioinf 2021;22.

[95] Yang ZD, Zhong WH, Zhao L, Chen CYCA. ML-DTI: Mutual Learning Mechanism for Interpretable Drug-Target Interaction Prediction. J Phys Chem Lett 2021;12:4247–61.

[96] Zhang SG, Jiang MJ, Wang S, Wang XF, Wei ZQ, Li ZSAG-DTA. Prediction of Drug-Target Affinity Using Self-Attention Graph Network. Int J Mol Sci 2021;22.

[97] Tanoori B, Jahromi MZ. Using drug-drug and protein-protein similarities as feature vector for drug-target binding prediction. Chemometr Intell Lab 2021;217.

[98] Mukherjee S, Ghosh M, Basuchowdhuri P. Deep Graph Convolutional Network and LSTM based approach for predicting drug-target binding affinity. 2022.

[99] Yang Z, Zhong W, Zhao L, Yu-Chian Chen C. MGraphDTA: deep multiscale graph neural network for explainable drug-target binding affinity prediction. Chem Sci 2022;13:816–33.

[100] Yuan WN, Chen GX, Chen CYC. FusionDTA: attention-based feature polymerizer and knowledge distillation for drug-target binding affinity prediction. Briefings Bioinf 2022;23.

[101] Wang J, Wen N, Wang C, Zhao L, Cheng L. ELECTRA-DTA: a new compound-protein binding affinity prediction model based on the contextualized sequence encoding. J Cheminform 2022;14:14.