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

A B S T R A C T

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.

© 2022 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

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,
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
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 and connectivity fingerprints with restart, and limits the learning capability. To more effectively represent compounds and proteins, graph-based features have also been 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 and connectivity fingerprints with restart, and limits the learning capability. To more effectively represent compounds and proteins, graph-based features have also been used by encoding sequences in numerical matrices via one-hot or word embedding (such as Prot2Vec and Mol2Vec) [38,44,45].
| Tool          | Date       | Input protein features           | Input compound features          | Protein feature extractor                  | Compound feature extractor                  | Methods                     |
|--------------|------------|----------------------------------|----------------------------------|-------------------------------------------|--------------------------------------------|----------------------------|
| SimBoost     | 04/2017    | Target similarity                | Drug similarity                  | –                                         | –                                         | Gradient boosting tree model |
| ACNN         | 2017       | Atomic coordinates               | Atomic coordinates               | Atomic convolution layer                  | Atomic convolution layer                  | Atomic fully connected layer |
| DeepDTA      | 09/2018    | Label encoding                   | Label encoding                   | CNN blocks                               | CNN blocks                               | Fully connected layers      |
| DeepAffinity | 02/2018    | Structural property sequence     | Structural property sequence     | Seq2seq autoencoders                      | Seq2seq autoencoders                      | Unified RNN-CNN             |
| WiDeDTA      | 02/2019    | Textual information              | Textual information              | CNN blocks                               | CNN blocks                               | Fully connected layers      |
| GraphDTA     | 06/2019    | One-hot encoding                 | Molecular graph                  | Convolutional layers                      | 4 graph neural network variants           | Fully connected layers      |
| RFscore      | 08/2019    | 36 intermolecular features       | 36 intermolecular features       | –                                         | –                                         | Random forest               |
| AttentionDTA | 11/2019    | Label encoding                   | Label encoding                   | CNN block                                | CNN block                                | Attention block- fully connected layers |
| Taba         | 01/2020    | The average distance between pairs of atoms | The average distance between pairs of atoms | –                                         | –                                         | Machine-learning model      |
| GAT_GCN      | 04/2020    | Peptide frequency                | Graph structure                  | CNN                                      | GCN                                      | Fully connected layers      |
| SAnDReS      | 05/2020    | Docking scores                   | Docking scores                   | –                                         | –                                         | Machine-learning model      |
| DeepCDA      | 05/2020    | N-gram embedding                 | SMILES sequence                  | CNN-LSTM-Two-sided attention mechanism    | CNN-LSTM-Two-sided attention mechanism    | Fully connected layers      |
| DGGraphDTA   | 06/2020    | Protein graph                    | Molecular graph                  | CNN                                       | GCN                                      | Fully connected layers      |
| JoVA         | 08/2020    | Multiple unimodal representations | Multiple unimodal representations | Joint view attention module              | Joint view attention module              | Prediction model            |
| Fusion       | 11/2020    | Atomic representation            | Atomic representation            | CNNs                                     | SG-GCNs                                  | Fully connected layers      |
| DeepGS       | 2020       | Symbolic sequences               | Molecular structure              | Prot2Vec-CNN-BiGRU blocks                | Smi2Vec-CNN-BiGRU blocks                  | Fully connected layers      |
| DeepDTAF     | 01/2021    | Sequence, structural property information | SMILES string                  | Dilated/traditional convolution layers    | Dilated convolution layers                | Fully connected layers      |
| GanDTI       | 03/2021    | Protein sequences                | Molecule fingerprints-adjacency matrix | Attention module                           | Residual graph neural network             | MLP                        |
| Multi-PLI    | 04/2021    | One-hot vectors                  | One-hot vectors                  | CNN blocks                               | CNN blocks                               | Fully connected layers      |
| ML-DTI       | 04/2021    | Protein sequences                | SMILES string                    | CNN block (mutual learning)              | CNN block (mutual learning)              | Linear transformation layers |
| DEELIG       | 06/2021    | Atomic level-structural information-sequences | Physical properties-fingerprints | CNN                                      | Fully connected layers                    | Fully connected layers      |
| GEFA         | 07/2021    | Sequence embedding features      | Graph representation             | GCN                                      | GCN                                      | Linear layers               |
| SAG-DTA      | 08/2021    | Label encoding                   | Molecular graph                  | CNN                                      | Graph convolutional layer-SAGPooling layer | Fully connected layers      |
| Tanoori et al. | 08/2021 | SW sequence similarity            | CS similarity                    | –                                         | –                                         | GBM                        |
| EmbedDTI     | 11/2021    | Amino acids                      | Structural information           | CNN                                      | Attention-GCNs                           | Fully connected layers      |
| DeepPLA      | 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    | 01/2022    | Amino acids                      | Adjacency representation         | BiLSTM                                   | GCN                                      | Fully connected layers      |
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 SMILES. 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. Xenatide-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, Lui TQ, Baitulak 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 Lang SA, Okun I, Ivanenko 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, Oliveri JR, Canduri T, Silva RC, 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, Pinto 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 2019;20.

[17] Wojcikowsky 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 Name H. Prediction of drug-target interactions via deep learning with convolution on protein sequences. Plos Comput Biol 2019;15.

[19] Tsubaki M, Tomiki 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. Critical Review. Curr Med Chem 2013;20:2839–60.

[21] Chen Z, Zhao P, Li FY, Marquez-Lago TT, Leier A, Revote J, et al. Learn: 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.

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

[23] Chen Z, Zhao P, Li FY, Marquez-Lago TT, Leier A, Revote J, et al. Learn: 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.
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.

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.

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

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.

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.

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.

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

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.