frontiers in Genetics

REVIEW
published: 21 June 2021
doi: 10.3389/fgene.2021.680117

Application of Machine Learning for Drug–Target Interaction Prediction

*Lei Xu[1], Xiaoqing Ru[2] and Rong Song[1]**

[1] School of Electronic and Communication Engineering, Shenzhen Polytechnic, Shenzhen, China, [2] Department of Computer Science, University of Tsukuba, Tsukuba, Japan

Exploring drug–target interactions by biomedical experiments requires a lot of human, financial, and material resources. To save time and cost to meet the needs of the present generation, machine learning methods have been introduced into the prediction of drug–target interactions. The large amount of available drug and target data in existing databases, the evolving and innovative computer technologies, and the inherent characteristics of various types of machine learning have made machine learning techniques the mainstream method for drug–target interaction prediction research. In this review, details of the specific applications of machine learning in drug–target interaction prediction are summarized, the characteristics of each algorithm are analyzed, and the issues that need to be further addressed and explored for future research are discussed. The aim of this review is to provide a sound basis for the construction of high-performance models.

Keywords: machine learning, drug–target interactions, data, features, task algorithms, drug development

OPEN ACCESS

Edited by:
Quan Zou,
University of Electronic Science and Technology of China, China

Reviewed by:
Ying Hong Li,
Chongqing University of Posts and Telecommunications, China
Changli Feng,
Taishan University, China

****Correspondence:***
Rong Song
sr1@szpt.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 March 2021*
Accepted: *28 May 2021*
Published: *21 June 2021*

Citation:
Xu L, Ru X and Song R (2021) Application of Machine Learning for Drug–Target Interaction Prediction. Front. Genet. 12:680117. doi: 10.3389/fgene.2021.680117

INTRODUCTION

Tens of thousands of known diseases threatening human health, and new ones are being added every year. They include emerging diseases (e.g., the currently prevalent COVID-19) and diseases that have plagued the public for many years and have no cure so far (e.g., Parkinson's disease and Alzheimer's disease) (Xu et al., 2018a, 2019). Rapidly and accurately discovering drugs that can effectively treat diseases is very important for the development of society. Long cycle and high cost are common phenomena in current drug development, but these fail to guarantee a high success rate. Many steps are required from drug development to final marketing, including drug discovery, preclinical and clinical trials, and marketing approval (Srivastava et al., 2019; Li Z. et al., 2020). The overall success rate of drug discovery and preclinical studies, which are part of the laboratory development phase, is approximately 0.05–0.1%, and less than 1% of the candidate compounds are likely to have the expected effect and proceed to the clinical trial phase. Investigating drug–target interactions is an important step in the drug discovery process and can improve the success rate of new drug discovery (Chen et al., 2019; Huang et al., 2020; Zeng et al., 2020b). These not only signal the need to expend significant resources to find and test candidate compounds one by one during the drug development phase to confirm that they meet expectations, but also demonstrate the importance of drug–target interaction prediction in the overall drug development process. Supplementally, an obvious drawback of biomedical experiment is that it does not allow for rapidly finding and solving problems, which can be detrimental to the treatment of emerging and highly infectious diseases. Therefore, machine learning methods have been introduced into the prediction of drug–target interactions.

Machine learning, a computer technology for data analysis designed to build predictive models using datasets, has become an important means of modern biological research (Xu et al., 2018b; Yang et al., 2018; Liu et al., 2019, 2020; Tang et al., 2020; Zeng et al., 2020a). It has become a mainstream technique for analyzing and solving problems involved in drug–target interaction prediction studies (Cai et al., 2018; Stephenson et al., 2019; Zeng et al., 2019; Fu et al., 2020; Wang J. et al., 2020).

THREE FACTORS

The existing data background, powerful toolkits, and current status and requirements have promoted machine learning to become the mainstream method of drug–target interaction prediction.

(1) Existing databases. With the emergence of sequencing technology, high-throughput technology and computer-aided drug design method, a large number of proteins have been sequenced and many compounds have been synthesized. On the basis of existing related works and accumulated experience, relevant data has been organized and various databases have been constructed. Most of the data in these databases are publicly available and free to download, which provides a good data foundation for solving drug–target interaction prediction problems by machine learning. Researchers can collect datasets from databases that cover different information according to their needs (Zheng et al., 2019, 2020). Some representative databases are briefly described here.

UniProt database[1]: UniProt is supported by many institutions, and is the most informative and comprehensive protein database (Consortium, 2015). It consists of five sub-databases: Swiss-Prot, TrEMBL, UniRef, UniParc, and Proteomes. Each sub-database has its own unique function. For example, Swiss-Prot is a high-quality, manually annotated, non-redundant database, in which protein annotations are derived mainly from the literature or E-value verification calculation analysis results. Proteomes is a database that provides proteomic information for species with fully sequenced genomes.

PubChem database[2]: PubChem is an open chemistry database that collects information including chemical structures, identifiers, physicochemical properties, and biological activities of chemical molecules (Kim et al., 2016, 2021). It is the world's largest database with free access to chemical information, and currently covers 109 million compounds. PubChem has become an important chemical information resource for scientists, students, and the public.

DrugBank database[3]: As a bioinformatics and cheminformatics resource, DrugBank combines detailed drug data (i.e., chemical, pharmacological, and pharmaceutical) with comprehensive target information (i.e., sequence, structure, and pathway) (Wishart et al., 2018). The latest DrugBank release (version 5.1.8.) contains 14,443 drug molecules and 5,244 non-redundant protein sequences associated with these drugs. The database describes not only clinical information on drugs, namely drug side effects and drug–drug interactions, but also contains molecular-level data, such as chemical structures of drugs and proteins targeted by drugs (Wishart et al., 2008). One significant function of DrugBank is that it supports comprehensive and complex searches, so it is used widely by the pharmaceutical industry, medicinal chemists, pharmacists, physicians, students, and the general public.

KEGG database[4]: KEGG was established in 1995 by the Kanehisa Laboratories at the Bioinformatics Center, Kyoto University, Japan, and is now one of the most commonly used international bioinformatics databases (Kanehisa and Goto, 2000). KEGG is a database used to understand the high-level functions and practicability of biological systems from molecular-level information (Li H. et al., 2020; Wang et al., 2021a) (especially large-scale molecular datasets generated by genome sequencing and other high-throughput techniques), of which the data information can be roughly classified into four major categories: system information, genetic information, chemical information, and medical information.

BindingDB database[5]: BindingDB is a publicly available, web-accessible database for measuring binding affinity, focusing on the interactions between proteins considered to be drug targets and drug-like small molecules (Liu et al., 2007). BindingDB currently contains 2,114,159 binding data between 8,202 protein targets and 928,022 small molecules.

(2) Powerful toolkits and web servers. Bioinformatics and cheminformatics are emerging interdisciplinary fields that use computers to solve biological and chemical problems. Many toolkits and web servers have been developed (Zuo et al., 2017; Zou et al., 2019; Lin et al., 2020; Pang and Liu, 2020; Shao et al., 2021), which can help to solve problems in drug–target interaction prediction.

STITCH[6]: STITCH not only includes experimentally validated drug–target interaction data, but also integrates predicted drug–target relationships (Kuhn et al., 2007). This website can clearly depict the protein–protein interactions, protein–compound interactions, and the strength of the interactions.

SwissTargetPrediction[7]: SwissTargetPrediction can estimate the most likely macromolecule to be targeted by a biologically active small molecule and count the percentage of each target type targeted by the small molecule (Gfeller et al., 2014).

RDkit[8]: RDkit is a powerful python toolkit for chemical information, which has functions such as acquiring molecule information from multiple formats, obtaining information about atoms, bonds, and rings in molecules, generating molecular descriptors and molecular fingerprints of compounds, and calculating similarities of compound structures (Landrum, 2013).

OpenChem[9]: OpenChem is a pytorch-based deep learning toolkit for computational chemistry and drug design,

[1]https://www.uniprot.org/

[2]https://pubchem.ncbi.nlm.nih.gov/

[3]https://go.drugbank.com/

[4]https://www.genome.jp/kegg/

[5]https://www.bindingdb.org/bind/index.jsp

[6]http://stitch.embl.de/

[7]http://www.swisstargetprediction.ch/

[8]https://www.rdkit.org/

[9]https://mariewelt.github.io/OpenChem/html/index.html

which contains Feature2Label, Smiles2Label, Graph2Label, SiameseModel, GenerativeRNN, and MolecularRNN (Korshunova et al., 2021). Users can train predictive models for classification, regression, and multi-task problems, and develop generative models for generating novel molecules with optimized properties. Its goal is to make deep learning an easy-to-use tool for researchers in computational chemistry and drug design.

iFeature[10]: iFeature is a python toolkit that can compute various structural and physicochemical property descriptors from protein and peptide sequences. iFeature can compute and extract comprehensive spectra for 18 major sequence coding schemes, including 53 different types of feature descriptors. In addition, iFeature integrates 12 different types of commonly used feature clustering, selection, and dimensionality reduction algorithms (Chen et al., 2018).

Pse-in-one[11]: Pse-in-one is a python toolkit that generates all possible pseudo-components of DNA, RNA, and protein sequences. It covers a total of 28 different patterns, 14 for DNA sequences, 6 for RNA sequences, and 8 for protein sequences (Liu et al., 2015, 2017). This toolkit is widely and increasingly used by researchers to tackle various problems in computational biology, and a more specific and detailed version BioSeq-Analysis (Liu, 2019) has recently been released.

(3) Current status and requirements. With the development of high-throughput technologies, many compounds and proteins have been mined. The human genome contains more than 20,000 genes, and approximately 80% of them can encode one or more proteins. Only a small number of proteins have been identified as pharmacologically active and are targets for currently approved drugs. The pharmacological functions of most proteins remain to be demonstrated. This is also true for most compounds. For example, there are currently 111 million compounds in the PubChem database, but proteins that could interact with many of these compounds are unknown. In addition, it is obvious that the traditional approach of wet experiments is not feasible for some emerging, highly infectious and destructive new pathogens, such as the SARS, H7N9, Ebola, Mers, and COVID-19 viruses (Cheng et al., 2021). Considering the huge amounts of available data and large numbers of diseases that cause serious social health risks, using computational chemistry-related theories and computer simulation methods to computationally predict drug–target interaction can effectively improve efficiency. Machine learning-based methods have become effective ways to compensate for the shortcomings of traditional biochemical experimental methods.

APPLICATIONS

The current drug–target interaction prediction procedures are shown in **Figure 1**. Existing studies on drug–target interaction prediction have shown that using different calculation or optimization methods in the steps of data set acquisition, feature extraction and processing, and task algorithm selection can build models with good performance.

(1) Dataset acquisition. Redundant data, unbalanced categories, and unrepresentative samples can lead to long experimental cycles, as well as inaccurate and biased experimental results. Different data acquisition methods have been used to avoid or reduce the impact of these problems on model construction. For example, Wang et al. (2010) collected negative examples by random selection to solve the data imbalance problem. Wang et al. (2018) also used random selection to extract negative examples, and this operation was performed five times to reduce the impact of the unverified negative samples. Pdti-EssB (Mahmud et al., 2020) used random under-sampling and under-sampling clustering to address the data imbalance problem.

Currently, most target molecules are proteins, of which four protein families [kinases, G protein-coupled receptors (GPCRs), ion channels, and nuclear receptors] account for 44% of the target molecules, and 70% of the currently developed drugs are targeted to these four protein families. Datasets established by Yamanishi et al. (2008), which contain the interactions between these four proteins and drugs,have been widely used (Öztürk et al., 2018; Mahmud et al., 2020). The relevant data can be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/. Most of the computational approaches based on these datasets have focused on binary classification, that is, they only explore whether a drug can interact with a particular protein. To further accelerate process and reduce cost, drug–target affinity has been explored in some studies. Drug–target affinity is a key property that determines the strength of the interaction between the small molecule drug and the target. The commonly used datasets for predicting drug–target affinity are the Kinase (Davis et al., 2011) and KIBA (Tang et al., 2014) datasets.

(2) Feature extraction and processing. Accurate and comprehensive descriptions of the biological or chemical functional information of drugs and targets in numerical form play an important role in the construction of high-performance models. Feature extraction of drugs and targets can be performed from different perspectives (Cheng, 2019; Zhao T. et al., 2020). For example, iGPCR-Drug (Xiao et al., 2013) obtains drug features by discrete Fourier transform of drug molecular fingerprints and extracts GPCR features according to pseudo amino acid compositions. DrugE-Rank (Yuan et al., 2016) represents drug features according to general descriptors and extracts target features according to amino acid composition, transformation, and distribution. TargetGDrug (Hu J. et al., 2016) extracts drug features by applying wavelet transform to drug molecular fingerprints and extracts GPCR features according to evolutionary information. Ru et al. (2020) extracted protein features using the distance-based top-n-gram algorithm and obtained drug features according to general descriptors. Chemical databases store information in a textual representation and the simplified molecular input line entry specification (SMILES) format is a common standard used in many cheminformatics software. Each SMILES string encodes structural information that can be used to predict complex chemical properties, and a large number of machine learning

[10]https://ifeature.erc.monash.edu/

[11]http://bioinformatics.hitsz.edu.cn/Pse-in-One/

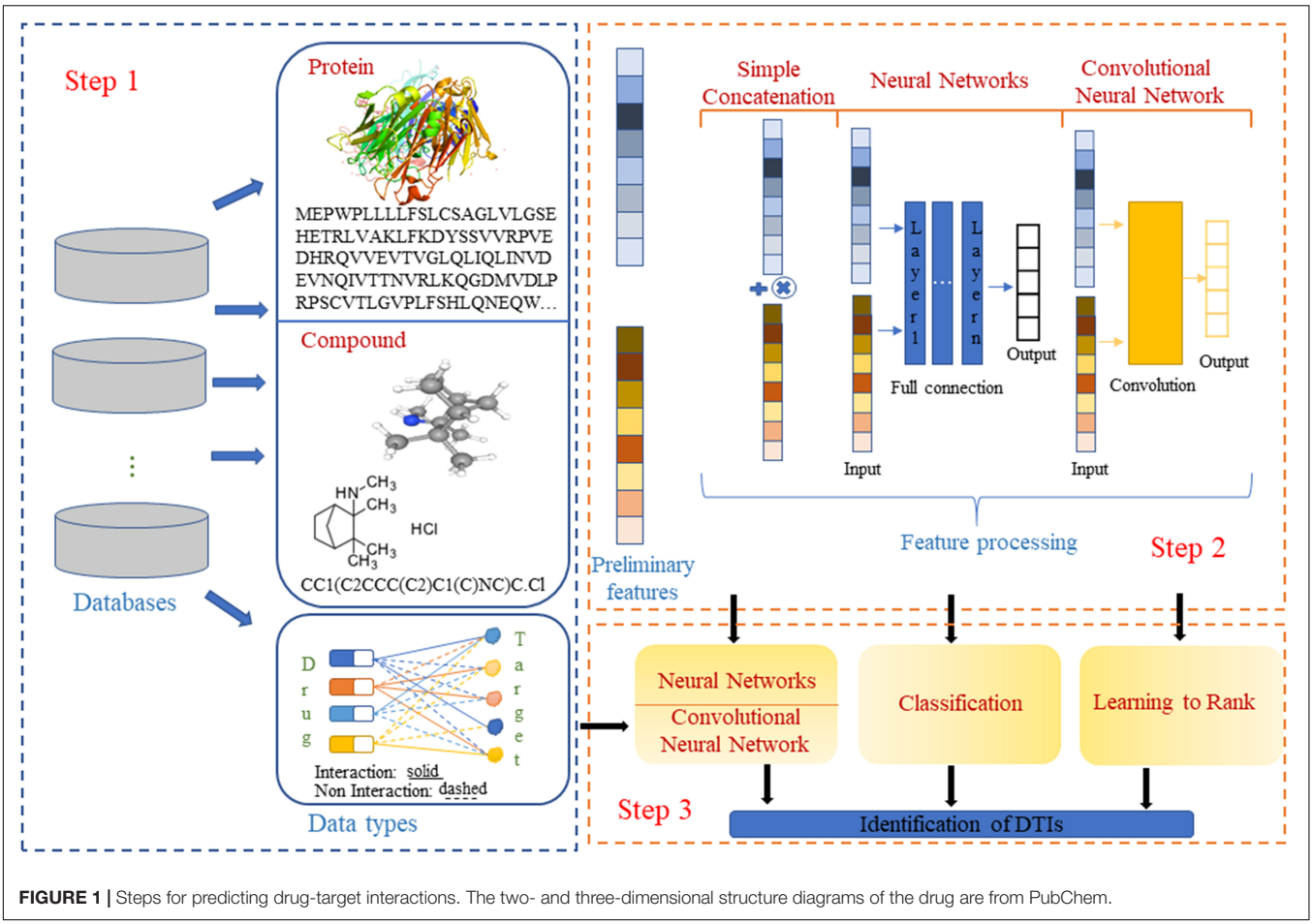

FIGURE 1 | Steps for predicting drug-target interactions. The two- and three-dimensional structure diagrams of the drug are from PubChem.

models can extract molecular features of compounds according to SMILES strings. Recently, convolutional neural networks (CNNs) and recurrent neural networks have been used for molecular feature extraction. Hirohara et al. (2018) transformed SMILES strings into two-dimensional matrices and used CNNs to extract molecular features. Goh et al. (2017) applied natural language processing to SMILES feature extraction and used recurrent neural networks for molecular strings.

The presence of invalid or redundant features not only reduces the accuracy of the experiment result but also lengthens the experimental period. Low-dimensional and comprehensive information feature sets are expected. Therefore, a variety of methods for processing features have been applied to related rearch (Zou et al., 2016a,b; Guo et al., 2020; Zhang G. et al., 2020; Zhao X. et al., 2020). For example, to reduce the noise between features, Li et al. (2017) used principal component analysis (PCA) to reduce the dimensionality of drugs and targets features. Tabei et al. (2012) combined 881 substructures of drugs and 876 Pfam domain structures of targets by tensor product to form feature vectors of drug–target pairs. MFDR (Hu P.-W. et al., 2016) used autoencoders as the building blocks of a deep network to reconstruct drug and protein features into a low-dimensional new representation. DeepConv-DT (Lee et al., 2019) used CNNs on raw protein sequences to capture local amino acid residue information by convolving amino acid subsequences of various lengths.

(3) Selection of task algorithms. Several task algorithms have been used for drug–target interaction prediction, such as classification algorithms, learning to rank algorithms, and deep learning algorithms (Cheng et al., 2019; Lv et al., 2019; Tao et al., 2020; Zhang Y. et al., 2020).

Most of the existing studies treat drug–target interaction prediction as binary tasks, and different classification algorithms have been applied. For example, Bleakley and Yamanishi (2009) proposed a bipartite local model (BLM) based on a support vector machine (SVM) kernel to predict drug–target relationships. LRF-DTI (Shi et al., 2019) is a drug–target interaction prediction method using Lasso for feature extraction and random forest for classification. Yamanishi et al. (2010) used a distance learning algorithm as a classifier. Pred-binding (Shar et al., 2016) extracted molecular structure and protein sequence features, and used support vector machines and random forests to classify whether drugs and targets can be docked.

Drug–target interaction prediction can be regarded as a ranking task. Exploring the strength of drug–target interactions can shorten the drug development process and save expenses. Zhang et al. (2015) applied six learning to rank algorithms (Prank, RankNet, RankBoost, SVMRank, AdaRank, and ListNet)

to virtual screening of drugs,their study showed that learning to rank is an effective computational strategy, especially because of its novel use in cross-target virtual screening and heterogeneous data integration. DrugE-Rank (Yuan et al., 2016) used protein amino acid composition, transformation and distribution information, compound descriptor information, and output information of six classifiers as features to be input into the learning to ranking algorithm to improve the performance of drug-target interaction prediction.

Neural networks have also been used to solve related problems in the prediction of drug–target interactions. Prado-Prado et al. (2011) used the entropy information of drug–protein complexes and neural networks to predict drug–target affinity values. DeepDTA (Öztürk et al., 2018) proposed a deep-learning based model that used only sequence information of both targets and drugs, One novel approach used in this work is the modeling of protein sequences and compound 1D representations with CNNs. GraphDTA (Nguyen et al., 2019) focused on the fact that molecules are by nature formed by chemical bonding of atoms, and used graph convolutional network to learn drug-target binding affinity.

DISCUSSION

Under the background of the existing chemical and biological computing theory, big data and rapid development of computer technology, the use of machine learning for drug-target interaction prediction does have many benefits, but there are still some problems that need to be further explored.

(1) Data heterogeneity. Most of the existing studies are based on publicly available data in databases that collect data with different focuses, and each database has its own criteria for judging the data. Drugs, targets, and related data from different databases often have different terminological descriptions and different organization structures, such inconsistencies make data integration difficult.

(2) Effective representation of biological and chemical features. Feature engineering is a key concern in building machine learning models. There are often technical difficulties in how to effectively extract key features and how to deal with data with high dimensionality. Existing studies have shown that the features of proteins and drugs can be extracted from a variety of angles, and the combination of information from these angles can achieve complementary effects. Most drug–target interaction prediction studies only extract relatively one-sided information, and do not comprehensively consider the information from multiple perspectives. In addition, most studies have focused on extracting drug molecule and protein features separately, ignoring the potentially valid association that may exist between drug and target. Moreover, the direct concatenation of biologically unrelated features may lead to a decrease in prediction accuracy.

(3) Characteristics of task algorithms. The classification, ranking, or deep learning methods used in drug–target interaction prediction all have their own characteristics. Different computational approaches can be used to solve different problems in drug–target interaction prediction, however, these algorithms also have shortcomings. Classification is the simplest and most understandable task. However, there is an obvious and long-standing defect in this task that it is necessary to collect negative samples. Most existing classification studies take experimentally validated drug–target pairs with known interactions as positive samples, and unvalidated or unknown drug–target pairs as negative examples. Among these negative examples, there may be positive samples that have not been accurately validated, the performance of a model that is based on such a dataset will be biased.

On the basis of the existence of one-to-many or many-to-many relationships between queries and documents, learning to rank can be used in multi-target drug discovery. Early drug development followed the "one drug, one target" principle, with the aim of finding high-affinity, high-selective drugs for a specific receptor associated with a particular disease. However, the number of complex diseases is increasing and the proteins associated with these diseases are not limited to one, therefore drug combinations are used to achieve the optimal therapeutic effect. Clinical pharmacology studies have shown that drug combinations greatly increase the incidence of adverse drug reactions, but because of the lack of multi-target drugs, such risks have to be taken. Multi-target drugs are undoubtedly an important area for future research. Therefore, using the characteristics of learning to rank to tackle the multi-target problem of drugs deserves to be explored further. Learning to rank was originally applied for information retrieval. Its output is a relative score of correlation between queries and documents (Cheng, 2020; Ru et al., 2021). This is not sufficient for studies that require accurate prediction of drug–target affinities.

The use of neural networks for predicting accurate drug–target affinity values has shown great potential in this research area. Neural networks can fuse drug and target features, which have changed the current situation of simple concatenation or tensor products of drug and target features. Deep learning contains more neural network structures with multiple implicit layers compared with traditional machine learning, which allows deep learning to handle large datasets and identify complex patterns from the learning process. But for the same reason, neural networks require much more execution time than classification or ranking algorithms. It will lead to overfitting when the drug and target feature dimensions are high.

Although existing machine learning methods have opened a new area in drug–target interaction prediction, they have not achieved satisfactory results so far. Therefore, there is still a need to develop new theoretical and computational methods for drug–target interaction prediction.

CONCLUSION

Drug–target interaction prediction can help to screen out unsuitable compounds and is an important step in the development of new drugs. In this review, we describe the importance of drug–target interaction prediction, analyze in

detail the three main reasons why machine learning has become a mainstream technique, summarize the specific applications of machine learning methods in each step of building machine learning models, analyze the shortcomings of existing research methods, and discuss several aspects that can be further explored (Wei et al., 2014, 2017a,b, 2018, 2019; Ding et al., 2017, 2019, 2020a,b; Jin Q. et al., 2019; Jin S. et al., 2019; Li J. et al., 2020; Su et al., 2020; Wang H. et al., 2020; Zeng et al., 2020c,d; Zhai et al., 2020; Wang et al., 2021b). This review provides meaningful perspectives for future drug–target interaction prediction studies, especially the application of learning to rank to deal with multi-target drug problems.

AUTHOR CONTRIBUTIONS

XR drafted the manuscript. LX and RS initiated the idea, conceived the whole process, and finalized the manuscript. All authors have read and approved the final manuscript.

FUNDING

This work was supported by the natural science foundation of Guangdong province (grant No. 2018A0303130084) and the Grant of Shenzhen Polytechnic (No. 6021310015K).

REFERENCES

Bleakley, K., and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics* 25, 2397–2403. doi: 10.1093/bioinformatics/btp433

Cai, J., Cai, H., Chen, J., and Yang, X. (2018). Identifying "many-to-many" relationships between gene-expression data and drug-response data via sparse binary matching. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 165–176.

Chen, J., Peng, H., Han, G., Cai, H., and Cai, J. (2019). HOGMMNC: a higher order graph matching with multiple network constraints model for gene–drug regulatory modules identification. *Bioinformatics* 35, 602–610. doi: 10.1093/bioinformatics/bty662

Chen, Z., Zhao, P., Li, F., Leier, A., Marquez-Lago, T. T., Wang, Y., et al. (2018). iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics* 34, 2499–2502. doi: 10.1093/bioinformatics/bty140

Cheng, L. (2019). Computational and biological methods for gene therapy. *Curr. Gene Ther.* 19, 210–210. doi: 10.2174/1566523219041910221133 07

Cheng, L. (2020). Omics Data and Artificial Intelligence: New Challenges for Gene Therapy. *Curr. Gene Ther.* 20:1. doi: 10.2174/1566523220012006041500 41

Cheng, L., Han, X., Zhu, Z., Qi, C., Wang, P., and Zhang, X. (2021). Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief. Bioinform.* 22, 1442–1450. doi: 10.1093/bib/bbab042

Cheng, L., Zhao, H., Wang, P., Zhou, W., Luo, M., Li, T., et al. (2019). Computational Methods for Identifying Similar Diseases. *Mol. Ther. Nucleic Acids* 18, 590–604. doi: 10.1016/j.omtn.2019.09.019

Consortium, U. (2015). UniProt: a hub for protein information. *Nucleic Acids Res.* 43, D204–D212.

Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., et al. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* 29, 1046–1051.

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Ding, Y., Tang, J., and Guo, F. (2020a). Identification of Drug-Target Interactions via Dual Laplacian Regularized Least Squares with Multiple Kernel Fusion. *Knowl. Based Syst.* 204:106254. doi: 10.1016/j.knosys.2020.106254

Ding, Y., Tang, J., and Guo, F. (2020b). Identification of drug-target interactions via fuzzy bipartite local model. *Neural Comput. Appli.* 23, 10303–10319. doi: 10.1007/s00521-019-04569-z

Fu, X., Cai, L., Zeng, X., and Zou, Q. J. B. (2020). StackCPPred: a stacking and pairwise energy content-based prediction of cell-penetrating peptides and their uptake efficiency. *Bioinformatics* 36, 3028–3034. doi: 10.1093/bioinformatics/btaa131

Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., and Zoete, V. (2014). SwissTargetPrediction: a web server for target prediction of bioactive small molecules. *Nucleic Acids Res.* 42, W32–W38.

Goh, G. B., Hodas, N. O., Siegel, C., and Vishnu, A. (2017). Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv* 171202034.

Guo, Z., Wang, P., Liu, Z., and Zhao, Y. (2020). Discrimination of Thermophilic Proteins and Non-thermophilic Proteins Using Feature Dimension Reduction. *Front. Bioeng. Biotechnol.* 8:584807. doi: 10.3389/fbioe.2020.584807

Hirohara, M., Saito, Y., Koda, Y., Sato, K., and Sakakibara, Y. (2018). Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC bioinformatics* 19:526. doi: 10.1186/s12859-018-2523-5

Hu, J., Li, Y., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. (2016). GPCR–drug interactions prediction using random forest with drug-association-matrix-based post-processing procedure. *Comput. Biol. Chem.* 60, 59–71. doi: 10.1016/j.compbiolchem.2015.11.007

Hu, P.-W., Chan, K. C., and You, Z.-H. (2016). "Large-scale prediction of drug-target interactions from deep representations," in *2016 International Joint Conference on Neural Networks (IJCNN* (Vancouver: IEEE), 1236–1243.

Huang, J., Chen, J., Zhang, B., Zhu, L., and Cai, H. (2020). Evaluation of gene–drug common module identification methods using pharmacogenomics data. *Brief. Bioinform.* 22:bbaa087.

Jin, Q., Meng, Z., Tuan, D. P., Chen, Q., Wei, L., and Su, R. (2019). DUNet: A deformable network for retinal vessel segmentation. *Knowl. Based Syst.* 178, 149–162. doi: 10.1016/j.knosys.2019.04.025

Jin, S., Zeng, X., Fang, J., Lin, J., Chan, S. Y., and Erzurum, S. C. (2019). Cheng FJNsb, applications: A network-based approach to uncover microRNA-mediated disease comorbidities and potential pathobiological implications. *NPJ Syst. Biol. Appl.* 5, 1–11.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* 49, D1388–D1395.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213.

Korshunova, M., Ginsburg, B., Tropsha, A., and Isayev, O. (2021). OpenChem: A Deep Learning Toolkit for Computational Chemistry and Drug Design. *J. Chem. Inform. Model.* 61, 7–13. doi: 10.1021/acs.jcim.0c00971

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2007). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 36, D684–D688.

Landrum, G. (2013). Rdkit documentation. *Release* 1:4.

Lee, I., Keum, J., and Nam, H. (2019). DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput. Biol.* 15:e1007129. doi: 0.1371/journal.pcbi.1007129

Li, H., Long, C., Xiang, J., Liang, P., Li, X., and Zuo, Y. (2020). Dppa2/4 as a trigger of signaling pathways to promote zygote genome activation by binding to CG-rich region. *Brief. Bioinform.* doi: 0.1093/bib/bbaa342 [Epub Online ahead of print].

Li, J., Pu, Y., Tang, J., Zou, Q., and Guo, F. (2020). DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences. *Brief. Bioinform.* 22:bbaa159.

Li, Z., Han, P., You, Z.-H., Li, X., Zhang, Y., Yu, H., et al. (2017). In silico prediction of drug-target interaction networks based on drug chemical structure and protein sequences. *Sci. Rep.* 7, 1–13.

Li, Z., Zhang, T., Lei, H., Wei, L., Liu, Y., Shi, Y., et al. (2020). Research on Gastric Cancer's Drug-resistant Gene Regulatory Network Model. *Curr. Bioinform.* 15, 225–234. doi: 10.2174/1574893614666190722102557

Lin, X., Quan, Z., Wang, Z., Huang, H., and Zeng, X. (2020). A novel molecular representation with BiGRU neural networks for learning atom. *Brief. Bioinform.* 21, 2099–2111. doi: 10.1093/bib/bbz125

Liu, B. (2019). BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.* 20, 1280–1294. doi: 10.1093/bib/bbx165

Liu, B., Gao, X., and Zhang, H. (2019). BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* 47:e127. doi: 10.1093/nar/gkz740

Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.* 43, W65–W71.

Liu, B., Wu, H., and Chou, K.-C. (2017). Pse-in-One 2.0: an improved package of web servers for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nat. Sci.* 9:67. doi: 10.4236/ns.2017.94007

Liu, T., Lin, Y., Wen, X., Jorissen, R. N., and Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201.

Liu, X., Hong, Z., Liu, J., Lin, Y., Rodríguez-Patón, A., and Zou, Q. (2020). Zeng XJBib: Computational methods for identifying the critical nodes in biological networks. *Brief. Bioinform.* 21, 486–497. doi: 10.1093/bib/bbz011

Lv, Z. B., Ao, C. Y., and Zou, Q. (2019). Protein Function Prediction: From Traditional Classifier to Deep Learning. *Proteomics* 19:2.

Mahmud, S. H., Chen, W., Meng, H., Jahan, H., Liu, Y., and Hasan, S. M. (2020). Prediction of drug-target interaction based on protein features using undersampling and feature selection techniques with boosting. *Anal. Biochem.* 589:13507.

Nguyen, T., Le, H., and Venkatesh, S. (2019). GraphDTA: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv* doi: 10.1101/684662

Öztürk, H., Özgür, A., and Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics* 34, i821–i829.

Pang, Y., and Liu, B. (2020). SelfAT-Fold: Protein Fold Recognition Based on Residue-Based and Motif-Based Self-Attention Networks. *IEEE/ACM Trans. Comput. Biol. Bioinform.* doi: 10.1109/TCBB.2020.3031888 [Epub Online ahead of print].

Prado-Prado, F., García-Mera, X., Abeijón, P., Alonso, N., Caamaño, O., Yáñez, M., et al. (2011). Using entropy of drug and protein graphs to predict FDA drug-target network: theoretic-experimental study of MAO inhibitors and hemoglobin peptides from Fasciola hepatica. *Eur. J. Med. Chem.* 46, 1074–1094. doi: 10.1016/j.ejmech.2011.01.023

Ru, X., Wang, L., Li, L., Ding, H., Ye, X., and Zou, Q. (2020). Exploration of the correlation between GPCRs and drugs based on a learning to rank algorithm. *Comput. Biol. Med.* 119:103660. doi: 10.1016/j.compbiomed.2020.103660

Ru, X., Ye, X., Sakurai, T., and Zou, Q. (2021). Application of learning to rank in bioinformatics tasks. *Brief. Bioinform.* doi: 10.1093/bib/bbaa1394 [Epub Online ahead of print].

Shao, J., Yan, K., and Liu, B. (2021). FoldRec-C2C: protein fold recognition by combining cluster-to-cluster model and protein similarity network. *Brief. Bioinform.* 22:bbaa144. doi: 10.1093/bib/bbaa144

Shar, P. A., Tao, W., Gao, S., Huang, C., Li, B., Zhang, W., et al. (2016). Pred-binding: large-scale protein–ligand binding affinity prediction. *J. Enzyme Inhib. Med. Chem.* 31, 1443–1450. doi: 10.3109/14756366.2016.1144594

Shi, H., Liu, S., Chen, J., Li, X., Ma, Q., and Yu, B. (2019). Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics* 111, 1839–1852. doi: 10.1016/j.ygeno.2018.12.007

Srivastava, N., Mishra, B. N., and Srivastava, P. (2019). In-Silico Identification of Drug Lead Molecule Against Pesticide Exposed-neurodevelopmental Disorders Through Network-based Computational Model Approach. *Curr. Bioinform.* 14, 460–467. doi: 10.2174/1574893613666181112130346

Stephenson, N., Shane, E., Chase, J., Rowland, J., Ries, D., Justice, N., et al. (2019). Survey of Machine Learning Techniques in Drug Discovery. *Curr. Drug Metab.* 20, 185–193. doi: 10.2174/1389200219666180820112457

Su, R., Hu, J., Zou, Q., Manavalan, B., and Wei, L. (2020). Empirical comparison and analysis of web-based cell-penetrating peptide prediction tools. *Brief. Bioinform.* 21, 408–420. doi: 10.1093/bib/bby124

Tabei, Y., Pauwels, E., Stoven, V., Takemoto, K., and Yamanishi, Y. (2012). Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. *Bioinformatics* 28, i487–i494.

Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., et al. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *J. Chem. Inform. Model.* 54, 735–743. doi: 10.1021/ci400709d

Tang, Y.-J., Pang, Y.-H., and Liu, B. (2020). IDP-Seq2Seq: Identification of Intrinsically Disordered Regions based on Sequence to Sequence Learning. *Bioinformaitcs* 36, 5177–5186. doi: 10.1093/bioinformatics/btaa667

Tao, Z., Li, Y., Teng, Z., and Zhao, Y. (2020). A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med* 2020:8926750.

Wang, H., Ding, Y., Tang, J., and Guo, F. (2020). Identification of membrane protein types via multivariate information fusion with Hilbert-Schmidt Independence Criterion. *Neurocomputing* 383, 257–269. doi: 10.1016/j.neucom.2019.11.103

Wang, H., Liang, P., Zheng, L., Long, C., Li, H., and Zuo, Y. (2021a). eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. *Bioinformatics*

Wang, H., Tang, J., Ding, Y., and Guo, F. (2021b). Exploring associations of non-coding RNAs in human diseases via three-matrix factorization with hypergraph-regular terms on center kernel alignment. *Brief. Bioinform.* doi: 10.1093/bib/bbaa409 [Epub Online ahead of print].

Wang, J., Wang, H., Wang, X., and Chang, H. (2020). Predicting drug-target interactions via FM-DNN learning. *Curr. Bioinform.* 15, 68–76. doi: 10.2174/1574893614666190227160538

Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., et al. (2018). A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *J. Comput. Biol.* 25, 361–373. doi: 10.1089/cmb.2017.0135

Wang, Y.-C., Yang, Z.-X., Wang, Y., and Deng, N.-Y. (2010). Computationally probing drug-protein interactions via support vector machine. *Lett. Drug Des. Discov.* 7, 370–378. doi: 10.2174/157018010791163433

Wei, L., Ding, Y., Su, R., Tang, J., and Zou, Q. (2018). Prediction of human protein subcellular localization using deep learning. *J. Parallel Distrib. Comput.* 117, 212–217.

Wei, L., Liao, M., Gao, Y., Ji, R., He, Z., and Zou, Q. (2014). Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 11, 192–201. doi: 10.1109/tcbb.2013.146

Wei, L., Wan, S., Guo, J., and Wong, K. K. L. (2017a). A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* 83, 82–90. doi: 10.1016/j.artmed.2017.02.005

Wei, L., Xing, P., Shi, G., Ji, Z., and Zou, Q. (2019). Fast Prediction of Protein Methylation Sites Using a Sequence-Based Feature Selection Technique. *IEEE-ACM Trans. Comput. Biol. Bioinform.* 16, 1264–1273. doi: 10.1109/tcbb.2017.2670558

Wei, L., Xing, P., Zeng, J., Chen, J., Su, R., and Guo, F. (2017b). Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* 83, 67–74. doi: 10.1016/j.artmed.2017.03.001

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082.

Wishart, D. S., Knox, C., Guo, A. C., Cheng, D., Shrivastava, S., Tzur, D., et al. (2008). DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36, D901–D906.

Xiao, X., Min, J.-L., Wang, P., and Chou, K.-C. (2013). iGPCR-Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS One* 8:e72234. doi: 10.1371/journal.pone.0072234

Xu, L., Liang, G., Shi, S., and Liao, C. (2018a). SeqSVM: A Sequence-Based Support Vector Machine Method for Identifying Antioxidant Proteins. *Int. J. Mol. Sci.* 19:1773. doi: 10.3390/ijms19061773

Xu, L., Liang, G., Liao, C., Chen, G. D., and Chang, C. C. (2018b). An Efficient Classifier for Alzheimer's Disease Genes Identification. *Molecules* 23:13.

Xu, L., Liang, G., Liao, C., Chen, G. D., and Chang, C. C. (2019). k-Skip-n-Gram-RF: A Random Forest Based Method for Alzheimer's Disease Protein Identification. *Front. Genet.* 10:7. doi: 10.3389/fgene.2019.00033

Yamanishi, Y., Araki, M., Gutteridge, A., Honda, W., and Kanehisa, M. (2008). Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24, i232–i240.

Yamanishi, Y., Kotera, M., Kanehisa, M., and Goto, S. (2010). Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26, i246–i254.

Yang, X., Han, G., Chen, J., and Cai, H. (2018). Finding correlated patterns via high-order matching for multiple sourced biological data. *IEEE Trans. Biomed. Eng.* 66, 1017–1025. doi: 10.1109/tbme.2018.2866266

Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., and Zhu, S. (2016). DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* 32, i18–i27.

Zeng, X., Lin, Y., He, Y., Lv, L., and Min, X. (2020a). Deep collaborative filtering for prediction of disease genes. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 17, 1639–1647.

Zeng, X., Song, X., Ma, T., Pan, X., Zhou, Y., Hou, Y., et al. (2020b). Cheng FJJopr: Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J. Proteome Res.* 19, 4624–4636. doi: 10.1021/acs.jproteome.0c00316

Zeng, X., Zhu, S., Hou, Y., Zhang, P., Li, L., Li, J., et al. (2020c). Network-based prediction of drug–target interactions using an arbitrary-order proximity embedded deep forest. *Bioinformatics* 36, 2805–2812. doi: 10.1093/bioinformatics/btaa010

Zeng, X., Zhu, S., Liu, X., Zhou, Y., Nussinov, R., and Cheng, F. (2019). deepDR: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics* 35, 5191–5198. doi: 10.1093/bioinformatics/btz418

Zeng, X., Zhu, S., Lu, W., Liu, Z., Huang, J., Zhou, Y., et al. (2020d). Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* 11, 1775–1797. doi: 10.1039/c9sc04336e

Zhai, Y., Chen, Y., Teng, Z., and Zhao, Y. (2020). Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cell Dev. Biol.* 8:591487. doi: 10.3389/fcell.2020.591487

Zhang, G., Yu, P., Wang, J., and Yan, C. (2020). Feature Selection Algorithm for High-dimensional Biomedical Data Using Information Gain and Improved Chemical Reaction Optimization. *Curr. Bioinform.* 15, 912–926. doi: 10.2174/1574893615666200204154358

Zhang, W., Ji, L., Chen, Y., Tang, K., Wang, H., Zhu, R., et al. (2015). When drug discovery meets web search: learning to rank for ligand-based virtual screening. *J. Cheminform.* 7, 1–13.

Zhang, Y., Yan, J., Chen, S., Gong, M., Gao, D., Zhu, M., et al. (2020). Review of the Applications of Deep Learning in Bioinformatics. *Curr. Bioinform.* 15, 898–911. doi: 10.2174/1574893615999200711165743

Zhao, T., Hu, Y., Peng, J., and Cheng, L. (2020). DeepLGP: a novel deep learning method for prioritizing lncRNA target genes. *Bioinformatics* 36, 4466–4472. doi: 10.1093/bioinformatics/btaa428

Zhao, X., Jiao, Q., Li, H., Wu, Y., Wang, H., Huang, S., et al. (2020). ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles. *BMC Bioinformatics* 21:43. doi: 10.1186/s12859-020-3388-y

Zheng, L., Huang, S., Mu, N., Zhang, H., Zhang, J., Chang, Y., et al. (2019). RAACBook: a web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* 2019:baz131.

Zheng, L., Liu, D., Yang, W., Yang, L., and Zuo, Y. (2020). RaacLogo: a new sequence logo generator by using reduced amino acid clusters. *Brief. Bioinform.* 22:bbaa096.

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016a). Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.* 10:114. doi: 10.1186/s12918-016-0353-5

Zou, Q., Xing, P., Wei, L., and Liu, B. (2019). Gene2vec: Gene Subsequence Embedding for Prediction of Mammalian N6-Methyladenosine Sites from mRNA. *RNA* 25, 205–218. doi: 10.1261/rna.069112.118

Zou, Q., Zeng, J., Cao, L., and Ji, R. (2016b). A novel features ranking metric with application to scalable visual and bioinformatics data classification. *Neurocomputing* 173, 346–354. doi: 10.1016/j.neucom.2014.12.123

Zuo, Y., Li, Y., Chen, Y., Li, G., Yan, Z., and Yang, L. (2017). PseKRAAC: a flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* 33, 122–124. doi: 10.1093/bioinformatics/btw564

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor declared a past co-authorship with one of the authors, LX.