Technology and Health Care 31 (2023) S487–S495
DOI 10.3233/THC-236042
IOS Press

Neural networks prediction of the protein-ligand binding affinity with circular fingerprints

Zuode Yin[a,1], Wei Song[b,1], Baiyi Li[a], Fengfei Wang[c], Liangxu Xie[a,*] and Xiaojun Xu[a,*]
[a]*Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, Jiangsu, China*
[b]*School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou, Jiangsu, China*
[c]*School of Mathematics and Physics, Jiangsu University of Technology, Changzhou, Jiangsu, China*

Abstract.
BACKGROUND: Protein-ligand binding affinity is of significant importance in structure-based drug design. Recently, the development of machine learning techniques has provided an efficient and accurate way to predict binding affinity. However, the prediction performance largely depends on how molecules are represented.
OBJECTIVE: Different molecular descriptors are designed to capture different features. The study aims to identify the optimal circular fingerprints for predicting protein-ligand binding affinity with matched neural network architectures.
METHODS: Extended-connectivity fingerprints (ECFP) and protein-ligand extended connectivity fingerprints (PLEC) encode circular atomic and bonding connectivity environments with the preference for intra- and inter-molecular features, respectively. Densely-connected neural networks are employed to map the circular fingerprints of protein-ligand complexes to binding affinities
RESULTS: The performance of neural networks is sensitive to the parameters used for ECFP and PLEC fingerprints. The R2_score of the evaluated ECFP and PLEC fingerprints reaches 0.52 and 0.49, higher than that of the improperly set ECFP and PLEC fingerprints with R2_score of 0.45 and 0.38, respectively. Additionally, compared to the predictions from the standalone fingerprints, the ECFP+PLEC conjoint ones slightly improve the prediction accuracy with R2_score of approximately 0.55.
CONCLUSION: Both intra- and inter-molecular structural features encoded in the circular fingerprints contribute to the protein-ligand binding affinity. Optimizing the parameters of ECFP and PLEC can enhance performance. The conjoint fingerprint scheme can be generally extended to other molecular descriptors for enhanced feature engineering and improved predictive performance.

Keywords: Protein-ligand binding affinity, molecular fingerprints, neural networks, ECFP, PLEC

1. Introduction

As biomolecules are essential components of living systems, understanding their biological, chemical, and physical properties is crucial in elucidating their functions. Quantitative structure-activity/property

[1]These authors contributed equally to this work.
[*]Corresponding authors: Xiaojun Xu and Liangxu Xie, Institute of Bioinformatics and Medical Engineering, Jiangsu University of Technology, Changzhou, Jiangsu, China. E-mails: xuxiaojun@jsut.edu.cn; and xieliangxu@jsut.edu.cn.

relationship (QSAR/QSPR) models have been developed to establish the link between molecular structures and properties [1,2]. In particular, protein-ligand binding affinity is of significant importance in QSAR/QSPR studies [3,4]. Although various approaches have been developed for predicting binding affinity, recent advances in machine-learning techniques have paved an alternative way that offers impressive efficiency and accuracy in prediction [5,6].

By converting molecular structures into computer-readable feature vectors, machine learning maps the input structural features through hierarchical non-linear functions to the output molecular properties. However, the prediction performance depends on how molecules are represented. Many types of molecular descriptors have been designed to meet the different demands of structure-property modeling [7–12]. Binding affinity of a protein-ligand complex is related to three types of interactions: intra-protein, intra-ligand, and inter-protein-ligand. In this study, two types of circular fingerprints are used to capture the intra-molecular and inter-molecular structural and interaction information. Specifically, the intra-protein and intra-ligand features are encoded by the extended-connectivity fingerprints (ECFP) [7], and the inter-molecular features are encoded by the protein-ligand extended connectivity fingerprints (PLEC) [10]. Neural networks are employed to learn features from the circular fingerprints to evaluate the effects of intra- and inter-molecular features on protein-ligand binding affinity. The improved performance from the conjoint (ECFP+PLEC) fingerprints further suggests that these two fingerprints can provide complementary information for predicting molecular properties of complex systems.

2. Materials and methods

2.1. Protein-ligand binding data

The PDBbind database [13] holds a PDB-wide collection of experimentally measured binding affinities for the protein-ligand complexes with known 3D structures. The refined set of PDBbind is selected in this study due to its higher quality binding affinity data in the form of dissociation constant K_d, inhibition constant K_i or -logK_d/K_i constant (pK). The current release of PDBbind refined set (version 2020) contains 5316 protein-ligand complexes, which is randomly split into training, validation and test subsets with a ratio of 3:1:1 for the subsequent neural networks learning and prediction.

2.2. ECFP fingerprints

As shown in Fig. 1A, the ECFP generation process assigns an integer identifier to each substructure feature, denoted by a central atom and a diameter, by a hashing procedure. The positions of "1" bits (representing the presence of particular substructures) in the fixed-length binary vector representation are derived from these integer identifiers. There are two major parameters: maximum diameter ($d_{\max}$) and number of bits (*nBits*). In this study, ECFP fingerprints are computed by the RDKit package (https://rdkit.org). For clarity, we use the notation of ECFP$^{d_{\max}}_{nBits}$ (e.g., ECFP$^{4}_{2048}$) to explicitly denote the ECFP fingerprints generated with the specified values of $d_{\max}$ and *nBits*.

2.3. PLEC fingerprints

The PLEC generation process, as depicted in Fig. 1B, involves the identification of a pair of interacting atoms (from the protein and ligand, respectively) if the distance between them is within a distance cut-off (in 3D). As with the ECFP generation process shown in Fig. 1A, the substructures denoted by the central

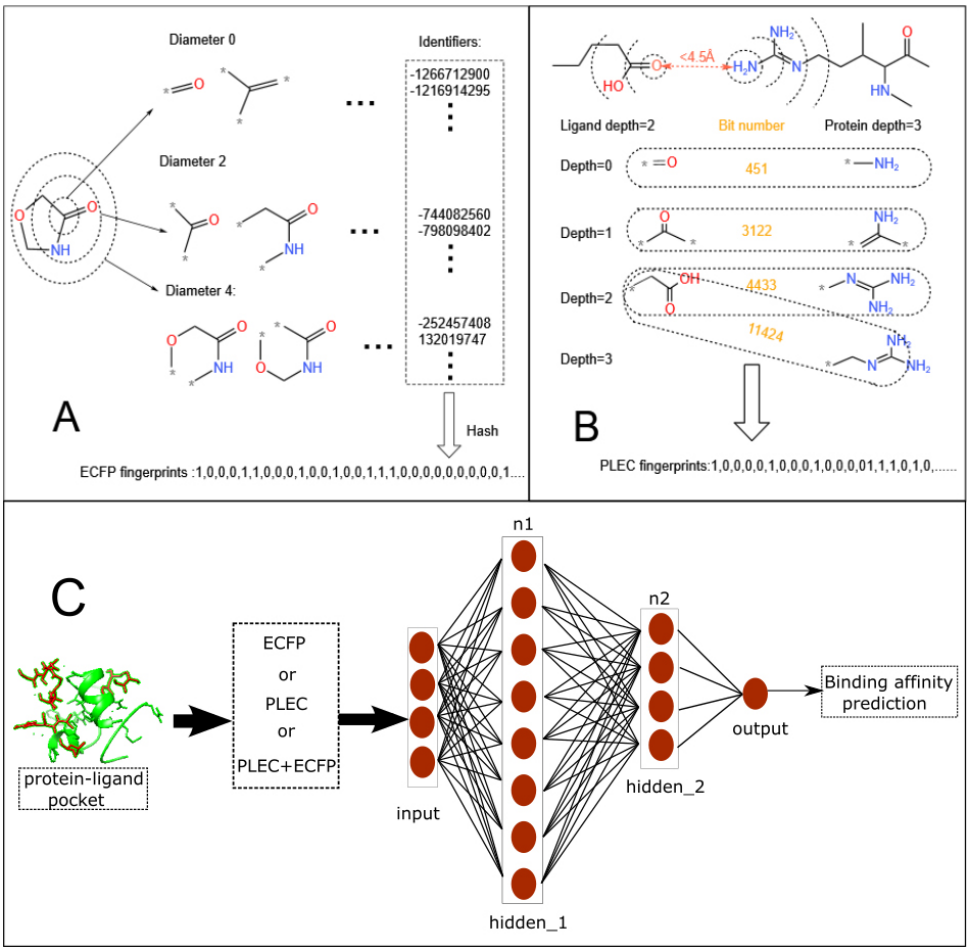

Fig. 1. Constructions of the ECFP (A) and PLEC (B) fingerprints. (C) Neural network architecture for the molecular fingerprints-based protein-ligand binding affinity prediction.

atom itself and its atomic neighbors and bonding connectivity (up to the predefined maximum diameter, i.e., depth) are generated for each atom. An integer identifier is then assigned to each pair of substructures and the positions of "1" bits in the fixed-length binary vector representation are derived from these identifiers. There are four major parameters: distance cut-off, protein depth ($d_{protein}$), ligand depth (d_{ligand}), and number of bits (*nBits*). In this study, the default value of 4.5 Å is applied for the distance cut-off and different values of $d_{protein}$, d_{ligand}, and *nBits* are used to generate different fingerprints. The PLEC fingerprints are computed by the Open Drug Discovery Toolkit (https://oddt.readthedocs.io). For clarity, we use the notation of $\text{PLEC}^{d_{ligand}-d_{protein}}_{nBits}$ (e.g., $\text{PLEC}^{2-10}_{4096}$) to explicitly denote the PLEC fingerprints generated with the specified values of d_{ligand}, $d_{protein}$, and *nBits*.

2.4. Neural networks and performance metrics

As shown in Fig. 1C, a densely-connected neural network (DNN) is adopted to predict the protein-ligand binding affinities. We set the number of hidden layers to be two and train the DNN with different number of neurons in hidden layers ($n1$ and $n2$ in Fig. 1C). The GridSearchCV is used to optimize

the hyperparameters. Specifically, eight different values of $n1$ and $n2$, including 32, 64, 128, 256, 512, 1024, 2048, and 4096 are assessed individually. The activation function is tuned with "Relu", "tanh", and "softmax". The "Adam" optimizer with a default learning rate of 0.001 is employed due to its adaptive learning momentum strategy and high efficiency [14]. To control the balance between overfitting and underfitting, the early stopping conditioned on the validation loss with a tolerance of 0.01 within 20 epochs is adopted.

Binding affinity prediction is a single-valued regression task. The mean squared error (MSE), mean absolute error (MAE), Pearson correlation coefficient (R), and R^2_score (R^2) are calculated to quantitatively assess the predictive performance.

3. Results

3.1. Intra- and inter-molecular features

To encode exclusively the intra-molecular (intra-protein and intra-ligand) structural features within the pocket region, we generate the ECFP fingerprints of protein-ligand complexes as follows: (i) ECFP fingerprints are calculated separately for the protein pocket ($ECFP_{protein}$) and the corresponding ligand molecule ($ECFP_{ligand}$); (ii) ECFP fingerprint of the protein-ligand complex is calculated by performing a bitwise "OR" operation, i.e., $ECFP_{complex} = ECFP_{protein}$ OR $ECFP_{ligand}$. For example, if the ECFPs for the pocket and ligand are "0100011000…" and "0011000011…", respectively. The ECFP for the protein-ligand complex after the bitwise "OR" operation would be "0111011011…". This approach ensures that the ECFP fingerprints do not encode any inter-molecular interactions.

By definition, PLEC fingerprints encode circular environments for all atom pairs in contact between a protein and its ligand. Unlike ECFP, PLEC fingerprints exclusively account for inter-molecular interaction features within the pocket region of a protein-ligand complex. It should be noted that while the ECFP environment is based solely on a molecule's topology to describe the physicochemical properties of a central atom, the selection of atom pairs in contact actually involves the 3D structural information of the protein-ligand complex.

The tested range of ($d_{\max}$, *nBits*) for ECFPs is selected based on trends in DNN performance. Specifically, for a fixed $d_{\max}$, we increase the value of *nBits* from 1024 by a factor of 2 until the DNN performance converges or starts to degrade, which determines the $d_{\max}$-dependent best DNN performance. Starting from 2, we gradually increase the value of $d_{\max}$until the $d_{\max}$-dependent best DNN performance converges or starts to degrade. Similarly, the tested range of ($d_{protein}$, d_{ligand}, *nBits*) for PELCs is also selected based on DNN performance.

3.2. ECFP performance

The loss (MSE) can be tracked during the DNN training process. Figure 2A gives the loss in the training and validation subsets for the $ECFP^8_{2048}$ fingerprint, which indicates a tendency toward reaching a balance point between overfitting and underfitting. Furthermore, the similar predictive performance in the training and validation subsets, as shown in the scatter plot of the predicted pK vs. experimental (true) pK (Fig. 2B) for the $ECFP^8_{2048}$ fingerprint, indicates that the DNN model is properly trained. It should be noted that the training and test results for other ECFP fingerprints are similar to those for $ECFP^8_{2048}$ (as shown in Fig. 2A and B).

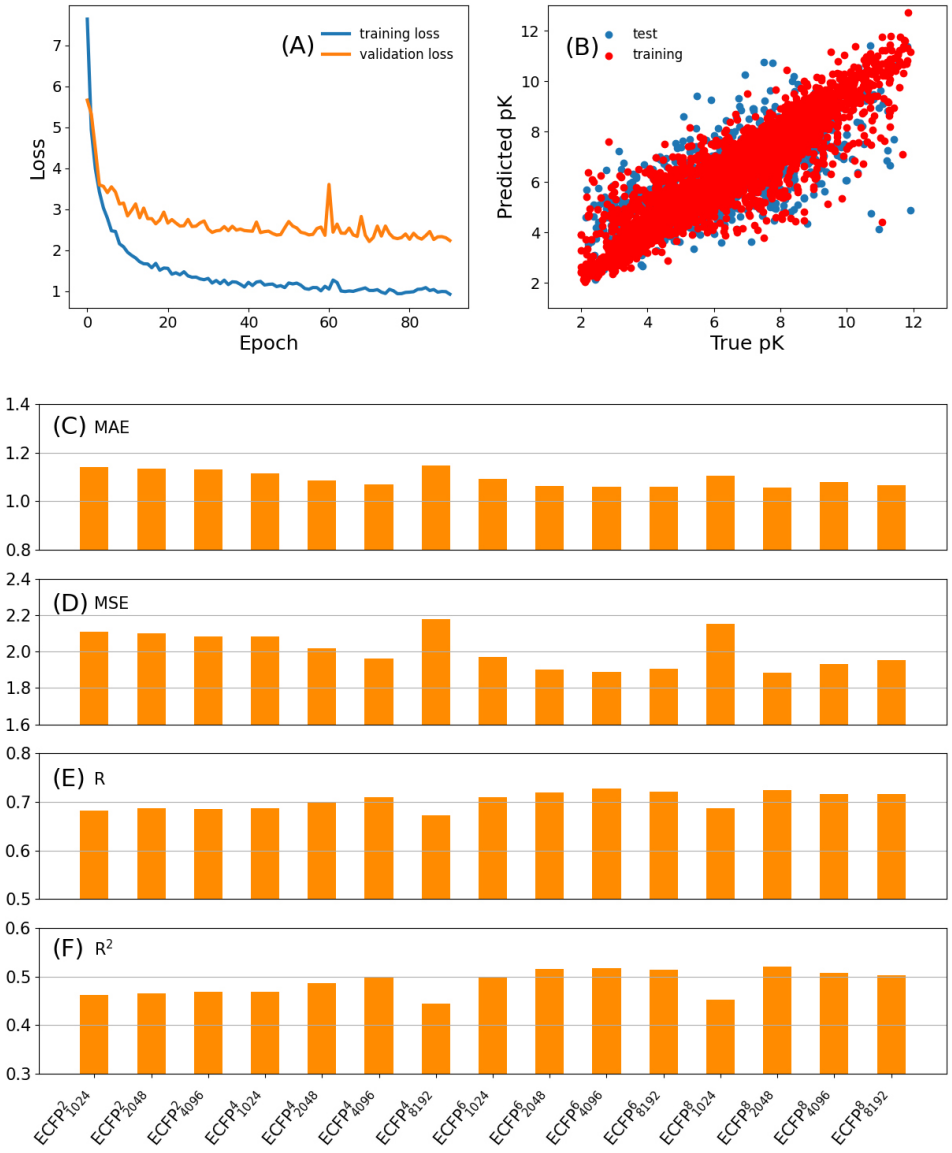

Fig. 2. The loss in the training and validation subsets (A) and the scatter plot of the predicted pK vs. experimentally determined pK (B) for ECFP^{8}_{2048} , respectively. The MAE (C), MSE (D), R (E), and R^2 (F) in the test subset for different ECFPs.

To quantitatively evaluate the predictive performance of different ECFP fingerprints, we calculate the MAE, MSE, R, and R^2, as shown in Fig. 2C, D, E, and F, based on the predicted and true pKs in the test subset. Overall, there are no significant differences in predictive performance caused by changes in $d_{\max}$ and *nBits*. The MAE, MSE, R, and R^2 of all the tested ECFP fingerprints are in the ranges of (1.06, 1.15), (1.88, 2.18), (0.67, 0.73), and (0.45, 0.52), respectively. For $d_{\max} = 2$, very slight differences in prediction at *nBits* $= 1024$, 2048, and 4096 suggest that the model is nearly independent of *nBits* values. However, for $d_{\max} = 4$, 6, and 8, predictive performance improves and reaches convergence as *nBits* increases. Among all the tested ECFP fingerprints, ECFP^{6}_{2048}, ECFP^{6}_{4096}, ECFP^{6}_{8192}, ECFP^{8}_{2048} and ECFP^{8}_{4096} show similar but better performances than the others.

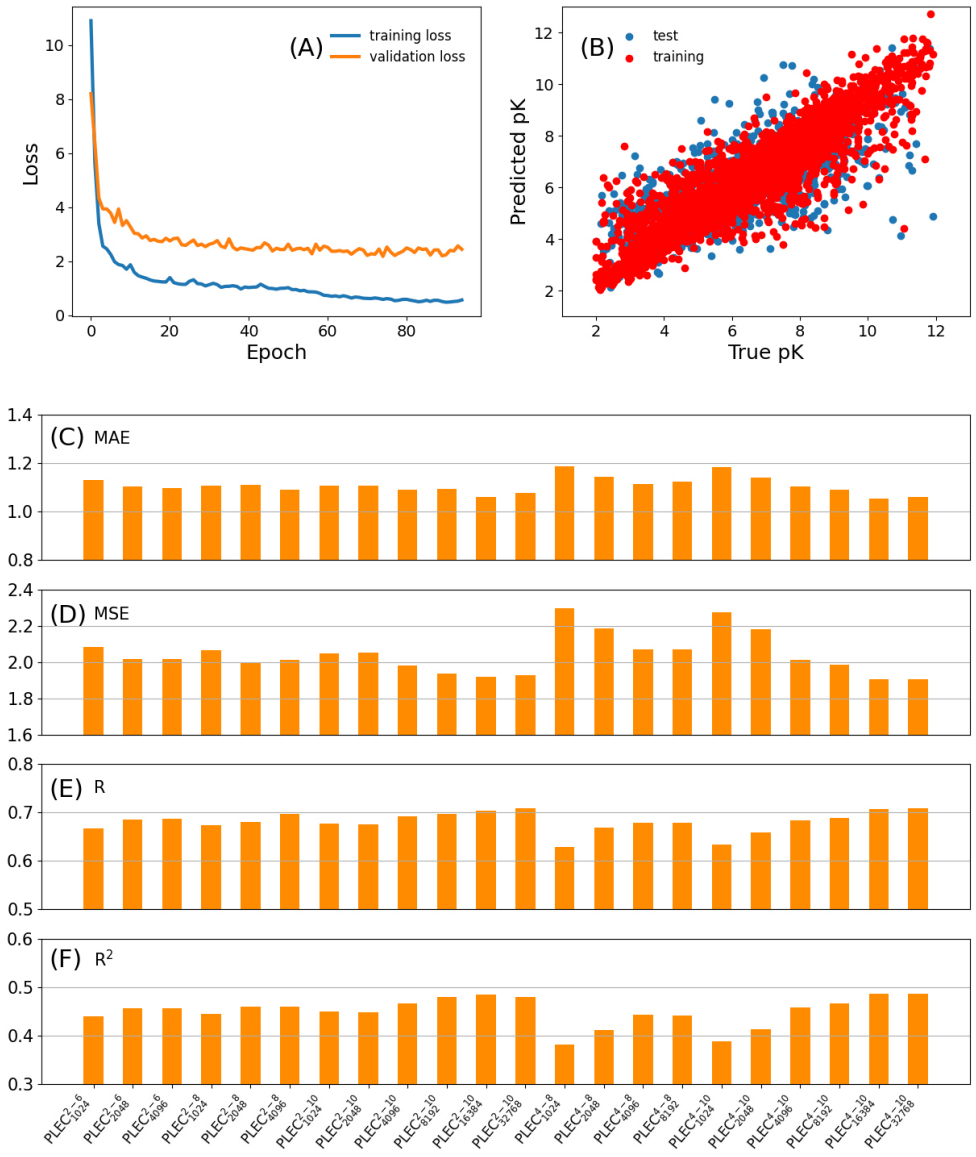

Fig. 3. The loss in the training and validation subsets (A) and the scatter plot of the predicted pK vs. experimentally determined pK (B) for $\text{PLEC}^{4-10}_{16384}$, respectively. The MAE (C), MSE (D), R (E), and R^2 (F) in the test subset for different PLECs.

3.3. PLEC performance

As depicted in Fig. 3A and B for $\text{PLEC}^{4-10}_{16384}$ (others not shown), the loss (MSE) in training and validation subsets and the scatter plot for the comparison between predicted and true pK values suggest that the DNN models for the PLEC fingerprints are well trained. The MAE, MSE, R, and R^2 for all the tested PLEC fingerprints, as shown in Fig. 3C, D, E, and F, are within the ranges of (1.05, 1.19), (1.91, 2.30), (0.63, 0.71), and (0.38, 0.49), respectively. Similar to the ECFP fingerprints, the values of $d_{protein}$, d_{ligand}, and *nBits* have an impact on the predictive performance of the PLEC-based DNN model. For a fixed set of $d_{protein}$ and d_{ligand} (e.g., $d_{protein} = 4$ and $d_{ligand} = 10$), increasing the value of *nBits* leads to an

Table 1
Performance (R^2) of three types of fingerprints for different models

Fingerprints	RF	XGBoost	LightGBM	DNN
$ECFP^{6}_{4096}$	0.53	0.53	0.53	0.52
$ECFP^{8}_{2048}$	0.52	0.53	0.53	0.52
$PLEC^{2-10}_{16384}$	0.49	0.48	0.48	0.48
$PLEC^{4-10}_{16384}$	0.47	0.48	0.47	0.49
$ECFP^{6}_{4096} + PLEC^{2-10}_{16384}$	0.57	0.56	0.55	0.55
$ECFP^{6}_{4096} + PLEC^{4-10}_{16384}$	0.55	0.54	0.54	0.55
$ECFP^{8}_{2048} + PLEC^{2-10}_{16384}$	0.57	0.56	0.55	0.54
$ECFP^{6}_{4096} + PLEC^{2-10}_{16384}$	0.56	0.56	0.55	0.54

increase in predictive performance until it reaches convergence. Among all the tested PLEC fingerprints, $PLEC^{2-10}_{8192}$, $PLEC^{2-10}_{16384}$, $PLEC^{2-10}_{32768}$, $PLEC^{4-10}_{16384}$, and $PLEC^{4-10}_{32768}$ exhibit similar but better performances than the others.

3.4. ECFP+PLEC performance

We select two of the top-performing ECFP fingerprints ($ECFP^{6}_{4096}$ and $ECFP^{8}_{2048}$) and two of the top-performing PLEC fingerprints ($PLEC^{2-10}_{16384}$ and $PLEC^{4-10}_{16384}$) to create four ECFP+PLEC fingerprints. As listed in Table 1, the performances of the four conjoint fingerprints are quite similar with R^2 values of approximately 0.55. Although the improvement in prediction accuracy compared to standalone fingerprints is modest, the results suggest that combining intra- and inter-molecular structural information can enhance the ability of DNN models to capture molecular features for binding affinity prediction.

In addition to the choice of molecular representation, the performance of machine learning is also influenced by the underlying algorithms used for feature engineering. As listed in Table 1, we test the performances of three typical machine learning algorithms, i.e., random forest (RF), XGBoost and LightGBM. However, the similar predictive performances of all four models indicate that altering the feature engineering algorithm may not significantly improve the prediction accuracy. The information of presence/absence of substructures alone is limited and not sufficient for accurately predicting the binding affinities of protein-ligand complexes. Additional information, such as the positions of substructures, may be beneficial for improving predictions. Therefore, properly encoding the spatial information of identified substructures remains a challenge for the further development of circular fingerprints-based molecular descriptors.

4. Discussion

The independent nature of the substructures represented by the "1" bits in both ECFP and PLEC fingerprints suggests that the DNN model is a suitable choice, as it can effectively learn useful information from independent features of input data. The number of hidden layers (n_{hidden}) in DNN is a crucial hyperparameter, and its performance is sensitive to the choice of n_{hidden}. In this study, we only test the DNN performance with $n_{hidden} = 2$. Nevertheless, the trends of DNN predictions with the ECFP and PLEC parameters (e.g., $d_{protein}$, d_{ligand}, $d_{\max}$, and *nBits*) should be relatively insensitive to n_{hidden}. Moreover, different machine learning algorithms, such as support vector machine (SVM), convolutional neural network (CNN), and graph neural network (GNN), map the input features to output targets in various aspects and capabilities. Combining DNN with other algorithms may enhance feature engineering for better performance.

Biomolecules are primarily composed of carbon, nitrogen, and oxygen atoms. Although their covalent bonding patterns differ, the substructures encoded by circular fingerprints (both ECFP and PLEC) are relatively limited, especially for proteins. To address this issue, various types of fingerprints are designed to encode specific aspects of molecular features. Conjoint fingerprints, which combine different types of fingerprints, are expected to be more informative and can achieve better performance than individual ones [15–18]. In fact, binding affinities are determined by the 3D structures of protein-ligand complexes. Combining circular fingerprints with 3D ones, such as molecular surface fingerprints [11], may provide a more comprehensive representation of the underlying molecular structure and enhance predictive performance.

Furthermore, it is important to note that protein-ligand binding data is collected from different experimental platforms with their 3D structures solved in different resolutions. Directly analyzing the integrated data may lead to issues such as the batch effect [19,20], resulting in limited predictive performance. To overcome these limitations, a promising approach is to combine molecular docking and molecular dynamics simulations [21,22] with machine learning algorithms. This can provide a more accurate representation of the protein-ligand binding interactions and improve predictive performance.

5. Conclusions

In this study, we exclusively use two types of circular fingerprints, ECFP and PLEC, to encode the intra-molecular and inter-molecular structural features of protein-ligand complexes. The neural networks learning from the standalone fingerprints illustrates that both types of features contribute to the protein-ligand binding affinities. Through systematic testing of different ECFP and PLEC fingerprints, we identified the optimal combinations of the fingerprints with neural network architectures, which can be applied to various applications such as virtual screening, structure-activity relationship modeling and compound library analysis. The improved predictive performance from the ECFP+PLEC conjoint fingerprints suggests that the combination of two complementary fingerprints can better capture the molecular features and interactions that determine the protein-ligand binding affinity. We anticipate that the conjoint fingerprint scheme can be generally extended to other molecular descriptors for enhanced feature engineering and improved predictive performance.

Funding

This research was funded by the National Natural Science Foundation of China (12074151, 22003020) and the Natural Science Foundation of Jiangsu Province (BK20191031).

Conflict of interest

The authors declare no conflict of interest.

References

[1] Danishuddin Khan AU. Descriptors and their selection methods in QSAR analysis: paradigm for drug design. Drug Discovery Today. 2016; 21(8): 1291-1302.
[2] Khan PM, Roy K. Current approaches for choosing feature selection and learning algorithms in quantitative structure – activity relationships (QSAR). Expert Opinion on Drug Discovery. 2018; 13(12): 1075-1089.
[3] Gilson MK, Zhou H. Calculation of Protein-Ligand Binding Affinities. Annual Review of Biophysics and Biomolecular

Structure. 2007; 36(1): 21-42.
[4] Fu H, Chen H, Blazhynska M, Lacam EGC, Szczepaniak F, Pavlova A, et al. Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations. Nature Protocols. 2022; 17(4): 1114-1141.
[5] Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug – target interaction: a survey paper. Briefings in Bioinformatics. 2021; 22(1): 247-269.
[6] Dhakal A, McKay C, Tanner JJ, Cheng J. Artificial intelligence in the prediction of protein – ligand interactions: recent advances and future directions. Briefings in Bioinformatics. 2022; 23(1): b476.
[7] Rogers D, Hahn M. Extended-Connectivity Fingerprints. Journal of Chemical Information and Modeling. 2010; 50(5): 742-754.
[8] Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY. Predicting Drug – Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. Journal of Chemical Information and Modeling. 2019; 59(9): 3981-3988.
[9] Da C, Kireev D. Structural Protein-Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. Journal of Chemical Information and Modeling. 2014; 54(9): 2555-2561.
[10] Wójcikowski M, Kukieka M, Stepniewska-Dziubinska MM, Siedlecki P. Development of a protein-ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. Bioinformatics. 2019; 35(8): 1334-1341.
[11] Gainza P, Sverrisson F, Monti F, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods. 2020; 17(2): 184-192.
[12] Wang DD, Chan M, Yan H. Structure-based protein-ligand interaction fingerprints for binding affinity prediction. Computational and Structural Biotechnology Journal. 2021; 19: 6291-6300.
[13] Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. Journal of Medicinal Chemistry. 2004; 47(12): 2977-2980.
[14] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2015; arXiv1412.6980.
[15] Tseng YJ, Hopfinger AJ, Esposito EX. The great descriptor melting pot: mixing descriptors for the common good of QSAR models. Journal of Computer-Aided Molecular Design. 2012; 26(1): 39-43.
[16] Xie L, Xu L, Kong R, Chang S, Xu X. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. Frontiers in Pharmacology. 2020; 11.606668.
[17] Rahaman O, Gagliardi A. Deep Learning Total Energies and Orbital Energies of Large Organic Molecules Using Hybridization of Molecular Fingerprints. Journal of Chemical Information and Modeling. 2020; 60(12): 5971-5983.
[18] Mendolia I, Contino S, De Simone G, Perricone U, Pirrone R. EMBER – Embedding Multiple Molecular Fingerprints for Virtual Screening. International Journal of Molecular Sciences. 2022; 23(4):2156.
[19] Huang H, Wu N, Liang Y, Peng X, Shu J. SLNL: A novel method for gene selection and phenotype classification. International Journal of Intelligent Systems. 2022; 37(9): 6283-6304.
[20] Huang H, Rao H, Miao R, Liang Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. BMC Bioinformatics. 2022; 23(10): 353.
[21] Morris CJ, Corte DD. Using molecular docking and molecular dynamics to investigate protein-ligand interactions. Modern Physics Letters B. 2021; 35(08): 2130002.
[22] Gu S, Shen C, Yu J, et al. Can molecular dynamics simulations improve predictions of protein-ligand binding affinity with machine learning? Briefings in Bioinformatics. 2023; bbad008. doi: 10.1093/bib/bbad008.