Mathematical Biosciences and Engineering

MBE, 21(2): 2608–2625.
DOI: 10.3934/mbe.2024115
Received: 25 October 2023
Revised: 05 January 2024
Accepted: 10 January 2024
Published: 18 January 2024

http://www.aimspress.com/journal/MBE

Research article

Integrative approach for predicting drug-target interactions via matrix factorization and broad learning systems

Wanying Xu[1], Xixin Yang[1,2,*], Yuanlin Guan[3,4], Xiaoqing Cheng[1] and Yu Wang[1]

[1] College of Computer Science & Technology, Qingdao University, Qingdao 266071, China
[2] School of Automation, Qingdao University, Qingdao 266071, China
[3] Key Lab of Industrial Fluid Energy Conservation and Pollution Control, Ministry of Education, Qingdao University of Technology, Qingdao 266520, China
[4] School of Mechanical & Automotive Engineering, Qingdao University of Technology, Qingdao 266520, China

* **Correspondence:** Email: yangxixin@qdu.edu.cn.

Abstract: In the drug discovery process, time and costs are the most typical problems resulting from the experimental screening of drug-target interactions (DTIs). To address these limitations, many computational methods have been developed to achieve more accurate predictions. However, identifying DTIs mostly rely on separate learning tasks with drug and target features that neglect interaction representation between drugs and target. In addition, the lack of these relationships may lead to a greatly impaired performance on the prediction of DTIs. Aiming at capturing comprehensive drug-target representations and simplifying the network structure, we propose an integrative approach with a convolution broad learning system for the DTI prediction (ConvBLS-DTI) to reduce the impact of the data sparsity and incompleteness. First, given the lack of known interactions for the drug and target, the weighted K-nearest known neighbors (WKNKN) method was used as a preprocessing strategy for unknown drug-target pairs. Second, a neighborhood regularized logistic matrix factorization (NRLMF) was applied to extract features of updated drug-target interaction information, which focused more on the known interaction pair parties. Then, a broad learning network incorporating a convolutional neural network was established to predict DTIs, which can make classification more effective using a different perspective. Finally, based on the four benchmark datasets in three scenarios, the ConvBLS-DTI's overall performance out-performed some mainstream methods. The test results demonstrate that our model achieves improved prediction effect on the area under the receiver operating characteristic curve and the precision-recall curve.

Keywords: drug-target interaction prediction; broad learning system; neighbor regularization logistic matrix factorization

1. Introduction

Drug-target interactions (DTIs) involve the binding of a drug to the relevant site of a target protein to trigger a biochemical reaction [1]. The efficacy is related to the biological activity of the protein. However, it is complicated for experiments to predict a drug's success and drug discovery is time-consuming and expensive [2,3], which is estimated to typically take 12–15 years and cost over $100 million [4]. For these reasons, in the past decades, computer-aided drug design (CADD) has been proposed to discriminate new drugs and consists of processes such as virtual screening, molecular docking, and QSAR methods [5]. Currently, due to limited ligand data and the limited information on the structure of novel target proteins [6], these approaches are inappropriate and inefficient given the growth of available biological and chemical data [2]. Recently, with the advent of various deep learning methods, a significant future trend in AI-based drug discovery has been identified [7]. It is essential for drug discovery to accurately predict the number of DTIs [8]. Therefore, it is urgent to devise richer and more compatible computational methods to differentiate between potential DTIs.

The concept of "guilt-by-association" [9] has been described in DTIs prediction. It is defined that if drug A has target proteins, and the action event between drug B is similar to drug A, targets interactions are likely to appear, and the reverse is also true. Machine learning methods are used for DTIs prediction and can successfully solve the assumption. For instance, Mei et al. [10] proposed bipartite local models (BLMs) that considered neighbors' interaction profiles where neighbor-based interaction-profile inferring (NII) can be effective in defining a new candidate problem. Luo et al. [11] used an inductive matrix completion method, in which seven kinds of drug/target-related similarities were included in an integrated network (e.g., drugs, proteins, diseases, and side-effects). Ezzat et al. [12] proposed graph regularized matrix factorization (GRMF) and weighted graph regularized matrix factorization (WGRMF) methods that introduced graph regularization into the matrix factorization in order to learn manifolds. Moreover, a preprocessing step (WKNKN) has been developed to rescore unknown drug-target pairs that were previously regarded as null values. Although these methods have been proven to be effective, there are challenges to overcome complex data structures such as interaction networks of drugs or targets. Furthermore, the rapid growth of drug/target-related data has outpaced their ability to process and analyze information. With the emergence of diverse and enriched feature representations, the efficacy of the above methods may limit the exploration of more comprehensive topological information and node characteristics between drugs and target proteins.

Network-based algorithms and feature-based algorithms become famous in the field. Generally, identifying DTIs is considered as a binary classification task by extracting features vectors of drugs as well as targets. Several number of heterogeneous data have been integrated into a heterogeneous network to boost the accuracy of DTIs prediction tasks [13]. The deep belief network (DBN) [14] has been proposed to build an end-to-end method for abstracting raw input samples. Moreover, sequence-based approaches are universal. Different architectures [15–18] have been developed for feature extraction of sequence information. DrugVQA [19] employs a bidirectional long-short time memory network to tackle the prediction problem. Furthermore, graph-based methods are suitable for the two-dimensional representation of structural information. Zhao et al. [20] utilized a combination of graph

convolutional network (GCN) and deep neural network (DNN) to enhance the identification of DTIs. GNN was coupled with CNN, which was designed as drug feature and target feature extraction method [21]. LASSO has been employed by You et al. [22] as a feature procession. Thafar et al. [23] constructed the DTi2Vec model including graph embedding which capture relationships between drugs and targets and then these features are fed into the ensemble classifier for prediction analyses. Huang et al. designed a molecular sub-structure representation and used massive unlabeled biomedical data through an augment transformer [24]. Peng et al. [25] introduced CNN to identify DTIs and trained the denoising autoencoder (DAE) as a feature selector. Although these methods can effectively predict DTIs, the problem of parameter count and computation amount need to be given more attention.

The broad learning system (BLS) [26] is characterized by a relatively simple neural network architecture comprising only three layers of neurons. Inspired by the concept of the random vector functional-link neural network (RVFLNN) [27,28], its training procedure is facilitated through pseudo-inverse calculations. Due to its training procedure and flat structure, BLS has the advantages of fast computing speed and few training parameters. Therefore, BLS has been widely applied to various disciplines including medicine [29]. For instance, Fan et al. [30] proposed a stacked ensemble classifier build by BLS for the prediction of interactions between lncRNA and proteins. Zheng et al. [31] designed a modified BLS-based model to predict miRNA-disease associations using sequence similarities of microRNA (miRNAs). The above applications of BLS in this area have been proven to be useful. However, there is a lack of related research for DTIs based on BLS. Additionally, since labeled data volumes are always sparse and insufficient, prediction modeling is to performance is inadequate. By fusing information from multiple aspects to overcome the limitations, the above methods can improve the performance, indicating that these combined models could solve the challenge of interaction matrix sparsity.

In this study, we developed a novel model called ConvBLS-DTI to predict DTIs. Compared with the previous DTI predictive methods, ConvBLS-DTI integrates matrix factorization with the broad learning system, yielding reliable DTI prediction results. The task of DTI prediction is formulated as a binary classification problem to determine whether a drug-target pair is a DTI. The major contributions of this paper are as follows:

1) We address the challenges of data sparsity and incompleteness by employing a WKNKN algorithm as a pre-processing step, which help to mitigate the adverse effects of a large number missing interaction value.

2) We propose a matrix factorization technique used on the interaction matrix to generate two latent feature matrices for drugs and targets, thereby enabling the learning of low-dimensional vector representations of features.

3) Based on the CNN algorithm, ConvBLS-DTI can handle the DTIs prediction, taking the extracted drug-target pairs feature vectors as inputs.

2. Methods

2.1. Overall framework

The architecture of the proposed ConvBLS-DTI method is depicted in Figure 1. It is primarily composed of three sessions: First, we utilized the WKNKN algorithm to alleviate the sparsity of the DTI matrix, thus enhancing the input information complement of the model and improving its

predictive performance. After construction of the DTI matrix, matrix factorization is used to decompose DTI matrix into two feature matrices of low ranks which obtains vector representation of the drug features and target features. Then, the drug feature vectors combine with the target feature vectors together to get the final feature vectors. Finally, ConvBLS is built for classification. A CNN is leveraged to enhance the nodes' representation, followed by a broad learning module, which further enable satisfactory results in effective identification of DTIs.

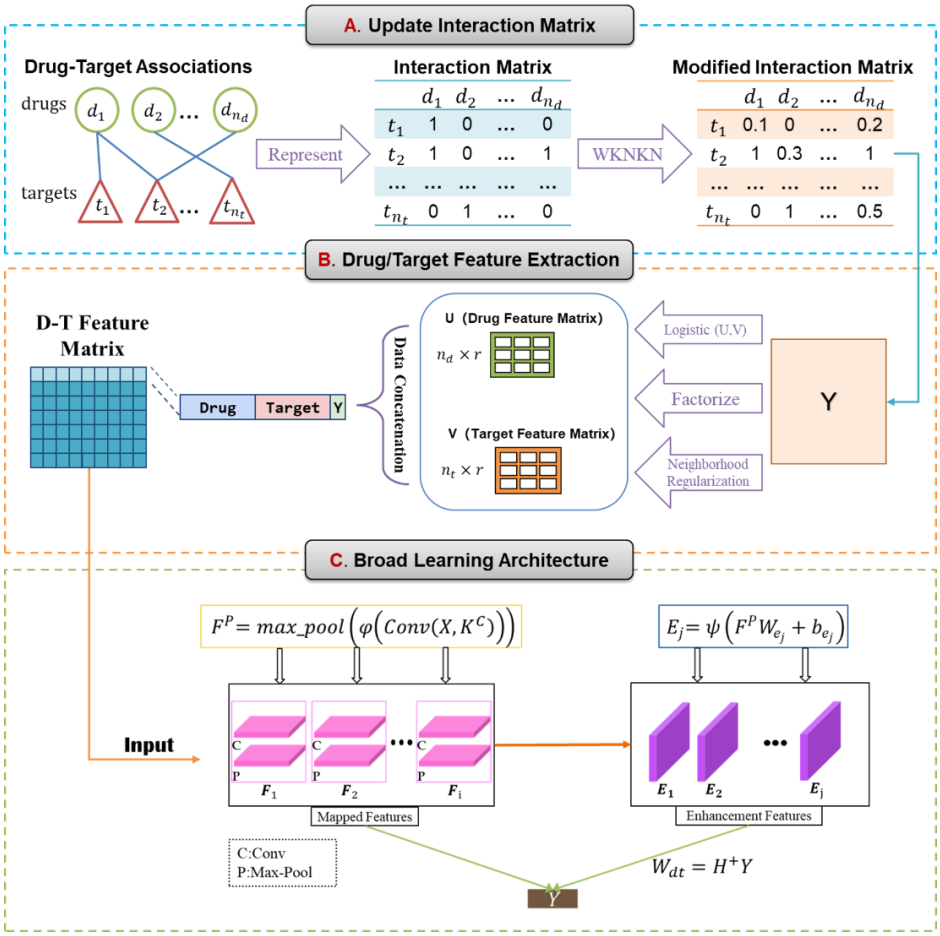

Figure 1. Overview of the ConvBLS-DTI predictive workflow.

2.2. WKNKN data processing method

We initially reconstruct the interaction matrix using the computational preprocessing technique, which can effectively complement the interaction matrix for the identification of DTIs and improve the known DTI samples. As shown on the left side of step A in Figure 1, the green circles, the red triangles, and the blue lines separately denote drugs, targets and the known interaction. $\boldsymbol{D} = \{d_i\}_{i=1}^{n_d}$ and $\boldsymbol{T} = \{t_j\}_{j=1}^{n_t}$ are separately described as each node for drugs and targets, where n_d is the number of drugs and n_t is the number of targets. The associations between n_d drugs and n_t targets are represented by an interaction matrix $\boldsymbol{Y} \in \{0,1\}$, in which $Y_{ij} = 1$ indicates a known interaction between drug d_i and target t_j, and $Y_{ij} = 0$ otherwise. In addition, the similarity matrix of both drugs and targets is represented as $\boldsymbol{SD} \in \boldsymbol{R}^{n_d \times n_d}$ and $\boldsymbol{ST} \in \boldsymbol{R}^{n_t \times n_t}$.

Numerous unknown interactions can significantly impact the evaluation outcome of the model and introduce prediction bias. In DTIs prediction, weighted k-nearest neighbors (k-NN) has been employed to leverage similarity measures to promote further prediction performance. Weighted k-NN considers both neighbor similarity and distances, incorporating distance weights to calculate likelihood values of unconfirmed drug-target interactions. Specifically, given a drug-target pair, the algorithm first identifies the k-NN and assigns weights to each neighbor based on their similarity and distance. Weight involved in WKNKN [12] is computed by Gaussian weighting method. The calculated weighted likelihood values can be used to predict the likelihood of unknown DTIs within the matrix. Here, the specific operation is achieved through the following three steps:

$$Y_d(d) = \frac{1}{M_d}\sum_{i=1}^{\mathrm{K}} \omega_{d_i}\,\mathrm{Y}(d_i) \tag{2.1}$$

where $Y_d(d)$ denotes the likelihood score of interaction for drug d_i. M_d is defined as a normalization term, ω coefficient represents the weights of the K nearest known neighbors of drug d_i. Similarly, the same terms are computed to estimate the interaction likelihood score of the target t_j:

$$Y_t(t) = \frac{1}{M_t}\sum_{i=1}^{\mathrm{K}} \omega_{t_j}\,\mathrm{Y}(t_j) \tag{2.2}$$

where $Y_t(t)$ denotes the likelihood score of interaction for target t_j. M_t is the normalization term and ω coefficient represents the weights of the K nearest known neighbors of target t_j. Finally, the derived formula is as follows:

$$Y_{WKNKN} = \max\left(\frac{Y_d+Y_t}{2}, \boldsymbol{Y}\right) \tag{2.3}$$

Therefore, if Y_{ij} is 0, Y_{WKNKN} replaces it with an average of the weighted interaction likelihood value. For the matrix representation, 0 and 1 denote the absence and presence of interactions between drugs and targets, respectively. Likelihood serves as a measure of the possibility of interaction between a drug and a target, typically ranging from 0 to 1. Higher likelihood values indicate a higher likelihood of interaction, while lower likelihood values suggest a lower likelihood.

2.3. NRLMF feature extraction method

Considering that most studies mainly concentrate on extracting features from drugs and targets individually and less on the relationships of the DTI, neighbor regularization logistic matrix factorization (NRLMF) [32] is used to represent drugs and targets in the right part of step B. NRLMF is an unsupervised learning strategy that mainly infers unknowns through known interactions and their similarities, so no negative samples are required. The valid connections are denoted as the modified interaction matrix made of known and unknown interactions. As shown in Figure 1, the DTI probabilities can be defined as a logistic function:

$$P_{i,j} = \frac{\exp\left(u_i v_j^{\mathrm{T}}\right)}{1+\exp\left(u_i v_j^{\mathrm{T}}\right)} \tag{2.4}$$

where each term $u_i \in \boldsymbol{R}^{1\times r}$ is denoted as the r-dimensional potential representation of each drug d_i. Similarly, each term $v_j \in \boldsymbol{R}^{1\times r}$ represents the r-dimensional potential representation of each target t_j.

In this way, the potential feature vectors for all drugs and all targets can be summarized as $\boldsymbol{U} = \left(u_1^{\mathrm{T}}, \ldots, u_{n_d}^{\mathrm{T}}\right)$ and $\boldsymbol{V} = \left(v_1^{\mathrm{T}}, \ldots, v_{n_t}^{\mathrm{T}}\right)$, where T refers to the transpose of the matrix.

The neighborhood regularization method proposes to add the nearest neighbors of drugs and targets to further increase information diversity and enable higher accuracy without overfitting. The neighborhood regularization is achieved by:

$$\frac{\alpha}{2}\sum_i^{n_d}\sum_j^{n_d}\boldsymbol{SP}_{i\mu}\left\|u_i - u_j\right\|_{\mathrm{f}}^2 \tag{2.5}$$

$$\frac{\alpha}{2}\sum_i^{n_t}\sum_j^{n_t}\boldsymbol{SQ}_{\vartheta j}\left\|v_i - v_j\right\|_{\mathrm{f}}^2 \tag{2.6}$$

where α is the Laplace regularization parameter, and $\|\cdot\|_{\mathrm{f}}$ is the Frobenius norm of the matrix, and the parameters $\boldsymbol{SP}$ and $\boldsymbol{SQ}$ represent neighbors similarity measure matrix are given by:

$$\boldsymbol{SP}_{i\mu} = \boldsymbol{SD}_{i\mu} \text{ if } d_\mu \in W_d(d_i) \text{ else } \boldsymbol{SP}_{i\mu} = 0 \tag{2.7}$$

$$\boldsymbol{SQ}_{\vartheta j} = \boldsymbol{ST}_{\vartheta j} \text{ if } t_\vartheta \in W_t\left(t_j\right) \text{ else } \boldsymbol{SQ}_{\vartheta j} = 0 \tag{2.8}$$

where $\boldsymbol{SD}$ and $\boldsymbol{ST}$ denote as similarity matrix, and $W_d(d_i)$ is defined as the nearest neighbors of a node d_i, and $W_t\left(t_j\right)$ is defined as the nearest neighbors of a node t_j.

The matrix factorization (MF) method decomposes the interaction matrix into two low-rank matrices. The MF is formulated as a feature extraction task to obtain the description of the drugs and their targets as features. The feature matrix is obtained by maximizing the objective function via the posterior probability distribution:

$$\max_{\boldsymbol{U},\boldsymbol{V}} P\left(\boldsymbol{U},\boldsymbol{V}|\boldsymbol{Y},\sigma_d^2,\sigma_t^2\right) \tag{2.9}$$

where $\boldsymbol{Y}$ denotes the interaction matrix, σ_d^2 and σ_t^2 are parameters that control the variance of Gaussian distribution of drug set and target set.

Thus, drugs and targets can be denoted as two r-dimensional feature representations. As illustrated in Figure 2, the drug feature is $\boldsymbol{U} - \boldsymbol{D} = [DF_1, DF_2, \ldots, DF_r]$, and the target feature is $\boldsymbol{V} - \boldsymbol{T} = [TF_1, TF_2, \ldots, TF_r]$. Then, the drug feature vectors and target feature vectors are merged and assigned the label based on the interaction matrix $\boldsymbol{Y}$. To ensure the quality of the model, the number of negative samples is equal to positive number in each dataset. Negative samples are randomly generated based on the interaction matrix $\boldsymbol{Y}$. The pairwise drug–target feature vector is used as input to the neural network, which can be expressed as $\boldsymbol{FV}$= $[DF_1, DF_2, \ldots, DF_r, TF_1, TF_2, \ldots, TF_r]$.

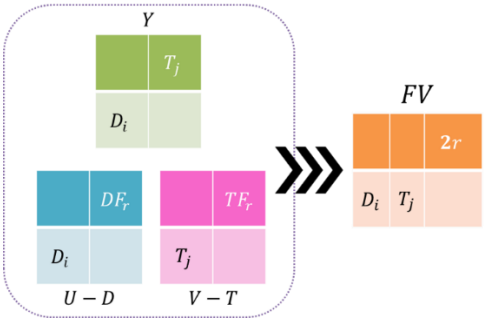

Figure 2. Construction of potential feature vectors for a drug-target pair.

2.4. ConvBLS prediction method

After concatenating the features of drugs and targets, in order to achieve better performance and train more effectively, the ConvBLS model is used as a classification method to determine the predictions of the DTIs. As shown in Figure 3, we developed a broad learning system that combined convolutional neural network to extract high-quality drug and target representations for better prediction.

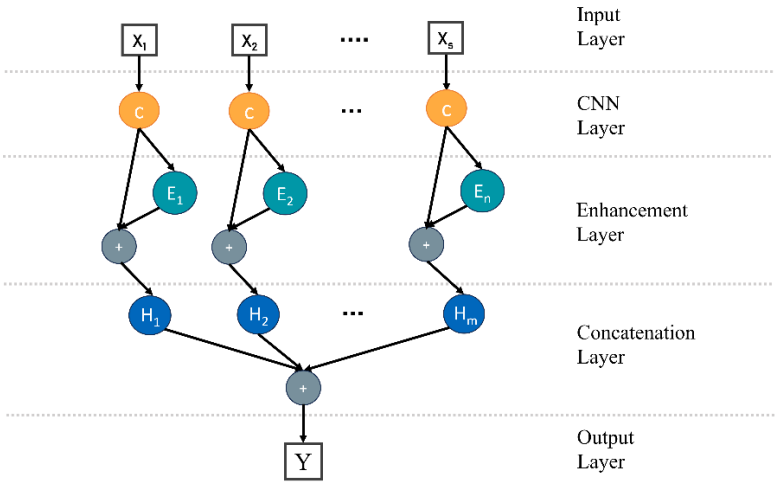

Figure 3. ConvBLS network structure.

ConvBLS mostly includes two parts: The 1D-CNN module and enhancement nodes. CNN block is used to learn a representative features of targets and drugs. The input data of 1D-CNN is a one-dimensional feature vectors, and the convolution kernel is also in one-dimensional form. The enhancement layer is responsible for further feature extracting. The detailed network structure is shown in Figure 3. This section completes the classification task with ConvBLS.

Given the lack of learning ability of the original feature mapping, the CNN block is selected for sequence data of drug-target features. It contains multiple groups of feature mapping nodes composed of a 1D-CNN layer and a max pooling layer. To solve complex tasks, learning models increasingly go deeply. The multiscale random convolution feature is expanded to improve robustness. The detailed computational procedure is as follows:

First, the drug–target predictive model ConvBLS is constructed based on previous obtained feature data FV. The input is connected to the mapping matrix by applying 1-D convolution kernels to generate the corresponding feature representation. All the random items can act as the convolution kernel so that we can achieve the output:

$$\boldsymbol{F^C} = \varphi\big(Conv(\boldsymbol{X}, K^C)\big) \tag{2.10}$$

where $\boldsymbol{X}$ is the input feature vectors, K^C is the convolution kernel, $Conv(.)$ is denoted as the convolution function, and $\varphi(.)$ is the activation function. The descriptor of the mapping feature is called $\boldsymbol{F^C}$. Then, the down-sampling method is used to allow the feature to be robust:

$$\boldsymbol{F^P} = max\_pool\big(\boldsymbol{F^C}\big) \tag{2.11}$$

where $\boldsymbol{F^P}$ is the result after max pooling function. Next, an enhancement layer is built. Using random weights and nonlinear transformation, the enhancement nodes are obtained:

$$E_j = \psi\left(\boldsymbol{F^P} W_{e_j} + b_{e_j}\right) j = 1,2, \dots, n \tag{2.12}$$

where $\psi(.)$ is the activation function, weights and bias represented as W_{e_j} and b_{e_j}, which are randomly initialized, and n is the group number of enhanced nodes. All of the enhancement nodes can be represented as $\boldsymbol{E^n} \equiv [E_1, E_2, \dots E_n]$. Finally, the improved feature layer and enhancement layer are concatenated into one matrix as a single neural network. Hence, featurization constitutes the outputs of weight of the BLS, based on $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{W_{dt}}$.

$$\boldsymbol{W_{dt}} = \boldsymbol{H^+ Y} = [\boldsymbol{F^P}|\boldsymbol{E^n}]^+ \boldsymbol{Y} \tag{2.13}$$

The ridge regression approximation algorithm [33] is utilized to determine the $[\boldsymbol{F^P}|\boldsymbol{E^n}]^+$:

$$[\boldsymbol{F^P}|\boldsymbol{E^n}]^+ = \lim_{\lambda \to 0}\left(\lambda \boldsymbol{I} + [\boldsymbol{F^P}|\boldsymbol{E^n}][\boldsymbol{F^P}|\boldsymbol{E^n}]^{\mathrm{T}}\right)^{-1}[\boldsymbol{F^P}|\boldsymbol{E^n}]^{\mathrm{T}} \tag{2.14}$$

3. Materials

Here, we describe the dataset used in this paper and provide the experiment setup and evaluation metrics for comparing model performance in subsequent experiments.

3.1. Dataset description

In this study, two benchmark datasets are used for evaluating our proposed model: the Yamanishi's dataset and Luo's dataset. The first one is the gold benchmark dataset created by Yamanishi et al. [34]. It is classified into four categories based on the target protein class, namely: (i) enzyme (E), (ii) ion channel (IC), (iii) G protein-coupled receptor (GPCR), and (iv) nuclear receptor (NR). Since the discovery of the interactions in these datasets 14 years ago, we implemented the completed version of the original golden standard datasets collected by Liu et al. [35]. The new datasets added information on the KEGG pathways [36], DrugBank [37], and ChEMBL [38] databases. The second one was developed by Luo et al. [11], consisting of four categories of nodes (drugs, proteins, diseases, and side-effects) and six types of connections (drug-target interaction, drug-drug interactions, protein-protein interactions, drug-disease associations, protein-disease associations, and drug-side-effect associations). Table 1 lists the detailed statistical entries of the complete datasets included in our analysis. Sparsity represents the proportion of known DTI numbers in all possible DTI combinations.

Table 1. Summary of the four benchmark datasets.

Dataset	Drugs	Targets	Interactions	Sparsity
NR	54	26	166	0.118
GPCR	223	95	1096	0.052
IC	210	204	2331	0.054
E	445	664	4256	0.014
Luo	708	1512	1923	0.002

3.2. Experimental setup

Table 2 lists the parameter settings in the experiments depending on datasets. The best parameters of ConvBLS-DTI were selected by performing a grid search. Some key parameters were set as follows: The number of the nearest known neighbors K is set to 5 for NR and 7 for others; the feature dimension r is set to 50 for a relatively small dataset NR, and 100 was an appropriate setting for GPCR, IC, and E datasets. The convolution kernel size is taken from {3–9}. The Tanh function was chosen as the activation function for every layer. A number of experiments are performed to determine the optimal classification parameters of BLS. Specifically, the shrinkage scale (sc) of the enhancement nodes plays a central role in this experiment. The parameters of all baseline methods were set based on the suggestions from the respective studies available in the literature.

Table 2. Hyper-parameters in the experiments.

Parameter	Value
K	$K \in \{1,2,3,5,7,9\}$
r	$r \in \{50,100\}$
sc	2
filter size	4
number of filters	5
enhancement nodes	$n \in [100,1000]$

3.3. Cross-validation strategy and assessment metrics

For the cross-validation experiments, there are three different experimental settings for comparison, depending on whether the drug and target involved in the test pair are training entities:

1) CV_d: Predicts the interactions between testing drugs and training targets;
2) CV_t: Predicts the interactions between training drugs and testing targets;
3) CV_{dt}: Predicts the interactions between testing drugs and testing targets.

The 10-fold cross-validation is one of the most widely available methods. All models were trained and tested using 10-fold cross-validation. In this study, the final results are given with the area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (AUPR) to judge the prediction performance. They are widely used in this field [39,40]. Since there are few true DTIs, AUPR is a more precise quality indicator than AUC because it punished those in which lots of false positive examples were found from the top-ranked prediction score [41], so we consider it as an evaluation sign. In addition, the Sen score was another metric used in this study. The average values are used for the results of each dataset.

4. Results

Experiments were run under the environment of Windows 10 Professional Edition and i5-7200H CPU. Our aim of this study was to construct an efficient computation method with excellent performance for DTI prediction. Therefore, we first observe the performance of two BLS-based models from different perspectives on the four datasets. Then, we compared the prediction results of our model with representative methods under three settings: NRLMF [32], DTINet [11], WKNNIR [42],

DTi2Vec [43], ADA-GRMFC [44], and BLS-DTI. Finally, the optimal of core parameter in the experiment was reported.

4.1. Comparisons of BLS-based methods

We first compared our model with the BLS-DTI. Tables 3 and 4 list the AUC and AUPR results on the prediction tasks. As shown in Table 3, our model is found to outperform BLS-DTI in the AUC and AUPR. It highlights the importance of the feature extraction ability in the BLS. The enhancement layer is included in the two networks. We considered the prediction performance of BLS is insufficient for DTI prediction tasks due to the lack of the ability to obtain deep features. ConvBLS-DTI provides the better performance in terms of CNN method. Specifically, ConvBLS-DTI exhibits higher results than BLS-DTI for the E dataset, providing 0.22 higher AUC score, an improvement of 28%, with a 0.152 greater AUPR value, an improvement of 19%. The same positive results are also found in the other three datasets, which indicates that the performance of the model ConvBLS-DTI is improved when the CNN method is added to the BLS network.

Table 3. AUC and AUPR of BLS-DTI and ConvBLS-DTI.

Method	E		IC		GPCR		NR	
	AUC	AUPR	AUC	AUPR	AUC	AUPR	AUC	AUPR
BLS-DTI	0.789	0.814	0.907	0.916	0.854	0.863	0.854	0.863
ConvBLS-DTI	**0.969**	**0.962**	**0.971**	**0.967**	**0.968**	**0.961**	**0.968**	**0.961**

For a more comprehensive evaluation, the following Table 4 shows the addition AUC and AUPR scores of each experimental setting on four datasets. Similarly, ConvBLS-DTI obtain higher performance in all scenarios, outperforming the other method BLS-DTI. Compared with NR and GPCR datasets, IC and E datasets contribute to higher AUC and AUPR scores, with AUPR values of 0.947 and 0.961, respectively. The possible reason is that the number of DTIs in the NR and GPCR categories is smaller than the other categories, especially for NR with only 166 drug-target pairs.

Table 4. Performances on three prediction experimental settings.

Dataset	Method	CV_d			CV_t			CV_{dt}		
		AUC	AUPR	SEN	AUC	AUPR	SEN	AUC	AUPR	SEN
NR	BLS-DTI	0.8882	0.8753	0.8510	0.8048	0.7942	0.8088	0.9274	0.9182	0.9403
	ConvBLS-DTI	0.9509	0.9531	0.9186	0.9693	0.9688	0.9313	0.9545	0.95099	0.9153
IC	BLS-DTI	0.8830	0.8723	0.6141	0.9572	0.9547	0.7505	0.6981	0.6461	0.9160
	ConvBLS-DTI	0.9747	0.9770	0.9496	0.9719	0.9732	0.9390	0.9541	0.9654	0.9377
GPCR	BLS-DTI	0.9077	0.8950	0.6758	0.9134	0.8980	0.6570	0.7262	0.6837	0.8750
	ConvBLS-DTI	0.9557	0.9632	0.9317	0.9242	0.9460	0.9277	0.8926	0.9170	0.9000
E	BLS-DTI	0.8553	0.8182	0.7191	0.8689	0.8527	0.6807	0.8874	0.8675	0.6444
	ConvBLS-DTI	0.9614	0.9677	0.9314	0.9588	0.9691	0.9413	0.9643	0.9681	0.9330

The green part is the best performance in comparison models.

4.2. Comparisons with representative models

In this section, under the same datasets, evaluation metrics and experimental scenarios (CV_d, CV_t, and CV_{dt}), six advanced methods, including NRLMF, DTINet, WKNNIR, DTi2Vec, ADA-GRMFC, and BLS-DTI, are involved into the performance comparison. Tables 5 and 6 show the AUC and AUPR of the methods participating in the CV_d and CV_t settings. In general, based on the main evaluation metrics, our method has overall better performance than the other methods under different scenarios. For CV_d, ConvBLS-DTI shows a high performance in all datasets. For CV_t, minimal difference is found in the AUC score obtained using IC and GPCR datasets, but the AUPR result achieved by our method increases by 2.05%, 4.41%, 3.6%, 5.54%, and 6.12% on NR, GPCR, IC, E, and Luo datasets, respectively, compared with that of the second-best model. In particular, ConvBLS-DTI performs better than BLS-DTI. In the results of predicting novel drugs and known targets, ConvBLS-DTI is better than other methods. In the experiment scenario of CV_d, ConvBLS-DTI achieves AUPR values of 0.917, 0.968, 0.972, 0.958, and 0.972 on NR, IC, GPCR, E, and Luo datasets, respectively. In the experiment scenario of CV_t, the AUPR values of ConvBLS-DTI are 0.846, 0.950, 0.946, 0.952, and 0.954 on NR, IC, GPCR, E, and Luo datasets, respectively. Overall, it can be concluded that the proposed ConvBLS-DTI is superior to all the compared methods and proves that broad learning system can also be a rational tool to help for predicting DTIs.

Table 5. AUC with different methods on all datasets in CV_d and CV_t.

Setting	Dataset	NRLMF	DTINet	WKNNIR	DTi2Vec	ADA-GRMFC	BLS-DTI	ConvBLS-DTI
CV_d	NR	0.842	0.701	0.817	0.917	0.866	0.856	0.937
	IC	0.904	0.842	0.929	0.897	0.802	0.861	0.968
	GPCR	0.831	0.752	0.834	0.955	0.827	0.886	0.973
	E	0.857	0.769	0.86	0.846	0.841	0.891	0.958
	Luo	0.92	0.881	0.902	0.861	0.859	0.901	0.979
CV_t	NR	0.813	0.756	0.82	0.654	0.814	0.833	0.865
	IC	0.938	0.879	0.949	0.908	0.938	0.802	0.947
	GPCR	0.958	0.907	0.956	0.866	0.896	0.897	0.951
	E	0.943	0.841	0.927	0.853	0.939	0.862	0.960
	Luo	0.835	0.838	0.851	0.911	0.952	0.753	0.969

Indicated in blue is the best result in each category compared with other models.

Table 6. AUPR with different methods on all datasets in CV_d and CV_t.

Setting	Dataset	NRLMF	DTINet	WKNNIR	DTi2Vec	ADA-GRMFC	BLS-DTI	ConvBLS-DTI
CV_d	NR	0.532	0.346	0.571	0.912	0.607	0.857	0.917
	IC	0.514	0.47	0.529	0.911	0.39	0.882	0.968
	GPCR	0.486	0.373	0.502	0.953	0.384	0.885	0.972
	E	0.371	0.215	0.423	0.863	0.426	0.834	0.958
	Luo	0.476	0.299	0.492	0.945	0.721	0.906	0.972
CV_t	NR	0.522	0.435	0.63	0.639	0.466	0.829	0.846
	IC	0.735	0.526	0.781	0.917	0.824	0.842	0.950
	GPCR	0.803	0.574	0.858	0.875	0.631	0.906	0.946
	E	0.724	0.379	0.719	0.876	0.825	0.902	0.952
	Luo	0.303	0.138	0.571	0.899	0.878	0.804	0.954

Indicated in blue is the best result for each category comparing all other models.

In particular, the ConvBLS-DTI has satisfactory performance under the CV_{dt} setting. The AUC and AUPR histograms for the different algorithms are shown in Figures 4 and 5, respectively. The results are entirely consistent across all datasets given in the CV_{dt} (specifically in terms of AUPR metric). For CV_{dt}, the AUC and AUPR values of ConvBLS-DTI are higher than other methods on all datasets, although DTi2Vec is very competitive compared with ConvBLS-DTI. Overall, our method improves the AUPR more than the AUC.

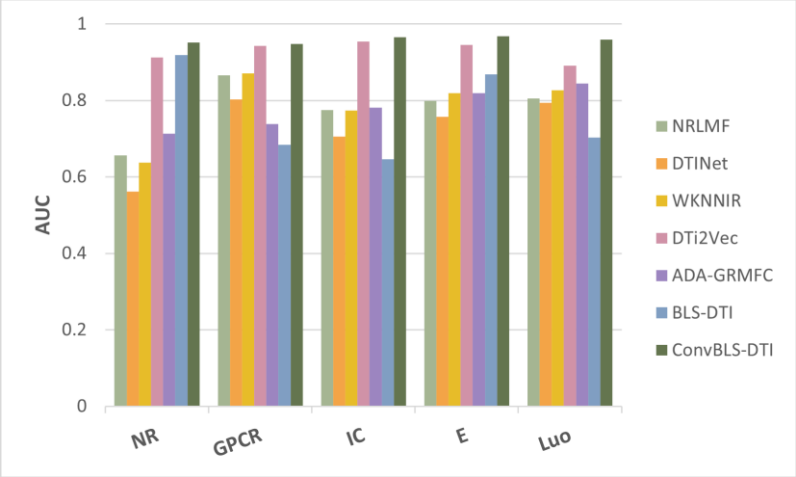

Figure 4. Comparison results of the AUC metric for the CV_{dt}.

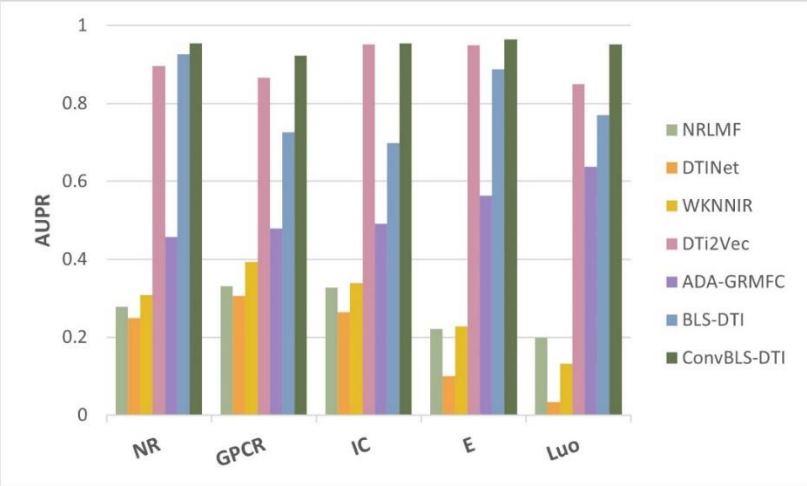

Figure 5. Comparison results of metric AUPR under the CV_{dt}.

4.3. Ablation experiment

To measure the impact of the WKNKN method on ConvBLS-DTI, ablation experiments were conducted by removing the WKNKN method under three different CV strategies using the Luo et al. dataset. The variant of ConvBLS-DTI without the WKNKN method is denoted as ConvBLS-DTI (without WKNKN). Performance comparisons between ConvBLS-DTI and the variant in terms of AUC and AUPR are presented in Tables 7 and 8, respectively. The findings in Tables 7 and 8 suggest that the utilization of the WKNKN method contributes to improve the performance of ConvBLS-DTI.

Table 7. Ablation results in terms of AUC on the Luo et al. dataset under three different CVs.

Model	CV_d	CV_t	CV_{dt}
ConvBLS-DTI	0.9785	0.9691	0.9590
ConvBLS-DTI (without WKNKN)	0.9758	0.9613	0.9516

Table 8. Ablation results in terms of AUPR on the Luo et al. dataset under three different CVs.

Model	CV_d	CV_t	CV_{dt}
ConvBLS-DTI	0.9718	0.9535	0.9522
ConvBLS-DTI (without WKNKN)	0.9636	0.9463	0.9478

4.4. Optimization of model parameters

In this study, the datasets IC and GPCR were applied to test the influence of the convolution kernel size. As illustrated in Figure 6, by varying the size of the convolutional kernel (3, 4, 5, 6, 7, 8, and 9), the AUPR value of the ConvBLS-DTI method progressively improved with an increase in kernel size and reached its optimal performance when kernel size was set to 5. Subsequently, the performance showed a decline. A kernel size set to 5 achieved good results.

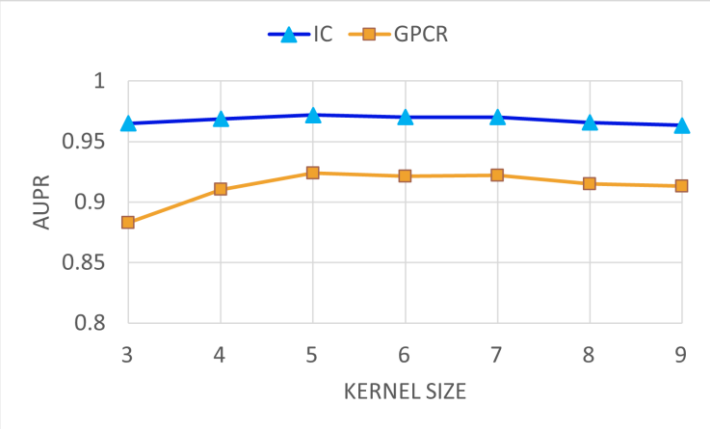

Figure 6. Comparison results of metric AUPR under the CV_{dt}.

5. Conclusions

In this paper, we aimed to solve the problem of sparsity and incompletion of the drug interaction data. A new framework called ConvBLS-DTI was proposed to predict DTIs by applying an advanced fusion of BLS approach. Our method integrates WKNKN, MF, and BLS to improve the DTI prediction results. The method takes advantage of matrix factorization for the latent low-dimensional feature representation and predicts DTIs based on the broad learning architecture. Moreover, the WKNKN algorithm was used as a preprocessing step to increase the availability of relevant information for a large number of missing correlations. Compared with the BLS-DTI, our model achieved AUC and AUPR values of 0.971 and 0.967, respectively, for the IC dataset under tenfold-cross-validation experiments. These findings illustrate that the combination of CNN and BLS could improve the prediction performance for DTIs. Additionally, compared with other previous methods， the best AUC and AUPR values of the proposed method were 0.9643 and 0.9681 for the E dataset and CV_{dt} setting, respectively. The results show that our model acquires improved prediction effect on AUC and AUPR using extensive experimental verification.

In future studies, greater emphasis will be placed on optimizing the BLS structure to enhance the feature extraction ability. In fact, the results of the present prediction model can be heavily influenced by the mapping algorithms and the effectiveness of the dataset. Therefore, our models might be further developed using other deep-learning models to increase the identifying power. Overall, with the availability of more data and the development of new approaches, it is expected that more applications of our model can be achieved.

Use of AI tools declaration

The authors that declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the project of the Natural Science Foundation of Shandong Province,

China (Natural Science Foundation of Shandong Province, No. ZR2019PEE018), Shandong Province Science and Technology SMES Innovation Ability Enhancement Project (Natural Science Foundation of Shandong Province, No. 2021TSGC1063), Major Scientific and Technological Innovation Projects of Shandong Province (Natural Science Foundation of Shandong Province, No. 2019JZZY020101), and the project of the Natural Science Foundation of Qingdao (No. 23-2-1-216-zyyd-jch).

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. X. Lin, S. Xu, X. Liu, X. Zhang, J. Hu, Detecting drug-target interactions with feature similarity fusion and molecular graphs, *Biology (Basel)*, **11** (2022), 967. https://doi.org/10.3390/biology11070967
2. N. R. C. Monteiro, B. Ribeiro, J. P. Arrais, Drug-target interaction prediction: End-to-end deep learning approach, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **18** (2021), 2364–2374. https://doi.org/10.1109/TCBB.2020.2977335
3. R. Chen, X. Liu, S. Jin, J. Lin, J. Liu, Machine learning for drug-target interaction prediction, *Molecules*, **23** (2018), 2208. https://doi.org/10.3390/molecules23092208
4. J. P. Hughes, S. Rees, S. B. Kalindjian, K. L. Philpott, Principles of early drug discovery, *Br. J. Pharmacol.*, **162** (2011), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x
5. M. Rudrapal, D. Chetia, Virtual screening, molecular docking and QSAR studies in drug discovery and development programme, *J. Drug Deliv. Sci. Ther.*, **10** (2020), 225–233. https://doi.org/10.22270/jddt.v10i4.4218
6. Q. Ye, C. Y. Hsieh, Z. Yang, Y. Kang, J. Chen, D. Cao, et al., A unified drug-target interaction prediction framework based on knowledge graph and recommendation system, *Nat. Commun.*, **12** (2021), 6775. https://doi.org/10.1038/s41467-021-27137-3
7. S. Luukkonen, H. W. van den Maagdenberg, M. T. M. Emmerich, G. J. P. van Westen, Artificial intelligence in multi-objective drug design, *Curr. Opin. Struct. Biol.*, **79** (2023), 102537. https://doi.org/10.1016/j.sbi.2023.102537
8. F. Li, Z. Zhang, J. Guan, S. Zhou, Effective drug-target interaction prediction with mutual interaction neural network, *Bioinformatics*, **38** (2022), 3582–3589. https://doi.org/10.1093/bioinformatics/btac377
9. M. Bagherian, E. Sabeti, K. Wang, M. A. Sartor, Z. Nikolovska-Coleska, K. Najarian, Machine learning approaches and databases for prediction of drug-target interaction: A survey paper, *Brief Bioinf.*, **22** (2021), 247–269. https://doi.org/10.1093/bib/bbz157
10. J. P. Mei, C. K. Kwoh, P. Yang, X. L. Li, J. Zheng, Drug-target interaction prediction by learning from local information and neighbors, *Bioinformatics*, **29** (2013), 238–245. https://doi.org/10.1093/bioinformatics/bts670
11. Y. Luo, X. Zhao, J. Zhou, J. Yang, Y. Zhang, W. Kuang, et al., A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information, *Nat. Commun.*, **8** (2017), 573. https://doi.org/10.1038/s41467-017-00680-8

12. A. Ezzat, P. Zhao, M. Wu, X. L. Li, C. K. Kwoh, Drug-target interaction prediction with graph regularized matrix factorization, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **14** (2017), 646–656. https://doi.org/10.1109/TCBB.2016.2530062
13. N. Zong, H. Kim, V. Ngo, O. Harismendy, Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations, *Bioinformatics*, **33** (2017), 2337–2344. https://doi.org/10.1093/bioinformatics/btx160
14. M. Wen, Z. Zhang, S. Niu, H. Sha, R. Yang, Y. Yun, et al., Deep-learning-based drug-target interaction prediction, *J. Proteome Res.*, **16** (2017), 1401–1409. https://doi.org/10.1021/acs.jproteome.6b00618
15. Y. B. Wang, Z. H. You, S. Yang, H. C. Yi, Z. H. Chen, K. Zheng, A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network, *BMC Med. Inf. Decis. Mak.*, **20** (2020), 49. https://doi.org/10.1186/s12911-020-1052-0
16. H. Öztürk, A. Özgür, E. Ozkirimli, DeepDTA: Deep drug-target binding affinity prediction, *Bioinformatics*, **34** (2018), i821–i829. https://doi.org/10.1093/bioinformatics/bty593
17. C. Chen, H. Shi, Z. Jiang, A. Salhi, R. Chen, X. Cui, et al., DNN-DTIs: Improved drug-target interactions prediction using XGBoost feature selection and deep neural network, *Comput. Biol. Med.*, **136** (2021), 104676. https://doi.org/10.1016/j.compbiomed.2021.104676
18. Q. Zhao, H. Zhao, K. Zheng, J. Wang, HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism, *Bioinformatics*, **38** (2022), 655–662. https://doi.org/10.1093/bioinformatics/btab715
19. S. Zheng, Y. Li, S. Chen, J. Xu, Y. Yang, Predicting drug–protein interaction using quasi-visual questionanswering system, *Nat. Mach. Intell.*, **2** (2020), 134–140. https://doi.org/10.1038/s42256-020-0152-y
20. T. Zhao, Y. Hu, L. R. Valsdottir, T. Zang, J. Peng, Identifying drug-target interactions based on graph convolutional network and deep neural network, *Brief Bioinf.*, **22** (2021), 2141–2150. https://doi.org/10.1093/bib/bbaa044
21. M. Tsubaki, K. Tomii, J. Sese, Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences, *Bioinformatics*, **35** (2019), 309–318. https://doi.org/10.1093/bioinformatics/bty535
22. J. You, R. D. McLeod, P. Hu, Predicting drug-target interaction network using deep learning model, *Comput. Biol. Chem.*, **80** (2019), 90–101. https://doi.org/10.1016/j.compbiolchem.2019.03.016
23. M. A. Thafar, R. S. Olayan, S. Albaradei, V. B. Bajic, T. Gojobori, M. Essack, et al., DTi2Vec: Drug-target interaction prediction using network embedding and ensemble learning, *J. Cheminf.*, **13** (2021), 71. https://doi.org/10.1186/s13321-021-00552-w
24. K. Huang, C. Xiao, L. M. Glass, J. Sun, MolTrans: Molecular interaction transformer for drug-target interaction prediction, *Bioinformatics*, **37** (2021), 830–836. https://doi.org/10.1093/bioinformatics/btaa880
25. J. Peng, J. Li, X. Shang, A learning-based method for drug-target interaction prediction based on feature representation learning and deep neural network, *BMC Bioinf.*, **21** (2020), 394. https://doi.org/10.1186/s12859-020-03677-1
26. C. L. P. Chen, Z. L. Liu, Broad learning system: An effective and efficient incremental learning system without the need for deep architecture, *EEE Trans. Neural Netw. Learn. Syst.,* **29** (2018), 10–24. https://doi.org/10.1109/Tnnls.2017.2716952

27. Y. H. Pao, Y. Takefuji, Functional-link net computing: theory, system architecture, and functionalities, *Computer*, **25** (1992), 76–79. https://doi.org/10.1109/2.144401
28. B. Igelnik, Y. H. Pao, Stochastic choice of basis functions in adaptive function approximation and the functional-link net, *IEEE Trans. Neural Netw*, **6** (1995), 1320–1329. https://doi.org/10.1109/72.471375
29. X. Gong, T. Zhang, C. L. P. Chen, Z. Liu, Research review for broad learning system: Algorithms, theory, and applications, *IEEE Trans. Cybern.*, **52** (2022), 8922–8950. https://doi.org/10.1109/TCYB.2021.3061094
30. X. N. Fan, S. W. Zhang, LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier, *Neurocomputing*, **370** (2019), 88–93. https://doi.org/https://doi.org/10.1016/j.neucom.2019.08.084
31. K. Zheng, Z. H. You, L. Wang, Y. R. Li, H. J. Jiang, MISSIM: Improved miRNA-disease association prediction model based on chaos game representation and broad learning system, *Intell. Comput. Methodol.*, **11645** (2019), 392–398. https://doi.org/10.1007/978-3-030-26766-7_36
32. Y. Liu, M. Wu, C. Miao, P. Zhao, X. L. Li, Neighborhood regularized logistic matrix factorization for drug-target interaction prediction, *PLOS Comput. Biol.*, **12** (2016), e1004760. https://doi.org/10.1371/journal.pcbi.1004760
33. A. E. Hoerl, R. W. Kennard, Ridge regression: Applications to nonorthogonal problems, *Technometrics*, **12** (2000), 55–67. https://doi.org/10.1080/00401706.1970.10488635
34. Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, M. Kanehisa, Prediction of drug-target interaction networks from the integration of chemical and genomic spaces, *Bioinformatics*, **24** (2008), 232–240. https://doi.org/10.1093/bioinformatics/btn162
35. B. Liu, D. Papadopoulos, F. D. Malliaros, G. Tsoumakas, A. N. Papadopoulos, Multiple similarity drug-target interaction prediction with random walks and matrix factorization, *Brief Bioinf.*, **23** (2022), 1–10. https://doi.org/10.1093/bib/bbac353
36. M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: New perspectives on genomes, pathways, diseases and drugs, *Nucleic Acids Res.*, **45** (2017), D353–D361. https://doi.org/10.1093/nar/gkw1092
37. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, et al., DrugBank 5.0: A major update to the drugbank database for 2018, *Oxford Univ. Press*, **46** (2018), D1074–D1082. https://doi.org/10.1093/nar/gkx1037
38. D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. de Veij, E. Félix, et al., ChEMBL: Towards direct deposition of bioassay data, *Nucleic Acids Res.*, **47** (2018), D930–D940. https://doi.org/10.1093/nar/gky1075
39. Q. H. Kha, V. H. Le, T. N. K. Hung, N. T. K. Nguyen, N. Q. K. Le, Development and validation of an explainable machine learning-based prediction model for drug-food interactions from chemical structures, *Sensors*, **23** (2023), 3962. https://doi.org/10.3390/s23083962
40. N. Q. K. Le, T. T. D. Nguyen, Y. Y. M. Ou, Identifying the molecular functions of electron transport proteins using radial basis function networks and biochemical properties, *J. Mol. Graphi. Model.*, **73** (2017), 166–178. https://doi.org/10.1016/j.jmgm.2017.01.003
41. M. Schrynemackers, R. Küffner, P. Geurts, On protocols and measures for the validation of supervised methods for the inference of biological networks, *Front. Gene.*, **4** (2013), 262. https://doi.org/10.3389/fgene.2013.00262

42. B. Liu, K. Pliakos, C. Vens, G. Tsoumakas, Drug-target interaction prediction via an ensemble of weighted nearest neighbors with interaction recovery, *Appl. Intell.*, **52** (2022), 3705–3727. https://doi.org/10.1007/s10489-021-02495-z
43. M. A. Thafar, R. S. Olayan, S. Albaradei, V. B. Bajic, T. Gojobori, M. Essack, et al., DTi2Vec: Drug-target interaction prediction using network embedding and ensemble learning, *J. Cheminf.*, **71** (2021), 1–18. https://doi.org/10.1186/S13321-021-00552-W
44. J. Zhang, M. Xie, Graph regularized non-negative matrix factorization with prior knowledge consistency constraint for drug-target interactions prediction, *BMC Bioinf.*, **23** (2022), 1–20. https://doi.org/10.1186/s12859-022-05119-6

AIMS Press