Wang *et al.*
Health Information Science and Systems (2023) 11:42
https://doi.org/10.1007/s13755-023-00243-w

Health Information Science and Systems

RESEARCH

LDS-CNN: a deep learning framework for drug-target interactions prediction based on large-scale drug screening

Yang Wang[1], Zuxian Zhang[2], Chenghong Piao[3], Ying Huang[2], Yihan Zhang[2], Chi Zhang[5], Yu-Jing Lu[2,4*] and Dongning Liu[1*]

Abstract

Background: Drug-target interaction (DTI) is a vital drug design strategy that plays a significant role in many processes of complex diseases and cellular events. In the face of challenges such as extensive protein data and experimental costs, it is suggested to apply bioinformatics approaches to exploit potential interactions to design new targeted medications. Different data and interaction types bring difficulties to study involving incompatible and heterology formats. The analysis of drug-target interactions in a comprehensive and unified model is a significant challenge.

Method: Here, we propose a general method for predicting interactions between small-molecule drugs and protein targets, Large-scale Drug target Screening Convolutional Neural Network (LDS-CNN), which used unified encoding to achieve the calculation of the different data formats in an integrated model to realize feature abstraction and potential object prediction.

Result: On 898,412 interaction data involving 1683 small-molecule compounds and 14,350 human proteins from 8.8 billion records, the proposed method achieved an area under the curve (AUC) of 0.96, an area under the precision-recall curve (AUPRC) of 0.95, and an accuracy of 90.13%. The experimental results illustrated that the proposed method attained high accuracy on the test set, indicating its high predictive ability in drug-target interaction prediction. LDS-CNN is effective for the prediction of large-scale datasets and datasets composed of data with different formats.

Conclusion: In this study, we propose a DTI prediction method to solve the problems of unified encoding of large-scale data in multiple formats. It provides a feasible way to efficiently abstract the features among different types of drug-related data, thus reducing experimental costs and time consumption. The proposed method can be used to identify potential drug targets and candidates for the treatment of complex diseases. This work provides a reference for DTI to process large-scale data and different formats with deep learning methods and provides certain suggestions for future research.

Keywords: Drug-target interaction prediction, Convolutional neural networks, United encoding, Large scale prediction

Introduction

Approved drugs are important research content for new drug discovery. Drug development based on approved drugs does not require consideration of the safety and efficacy of the original drug, effectively reducing the time and cost of the drug development process [1]. Inferring potential regulatory pathways based on drug-target

*Correspondence: luyj@gdut.edu.cn; liudn@gdut.edu.cn
[1] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China
[2] School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou 510006, China
Full list of author information is available at the end of the article

interactions is of great significance for drug design [2, 3]. Small molecule drugs can affect protein targets through physical binding, these proteins can interact with other targets and form a vast interaction network together [4–6]. The known small molecule drugs can be associated with more targets based on the type and property of these targets, and hence it is important to use the available data to explore potential regulatory relationships. It is necessary to search for compounds that can interact with target proteins in the vast network when developing new drugs. Traditional pharmaceutical methods in the field of drug-target interaction (DTI) prediction are time-consuming and laborious, such as Gaussian methods [7] and density functional theory (DFT) methods [8]. Companies and researchers are in urgent need of efficient computational methods to inform and advise traditional pharmaceuticals [9, 10].

In recent years, as the field of DTI requires efficient computational methods to improve efficiency, virtual drug screening on computers has become an increasingly popular research topic. Researchers have attempted to obtain solutions using molecular docking techniques, which is a well-established virtual drug screening method that can visualize small molecule-protein interactions and provide reliable docking conformations [11]. However, the efficiency of these methods is significantly limited due to the one-to-one speed of each docking process [12]. Several studies pointed out that the accuracy of docking methods still needs to be improved, and the accuracy of docking methods needs to be higher [13, 14]. Furthermore, docking methods predict binding sites by calculating the 3-D structure of drugs and proteins, which requires specific structural details [15]. However, there still remain a lot of proteins without structure details. And due to clinical trials being costly and cannot be tested on a large scale, traditional methods cannot effectively utilize interaction network data [16]. In view of the above problems, many researchers predict DTI using machine learning (ML) methods [17, 18], which allow the screening of potential drug-target combinations by simple model work. These methods help overcome the disadvantages of traditional drug discovery methods, such as high cost, low success rate, and long study time, dramatically reducing the cost of drug development.

To address these issues, we proposed the large-scale drug target screening convolutional neural network (LDS-CNN), a novel method for predicting prospective drug-target interactions using a convolutional neural network (CNN) with unified probability encoding. This method enables data compatibility by unifying the SMILES format for small molecule drugs and amino acid format for proteins. The LDS-CNN model uses a one-dimensional CNN to extract features and predicts drug or target data by uniform probability encoding. With stacked convolutional layers and pooling layers, it can extract and downscale features from 1000-length sequences, and eventually link fully connected layers to learn hierarchical representations of the data and output classification results. The global maximum pooling layer is adopted to extract the most significant features, while the reshaping layer is used to adjust the data shape. The final output layer uses a sigmoid activation function for dichotomous prediction. In further, sufficient data can help the LDS-CNN model provide more reliable conclusions over a larger range of data. A series of experiments were conducted to validate the stability and overall performance of this encoding. This work attempted to provide an efficient DTI identification method and reduce the experimental time and material cost on accomplishing the DTI tasks. It is anticipated to be used as a reference for deep learning research on large-scale data in the field of DTI. The contributions of this work include:

1. We propose a new encoding method for DTI study through combining the general sequence feature of drug, protein and gene data. This encoding allows a unified analysis for different data types, the results suggest that the proper encoding could bring new insight into identifying of potential drug target.
2. We enhance the performance of deep learning model by analyzing the quality of dataset, which involves about 1 million drug-related interactions from over 8 billion database records. The chemical space of compounds in the dataset are characterized as molecular weight and lipid-water partition coefficient, indicating the compounds holds a wide chemical space to allow a broader chemical exploration space.
3. We illustrate the effectiveness of deep learning method using unified encoding. The convolutional neural network is designed and optimized based on the classic linear-convolution architecture, that can effectively process DTI Big data while ensuring calculation efficiency and accuracy. It is anticipated to be a useful tool to identify potential drug targets in DTI research.
4. We identify several potential drug targets and validate these predictions by utilizing the AutoDock program and DS visualizer software. The molecular simulations show theoretical interactions between the drugs and their targets, suggesting further investigations on these predictions.

Related works

Traditional machine learning methods typically utilize small-scale drug-target interaction data for prediction [19, 20]. Bleakley et al. [21] utilized a support vector

machine (SVM) framework to predict DTIs. This framework applied the method known as bipartite local models (BLM), which first predicted the target protein of a given drug and then predicts the drug against the given protein. BLM method may not make correct predictions when those new drug candidates involved, and to solve this problem, Mei et al. [22] complemented the BLM with neighbor-based interaction profile inference (NII). Buza et al. [23] predicted DTIs using the BLM and the hub-aware regression technique ECkNN. In addition, Cheng et al. [24] predicted drug-drug interactions through decision trees, straightforward Bayesian, and other machine learning-based methods, and Bull et al. [25] utilized random forests to measure the similarity of non-targeted drug targets to develop drug development programs. Zhou et al. [26] also employed machine learning-based techniques, primarily dichotomous local models, matrix decomposition, and regularized least squares, to enhance the DTI prediction process. These works provided multiple solutions, but machine learning methods still remain difficulties to achieve efficient and accurate large-scale drug screening.

To improve large-scale screening accuracy, the researchers introduced a big data-driven deep learning approach to facilitate model extract more valuable features from large samples, and to ensure accurate prediction of unknown data when large samples are present. The high compatibility and reliable predictive capability of deep learning in dealing with big data provide many solutions to solve DTI domain problems [27–29]. These methods are very applicable to large-scale data and help research the most suitable candidate drug molecules, which can reduce experimental time compared to molecular docking [30]. Deep learning method is an end-to-end method for directly extracting features from protein and drug sequences and predicting the binding affinity of drug-protein interactions [31–33]. These methods are less dependent on specific data, usually using structural information, and have fast computational speed. In contrast to traditional machine learning methods which require expertise [34, 35], deep learning methods have been applied on automatic tasks such as interaction network inference and drug design to analyze large-scale and complex interaction relationships. For example, the deep learning methods are used to select candidate drugs based on data characteristics in the database and predict potential interactions between proteins and targets [36, 37].

In these methods, encoding of protein/drug sequences is a crucial step before designing models. The one-hot encoding method, as commonly used in various works, could represent molecules in amino acid sequences and drugs but this encoding may dilute the features due to the feature vector. Moreover, one-hot encoding may lead to a large feature vector that would be too sparse according to the task [38]. It is necessary to improve traditional encoding methods or design a novel efficient encoding strategy for adapting to the characteristics of DTI task, and further improve the performance of current analysis methods.

Furthermore, there are many effective methods working on DTI prediction were developed. For example, Huang et al. proposed a molecular interaction transducer (MolTrans) to improve DTI prediction performance [39]. Chu et al. developed a new DTI prediction method to improve the prediction performance of a cascade deep forest (CDF) based model, called DTI-CDF, which attempts to uncover drugs with multiple similarity-based features as well as similarity features between target proteins extracted from heterogeneous graphs containing known DTIs [40]. Lee et al. constructed a new DTI prediction model using a CNN-based deep learning approach to extract local residue patterns of target protein sequences [41], and Bagherian et al. described the data required for the DTI prediction task, containing a comprehensive catalog of machine learning methods and databases, highlighting the possible challenges of using machine learning methods for DTI prediction [42]. These methods have solved the problem of DTI to varying degrees, but issues such as computational accuracy and overhead still need to be urgently addressed.

Materials and methods

Collection of drug-target interaction data

The following databases are utilized in this work which are commonly used in recent research: (1) the PubChem database, which contains over 160 million compounds and small molecule drugs [43]; (2) the ChEMBL database, which provides drug-protein interaction data and contains over 2 million compounds [44]; (3) the DrugBank database, which provides information on drug-target interactions as well as detailed information on the structure, indications, metabolic pathways, and other properties of drugs [45]; (4) the STITCH database, which provides information on interactions between compounds and proteins [46]; (5) the STRING database, which provides information on interactions between proteins and proteins [47]. The detailed information of dataset downloaded from this database was shown in Table 1. The approved small molecule drugs data were obtained from the DrugBank database which contained 2739 records and their corresponding CIDs were obtained from the PubChem database. The compound-protein interactions data from the STITCH database contained as many as 8,863,842,013 records. The small molecule drug-protein interactions data were obtained from the

Table 1 Details of each database

Database	Data content	Dataset size
STITCH	Available interaction records	8,863,842,013
STRING	Proteins involved	67,592,465
Uniport	Human proteins involved	14,530
DrugBank	Small molecule drugs involved	2739
PubChem	SMILES data for small molecule drugs	2739

STITCH database, in which contained 638,547 records of human protein data according to the Uniport database. Due to the various data sources of the STITCH database, including the literature as well as from multiple biological pathways, it essentially covers the main objects of current research.

Dataset preprocessing

The target small molecule drugs were obtained from the DrugBank database, these drugs were listed as SMILES format by using the PubChem API. Then, the SMILES data were selected with length limitation of 400, those drugs with lengths over 400 would be removed. To obtain interaction data about the target small molecule drugs, it is also necessary to restrict the length of the amino acid sequence. The amino acid sequences of human proteins were collected from the UniProt database. In order to ensure the overall similarity of the data length, while considering the average length distribution of the protein data, the amino acid sequence length of the selected protein is limited to 600 characters, and proteins longer than 600 will be removed. Finally, 898,412 drug-small molecule-protein interactions involving 1683 small molecule pharmaceuticals and 14,530 human proteins were utilized in experiment training and testing.

To improve the quality of experimental data, negative samples should provide sufficient features for analysis models. Therefore, to provide more effective negative features, negative samples are generated through the following strategy: For drug-protein interactions data, the interactions between unrelated drugs and proteins are clear negative samples, which are easy to classify. However, interactions between drugs and proteins related to known results are difficult to classify. Considering these similar unknown interactions, negative samples are generated through random unrelated relationships and modifications of known interactions. Due to the lack of a standard dataset to build comprehensive and complete drug-protein interaction network, all interactions appearing in the dataset are considered positive samples based on known interactions. Negative samples are generated by randomly combining proteins and drug molecules and deleting samples that appear repeatedly in positive samples.

The final positive and negative samples in the datasets were randomly sorted, and then the datasets were divided into training set, validation set, and testing set with ratio of 3:1:1. Positive and negative samples in each dataset were roughly equal to balance the classification capability of model. Dataset_train1 and Dataset_train2 are fast test datasets for checking whether the parameters of the model work properly and the model acc respectively; Dataset_train3 is the optimization test dataset for verifying the effect of the parameters at different data sizes; Dataset_4 is the complete dataset for final performance test and model training. Each dataset is randomly selected and reordered from the whole datasets to avoid the duplication and correlation among datasets. Details of each dataset are shown in Table 2.

We characterized the chemical space of compounds for the three training datasets into two dimensions which were molecular weight (MW) and lipid-water partition coefficient (AlogP). As shown in Fig. 1, the compounds in the training, validation, and test sets have a wide range of molecular weights (12.011 to 3931.48) and lipid-water partition coefficients (− 25.1791 to 18.470), indicating that the compounds in each dataset held a wide chemical space to allow a broader chemical exploration space. Also, most of the chemical space of the test set (pink part) is distributed within the area of the training set (blue part) and the validation set (yellow part), suggesting that the training datasets were suitable for extracting data features and testing datasets can be used to evaluate the prediction performance.

Table 2 The scale of each dataset

Drug-protein interaction	Training dataset	Validation dataset	Test dataset
Dataset_train1	1282	372	346
Dataset_train2	11,691	3482	3963
Dataset_train3	125,823	47,132	43,872
Dataset_4	539,046	179,683	179,683

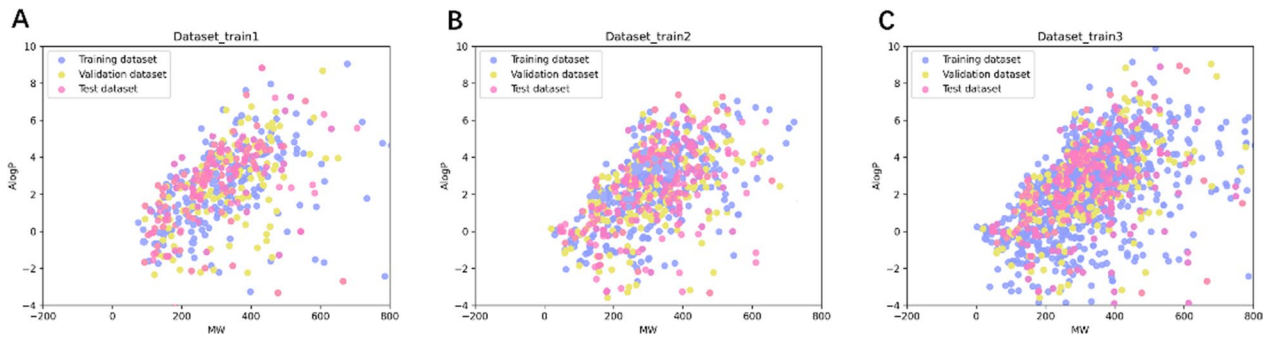

Fig. 1 Scatterplots of molecular weight (MW) and lipid-water partition coefficient (AlogP) in three datasets. **A**, **B** and **C** was the scatterplot of each dataset

Unified encoding

To ensure the applicability of the data and processing efficiency, input data were obtained with SMILES format Compared with other encoding methods (e.g., graph vectors), the SMILES format is almost applicable to current data, and thus avoided the compatibility issues of some data and of some methods. In addition, the data scale and model computation based on encoded vectors are relatively lower than 3D-molecular structure data and graph vector data, which can help more efficient task-solving. In this task, it would calculate as many as 8.8 billion records of DTI data, the computational performance should be considered first. An effective but simple data encoding was necessary.

One-hot encoding transforms protein sequences and SMILES sequences from a sequence format (size $L \times 1$, L refers to the sequence length) into a vector format (size $L \times 20$, 20 refers to the amount of amino acid types), effectively expanding the feature with characteristic information of original sequence in matrix form. For the redundancy of one-hot coding, the number of characters in SMILES is known to be 64, so only 1/64 (about 1.56%) information in one hot encoding is valid, and the remaining 98.4% information is invalid since all of these values are zero. This may lead to a particularly redundant in the calculation and significantly increase the complexity of model. In this task and other large-scale data task, redundancy of 98.44% is too sparse for model optimization and training due to the computational burden of training from large dimension of the vector.

The text data used contains not only capital letters in the amino acid sequence, but also special characters and lowercase letters in the SMILES formula. In order to distinguish similar atomic symbols (such as Carbon short for C and Calcium short for Ca) and enhance feature expression, atomic symbols that contain only one uppercase letter are extended to provide extra feature, for example, the C atom become C*. This not only achieves uniformity in terms of number of characters, but also increases the SMILES formula that is short compared to the amino acid sequence to prevent the model from overly adopting amino acid sequence features. For better feature abstraction from sequences data, all characters in the dataset were counted, and a character probability dictionary was calculated based on the occurrence of characters. This dictionary is used to provide overall character digitization and probability encoding calculation. The main steps of unified encoding process are shown in Fig. 2

The unified encoding design can provide a standard input for analysis model. In Fig. 2A, the SMILES sequences for small molecule drugs use the same probabilistic coding dictionary to encode atoms and the length of each SMILES sequence is limited to 400. In Fig. 2B, the amino acid sequences for proteins use the same probabilistic coding dictionary to encode amino acids and the length of each amino acid sequence is limited to 600. This process also normalizes the sequence length while probabilistically encoding all protein amino acid sequences. In Fig. 2C, the probability-encoded sequences of small molecule drug SMILES and protein amino acids are concatenated, and the resulting sequence is padded to 1000 characters. The sequence preserves the features of the original sequence and the features of the interaction, and the connection sites are represented by squares in gradient colors in the figure. Finally, an interaction with a fixed length of 1000 is generated. By this encoding method, the interaction data were calculated to the matrix of $L \times 1000$ for model training and testing.

Model design and parameter details

The effectiveness of the CNN model on many different tasks is based on local perception and weight sharing [48]. In DTI task, the analysis model should have the capability to abstract enough effective features from interaction data, and do not occupy too much expenses. Due to the large amount of data and

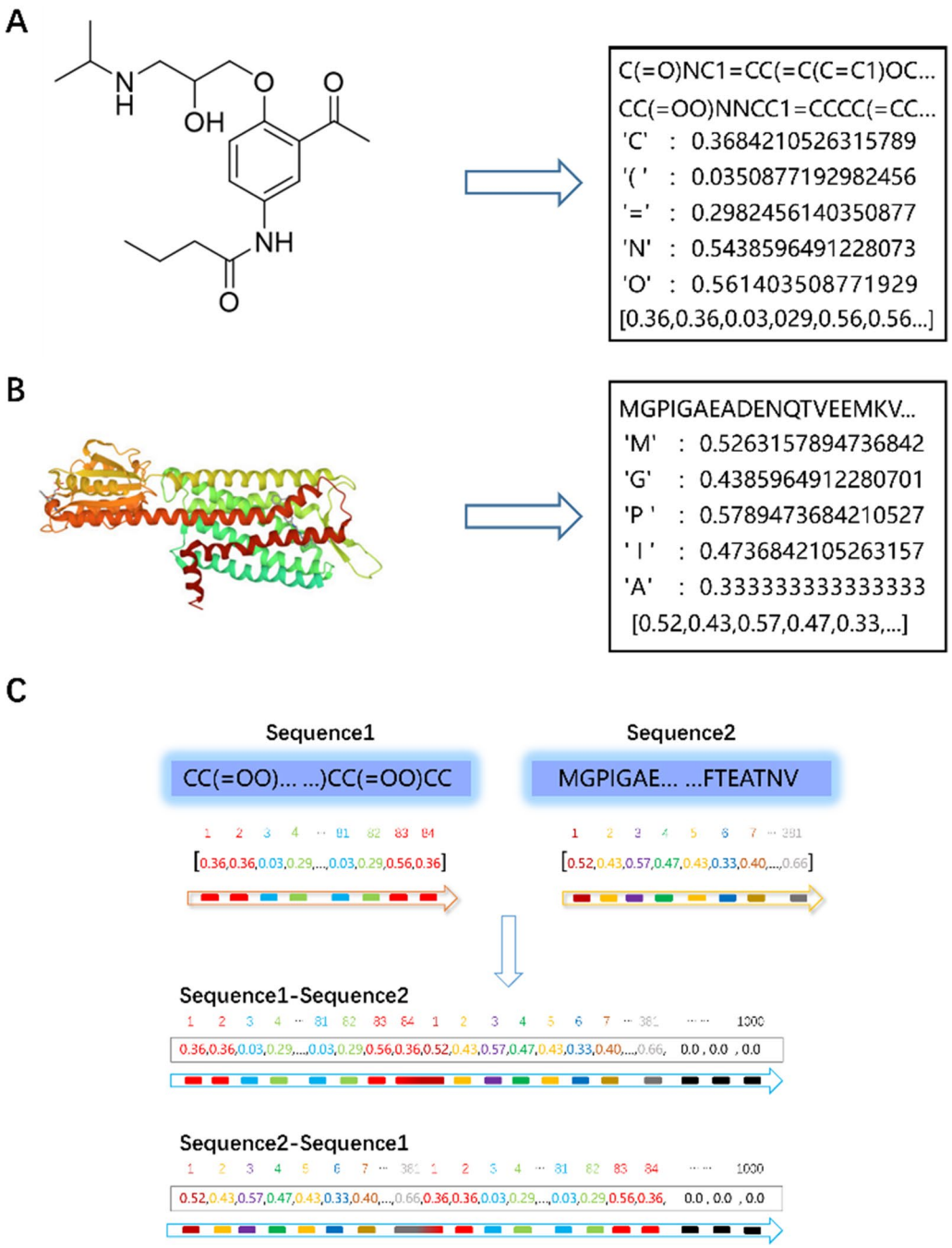

Fig. 2 The main steps of unified encoding. **A** The unified encoded sequence of drug small molecule SMILES. **B** The unified encoded sequence of protein amino acid. **C** Padding the concatenated unified encoded sequences of drug small molecule SMILES and protein amino acid sequences. The final encoding result is a vector with 1000 values

numerous interaction objectives, the SMILES data format of model input was selected for acceptable calculation pressure. The SMILES format was widely used with CNN model in various chemical molecular related tasks for its lower computational complexity and easier optimization and application. By exploiting the feature extraction capability of CNNs in local regions, the DTI screening tasks can be efficiently analyzed, which is advantageous in solving the large scale and diverse data pattern issues.

Therefore, the SMILES-encoded convolutional neural network (CNN) model was designed according to these task characteristics. The SMILES-based data format led to a framework targeted at processing sequences, rather than the graph vector graph convolutional network (GCN) models. For effective feature extraction and appropriate computation cost, a substructure of two convolutional layers and one max-pooling lay was adopted and two such substructures constitute the main structure of CNN model. The model details are shown in Fig. 3. The overall framework of this model includes: (1) The input layer (input) was used to receive the shape of the input data as input to the neural network. (2) To extract interactive features between small molecule drugs and proteins by applying a convolutional operation with uniform probability encoding, the convolutional layer (Conv1D) was arranged for four times. There were two convolutional layers that be set after input layer. (3) The output of feature map from the second convolutional layer was then abstracted by down-sampling, which the max-pooling layer (GlobalMaxPooling1D) would reduce the dimensionality of the features and retain the most significant features. (4) The output of the pooling layer was reshaped by using the *Reshape()* function from Keras to modify its shape to (Conv1D_filters, 1) for next step (Reshape). (5) The first dense layer (Dense) contained 256 neurons and Rectified Linear Unit (ReLU) was used as the activation function, for further processing the features extracted from the previous convolutional layers. (6) The second dense layer contained 2 neurons to output the two-category results with the Sigmoid activation function, for predicting the probability of potential drug-target interactions. (7) The key parameters such as loss function and optimizer were pre-set following the literature recommendations, which were binary cross-entropy and Adam in this work.

The 1-d convolutional layers were set to implement convolutional operations with size [None, 1000, 1]. The first dimension was the batch size (None was the default parameter in Keras), the second dimension of 1000 was the length of sequence encoding, and the third dimension of 1 was the dimensionality of input data. The input of model was formulated as:

$$A = \left[a_1^{(1)}, a_2^{(1)}, \ldots, a_q^{(1)}\right] \in D^{q \times d} \tag{1}$$

where D was the dataset in this work, q was the length of data, and d represented the dimensionality of the input data. In this task, q equaled 1000 and d equaled 1. The input of the first convolutional layer was as follows:

$$X = \left[x_{i:i+2,j}^{(0)}\right] \in D^{(q-2) \times m \times n} \tag{2}$$

where m represented the number of data channels, n represented the size of the convolution kernel, m equaled 1 and n equaled 3 here. The $x_{i,i+2,j}^{(0)}$ represented one data item of the ith convolution window in the jth channel. Since the valid-fill method was adopted, the output length of the convolutional layer was $q - 2 = 998$. The output of the convolutional layer was as:

$$\mathcal{C}^{(1)} = \left[c_i^{(1,k)}\right] \in D^{(q-2) \times Conv1D\ filters} \tag{3}$$

where $c_i^{(1,k)}$ represented the output of the ith convolution window in the kth convolution kernel. Each result $c_i^{(1,k)}$ was calculated by the following formula:

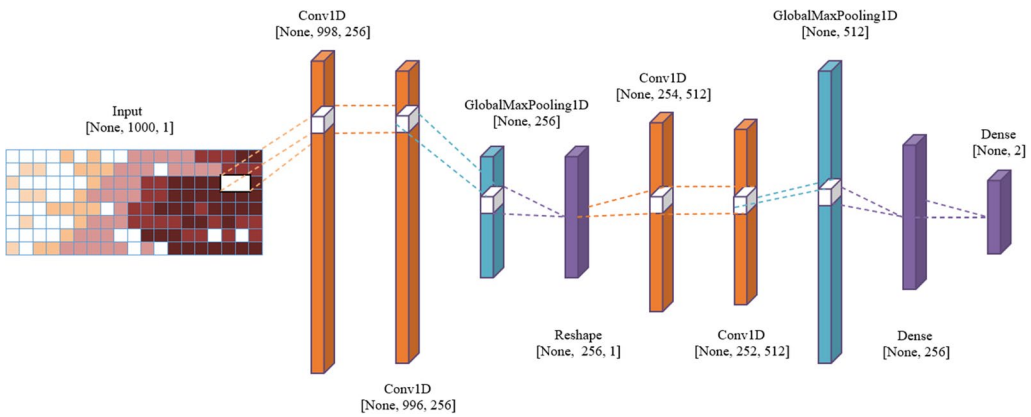

Fig. 3 Details of large-scale drug target screening convolution neural model. In LDS-CNN, there are two functional I substructures. Each substructure comprises a set of two convolutional layers and one global max pooling layer. The convolutional kernels in each substructure are the same to extract spatial information. All convolutional layers are appropriately designed to ensure that the input and output sizes of each layer are identical. The classification results are finally output from the fully connected layer

$$c_i^{(1,k)} = relu(w_k \cdot [a_i, a_{i+1}, a_{i+2}] + b_k) \quad (4)$$

where $w_k \in R^3$ was the weight vector and $b_k \in D$ was the bias term of the kth convolution kernel. The ReLU activation function was utilized in consideration of the characteristics of DTI task. The input and output of the second convolutional layer was as follows:

$$X = C^{(1)} = \left[c_i^{(1,k)}\right] \in D^{998\times 256} \quad (5)$$

$$C^{(2)} = \left[c_i^{(2,k)}\right] \in D^{996\times 256} \quad (6)$$

The global maximum pooling layer was utilized to reduce the number of computational parameters to avoid over-fitting problems. The input of this layer was the output from the second convolutional layer, and the output of the global maximum pooling layer was $P = [p_k] \in D^{256}$, where p_k equaling $max_{i=1,\ldots,996} c_i^{(k)}$ was the maximum value among all elements of the k-th channel. Then, the next following layer converted the output shape to $R = [r_i] \in D^{256\times 1}$, where r_i represented the i-th element to modify the data dimensions for the following layers. As shown in Fig. 3, the dimension after the global maximum pooling layer was [None, 256]. It needs to be modified to [None, 256, 1] for the subsequent calculation. After max-pooling calculation, the first dense layer had 256 neurons with ReLU activation function, and the second dense layer had 2 neurons with SoftMax activation function to output the prediction of binary classification. The final output of the model was a tuple $y = [y_1, y_2]$, to provide the predicted probability of both categories.

Results

Performance improvement from optimizers

In order to obtain stable results on dealing with large datasets, fivefold cross-validation tests were used to train the model. Figure 4 shows the effectiveness of each optimization algorithm including AUC (short for Area under curve), AUPRC (short for Area under precision-recall curve), accuracy, F-measure, and Sensitivity. As shown in Fig. 4, the Adam Optimizer performed better than other methods. Meanwhile, the Nadam, RMSprop, Adamax, and Adam optimizers also performed well with similar results. The Adam method can achieve an AUC of 0.9075, which was higher than the other methods of 0.8816, 0.8912, 0.8940, respectively. Similarly, the AUPRC value of Adam method was 0.8921, which was higher than the other three methods of 0.8687, 0.8687, 0.8908, respectively. The F-measure value of Adam method (0.8303) was also higher than other three methods (0.8121, 0.8292, 0.8103, respectively). Although the sensitivity of Adam method was not the highest (0.8291, lower than 0.9090 of Nadam method), due to the main objective of identification of potential DTIs, we still chose Adam rather than other three methods. There methods are all based on improved versions of the gradient descent method, with capability of adaptively adjusting the learning rate and momentum to speed up convergence and stability. They both utilize exponential moving averages of first- and second-order moments to estimate the direction and magnitude of the gradient, thus avoiding the problem of vanishing or exploding gradients. For the other methods such as Adagrad, Adadelta, and SGD, they have in common that they do not take into account the historical information of the gradient and momentum effects,

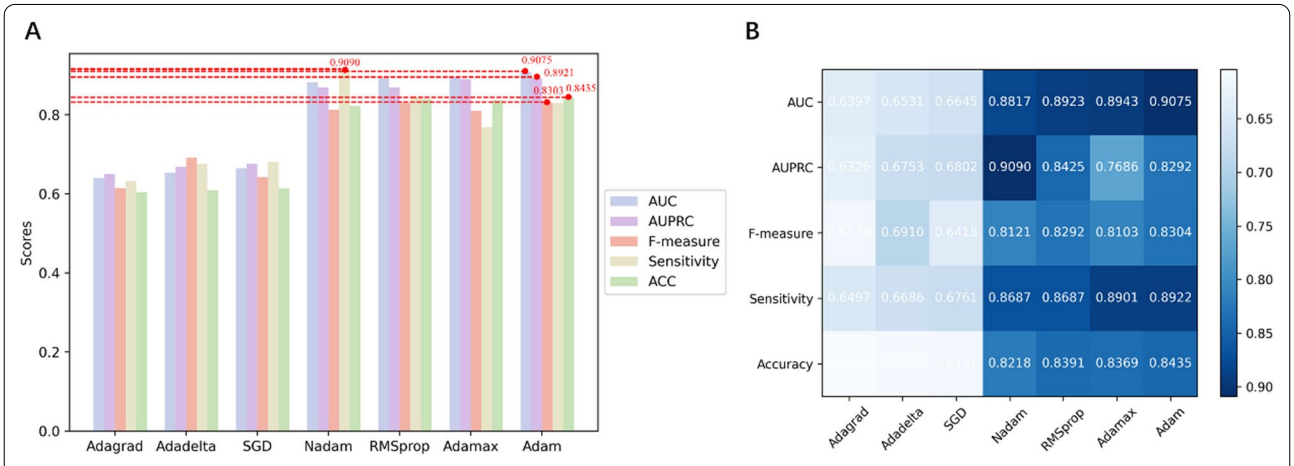

Fig. 4 Performance of different optimizers on Dataset_train3 with 200,000 drug-target interactions. **A** The score on the left side of figure is the value unit for the AUC, AUPRC, F-measure, Sensitivity, and Acc metrics on the right side. On the x-axis, the optimizers are divided into seven groups by region. **B** The heat map of each performance from with different optimizers

leading to a slow or unstable optimization process. They may be susceptible to noise or outliers, thus deviating from the direction of the optimal solution. Also, they all need to set the appropriate learning rate manually, to avoid the problem of failure to converge or overfitting. Adam combines the advantages of both Adagrad and RMSProp optimization methods and is able to adaptively adjust the learning rate and momentum to speed up the convergence and stability, so Adam is used as an optimizer for the LDS-CNN model.

Performance improvement from kernels

To extract more features, the proper number of kernels need to be tested under certain calculated pressure. When calculating protein and small molecule sequence data, the number of kernels is typically set to 32–64, in particular using Adam optimizer to train the model [49, 50]. Considering that uniform encoding is designed to provide more features, we can infer that the CNN model may extract more features by utilizing more kernels. Therefore, the number of kernels in this task is increased to 256–512, namely there are 256 kernels in the first and second convolutional layer, and 512 kernels in the third and fourth convolutional layer, respectively. The performance of LDS-CNN model using different kernel combination is shown in Fig. 5.

As shown is Fig. 5, when adopted the most kernels (256/512), the performance of each index was higher than other group, namely, AUC was 0.9175 while other three group were 0.7902 (32/64), 0.8370 (64/128), 0.8764 (128/256), respectively. The AUPRC and F-measure value were also similar to this conclusion: the AUPRC of group 256/512 was 0.9021, higher than other three methods of 0.8044, 0.8471, 0.8621; the F-measure of group 256/512 was 0.8503, higher than other three methods of 0.7240, 0.7810, 0.8105. The Acc and sensitivity values coincided with this conclusion.

The result showed with the increase of the number of convolutional kernels, the performance of LDS-CNN model also improved with a significant trend. The performance growth from 128/256 kernels to 256/512 kernels slowed down, indicating the 128/256 kernels could extract most features and there was a limitation on improvement from kernel increase. Moreover, the time cost of setting more kernels will also increase exponentially. When 128/256 kernels were set in the model, the calculation efficiency is about 35 ms per interaction. Adding more kernels would further bring extra calculation cost. The calculation efficiency declined sharply to 200 ms per interaction when set 256/512 kernels in model. Although the increasing number of kernels brought better prediction result, but calculation efficiency was also an important factor to be considered.

In Fig. 6A, it was showed that when training model with the same other parameters, more kernels brought better Acc and smaller loss value. In Fig. 6B, it showed that the performance on validation dataset also coincident with this conclusion. Figure 6C represented the model performance of different kernels. The AUC value indicated that more kernels bring better overall classification capability (from 0.9536 to 0.9621 while kernels increase from 128/256 to 256/512), the AUPRC and sensitivity values were also the similar trend. However, the f-measure value of 128/256 kernels (0.9092) was higher than 256/512 kernels (0.9058), and the sensitivity values differed from this situation, which indicating the classification capability on predicting the negative sample may be affected. This might be a little disturbance caused by

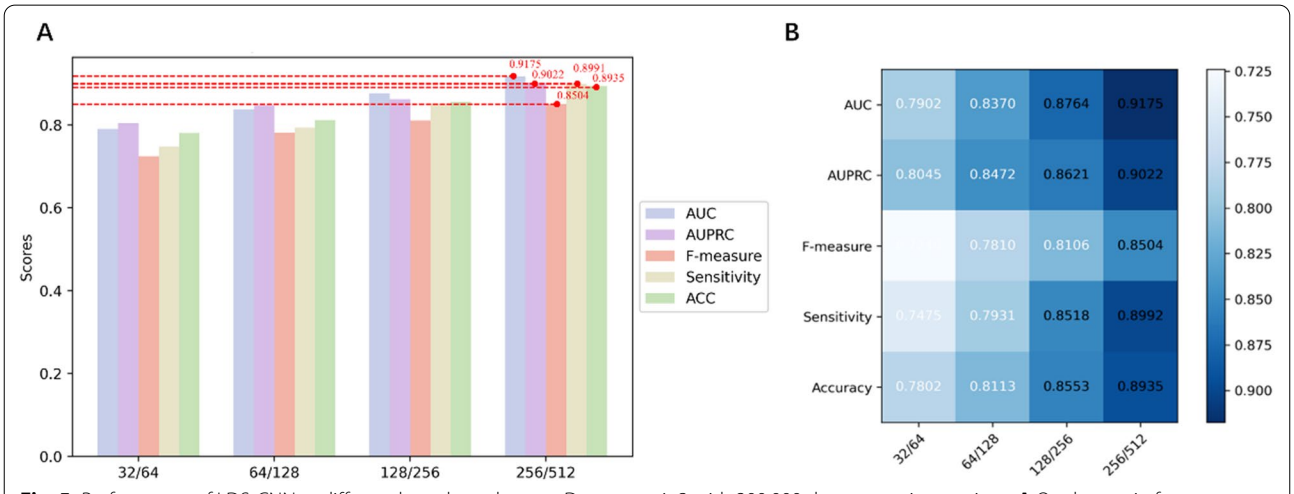

Fig. 5 Performance of LDS-CNN at different kernel numbers on Dataset_train3 with 200,000 drug-target interactions. **A** On the x-axis, four group are arranged according to different number of kernels, respectively. **B** The heat map of each performance in four groups

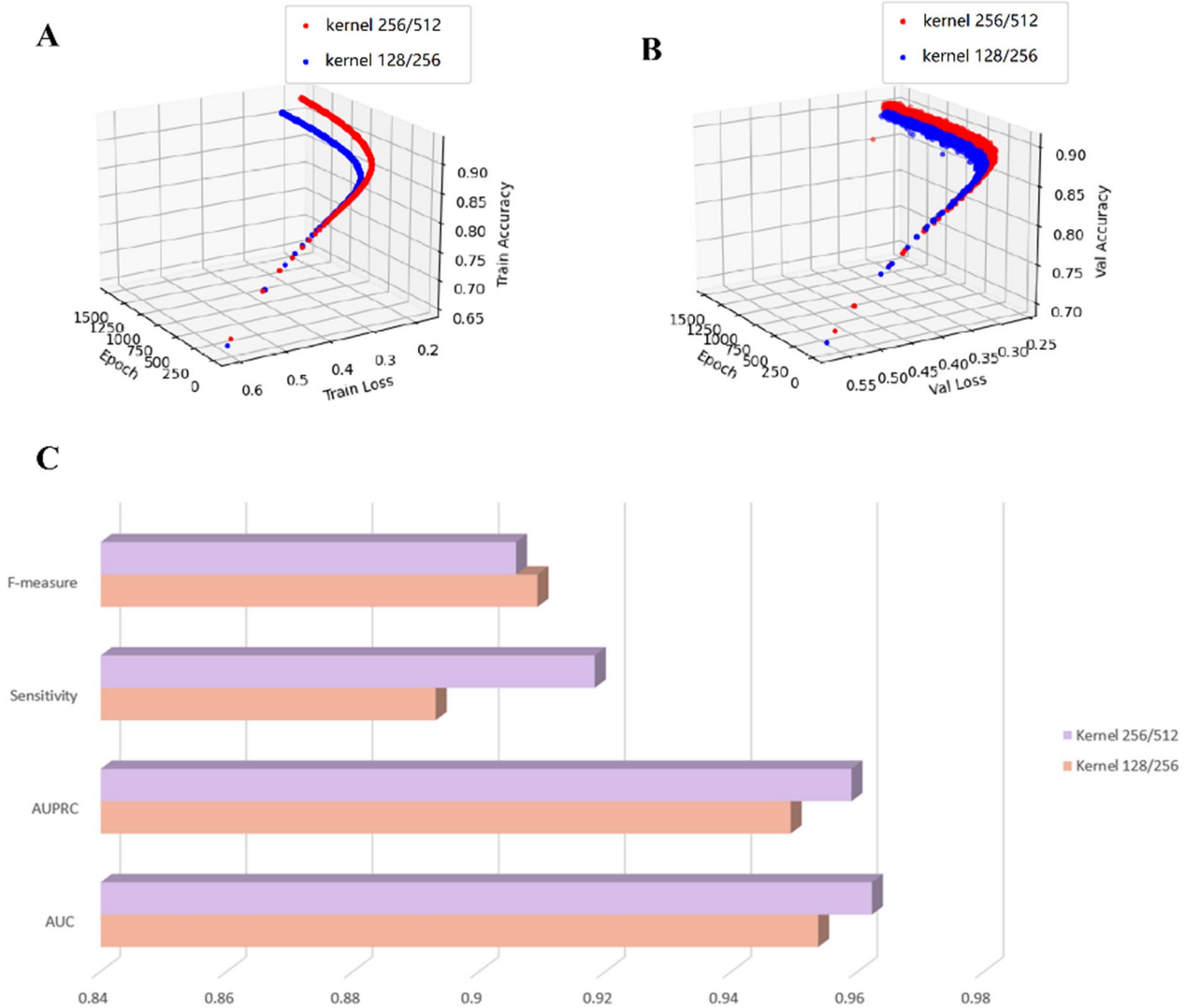

Fig. 6 Model evaluation metrics for kernel 128/256 and 256/512. **A** Performance curves of training loss and Acc value. **B** Performance curves of validation loss and Acc value. **C** Comparison of AUC, AUPRC, Sensitivity, and F-measure metrics on two different kernel group

errors, but due to the identification of potential DTIs in negative sample of being the main objective in DTI task, using 128/256 kernels may generate more satisfactory results. Further performance improvement can be considered by introducing more additional information to improve this situation. Finally, the all 800,000 samples were used to train the model, and the proposed method can achieve an accuracy of 90.13% on validation dataset. As a performance reference for the model.

Independent validation tests were conducted by downloaded three datasets from the ChEMBL database: GLP1R dataset, SHBG dataset, and IMDH1 dataset. The GLP1R dataset contained 3, 252 records, the SHBG dataset contained 204 records, and the IMDH1 dataset contained 64 records. ROC results of each dataset were shown in Fig. 7.

As shown in Fig. 7, the LDS-CNN model achieved a ROC of 0.91 on the GLP1R dataset, 0.88 on the SHBG dataset, and 0.76 on the IMDH1 dataset. The performance of LDS-CNN on the IMDH1 dataset was lower than average performance of model and other datasets, which may be mainly due to insufficient samples in the drug-protein interaction dataset.

The structural diversity of molecules in the modeling dataset and proper chemical space can help improve accurate and robust prediction models. More backbone structures indicated that the dataset encompassed a diverse chemical space which enhanced the screening

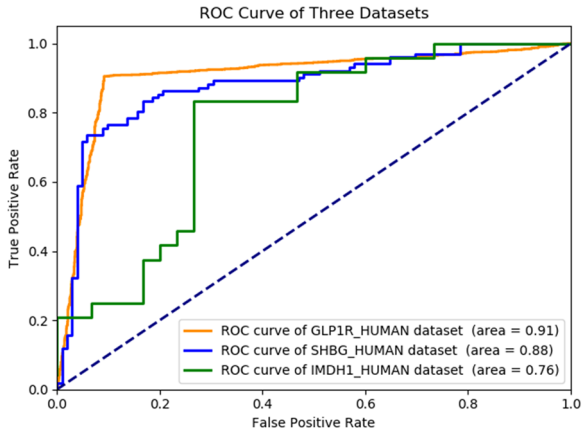

Fig. 7 External validation cohort evaluation model ROC diagram

Table 3 The details of three validation datasets

	ID	Compounds	Skeletons	Ratio (%)
GLP1R	P43220	1605	895	0.55
SHBG	P04278	102	40	0.39
IMDH1	P20839	24	10	0.41

accuracy of the model. Therefore, the rdkit tool was used to extract and identify molecules with defined Bemis-Murcko Scaffold (BMS) backbones for molecular backbone analysis. Commonly the Bemis-Murcko structure was one-to-one with the molecular structure of the compound. It can be generated by removing the side chain and identifying the ring structure connected to the linkage structure. Smaller ratio between skeletons of the compounds would bring higher diversity of the dataset. According to the Bemis Murcko backbone analysis, the percentage of backbones in the three external independent test datasets ranged from 41 to 55%, indicating that the structures of the compounds in the external test set were highly diverse, which was beneficial for enhancing the test confidence. Detailed information was listed in Table 3.

To further validate the efficacy of the proposed method, five other methods were utilized to compare with the proposed method using benchmark datasets. These five methods were: (1) the AEFS method [51], (2) the NGDTP method [52], (3) the MolTrans method [38], (4) the Watanabe's network method (W_1-method) and (5) Watanabe's molecular method (W_2-method) [53]. The dataset is the small molecule drugs-protein interaction data from STITCH database, and three evaluation indexes were utilized to comprehensively compare the performance of each method. These methods were used for comparison. The AUC results showed that LDS-CNN (AUC of 0.962199) was higher than other methods.

The STITCH database-constructed drug small molecule-protein interaction dataset is compared to the DPI dataset using a variety of methodologies. Table 4 lists the outcomes of the comparisons, illustrating that the AUC and AUPRC for each method. Our model excels in AUC and AUPR (AUC: 0.962 and AUPR: 0.959). LDS-CNN improved AUC by 5.6%, 2.9%, 5.2%, 1.5%, and 0.6% relative to AEFS, NGDTP, MolTrans, and Watanabe's network and molecular work, while AUPRC increased by 48.2%, 63.45%, 85.67%, 3.9%, and 3.2%. Our method enhances AUPRC more than AUC, as demonstrated by the experimental findings.

Validation results

According to the prediction result, the interaction between LYN and IL4RA suggests a potential drug-target interaction. Therefore, molecular simulation was used to validate this hypothesis. According to retrieval at the Drugbank database, the 1-Tert-Butyl-3-(4-Chloro-Phenyl)-1 h-Pyrazolo[3,4-D] Pyrimidin-4-Ylamine (DrugBank AC: DB03023) was an effective drug molecule targeting LYN. Docking simulations between the compound and the protein were executed using the AutoDock program and DS visualizer software. The 3-dimensional structural protein information was downloaded from the RCSB Protein Data Bank. The Lamarckian genetic algorithm was adopted to search the docking conformation. Finally, the optimized docking model of LYN_HUMAN and DB03023 with binding energy − 6.22 kcal/mol and inhibition constant (K_i) 27.70 μM was obtained. Complex models of protein and compound, as well as their interactions, were displayed in Fig. 8A, B the optimized docking model of IL4RA_HUMAN and DB03023 with binding energy − 5.41 kcal/mol and inhibition constant (K_i) 108.13 μM was obtained. Complex models of protein

Table 4 Performance comparison with other methods

	AEFS	NGDTP	MolTrans	W_1-method	W_2-method	LDS-CNN
AUC	0.906	0.933	0.910	0.947	0.956	0.962
AUPRC	0.477	0.324	0.102	0.920	0.927	0.959
F-measure	0.507	0.400	0.104	0.853	0.868	0.906

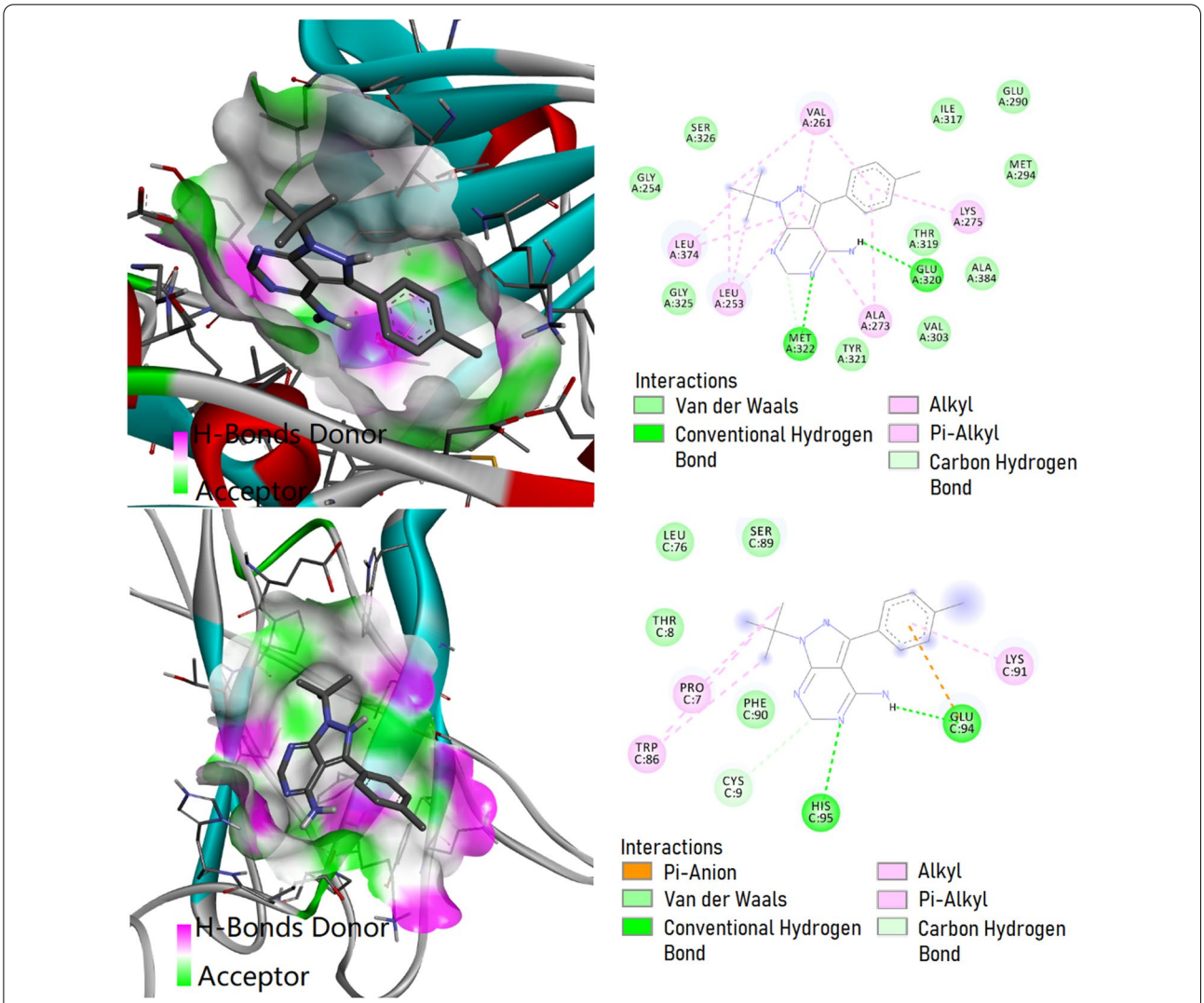

Fig. 8 **A** The complex model of ligand compound (DrugBank AC: DB03023) and receptor protein (UniProtKB: LYN_HUMAN) and their interactions. **B** The interactions between ligand compound and receptor protein. **C** The complex model of ligand compound (DrugBank AC: DB03023) and receptor protein (UniProtKB: IL4RA_HUMAN) and their interactions. **D** The interactions between ligand compound and receptor protein

and compound, as well as their interactions, were displayed in Fig. 8C, D.

It can observe from the complex model (Fig. 8A) that Var der Waals interactions exist between the compound and amino acid residues Gly254, Glu290, Met294, Val303, Ile317, Thr319, Tyr321, Gly325, Ser326, and Ala384. Moreover, some hydrophobic interactions exist, such as alkyl interactions between the compound and residues Leu253, Val261, Ala273, Lys275, and Leu374. Furthermore, it can observe from the complex model (Fig. 8C) that Var der Waals interactions exist between the compound and amino acid residues Pro7, Thr8, Cys9, Leu76, Ser89, and Phe90. Also, some hydrophobic interactions exist, such as π-alkyl interactions between the compound and residues Trp86 and Lys91. The drug (AC: DB03023) has a similar interaction way. Thus, we can infer that this drug may interact with IL4RA to perform a treating effect. In addition, the interaction between IL4RA and LYN is worth further experimental verification.

Conclusion

In this paper, we report on the LDS-CNN, a neural network model for the prediction of potential drug-target interactions. Compared with the current common methods in the field of DTI, the proposed method uses unified encoding and large-scale data for training to achieve feature abstraction and potential object prediction in different data formats within the integrated model. The

proposed method has the advantages of low experimental cost, appropriate time consumption and stable accuracy. Meanwhile, the overall performance can be further improved with an increasing scale of training data.

Results show that the proposed method achieved an AUC of 0.96 and AUPRC of 0.95, and an accuracy of 90.13% on a total of 898,412 drug-target protein interactions by using unified encoding. By validating the performance of the proposed method on different scale of dataset, the stability is proved, which suggests the potential application ability in predicting drug-target interactions. The combined energy calculation and module simulation results indicate that there are possibilities of actual existence for the prediction conclusion. The experimental conclusion also points out that model performance relates to the quality of the dataset, which suggests a future direction for model improvement through feature enhancement of the dataset.

In summary, we propose a novel deep learning model for identification of potential drug-target interactions by designing the unified encoding, which has advantages including low cost, high precision, and comprehensive data coverage. The proposed method is easy to be applied on specific research target by utilizing transfer learning. In the following work, we will further optimize the feature design to improve the efficiency of data encoding. Meanwhile, the proposed method provides a feasible application direction for designing different feature combinations to enhance the model performance. In order to further provides certain suggestion for future research, we will continue to improve it to address other analysis tasks based on unified encoding optimization, providing effective insights for drug-target interactions analysis and drug development.

Funding
This work was supported in part by the National Key R&D Program of China under Grant No. 2022YFB3304400, in part by the China University Industry, University and Research Innovation Fund under Grant No. 2021FNA03002, and in part by the Shanghai Pujiang Programme under Grant No. 22PJD104.

Data availability
All datasets used in this study are listed in the article. The code and data of the method are freely available at: https://github.com/ZuxianZhang/LDS-CNN.

Declarations

Conflict of interest
The authors declare that they have no competing interests or potential conflict.

Author details
[1]School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China. [2]School of Biomedical and Pharmaceutical Sciences, Guangdong University of Technology, Guangzhou 510006, China. [3]The First Affiliated Hospital of Ningbo University, Ningbo 315010, China. [4]Smart Medical Innovation Technology Center, Guangdong University of Technology, Guangzhou 510006, China. [5]Shanghai Institute of Biological Products, Shanghai 201403, China.

Received: 26 May 2023 Accepted: 14 August 2023
Published: 2 September 2023

References

1. Ye Y, et al. Drug-target interaction prediction based on adversarial Bayesian personalized ranking. Biomed Res Int. 2021;2021:6690154.
2. Yang Z, et al. FragDPI: a novel drug-protein interaction prediction model based on fragment understanding and unified coding. Front Comp Sci. 2022;17(5): 175903.
3. Huang K, et al. DeepPurpose: a deep learning library for drug–target interaction prediction. Bioinformatics. 2020;36(22–23):5545–7.
4. Cowen L, et al. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet. 2017;18(9):551–62.
5. Cheng FX, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018;9:2691.
6. Wei BM, Zhang Y, Gong X. DeepLPI: a novel deep learning-based model for protein-ligand interaction prediction for drug repurposing. Sci Rep. 2022;12(1):18200.
7. Perez-Nueno VI, et al. Detecting drug promiscuity using Gaussian ensemble screening. J Chem Inf Model. 2012;52(8):1948–61.
8. Rao L, et al. Nonfitting protein-ligand interaction scoring function based on first-principles theoretical chemistry methods: development and application on kinase inhibitors. J Comput Chem. 2013;34(19):1636–46.
9. Sajadi SZ, et al. AutoDTI plus plus: deep unsupervised learning for DTI prediction by autoencoders. BMC Bioinform. 2021;22(1):1–19.
10. Huang L, et al. CoaDTI: multi-modal co-attention based framework for drug-target interaction annotation. Brief Bioinform. 2022;23(6):bbac446.
11. Chavan G, Das D. Design and characterizations of pH-responsive drug delivery vehicles using molecular docking. Mater Technol. 2023;38(1):2196490.
12. Wang YB, et al. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. BMC Med Inform Decis Mak. 2020;20(1):1–9.
13. Varela D, Karlin V, Andre I. A memetic algorithm enables efficient local and global all-atom protein-protein docking with backbone and side-chain flexibility. Structure. 2022;30(11):1550+.
14. Li KQ, et al. Identification of a potential structure-based GPCR drug for interstitial cystitis/bladder pain syndrome: in silico protein structure analysis and molecular docking. Int Urogynecol J. 2023;34:1559–65.
15. Zeng M, et al. A deep learning framework for identifying essential proteins based on protein-protein interaction network and gene expression data. In: Proceedings 2018 IEEE international conference on bioinformatics and biomedicine. 2018. pp. 583–8.
16. Yu Z, et al. Predicting drug-disease associations through layer attention graph convolutional network. Brief Bioinform. 2021;22(4):bba243.
17. Vamathevan J, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discovery. 2019;18(6):463–77.
18. Schneider G. Automating drug discovery. Nat Rev Drug Discovery. 2018;17(2):97–113.
19. Chen RL, et al. Machine learning for drug-target interaction prediction. Molecules. 2018;23(9):2208.
20. Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018;34(17):821–9.
21. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. Bioinformatics. 2009;25(18):2397–403.
22. Mei JP, et al. Drug-target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45.
23. Buza K, Peska L. Drug-target interaction prediction with bipartite local models and hubness-aware regression. Neurocomputing. 2017;260:284–93.
24. Cheng FX, Zhao ZM. Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. J Am Med Inform Assoc. 2014;21(E2):E278–86.
25. Bull SC, Doig AJ. Properties of protein drug target classes. PLoS ONE. 2015;10(3):e0117955.

26. Zhou LQ, et al. Revealing drug-target interactions with computational models and algorithms. Molecules. 2019;24(9):1714.
27. Kwon S, Yoon S. DeepCCI: end-to-end deep learning for chemical-chemical interaction prediction. In: ACM-BCB' 2017: proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics. 2017. pp. 203–12.
28. Tran HN, Xuan QNP, Nguyen TT. DeepCF-PPI: improved prediction of protein-protein interactions by combining learned and handcrafted features based on attention mechanisms. Appl Intell. 2023;53:17887–902.
29. Wan FP, et al. DeepCPI: a deep learning-based framework for large-scale in silico drug screening. Genomics Proteomics Bioinform. 2019;17(5):478–95.
30. Dorahy G, Chen JZ, Balle T. Computer-aided drug design towards new psychotropic and neurological drugs. Molecules. 2023;28(3):1324.
31. Zhao QC, et al. AttentionDTA: prediction of drug-target binding affinity using attention model. In: IEEE international conference on bioinformatics and biomedicine (BIBM). 2019. pp. 64–9.
32. Lin X, et al. DeepGS: deep representation learning of graphs and sequences for drug-target binding affinity prediction. In: ECAI 2020: 24th European conference on artificial intelligence, vol. 325. 2020. pp. 1301–8.
33. Wen M, et al. Deep-learning-based drug-target interaction prediction. J Proteome Res. 2017;16(4):1401–9.
34. Wong A, et al. Amino acid motifs for the identification of novel protein interactants. Comput Struct Biotechnol J. 2023;21:326–34.
35. Khiar-Fernandez N, et al. Chemistry for the identification of therapeutic targets: recent advances and future directions. Eur J Org Chem. 2021;2021(9):1307–20.
36. Ye JH, et al. Machine learning advances in predicting peptide/protein-protein interactions based on sequence information for lead peptides discovery. Adv Biol. 2023. https://doi.org/10.1002/adbi.202200232.
37. Anusuya S, et al. Drug-target interactions: prediction methods and applications. Curr Protein Pept Sci. 2018;19(6):537–61.
38. Huang KX, et al. MolTrans: molecular interaction transformer for drug-target interaction prediction. Bioinformatics. 2021;37(6):830–6.
39. Chu Y, et al. DTI-CDF: a cascade deep forest model towards the prediction of drug-target interactions based on hybrid features. Brief Bioinform. 2021;22(1):451–62.
40. Lee I, et al. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15(6):e1007129.
41. Bagherian, et al. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform. 2020;22:247–69.
42. Otovic E, et al. Sequential properties representation scheme for recurrent neural network-based prediction of therapeutic peptides. J Chem Inf Model. 2022;62(12):2961–72.
43. Kim S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373–80.
44. Gaulton A, et al. ChEMBL: a large-scale bioactivity database for chemical biology and drug discovery. Nucleic Acids Res Database Issue. 2012;40:D1100–7.
45. Wishart DS, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017;46(D1):D1074–82.
46. Szklarczyk D, et al. STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380–4.
47. Szklarczyk D, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–46.
48. Zeng HY, et al. Convolutional neural network architectures for predicting DNA-protein binding. Bioinformatics. 2016;32(12):121–7.
49. Campana PA, Nikoloski Z. Self- and cross-attention accurately predicts metabolite-protein interactions. NAR Genom Bioinform. 2023;5(1):lqad008.
50. Zhao HC, Li YH, Wang JX. A convolutional neural network and graph convolutional network-based method for predicting the classification of anatomical therapeutic chemicals. Bioinformatics. 2021;37(18):2841–7.
51. Sun C, et al. Autoencoder-based drug–target interaction prediction by preserving the consistency of chemical properties and functions of drugs. Bioinformatics. 2021;37(20):3618–25.
52. Xuan P, et al. Prediction of drug-target interactions based on network representation learning and ensemble learning. IEEE/ACM Trans Comput Biol Bioinform. 2021;18(6):2671–81.
53. Watanabe N, Ohnuki Y, Sakakibara Y. Deep learning integration of molecular and interactome data for protein-compound interaction prediction. J Cheminform. 2021;13(1):36.