BMC Bioinformatics

RESEARCH

Open Access

AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders

Seyedeh Zahra Sajadi[1], Mohammad Ali Zare Chahooki[1*], Sajjad Gharaghani[2] and Karim Abbasi[2]

*Correspondence:
chahooki@yazd.ac.ir
[1] Department of Computer Engineering, Yazd University, Yazd, Iran
Full list of author information is available at the end of the article

Abstract

Background: Drug–target interaction (DTI) plays a vital role in drug discovery. Identifying drug–target interactions related to wet-lab experiments are costly, laborious, and time-consuming. Therefore, computational methods to predict drug–target interactions are an essential task in the drug discovery process. Meanwhile, computational methods can reduce search space by proposing potential drugs already validated on wet-lab experiments. Recently, deep learning-based methods in drug-target interaction prediction have gotten more attention. Traditionally, DTI prediction methods' performance heavily depends on additional information, such as protein sequence and molecular structure of the drug, as well as deep supervised learning.

Results: This paper proposes a method based on deep unsupervised learning for drug-target interaction prediction called AutoDTI++. The proposed method includes three steps. The first step is to pre-process the interaction matrix. Since the interaction matrix is sparse, we solved the sparsity of the interaction matrix with drug fingerprints. Then, in the second step, the AutoDTI approach is introduced. In the third step, we post-preprocess the output of the AutoDTI model.

Conclusions: Experimental results have shown that we were able to improve the prediction performance. To this end, the proposed method has been compared to other algorithms using the same reference datasets. The proposed method indicates that the experimental results of running five repetitions of tenfold cross-validation on golden standard datasets (Nuclear Receptors, GPCRs, Ion channels, and Enzymes) achieve good performance with high accuracy.

Keywords: Drug-target interactions, Deep learning, Unsupervised learning, Latent feature, Denoising autoencoder

Background

Protein targets are strictly related to some diseases. The target's biological activities reveal due to the therapeutic impact of drugs on these diseases. Therefore, to animate or repress a target's biological process in the drug discovery process, we consider a drug's interaction with the target proteins [1]. Thus, drug–target interactions (DTIs) play a prominent role in drug discovery. However, identifying and validating drug candidates via biological assays, from introducing the abstract concept to release it into the market, usually take 10–15 years and costs 0.8–1.5 billion dollars [2]. Therefore, various

computational methods to predict drug–target interactions are being used to aid the drug discovery process. Computational methods have some advantages, including low drug development costs, short time, low drug safety risk, and exploring a wide range of potential drug–target interactions. The computational approaches received more attention in recent years. Chen et al. [3], for DTI prediction, introduced some state-of-the-art computational models, including network-based approach and machine learning-based approach. Bagherian et al. [4] described data and databases required and broad category consisting of a machine learning approach for DTI prediction. Ding et al. [5] concentrated on machine learning-based methods, especially similarity-based methods that use drug and target similarities. Abbasi et al. [6] reviewed the deep learning-based approach in DTI, and they give some perspective on the future approaches.

In DTI prediction, computational approaches are divided into three major groups. The first group is called the ligand-based approach, which uses similar molecules and the similarity between the target proteins' ligands [7]. However, the results obtained from ligand-based methods might be incorrect when the number of target's known ligands are insufficient [8]. The second group comprises the docking approach. In this approach, the 3D structures of drug and protein are taken into account and used to determine their interaction tendency. One of the limitations of this approach is that they require the 3D structure of the target proteins [9, 10]. Hence, these methods could not be applied to new drug-target pairs that the 3D structures of proteins are unavailable [11]. For example, predicting the 3D structure for targets like GPCRs is still challenging [12]. The third group comprises the chemogenomics approaches that utilize information of drug and target concurrently to predict DTI. One of the advantages of chemogenomics approaches is that many online public databases can access their available data. For example, information such as the genomic sequences of targets and the chemical structure of drugs are used for DTI prediction [13]. This approach doesn't have the limitations mentioned in the previous two groups. The chemogenomics approach usually uses machine learning and deep learning methods for DTI predictions. This paper concentrates on computational methods that belong to the chemogenomics approach.

The proposed method by Chen et al. [14] integrated three different networks, such as protein–protein similarity network, drug-drug similarity network, and known drug-target interaction networks, into a heterogeneous network by known drug–target interactions and performed the random walk on this heterogeneous network. Mazharul Islam et al. [15] proposed a DTI-SNNFRA framework for DTI prediction based on shared nearest neighbor (SNN) by a partitioning clustering for sampling the search space in the first stage and fuzzy-rough approximation (FRA) in the second stage. Zeng et al. [16] proposed a network-based deep-learning method for DTI prediction by integrating ten networks called DeepDR. Then the low-dimensional representation of drugs and drug-disease pairs by a variational autoencoder were learned from the heterogeneous networks. Lim et al. [17] introduced a novel approach for predicting DTI based on a graph neural network that directly organized the 3D structural information on a protein–ligand binding posed into an adjacency matrix. A distance-aware graph attention mechanism was also devised to increase the performance of the model. Zong et al. [18] proposed a DeepWalk deep learning method for drug-target interaction prediction based on network topology similarity measures. Firstly, a heterogeneous network

created from biomedical linked datasets. After that DeepWalk was selected to measure the similarities within linked tripartite network (LTN).

With the increase of experimental data, the use of deep learning methods to predict DTIs has been increasing. Deep learning methods learn the input data's hierarchical features, leading to better performance than other standard machine learning methods. In deep learning-based DTI prediction, a drug-target pair has taken as input, and then the affinity of interaction is predicted as output. Wen et al. [19] adapted a deep learning method named DeepDTI that used a deep belief network (DBN). Their approach predicted the affinity value for a pair of FDA-approved drugs and targets. In their work, protein targets were not divided into different classes. The features of drugs were automatically extracted from extended-connectivity fingerprints (ECFP), and the features of target proteins were extracted from the composition of amino acids, dipeptides, and tripeptides [20]. Peng et al. [21] used sparse autoencoders to reduce the original features' dimension into a hidden representation, and then they trained a support vector machine (SVM) with hidden representation. In another study called DL-CPI [22], which used protein domain information, domain binary vectors were employed to represent the domains used to describe proteins. Ozturk et al. [23] introduced a DTI prediction approach which used the convolutional neural network (CNN) to learn the feature vectors for drug and protein target. On a kinase family bioassay dataset, their approach performed better [24, 25] than the conventional models like kronRLS-MKL [26] and SimBoost [27]. In a paper by Lee et al. [1], their DeepConv-DTI model predicted massive-scale DTIs using raw protein sequences for various target protein classes and diverse protein lengths. New protein features were generated with convolution filters on the entire protein sequence to capture local residue patterns. Then protein features and the drug features were concatenated and fed into the subsequent layers to predict the affinity value. Finally, their model was optimized with DTIs from MATADOR [28]. Abbasi et al. [29] combined convolutional layers and Long Short-Term Memory (LSTM) layers to learn more effective local substructures through a compound and a protein. Then they utilized a two-sided attention mechanism to weight each local substructure of the compound and protein sequence.

As an unsupervised approach to DTI prediction, matrix factorization (MF) techniques learn the latent feature matrices of drugs and targets from the DTI matrix. These two latent feature matrices are multiplied to reconstruct the interaction matrix for prediction. Among various unsupervised methods in DTI, regularized matrix factorization methods achieve a higher performance among the previous DTI prediction methods [30, 31]. Matrix factorization techniques suffer from the cold start problem as well as the sparsity. In this study, to overcome the issues mentioned above, the unsupervised approach of deep learning is utilized to extract latent factors of input data. To this end, in this paper, we have developed a new drug-target interaction prediction method named AutoDTI++, an unsupervised deep learning model by using denoising autoencoder. Denoising autoencoder is an unsupervised deep neural network that learns the latent factors from the matrix interaction. However, the learned latent factors are not very effective due to the sparse nature of the drug-target interaction matrix. Additional information such as drug fingerprints information has been utilized to address the drug-target interaction matrix sparsity problem.

To evaluate our proposed method, we have used cross-validation to compare it with six other state-of-the-art methods, namely DDR [32], DNILMF [33], NRLMF [34], KronRLS-MKL [26], BLM-NII [35], and COSINE [36]. We have evaluated the ability of AutoDTI++ using new drug cross-validation, new interaction cross-validation, and new target cross-validation. We computationally simulated a new target case and a new drug case (by leaving out their respective interactions) and tested our proposed method on these cases to investigate its ability to predict the left-out interactions. Finally, our model achieved better performance than most previous models.

In section methods, firstly, we describe the dataset used in our work in "Dataset" section. Our notations are described in "Notations" section. An overview is done on the neural network of denoising autoencoder (DAE) in "Denoising autoencoder" section. Then, our proposed method is described in "Workflow" section. The experimental results of our work, relevant discussion, and conclusion are given in the next sections, respectively.

Methods

Dataset

This study used the introduced benchmark dataset in [9] to evaluate our proposed approach. This dataset contains four different target protein types, namely nuclear receptors (NR), G protein-coupled receptors (GPCR), ion channels (IC), and enzymes (E). Table 1 shows some statistics, including the number of unique proteins, number of unique drugs, number of interactions, and the sparsity ratio for each dataset. The variable $Y \in \mathbb{R}^{n \times m}$ denotes the interaction matrix where n represents the number of drugs and m denotes the number of targets. Suppose the drug d_i and the target t_j interact, then $Y_{ij} = 1$, otherwise $Y_{ij} = 0$. Rows and columns of Y show the profiles of drugs and targets, respectively. The interaction profile for each drug or target is determined by Y_d and Y_t, respectively. Sparsity denotes the ratio between the number of DTIs and the number of all possible DTIs.

Preliminaries

In this section, first, we define the notations used in this paper. Then, we simply introduce denoising autoencoder.

Notations

The notation used in this paper is listed as follows:

Y_d, Y_t are the sparse row/columns of Y

$\tilde{Y}_d, \tilde{Y}_t$ are corrupted versions of Y_d, Y_t

Table 1 Drugs, targets, interactions, and sparsity in each dataset

Datasets	NR	GPCR	IC	E
No. of drugs	54	223	210	445
No. of targets	26	95	204	664
No. of interactions	90	635	1476	2926
Sparsity	0.064	0.030	0.034	0.01

$\widehat{Y}_d, \widehat{Y}_t$ are dense estimates of Y_d, Y_t
$\overline{Y}_d, \overline{Y}_t$ are dense low-rank representations of Y_d, Y_t

Denoising autoencoder

An autoencoder is an unsupervised neural network that includes two networks: an encoder and a decoder aiming to reconstruct the input domain. The encoding network maps the input to a hidden representation [37]. The decoding network reconstructs the original inputs from the hidden representation [38]. As a result, autoencoder is used to learn feature representation in an unsupervised manner. An autoencoder is considered a neural network that obtains higher-level representations of input data without requiring ground-truth label information. Given a training sample x ($x \in R^{d_0}$), it is encoded into the hidden representation $y \in R^{d_1}$ by the mapping f_c:

$$\textbf{Encoder} : y = f_c(x) = S_c(V^T x + b_1) \tag{1}$$

where S_c is the non-linear activation function of the encoder. Also, V and b_1 are respectively the weight matrix and the bias vector. After that, the representation of the hidden layer y is mapped to the reconstructed output $x^{'}$ of the same shape as x by function f_d:

$$\textbf{Decoder} : x^{'} = f_d(y) = S_d\left(W^T y + b_2\right) \tag{2}$$

where S_d, W, and b_2 are the same parameters of the decoder network. The full autoencoder is indicated by $\boldsymbol{nn}(\boldsymbol{x}) \stackrel{\text{def}}{=} \boldsymbol{f_d}(\boldsymbol{f_c}(\boldsymbol{x}))$.

Recently, many autoencoders have been introduced, like denoising autoencoder, sparse autoencoder, and variational autoencoder [29]. Denoising autoencoders add some noise to the input and then force the network to reconstruct the denoised input. One method to add some noise is to mask a random fraction of the input by replacing them with zero. In this case, we use the modified loss function to emphasize the denoising aspect of the network. To this end, two weight hyperparameters α and β are used to weight the terms as follows:

$$L_{\alpha,\beta}(x, \tilde{x}) = \alpha\left(\sum_{j \epsilon \mathcal{C}(\tilde{x})}\left[nn(\tilde{x})_j - x_j\right]^2\right) + \beta\left(\sum_{j \notin \mathcal{C}(\tilde{x})}\left[nn(\tilde{x})_j - x_j\right]^2\right) \tag{3}$$

where $\tilde{x} \in \mathbb{R}^N$ is a corrupted version of the input x, $\mathcal{C}$ is the set of corrupted elements in $\tilde{x}$, $0 < \alpha, \beta < 1$, and $nn(x)_j$ is the j^{th} the output of the network while fed with x.

Workflow

In this section, the proposed drug-target interaction prediction method called AutoDTI++ is presented, which consists of three steps:

(i) The first step includes a pre-processing step that transforms the binary values in the given drug-target matrix, Y, into the binary values in the drug fingerprint-target interaction matrix for filling missing values based on drug fingerprint.
(ii) The second step is to propose an AutoDTI model that uses an unsupervised deep learning technique based on denoising autoencoders to predict drug–target interactions.

(iii) The third step includes a post-processing step in which the drug-target interaction matrix is predicted from the output of the second step.

After presenting these three steps, we will present the proposed approach.

Pre-processing step

While deep learning has many successes in image and speech recognition [39], sparse data has received less attention and remains a challenging problem for neural networks. Therefore, there is no standard approach for using the sparse matrix as inputs of deep neural networks yet. Most papers on sparse inputs are obtained by pre-calculating estimates of missing values [40]. Sparse inputs have already been studied in the industry [41], where 5% of the values are missing. However, datasets in DTI often face more than 95% missing values. Since the drug-target interaction matrix relies on only interactions between drugs and targets, when additional information is available for the drugs and the targets, only using the interaction matrix can sound restrictive. Therefore in our case, we want to handle this issue by adding information on drugs fingerprint to the interaction matrix. Our approach uses the fingerprint of drugs to handle autoencoders' sparse input. To this end, the following steps are done:

(1) The first step represents the drug molecule by SMILES (simplified molecular-input line-entry system): each drug is represented by SMILES [42] strings, a sequential encoding of chemical structures.
(2) The second step, create the fingerprint-drug matrix (Z): utilize the PaDEL-descriptor software to transform SMILES string to fingerprints. PaDEL-descriptor software is used for calculating molecular descriptors (1D, 2D descriptors, and 3D descriptors) and ten types of fingerprints [43]. Each drug can be represented as a binary vector with a length of 800, in which indices indicate the existence of the specific substructures.
(3) The third step, create the fingerprint-target matrix ($W = Z.Y$): We multiply the fingerprint-drug matrix (Z) by the drug-target interaction matrix (Y). The result is a fingerprint-target matrix (W).
(4) The fourth step, normalization: normalize the fingerprint-target matrix with the min−max method.
(5) The fifth step, convert to the binary matrix: Since values greater than zero in this matrix represent an interaction between the target and the drug fingerprint, these values are replaced by one.

By performing these five steps, the obtained matrix is not sparse like the raw drug-target interaction matrix. With these pre-processing steps, almost half of the fingerprint–target interactions matrix is known.

The AutoDTI model

In the AutoDTI model, if it is assumed that the model's input is a drug-target interaction matrix, then drug-target known interactions can be encoded as a partially drug-target interaction matrix $Y \in \mathbb{R}^{n \times m}$. Each drug $d \in D = \{1 \ldots n\}$ can be represented by a

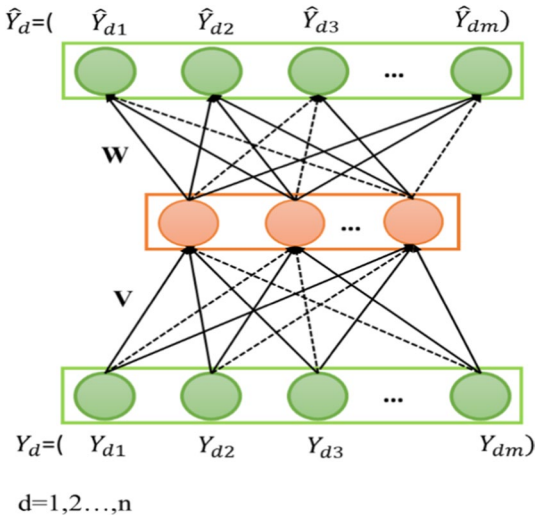

Fig. 1 The overall schematic of the D-AutoDTI model. The shaded nodes show the known interactions, and the solid connections show the weights that are updated for the input $\boldsymbol{Y_d}$. There are n copies of the neural network for each drug

partially observed vector $Y_d = (Y_{d1}, \ldots Y_{dm}) \in \mathbb{R}^m$. Similarly, each target $t \in T = \{1 \ldots m\}$ can be represented by a partially observed vector $Y_t = (Y_{1t}, \ldots Y_{nt}) \in \mathbb{R}^n$. Our aim in this work is to design a drug-based (target-based) autoencoder which can take each partially observed Y_d (Y_t) as input, project it into a low-dimensional latent space and then reconstruct Y_d (Y_t) in the output space to predict unknown interactions. We reconstruct the sparse vectors $Y_d(Y_t)$, into dense vectors $\widehat{Y}_d\left(\widehat{Y}_t\right)$. In this case, it is needed to define two types of autoencoders:

- D-AutoDTI is defined as $\widehat{Y}_d \stackrel{\text{def}}{=} nn(Y_d)$
- T-AutoDTI is defined as $\widehat{Y}_t \stackrel{\text{def}}{=} nn(Y_t)$

The learned parameters are regularized to prevent the over-fitting of the observed interactions. Formally, the objective function for the D-AutoDTI model is:

$$\min_{\theta} \sum_{d=1}^{n} \|Y_d - nn(Y_d, \theta)\|_o^2 + \frac{\lambda}{2} \cdot \left(\|W\|_F^2 + \|V\|_F^2\right) \tag{4}$$

where $\|.\|_F^2$ means that we only consider the contribution of the known interactions and regularization strength $\lambda > 0$. The proposed approach's training loss differs from the classic autoencoders, which only aim to reconstruct the input. Given the learned parameters$\widehat{\theta}$, D-AutoDTI's predicted interactions for drug d and target t are:

$$\widehat{Y}_{dt} = \left(nn\left(Y_d; \widehat{\theta}\right)\right)_t \tag{5}$$

Figure 1 shows the overall schematic of the utilized autoencoder. The shaded nodes illustrate the known interactions, and the solid connections show the weights that are updated for the input Y_d.

To train the autoencoders, the following three steps are performed:

i)Assign zero to unknown interactions in the edges of input layers,
ii) back-propagated values in the edges of the output layers are replaced by zero values,
iii)use a denoising loss to emphasize interaction prediction over interaction reconstruction.

One way to restrain the edges of the input is to turn the missing values to zero. We utilize an empirical loss that ignores the loss of unknown values to preserve the autoencoder from always returning zero. Missing values do not bring information to the network. The error is discarded for missing values. Therefore, the empirical loss back-propagates the error for known values while no error is back-propagated for missing values. In other words, this operation is equivalent to removing the neurons with missing values described in [44, 45]. Finally, masking noise is used from the denoising autoencoders empirical loss. Autoencoders in the training process are trained to predict missing values by simulating them. The final target is the prediction of these missing values. Thus, the classic unsupervised training of autoencoders converts to simulated supervised learning by emphasizing the prediction criterion. The training can be turned into pseudo-semi-supervised learning by mixing both criteria of reconstruction and prediction. The denoising autoencoders' loss becomes an assuring objective function. The final training loss function after regularization is:

$$
\begin{aligned}
L_{\alpha,\beta}\left(Y_d, \tilde{Y}_d\right) = \alpha & \left(\sum_{j \epsilon \mathcal{C}\left(\tilde{Y}_d\right)} \left\| (Y_d)_j - nn\left(\tilde{Y}_d\right)_j \right\|_o^2 \right) \\
& + \beta \left(\sum_{j \notin \mathcal{C}\left(\tilde{Y}_d\right)} \left\| (Y_d)_j - nn\left(\tilde{Y}_d\right)_j \right\|_o^2 \right) + \frac{\lambda}{2} \cdot \left(\|W\|_2^f + \|V\|_2^f \right)
\end{aligned}
\tag{6}
$$

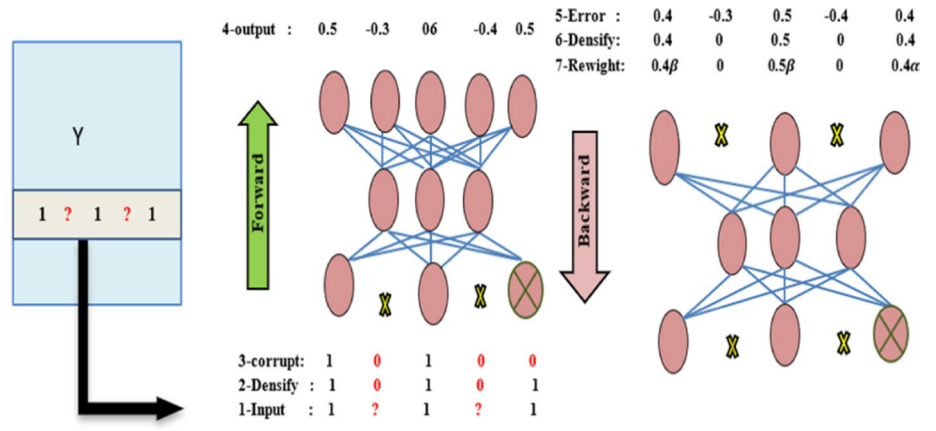

Fig. 2 Feed-Forward/Backward process is shown for denoising autoencoder. The input is obtained from the matrix of interactions, unknown values are turned to zero, some interactions input are corrupted, and a dense estimate is finally constructed. Before back-propagation, unknown interactions are converted to zero error. Use $\boldsymbol{\beta}, \boldsymbol{\alpha}$ hyper-parameters, reconstruction, and prediction errors are reweighed

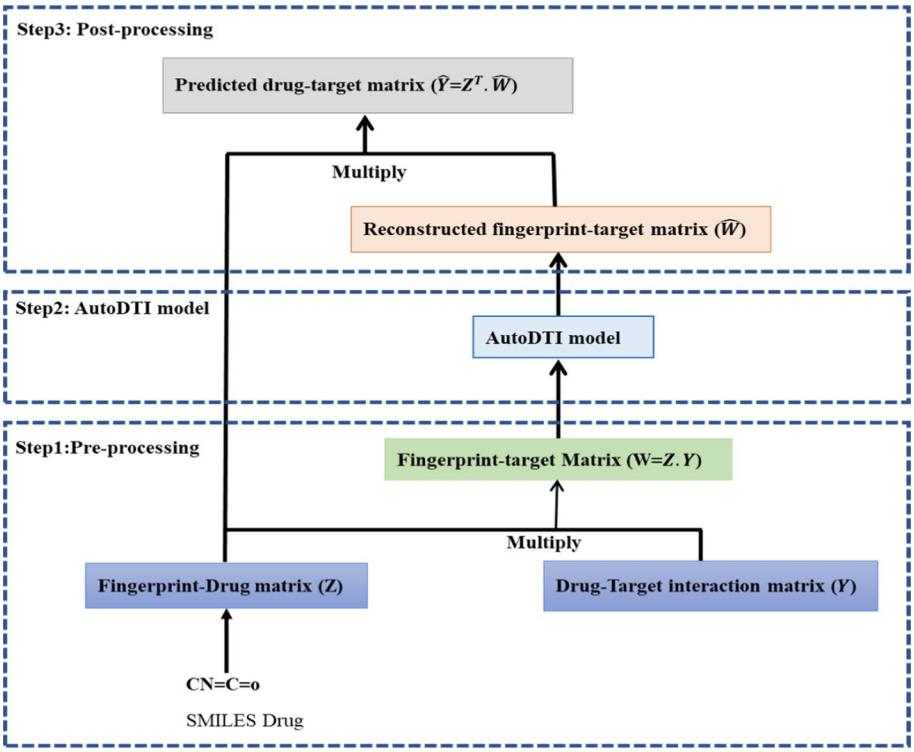

Fig. 3 The AutoDTI++ method includes three steps: pre-processing, AutoDTI model, post-processing

W and V are the vectors of weights of the network, and λ is the regularization hyper-parameter. The full-forward/backward process is explained in Fig. 2.

Post-processing

In the post-processing step, the drug-fingerprint matrix (Z^T) is multiplied by the output of the AutoDTI model ($\widehat{W}$). The product of multiplication is equivalent to the predicted drug-target interaction matrix.

AutoDTI++ proposed method

As shown in Fig. 3, the AutoDTI++ proposed method is performed in three steps which include: the first step is pre-processing, which explained in "Pre-processing step" section. The second step uses the AutoDTI model explained in "The AutoDTI model" section. In AutoDTI++ proposed method, the fingerprint-target matrix is applied as the AutoDTI model input instead of the drug-target interaction matrix. The third step is post-processing that explained in "Post-processing" section. Fingerprint-target reconstructed matrix ($\widehat{W}$) is calculated as follows:

$$\mathrm{W} = \mathrm{Z} \cdot \mathrm{Y} \tag{7}$$

$$\widehat{W} = \boldsymbol{nn}\left(W; \widehat{\theta}\right) \tag{8}$$

Z is a fingerprint-drug matrix, and Y is a drug-target matrix. Predicted interactions of AutoDTI++ for drug d and target t are:

$$\widehat{Y}_{dt} = {Z^T}_d \cdot \widehat{W}_t \tag{9}$$

where ${Z^T}_d$ is d^{th} row of Z^T matrix, $\widehat{W}_t$ is t^{th} the column of W reconstructed matrix.

Results

First, we introduce the cross-validation (CV) and the metric we used to evaluate our models. Second, we present the parameter settings. Then, we present some baseline approaches which are compared with our model. Finally, we compare our models with the baselines to illustrate the performance of our model.

Cross-validation experiments

We performed cross-validation under three scenarios described in [46] to perform a comprehensive empirical comparison among various methods as follows:

(1)S_p, denote the random drug–target pairs that are left out to be used as the test set;

(2)S_d, denote the entire drug interaction profiles that are left out to be used as the test set; and.

(3)S_t, denote the entire target interaction profiles that are left out to be used as the test set.

S_p is the traditional method for performance evaluation. Meanwhile, various approaches to predict interactions for new drugs and targets are evaluated using S_d, and S_t test sets. Here, new drugs and targets are those for which no interaction information is available in the training set. As such, conducting experiments under S_d and S_t provides information about the proposed approach's generalizability.

Such as previous works, we employed the area under the receiver operating characteristic (AUC) curve and the area under the precision-recall (AUPR) curve to evaluate prediction performance. We performed experiments to compare our proposed method with the existing techniques, including DDR, DNILMF, NRLMF, KRONRLS-MKL, BLM-NII, and COSINE. Specifically, we conducted five repetitions of the tenfold CV for each of the methods under each of the above scenarios using AUPR [47] as the evaluation metric. That is, the interaction data set was divided into ten folds, and each fold, in turn, was left out as the test set while the remaining nine folds were treated as the training set. The prediction performance for each of the folds is evaluated in terms of AUPR. This process is repeated five times, and the final AUPR score was the average over five such repetitions. For all experiments, AUPR was used as the main metric for performance evaluation. AUPR is more adequate because it heavily penalizes incorrect predictions of interactions [48], which is desirable here. After all, we do not want false predictions recommended by the prediction algorithm in practice.

Parameter settings

Experiments are conducted on the benchmark database [9]. We repeated this splitting procedure 5 times and reported average AUPR and AUC. First, we calculated AUC and

Table 2 AUC and AUPR scores of AutoDTI++ approach obtained under three prediction tasks (S_p, S_d, and S_t) overall datasets (NR, GPCR, IC, and E) by 5 repeats of tenfold CV

AutoDTI++	NR	GPCR	IC	E
S_p				
AUPR	0.84	0.85	0.90	0.82
AUC	0.87	0.86	0.91	0.90
S_d				
AUPR	0.62	0.47	0.50	0.33
AUC	0.60	0.47	0.49	0.50
S_t				
AUPR	0.84	0.83	0.86	0.77
AUC	0.87	0.85	0.86	0.84

AUPR on NR, GPCR, IC, and E datasets for the AutoDTI method without pre-processing. The obtained results are not acceptable. Then, we applied a pre-processing step on the AutoDTI method and called that AutoDTI++. Interestingly, after a pre-processing step, AutoDTI significantly improved the results of AUC and AUPR on all datasets.

We evaluated the performance of the AutoDTI++ model as the number of hidden units and the number of hidden layers varied. We observed that performance steadily increases with two hidden layers of (15, 5) units. We used sigmoid activation functions in each layer. Using a non-linear activation function in the hidden layer is critical for the excellent performance of AutoDTI++. We did fine-tuning by gradient-based back-propagation with a minibatch of size 100. We set the regularization strength to 10 for IC, GPCR, and E datasets, and we set it to 1 for the NR dataset.

Impact of the loss: we investigated the effects of hyper-parameters α, β on denoising loss. To this end, we used a greedy search, and the best performance is achieved with $\alpha = 0.4$ and β=0.6.

Comparisons with the state-of-the-art algorithms

AutoDTI++ method calculates AUC and AUPR on NR, GPCR, IC, and E datasets. For NR, GPCR, IC, and E datasets, AUPR and AUC scores for S_p,S_d, and S_t test sets show in Table 2. Figure 4 shows the ROC curve and precision-recall curve of the first repeat of tenfold cross-validation on four datasets. The mean-AUC and mean-AUPR are the average AUC and average AUPR of AutoDTI++ in the first repeat of tenfold CV.

Baseline approaches

To measure the prediction performance, six existing state-of-the-art DTI prediction methods are used to compare with our AutoDTI++ model on NR, GPCR, IC, and E datasets under three different CV settings, including DDR, DNILMF, NRLMF, Kron-RLS-MKL, and BLM-NII, and COSINE.

DDR

First, it is based on using a heterogeneous graph that applies a similarity selection procedure to select a set of informative and less-redundant similarities for drugs and target

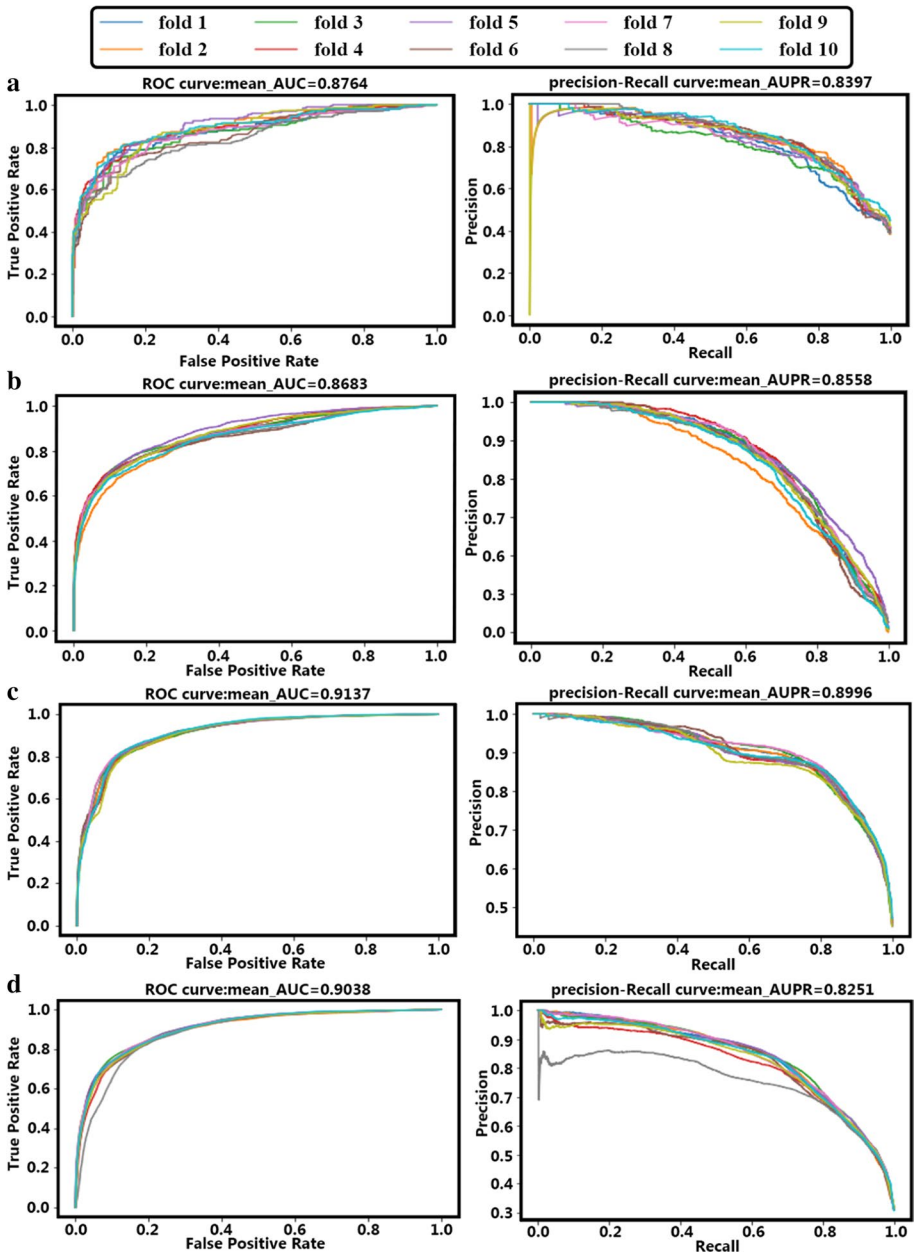

Fig. 4 The ROC and precision-recall curves of the first repeat of tenfold CV for four datasets; the right side is the precision-recall curve, and the left is the ROC curve. **a** The precision-recall and ROC curves for NR dataset; **b** The precision-recall and ROC curves for GPCR dataset; **c** The precision-recall and ROC curves for IC dataset; **d** The precision-recall and ROC curves for E dataset

proteins. DDR combines different similarities using the non-linear similarity fusion method. Then, manually, 12 different path-category-based feature patterns from the heterogeneous network are extracted. Finally, DDR applies a random forest model to predict DTIs.

KronRLS-MKL

First, it applies the weighted combination of multiple drug kernels and target kernels to get the final drug kernel and target kernel, and then KronRLS uses Kronecker product algebraic properties as the drug-target pairwise kernel. Finally, it uses Kronecker regularized least squares to predict DTIs.

NRLMF

NRLMF method focuses on modeling the probability. The interaction probability of a drug with a target is calculated by a logistic function of the drug-specific and target-specific latent vectors. Furthermore, the neighborhood regularization based on the local structure of the drug-target interaction data is utilized to improve the model's prediction ability.

DNILMF

DNILMF method is followed by the non-linear combination technique of multiple similarity measures for drugs and target proteins, as well as smoothing new drug-target predictions based on their neighbors.

BLM-NII

in BLM–NII, the neighbor-based interaction-profile inferring (NII) procedure is integrated into the bipartite local model (BLM) framework to form a DTI prediction approach, where the RLS classifier with GIP kernel was used as the local model.

We used 5-repeats of tenfold cross-validation to evaluate the predictive performance of DDR, KronRLS-MKL, NRLMF, DNILMF, BLM-NII, and COSINE for comparison with the AutoDTI++ method under the S_p CV setting. Figure 5 shows the comparison AUPR of AutoDTI++, DDR, KronRLS-MKL, NRLMF, DNILMF, BLM-NII, and COSINE on four datasets under the S_p CV setting.

We have shown that AutoDTI++, using 5-repeats of tenfold CV, achieves acceptable AUPR values than the other methods. From Fig. 5, we can see that, in terms of AUPR, under S_t setting on three datasets of NR, GPCR, and IC, the performance of the AutoDTI++ model is improved. The AutoDTI++ model on NR, GPCRs, and IC data sets performs better than DDR that is the best baseline method. The AutoDTI++ model, on the E dataset, performs better than all approaches except the DDR method. AutoDTI++ model achieves results for NR, GPCR, and IC, which respectively are 20%, 22%, and 6% higher than DDR. In terms of AUPR, under S_d the setting, the AutoDTI++ model is better than all other approaches except the DDR approach on all datasets. In terms of AUPR, under S_p the setting, the AutoDTI++ model performs better than DDR on NR and GPCRs datasets. AutoDTI++ model achieves results for NR and GPCR which are 1% and 6%, higher than DDR but for E and IC datasets, DDR method which are 10% and 2%, higher than AutoDTI++.

Case study

To evaluate the practical ability of AutoDTI++, we applied it to predict novel DTIs that are unknown in NR, GPCR, IC, and E datasets. For the prediction of novel interactions, we applied the trained model in all datasets. Then we used from the output the

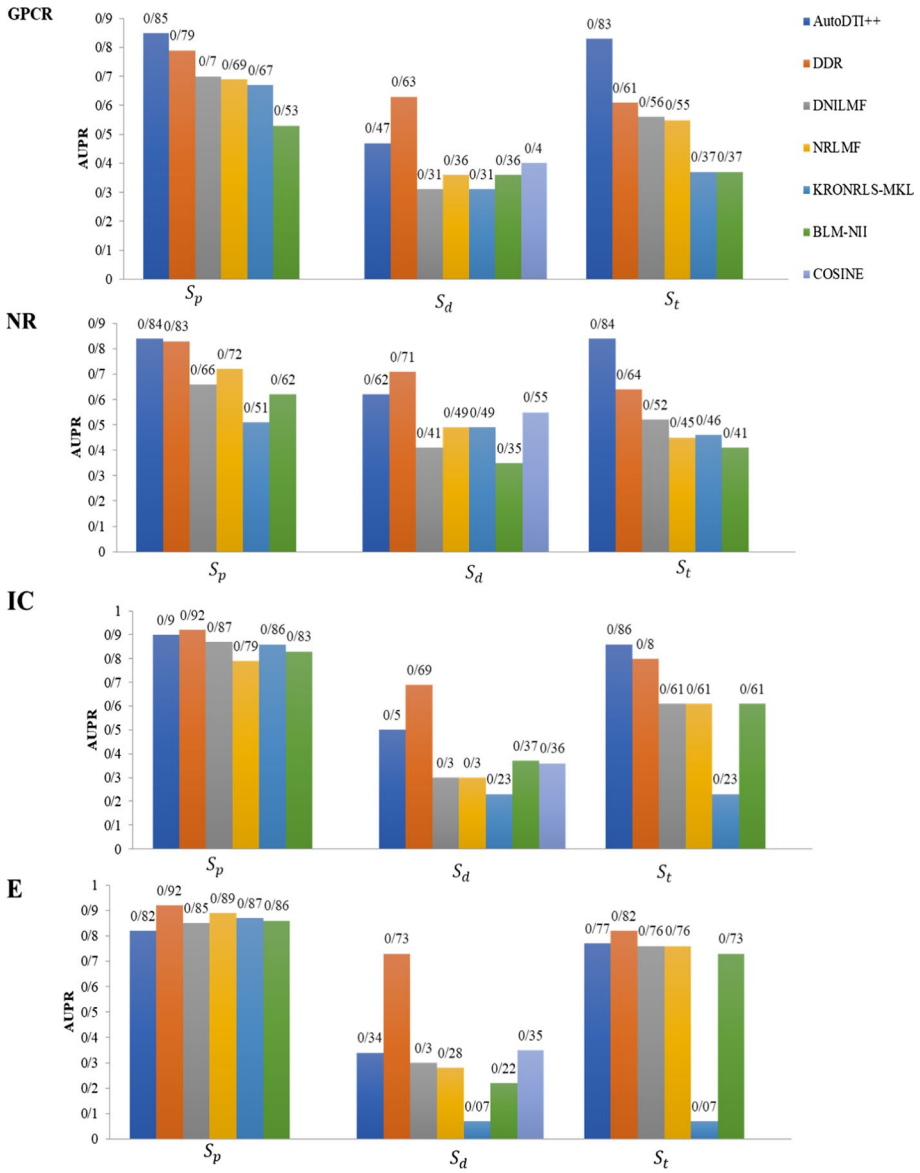

Fig. 5 Comparison results of AutoDTI++ method with the six states-of-the-art methods (DDR, DNILMF, NRLMF, KRONRLS-MKL, BLM-NII, and COSINE) in terms of AUPR scores, using 5-repeats of tenfold CV. Results are obtained under S_p, S_d, and S_t settings on NR, GPCR, IC, and E datasets. The results for DDR, DNILMF, NRLMF, KRONRLS-MKL, BLM-NII, and COSINE are captured using the best parameters published

interaction probability. The predicted probability is ranked in descending order. The high-probability drug-target pairs are predicted as novel DTIs in NR, GPCR, IC, and E datasets. We selected the top-ranked unknown DTI interaction for each dataset. To validate these new interactions, we selected several reference databases that included ChEMBL [49], DrugBank [50], KEGG [51], CTD [52], and STITCH [53]. These reference databases included many validated known DTIs obtained from experimental and published results on drug–target interactions.

The CTD reference database found drug D00217 represents acetaminophen, strongly inhibiting the enzyme cytochrome P450 2C8. AutoDTI++ also identified an interaction between D00217 and hsa1558 without a known interaction in the E dataset.

The KEGG reference database found drug D00636 that represents amiodarone hydrochloride, strongly inhibiting the target sodium voltage-gated channel alpha subunit 5. AutoDTI++ also identified the interaction between D00636 and hsa6331 without a known interaction in the IC dataset.

The DrugBank reference database found drug D02340 that represents loxapine, strongly inhibited the target dopamine receptor D1. AutoDTI++ also identified the interaction between D02340 and hsa1812 without a known interaction in the GPCR dataset.

In the ChEMBL reference database, found drug D00585 represents mifepristone strongly inhibited the target estrogen receptor 1. AutoDTI++ also identified the interaction between D00585 and hsa2099 without a known interaction in the NR dataset.

Discussion

This study introduces a novel DTI prediction method, AutoDTI++, which utilizes a denoising autoencoder for DTI prediction using a drug fingerprint-target interaction matrix. We have shown that we can achieve a more accurate prediction for different datasets by pre-processing the drug-target interaction matrix and applying it to the AutoDTI prediction model. To evaluate the proposed work, on different representative datasets, under various cross-validation settings, and using AUPR and AUC as the performance measures, we have shown that AutoDTI++ outperforms the other state-of-the-art methods that we used in the comparison. We also demonstrated that AutoDTI++ performs significantly better than the other existing methods when known DTIs are missing in the training data. We can see that AutoDTI performs worse because of the lack of additional side information and sparsity of the interaction matrix. In the proposed method, we used the drug fingerprint, which analyzes molecules as a graph and retrieves the molecular substructures from the whole molecular graph's subgraphs. Specifically, we used PaDEL-descriptor to extract a fingerprint from a raw SMILES string. Finally, each drug can be represented as a binary vector with a length of 800 whose indices indicate specific substructures' existence. In our model, the drug fingerprint provides additional information to build an interaction matrix without sparsity. Actually, if a drug interacts with a target, that target probably interacts with the substructure of that drug. Therefore, if the drug-target matrix, which is a sparse matrix, is multiplied by the drug-fingerprint matrix, which contains the drug substructure and is non-sparse, is obtained the fingerprint-target matrix, which is a non-sparse matrix and solves the problem of the sparse interaction matrix. Also, drug fingerprint adds additional information to the interaction matrix to build a more accurate model. Therefore, the AutoDTI++ model can handle the sparse interaction matrix and learn a much more effective feature vector for each drug, and our proposed model achieves much better performance. We observed that the best second method in predicting DTI in the S_p and S_t cross-validation settings and the first method in S_d cross-validation setting, in terms of the AUPR metric over the four different datasets, is the DDR method. The DDR approach utilizes a heterogeneous drug–target graph that contains information about various similarities between drugs

and similarities between proteins as drug targets. The DDR gives better results than the AutoDTI++ model, in the S_d setting. Possibly, one reason is that it uses the similarity between drugs while smoothing the predictions of new drugs by incorporating neighbor information based on the assumption that similarity may contribute to the accuracy of the predictions for their neighbors. As a result, the DDR model achieves better results in S_d cross-validation setting.

Approaches based on MF (NRLMF, DNILMF) perform worse than the AutoDTI++ model, especially in AUPR. Possibly, one reason is that AutoDTI++ can learn a non-linear latent representation through sigmoid activation function while MF models learn a linear latent representation. Therefore our proposed method learns sufficient and effective features by autoencoders neural networks to detect true DTIs. Also, a good advantage of using autoencoders in the AutoDTI++ approach is that they can fill in every vector that is not present in training data that leads to the superiority of the AuoDTI++ over the MF method. Another reason might be that MF approaches embed both drugs and targets into a shared latent space, but the AutoDTI++ model only embeds the target into latent space and uses the drug fingerprint feature.

In terms of AUPR, AutoDTI++ performs on IC better than E, NR, and GPCR datasets, possibly because IC has less sparsity than other datasets on matrix interaction. GPCR and NR have sparsity approximately the same, but NR is a little better than GPCRs, possibly because the number of targets affects results. Regarding a dataset, the input vector with a less number of targets is more suitable. Because the input vector with a larger number of targets is more sparsity difference, that results in an imbalance model. E dataset performs wost than other datasets because it has more sparsity in between all datasets.

Conclusions

We proposed a novel method called AutoDTI++ to predict DTIs based on autoencoders. Our proposed approach includes three steps. The first step consists of a pre-processing step that transforms the binary values in the given drug-target matrix to the binary values in the drug fingerprint-target interaction matrix for filling missing values based on drug fingerprint. The second step proposed an AutoDTI model that uses an unsupervised deep learning technique based on denoising autoencoders to predict interactions, and the third step is post-preprocessing. Subsequent pre-processing is applied to the AutoDTI model, and it achieves better performance. Experimental results show that the AutoDTI++ model achieves significantly more accurate results than the other state-of-the-art methods under cross-validations S_p, S_d, and S_t on NR, GPCR, IC, and E datasets, and different metrics of performance evaluation. As future work, first, we plan to expand our model by adding some additional information, such as amino acid sequences of target proteins. Second, we will develop our models to incorporate some additional information, such as similarity drugs and targets matrix, to solve the interaction matrix's sparsity problem. Finally, we will combine our models with other models of autoencoders.

Abbreviations
DTI: Drug-target interaction; CV: Cross-validation; MF: Matrix factorization; DBN: Deep belief network; SVM: Support vector machine; ECFP: Extended-connectivity fingerprints; CNN: Convolutional neural network; DAE: Denoising autoencoder; NR: Namely nuclear receptors; GPCR: G protein-coupled receptors; IC: Ion channels; E: Enzymes; AUC: Area under the receiver operating characteristic curve; ROC: Receiver operating characteristic; AUPR: Area under precision-recall curve; NII: Neighbor-based interaction-profile inferring; BLM: Bipartite local model; LTN: Linked tripartite network; SNN: Shared nearest neighbor; FRA: Fuzzy-rough approximation; LSTM: Long short-term memory.

Acknowledgements
Not applicable.

Authors' contributions
SZS developed and implemented the method, conducted the experiments, and wrote the manuscript. MAZCH and SGH conceptualized the study, interpreted the results, supervised the work, administered the project, and edited the manuscript. KA validated the work and edited the manuscript. All authors have read and approved the final manuscript.

Funding
This work was not supported by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets used in this project can be found in http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Computer Engineering, Yazd University, Yazd, Iran. [2] Laboratory of Bioinformatics and Drug Design (LBD), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

Received: 17 February 2021 Accepted: 9 April 2021
Published online: 20 April 2021

References

1. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug–target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol. 2019;15(6):e1007129.
2. Zhou L, Li Z, Yang J, Tian G, Liu F, Wen H, Peng L, Chen M, Xiang J, Peng L. Revealing drug–target interactions with computational models and algorithms. Molecules. 2019;24(9):1714.
3. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y. Drug–target interaction prediction: databases, web servers and computational models. Brief Bioinform. 2016;17(4):696–712.
4. Bagherian M, Sabeti E, Wang K, Sartor MA, Nikolovska-Coleska Z, Najarian K. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. Brief Bioinform. 2021;22(1):247–69.
5. Ding H, Takigawa I, Mamitsuka H, Zhu S. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Brief Bioinform. 2014;15(5):734–47.
6. Abbasi K, Razzaghi P, Poso A, Ghanbari-Ara S, Masoudi-Nejad A, Deep learning in drug target interaction prediction: current and future perspective. Curr Med Chem 2020.
7. Hendrickson JB. Concepts and applications of molecular similarity. Science. 1991;252(5009):1189–90.
8. Jacob L, Vert J-P. Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics. 2008;24(19):2149–56.
9. Chen Y, Zhi D. Ligand–protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins Struct Funct Bioinform. 2001;43(2):217–26.
10. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat Rev Drug Discov. 2004;3(11):935–49.
11. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M. Drug—target network. Nat Biotechnol. 2007;25(10):1119–26.
12. Opella SJ. Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. Annu Rev Anal Chem. 2013;6:305–28.
13. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):i232–40.
14. Chen X, Liu M-X, Yan G-Y. Drug–target interaction prediction by random walk on the heterogeneous network. Mol BioSyst. 2012;8(7):1970–8.
15. Islam SM, Hossain SMM, Ray S. DTI-SNNFRA: Drug-target interaction prediction by shared nearest neighbors and fuzzy-rough approximation. PLoS ONE. 2021;16(2):e0246920.

16. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–8.
17. Lim J, Ryu S, Park K, Choe YJ, Ham J, Kim WY. Predicting drug–target interaction using a novel graph neural network with 3D structure-embedded graph representation. J Chem Inf Model. 2019;59(9):3981–8.
18. Zong N, Kim H, Ngo V, Harismendy O. Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics. 2017;33(15):2337–44.
19. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H. Deep-learning-based drug–target interaction prediction. J Proteome Res. 2017;16(4):1401–9.
20. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.
21. Hu P-W, Chan KC, You Z-H, Large-scale prediction of drug–target interactions from deep representations. In: 2016 international joint conference on neural networks (IJCNN): 2016. IEEE: pp. 1236–1243.
22. Tian K, Shao M, Wang Y, Guan J, Zhou S. Boosting compound-protein interaction prediction by deep learning. Methods. 2016;110:64–72.
23. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug–target binding affinity prediction. Bioinformatics. 2018;34(17):i821–9.
24. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. J Chem Inf Model. 2014;54(3):735–43.
25. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol. 2011;29(11):1046–51.
26. Nascimento AC, Prudêncio RB, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinform. 2016;17(1):46.
27. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M. SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminform. 2017;9(1):1–14.
28. Günther S, Kuhn M, Dunkel M, Campillos M, Senger C, Petsalaki E, Ahmed J, Urdiales EG, Gewiess A, Jensen LJ. SuperTarget and Matador: resources for exploring drug-target relationships. Nucl Acids Res. 2007;36(suppl_1):D919–22.
29. Abbasi K, Razzaghi P, Poso A, Amanlou M, Ghasemi JB, Masoudi-Nejad A. DeepCDA: deep cross-domain compound–protein affinity prediction through LSTM and convolutional neural networks. Bioinformatics. 2020;36(17):4633–42.
30. Zheng X, Ding H, Mamitsuka H, Zhu S: Collaborative matrix factorization with multiple similarities for predicting drug–target interactions. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining: 2013, pp. 1025–1033.
31. Ezzat A, Zhao P, Wu M, Li X-L, Kwoh C-K. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinf. 2016;14(3):646–56.
32. Olayan RS, Ashoor H, Bajic VB. DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics. 2018;34(7):1164–73.
33. Hao M, Bryant SH, Wang Y. Predicting drug–target interactions by dual-network integrated logistic matrix factorization. Sci Rep. 2017;7(1):1–11.
34. Liu Y, Wu M, Miao C, Zhao P, Li X-L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput Biol. 2016;12(2):e1004760.
35. Mei J-P, Kwoh C-K, Yang P, Li X-L, Zheng J. Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics. 2013;29(2):238–45.
36. Lim H, Gray P, Xie L, Poleksic A. Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Sci Rep. 2016;6(1):1–11.
37. Bahi M, Batouche M: Deep semi-supervised learning for DTI prediction using large datasets and H2O-spark platform. In: 2018 international conference on intelligent systems and computer vision (ISCV): 2018. IEEE: 1–7.
38. Zhou Y, Arpit D, Nwogu I, Govindaraju V: Is joint training better for deep auto-encoders? https://arxiv.org/abs/1405.1380 2014.
39. Goodfellow I, Bengio Y, Courville A, Bengio Y. Deep learning, vol. 1. Cambridge: MIT Press; 2016.
40. Bishop CM. Neural networks for pattern recognition. Oxford: Oxford University Press; 1995.
41. Miranda V, Krstulovic J, Keko H, Moreira C, Pereira J. Reconstructing missing data in state estimation with autoencoders. IEEE Trans Power Syst. 2011;27(2):604–11.
42. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inform Comput Sci. 1988;28(1):31–6.
43. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
44. Salakhutdinov R, Mnih A, Hinton G: Restricted Boltzmann machines for collaborative filtering. In: Proceedings of the 24th international conference on Machine learning: 2007, pp 791–798.
45. Sedhain S, Menon AK, Sanner S, Xie L: Autorec: Autoencoders meet collaborative filtering. In: Proceedings of the 24th international conference on World Wide Web: 2015, pp 111–112.
46. Pahikkala T, Airola A, Pietilä S, Shakyawar S, Szwajda A, Tang J, Aittokallio T. Toward more realistic drug–target interaction predictions. Brief Bioinform. 2015;16(2):325–37.
47. Raghavan V, Bollmann P, Jung GS. A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inform Syst (TOIS). 1989;7(3):205–29.
48. Davis J, Goadrich M: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd international conference on Machine learning: 2006, pp 233–240.
49. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B. ChEMBL: a large-scale bioactivity database for drug discovery. Nucl Acids Res. 2012;40(D1):D1100–7.
50. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucl Acids Res. 2010;39(suppl_1):D1035–41.
51. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucl Acids Res. 2017;45(D1):D353–61.

52. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC, Mattingly CJ. The comparative toxicogenomics database: update 2017. Nucl Acids Res. 2017;45(D1):D972–8.
53. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. Nucl Acids Res. 2007;36(suppl_1):D684–8.

Publisher's Note