Molecular Diversity
https://doi.org/10.1007/s11030-022-10492-8

ORIGINAL ARTICLE

Matrix factorization with denoising autoencoders for prediction of drug–target interactions

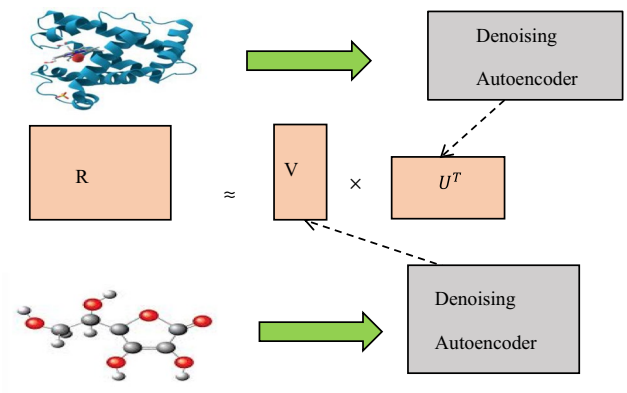
Seyedeh Zahra Sajadi[1] · Mohammad Ali Zare Chahooki[1] · Maryam Tavakol[2] · Sajjad Gharaghani[3]

Received: 16 April 2022 / Accepted: 1 July 2022

Abstract

Drug–target interaction is crucial in the discovery of new drugs. Computational methods can be used to identify new drug–target interactions at low costs and with reasonable accuracy. Recent studies pay more attention to machine-learning methods, ranging from matrix factorization to deep learning, in the DTI prediction. Since the interaction matrix is often extremely sparse, DTI prediction performance is significantly decreased with matrix factorization-based methods. Therefore, some matrix factorization methods utilize side information to address both the sparsity issue of the interaction matrix and the cold-start issue. By combining matrix factorization and autoencoders, we propose a hybrid DTI prediction model that simultaneously learn the hidden factors of drugs and targets from their side information and interaction matrix. The proposed method is composed of two steps: the pre-processing of the interaction matrix, and the hybrid model. We leverage the similarity matrices of both drugs and targets to address the sparsity problem of the interaction matrix. The comparison of our approach against other algorithms on the same reference datasets has shown good results regarding area under receiver operating characteristic curve and the area under precision–recall curve. More specifically, experimental results achieve high accuracy on golden standard datasets (e.g., Nuclear Receptors, GPCRs, Ion Channels, and Enzymes) when performed with five repetitions of tenfold cross-validation.

Graphical abstract



Display graphical of the hybrid model of Matrix Factorization with Denoising Autoencoderswith the help side information of drugs and targets for Prediction of Drug-Target Interactions

Keywords Drug–target interactions prediction · Deep learning · Hybrid model · Latent feature · Denoising autoencoder

Introduction

One of the main areas of drug discovery and repositioning is identifying drug–target interactions [1]. Re-using drugs, which have been approved by the FDA and whose safety

✉ Mohammad Ali Zare Chahooki
chahooki@yazd.ac.ir

Extended author information available on the last page of the article

Published online: 23 July 2022

profiles are readily available, for new interactions not only reduce costs but also decrease safety risks [2]. However, despite various technologies for biological assays, there are still limitations related to large-scale drug–target interactions (DTI). In addition, due to its high cost and lack of public experiments for drug repositioning, developing effective computational methods that can precisely detect drug–target interactions is essential.

There are three basic categories of computational approaches that have been developed for predicting new DTIs today [3]. The first category makes use of similar molecules and similar ligands of target proteins [4], known as the ligand-based approach. In some cases, ligand-based methods may not provide accurate results if there are not enough known ligands for a target [5]. The second category, namely molecular docking, leverages the 3D structure of the target proteins for the small molecule screening [6]. This approach is limited by the requirement of the target protein's 3D structures [7]. These methods cannot predict new drug–target pairs if 3D structures of proteins cannot be derived. Predicting the 3D structure of most targets, especially those related to membrane proteins, such as GPCRs, is a challenge [8]. The third category, denoted by chemogenomic approaches, utilizes the information of both target and drug together to predict DTI. A benefit of chemogenomic is the ability to access data from many online public databases. For instance, Wen et al. [9] conducted DTI prediction using the chemical structure graphs of drugs and genomic sequences of targets, which can be easily obtained from online databases. Unlike the other two approaches, this approach is free of the mentioned limitations. For DTI predictions, methods based on chemogenomic typically use machine-learning and deep-learning methods.

As experimental data have grown, deep-learning methods have gained popularity for predicting DTIs [10]. Due to the ability of deep-learning approaches to extract useful features derived from the input data and build complex models that can capture even difficult patterns in DTIs, they are preferred over alternative methods. Deep learning has been used in several studies to learn automatically high-level feature representation from the training data and is beneficial in many bioinformatics tasks [11–14]. Deep-learning approaches are commonly used to solve the problem of DTI prediction, which is modeled as a supervised classification problem. The features extracted from a drug–target pair are taken as the input, and then the interaction between the drug–target pair (DTP) is predicted as the output. In a paper by Wen et al. [9], a deep-learning method, named DeepDTI, is adapted using a deep belief network (DBN) for predicting the affinity value for pairs of drugs and targets. The features of drugs can be extracted automatically from extended-connectivity fingerprints (ECFP), and the features of target proteins have been extracted from amino acids, dipeptides, and tripeptides. Subsequently, Zeng et al. [15] utilize ten networks to predict DTI using a deep-learning-based method called DeepDR. Peng et al. [16] present a method for learning hidden latent features from RNA and protein sequences, using stacked autoencoders, and then training a support vector machine (SVM) on this representation. In addition, Fu et al. [13] employ stacked autoencoders to automatically learn high-level features of miRNAs and diseases, which are used in Deep Neural Network (DNN) to predict miRNA disease associations. Ozturk et al. [17] develop an approach to predict DTI using a convolutional neural network (CNN) to learn the features of drugs and proteins. Gligorijević et al. [14] propose a multimodal deep autoencoder (MDA) based on deep learning, in which multiple networks are merged to learn low-dimensional protein features using MDA. They train an SVM to predict protein functions from low-dimensional protein features. Moreover, Lee et al. [18] present the DeepConv-DTI model that predicts massive-scale DTIs based on raw protein sequences for several target protein classes and varying lengths. In their approach, convolution filters are applied to the whole sequence of the protein to catch patterns of local residues. Subsequently, to predict the affinity values, protein features and drug features are concatenated and then fed into fully connected layers. The mentioned work confirms that deep learning is capable of learning high-level features from original data in a very efficient and effective manner, which improves the performance of the methods and enabled them to achieve acceptable results.

Matrix Factorization (MF) approaches to DTI prediction learn effective latent factors directly from the drug–target interaction matrix. However, due to the high sparsity of the interaction matrix learning, the appropriate latent factors are significantly compromised in such methods. Additionally, the cold-start problem can limit the use of MF-based methods for predicting interactions when a new drug or target arrives in the system. In this paper, we benefit from additional side information to overcome these issues in MF-based methods. The side information can be gained from drug and target content information, such as the drug chemical structure, protein amino acid sequence, etc. Other works have already integrated side information into matrix factorization to determine effective latent factors via a hybrid MF setup [1, 19–21]. Even so, these approaches use the side information as regularizations and learned latent factors often are insufficient, specifically when the interaction matrix is very sparse. Thus, such information can be valuable in resolving the latent factor problem. Therefore, we extend the hybrid matrix factorization model with a deep structure to fully explore the latent space of the features and address the above-mentioned challenges. The deep-learning model in this framework is called the additional

stacked denoising autoencoder (aSDAE) [22], which incorporates the side information into the deep structure input for addressing the cold-start and data sparsity problems. Using this approach, we develop the hybrid model that couples deep-learning representation from the additional side information and matrix factorization from the interaction matrix. We future evaluate our proposed method compared to six other state-of-the-art methods, namely DDR [23], DNILMF [24], NRLMF [21], KronRLS-MKL [25], BLM-NII [26], and COSINE [27] via cross-validation (CV), in which, the performance of approaches are assessed three settings: new drugs, new interactions, and new targets. Our proposed method shows improved results compared to most previous models when applied to new target and drug cases (by excluding their interaction) to test.

Methods

Preliminaries

In this section, we first discuss how to formulate the problem discussed in this paper, and then followed by a brief review of Matrix Factorization and Additional Stacked Denoising Autoencoder.

Problem definition

The datasets are composed of three matrices : $R \in \mathbb{R}^{m \times n}$,$S^d \in \mathbb{R}^{m \times m}$, and $S^t \in \mathbb{R}^{n \times n}$. Drug–target interactions are encoded by the sparse interaction matrix R, which is made up of m drugs as rows and n targets as columns. Each entry $R_{ij}= 1$ of$R$, means that the drug i has an interaction with target$j$, otherwise $R_{ij}= 0$. Each drug or target interaction profile is specified by R_d andR_t, respectively. In the case of each drug d, a partially observed vector $R_d = (R_{d1}, \ldots ..R_{dn}) \in \mathbb{R}^n$ can be described. Identically, In the case of each target t, a partially observed vector $R_t = \left(R_{1t}, \ldots ..R_{mt}\right) \in \mathbb{R}^m$ can be described. The matrix S^d expresses the similarities between drug pairwise chemical structures, and the matrix S^t expresses the similarities between target pairwise genomic sequences. For drugs similarity, SIMCOMP scores are used [28], whereas Smith-Waterman scores are used for targets similarity [29]. Moreover, the side information matrix of drugs and targets are indicated by $X \in \mathbb{R}^{m \times p}$ and$Y \in \mathbb{R}^{n \times q}$, respectively.

Let u_i, $v_j \in \mathbb{R}^k$ be latent factor vector of drug i and target$j$, respectively, and latent space is characterized by k dimensions. Thus, to determine the matrix forms of latent factors for drugs and targets, we have $U = u_{1:m}$ and$V = v_{1:n}$, respectively. By learning drug and target latent factors U and V from the sparse interaction matrix R and the side information matrix X and$Y$, it is possible to predict the missing interaction in$R$.

Matrix factorization

By factorizing the interaction matrix, matrix factorization can map both drugs and targets to a joint latent factor space [30]. Therefore, drug–target interactions are modeled as inner products in that latent factor space. In the case of DTI prediction, when matrix factorization is performed, the original interaction matrix R is split into two low-rank matrices $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$, consisting of the drug and target latent factor vectors, respectively, such that $R \approx UV^T$. Matrix factorization can identify latent features of drugs and targets in an unsupervised manner. A matrix factorization technique may be effective for finding missing interactions in a matrix R, making it appropriate for DTI prediction. The objective function of matrix factorization is

$$\text{argmin}_{U,V}\mathcal{L}\left(R, UV^T\right) + \lambda\left(\|U\|_f^2 + \|V\|_f^2\right), \tag{1}$$

Loss function $\mathcal{L}(.,.)$ measures the distance between two matrices with the same size, the Regularizations are used to prevent overfitting by the two last terms and $\|.\|_f$ denotes the Frobenius norm.

Additional stacked denoising autoencoder

Denoising autoencoders (DAE) are neural networks that unsupervised encode input data without requiring labels as ground truths [31].The autoencoder is composed of two networks: one encoder and one decoder that take an input domain as input and then reconstruct it [32]. As the encoder $g(.)$ converts the input Y to a hidden factor $g(Y)$, the decoder $f(\cdot)$ converts this hidden factor back to a reconstructed version of Y, such that $f(g(Y)) \approx \widehat{Y}$. Several autoencoders have recently been presented, including the denoising autoencoder, sparse autoencoder, and variational autoencoder [33]. Noise is injected to the input in denoising autoencoders, forcing the network to reconstruct the denoised input [34]. One approach for adding noise is to replace random fractions of the input with zeros. Here, we use from additional denoising autoencoder (aDAE) that extends the denoising autoencoder by adding additional side information into the input data [22]. The additional denoising autoencoder (aDAE) considers random corruptions over Y and X to obtain $\widehat{Y}$ and $\widehat{X}$ if $Y = [y_1, \ldots y_n]$ and $X = [x_1, \ldots .x_n]$. A stacked denoising autoencoder (SDAE) stacks several denoising autoencoders together to form a higher-level representation [35]. We stacked multiple aDAE together to form an additional

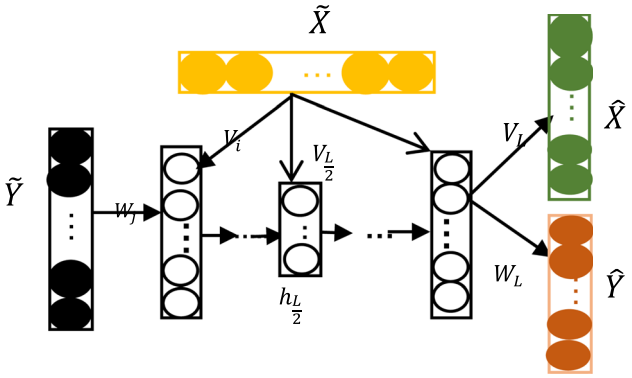


Fig. 1 The model of aSDAE [22]

stacked denoising autoencoder (aSDAE). The inputs are encoded and decoded for aDAE model as follows:

$$\begin{aligned} h &= g\left(W_1\tilde{Y} + V_1\tilde{X} + b_1\right) \\ \hat{Y} &= f\left(W_1 h + b_{\hat{s}}\right) \\ \hat{X} &= f(V_1 h + b_{\hat{x}}) \end{aligned} \quad . \tag{2}$$

A hidden representation h_l is computed for each hidden layer $L\epsilon\{1,\ldots,L-1\}$ of the aSDAE model as follows:

$$\begin{aligned} h &= g\left(W_1\tilde{Y} + V_1\tilde{X} + b_1\right) \\ \hat{Y} &= f\left(W_l h_l + b_{\hat{s}}\right) \\ \hat{X} &= f(V_l h_l + b_{\hat{x}}) \end{aligned} \quad . \tag{3}$$

$\widetilde{Y}$ and $\widetilde{X}$ stand for the corrupted version of Y and X, $\widehat{Y}$ and $\widehat{X}$ represent the reconstructions of Y and X, h stands for the latent representation of the inputs, b is bias vector and W and V are weight matrices, and $g(\cdot)$ and $f(\cdot)$ are activation functions such as $sigmoid(\cdot)$. Figure 1 shows the model of aSDA. Theoretically, the aSDAE model's objective function is formulated as follows:

$$\text{argmin}_{W,V,b}\alpha\|Y-\hat{Y}\|_f^2 + (1-\alpha)\|X-\hat{X}\|_f^2 + \lambda\left(\sum\|w_l\|_f^2 + \|V_l\|_f^2\right). \tag{4}$$

Here α is a trade-off parameter used to balance the outputs, while λ is a regularization parameter. Using the back-propagation algorithm, we can learn W_l, V_l, and b_l for each layer. The aSDAE reconstructs inputs using a deep network and minimizes the squared loss between inputs and their reconstructions. Since there are L layers in total, latent factors are derived from the $L/2$ layer.

Proposed method

The proposed method for predicting DTI is explained in this section, which involves two steps:

1. Pre-processing is the first step, which involves converting binary values in DTI matrix, R, into interaction likelihood values;
2. In the second step, we will propose our hybrid model to predict DTIs that uses matrix factorization coupled with stacked denoising autoencoders.

Pre-processing step

Creating a model for DTI can be challenging since, although interactions (positive values $R_{ij} = 1$) are known, some of the non-interactions (or 0's) in R may actually be true interactions. Therefore, in most approaches [36], negative samples are selected at random from the unknown data. However, this might lead to inaccurate results. Due to the fact that the drug–target interaction matrix is based only on interactions between drugs and targets, only using the interaction matrix when additional information regarding the drugs and the targets is available can sound restrictive. As a result, we intend to solve this issue by adding information on similarity matrix of drugs and targets to the interaction matrix. In order to calculate interaction likelihood values for these unknown instances, we adapt the WKNKN approach, which was based on the procedure described in [36] as a pre-processing step that uses similarity matrices of drugs and targets. Therefore, if R_{ij} is 0, WKNKN replaces it with a continuous value between 0 to 1. It is important to note, WKNKN uses K nearest known neighbors to infer the likelihood value of interactions between drug and target pairs.

The hybrid model

We develop a hybrid model with a combination aSDAE and matrix factorization that using from both interaction matrix and side information of drug and target for DTI prediction. Matrix factorization is one of the most widely used collaboration filtering methods, with good scalability and accuracy, and SDAE extracts high-level representations from

raw inputs. These two models are merged to create a more expressive learning model leveraging their benefits.

It is assumed that the input of model is a drug–target interaction matrix R, we first encode R into the set R^d containing m instances $\{R_1^{(d)}, \ldots, R_m^{(d)}\}$, where $R_i^{(d)} = \{R_{i1}, \ldots, R_{in}\}$ represent interaction of drug i on all the targets. Similarly, we can encode set $R^{(t)}$ with n instances $\left\{R_1^{(t)}, \ldots R_n^{(t)}\right\}$, where $R_j^t = \left\{R_{1j}, \ldots .R_{mj}\right\}$ represent interaction of target j on all the drugs. Let $\widetilde{R}^d$ and $\widetilde{R}^t$ denote their corrupted versions, respectively. In addition, $X \in \mathbb{R}^{m \times p}$ and $Y \in \mathbb{R}^{n \times q}$ are the additional side information also drug chemical structure and protein sequence composition descriptors (PSC) matrices, respectively, and corrupted versions are $\widetilde{X}$ and $\widetilde{Y}$.

We select the most common and simple features of drugs and targets in the present paper, representing the drugs with molecules by SMILES (simplified molecular-input line-entry system) and targets with sequence composition descriptors. The fingerprints of drugs can be used as additional information about drugs to the proposed method. SMILES [37] strings, a sequential encoding of chemical structures, are used to represent each drug in the first step. This is followed by using the PaDEL-descriptor software to create fingerprints from SMILES strings. With PaDEL-descriptor, molecular descriptors (1D, 2D, and 3D) and ten different kinds of fingerprints can be computed [38]. Drugs can be described by a binary vector with an index indicating the existence of specific substructures, with a length of 800. Also, our approach uses protein sequence composition as additional side information of targets. There are three major components to PSC: amino acid composition (AAC), dipeptide composition (DC), and tripeptide composition (TC). Every frequency of amino acids is called the AAC. Every two amino acid combinations have a statistical frequency called the DC. Every three amino acid combinations have a statistical frequency called the TC. Open-Source Software Propy [39] is used to calculate the protein descriptors. Each protein sequence composition is described by a 567-dimensional feature vector.

The inputs of the hybrid model are $\widetilde{R}^d$, $\widetilde{R}^t$, $\widetilde{X}$, $\widetilde{Y}$, and R. In the hybrid model, drug and target latent factors (i.e., U and V) are learned from R, $\widetilde{R}^d$, $\widetilde{R}^t$, $\widetilde{X}$ and $\widetilde{Y}$.

Our hybrid model is shown in Fig. 2 and then is formulated as follows:

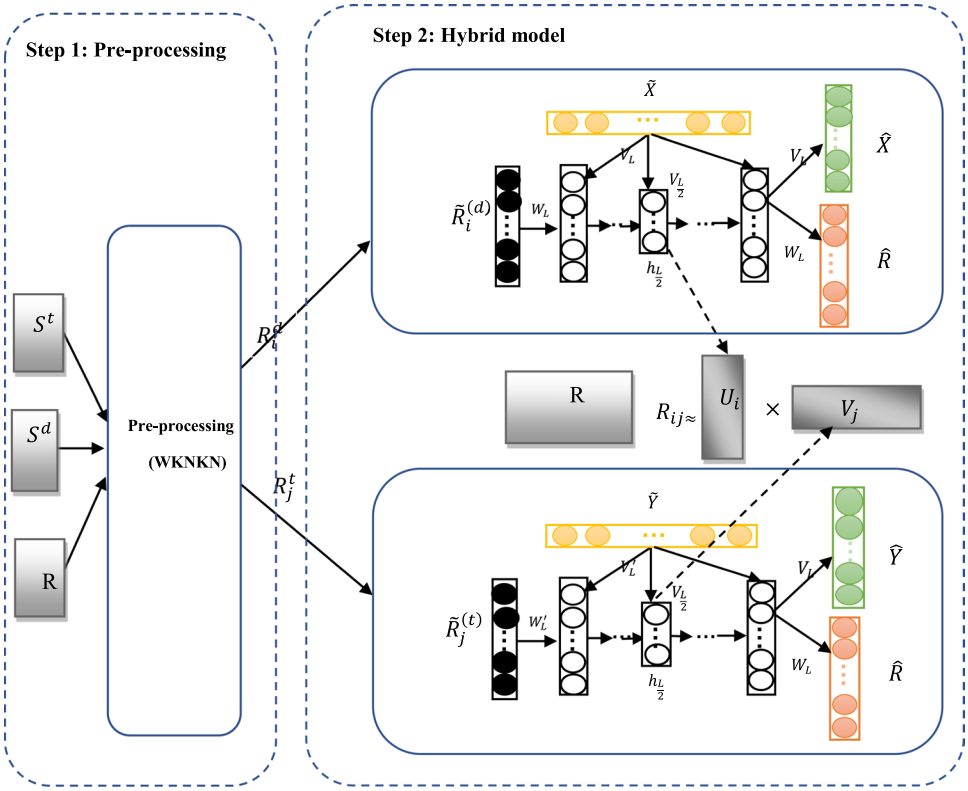


Fig. 2 The structure of proposed hybrid model and WKNKN pre-processing method. Three components make up the hybrid model: the upper component, the lower component, and middle component. The upper component and a lower component, which extract latent factors from drugs and targets, respectively; the middle component decomposes R into two latent factors

Table 1 Drugs, targets, interactions, and sparsity in each benchmark 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	0.03	0.034	0.01

$$
\begin{aligned}
L = & \sum_{i,j} I_{ij}\left(R_{ij} - U_i V_j^T\right)^2 + \alpha_1 \sum_i \left(R_i^{(d)} - \hat{R}_i^{(d)}\right)^2 \\
& quad + \left(1-\alpha_1\right) \sum_i \left(X_i - \hat{X}_i\right) + \alpha_2 \sum_j \left(R_j^{(t)} - \hat{R}_j^{(t)}\right)^2 \\
& + \left(1-\alpha_2\right) \sum_j \left(Y_j - \hat{Y}_j\right)^2 + \lambda \cdot \text{freg} \quad . \\
f_{\text{reg}} = & \sum \|U_i\|_f^2 + \sum \left\|V_j\right\|_f^2 \\
& + \sum_l \left(\|W_l\|_f^2 + \|V_l\|_f^2 + \|b_l\| + \left\|W_l'\right\|_f^2 + \left\|V_l'\right\|_f^2 + \left\|b_l'\right\|_f^2 \right)
\end{aligned}
\tag{5}
$$

The first term of the loss function is applied to decompose the interaction matrix R into V and U latent factor matrices. Here, I is an indicator matrix that shows the non-empty entities in R. Our aSDAE model extracts latent factor matrices from the interaction matrix, along with drug and target features, as in the last four terms of loss functions are represented, respectively.

Here,α_1, α_2 are the trade-off parameters, and f_{reg} is the regularizing terms that prevent the model from overfitting.W_l, V_l and W_l', V_l' represent the weight matrices of aSDAE at layer$l$. Also, b_l and b_l' are the bias vectors. For the regularization parameter is used fromλ.

In most cases, the middle layers of two aSDAEs act as a link between the interaction matrix and the feature of drugs and targets. These two middle layers are the key ability of our hybrid model to learn latent factor variables while also capturing the relationship between drug and target.

Prediction

We estimate the predicted interaction R_{ij} as $R_{ij\approx} U_i V_j^T$ after learning the latent factors for each drug and target and then construct a list of targets for each drug based on these prediction interactions.

Results

In this section, initially, we introduce the dataset. Second, we describe CV and the metrics to evaluate our model. Third, parameter settings are presented. And then we compare our model with some baseline approaches. Ultimately, the performance of the model is presented by comparing it with the baselines.

Dataset

We evaluate our proposed approach by using the benchmark dataset introduced in [40]. The target proteins in this dataset are nuclear receptors (NR), G protein-coupled receptors (GPCR), ion channels (IC), and enzymes (E). For each dataset, we present a few statistics in Table 1. These include the number of unique proteins, number of unique drugs, and number of interactions and as well as the sparsity coefficient, which relates the number of known DTIs to the total number of DTIs.

Cross-validation experiments

Using the three procedures represented in [41], we perform a cross-validation investigation to make a complete evaluation of different methods:

(1) S_p refer to random pairs of target and drug which are ignored and considered as the test set;
(2) S_d, To refer to the whole drug interaction profiles, which are ignored and considered as the test set; and.
(3) S_t refer to the whole target interaction profiles, which are ignored and considered as the test set.

Traditionally, performance is evaluated using S_p. In addition, we assess various approaches for predicting new

Table 2 Results of the proposed approach based on AUC and AUPR validation metrics, under three different CV settings S_p, S_d, S_t and datasets NR, GPCR, IC, and E by five repeats of tenfold CV

Hybrid model	NR	GPCR	IC	E
S_p				
AUPR	0.93	0.94	0.98	0.97
AUC	0.90	0.98	0.99	0.99
S_d				
AUPR	0.65	0.55	0.56	0.54
AUC	0.69	0.56	0.58	0.63
S_t				
AUPR	0.57	0.53	0.60	0.77
AUC	0.67	0.63	0.61	0.79

interactions drugs and targets using S_d and S_t CV. In this case, drugs and targets that have no interaction information in the training set are considered new ones. So, conducting an experiment under S_d and S_t provide insight into the generalizability of the proposed approach. Similar to previous studies, for evaluation of prediction performance, we use area under receiver operating characteristic (AUC) and area under precision–recall curve (AUPR). To compare our proposed method with current methods, we conduct experiments to compare it to DDR, DNILMF, NRLMF, KRON-RLS-MKL, BLM-NII, and COSINE. We evaluate the performance of DTIs prediction methods by performing five repeats of tenfolds cross-validation, and we use AUC as well as AUPR [42] as evaluation metrics. Each fold in the interaction dataset was treated as the test set, whereas the rest of the nine folds were applied as the training set. The AUPR scores are calculated by averaging the results of five repetitions. To evaluate performance, we use AUPR as our primary metric in all experiments. Since the AUPR punishes wrong interactions strongly, it is an appropriate measure [43].

Parameter settings

A benchmark database [40] is used to conduct experiments. We modify the number of hidden units and layers in our proposed model to evaluate its performance. Our study found that performance increases regularly with two hidden layers. From a sigmoid activation function is used in each layer. Using a nonlinear activation function in the hidden layer is necessary for the hybrid model to perform well. We set the parameters α_1 and α_2 to 0.6 and 0.4, respectively. We set the regularization coefficient (λ) to 10^{-6} and also the learning rate to 0.001. We do finetune by gradient-based backpropagation with a minibatch of 100 samples. We select the number of latent features (k) in the hybrid model for IC, GPCR, and E datasets to 25, and the NR dataset to 15. The epochs of neural network frameworks are at 200.

Comparisons with the state-of-the-art algorithms

The proposed hybrid method computes AUC and AUPR scores on NR, GPCR, IC, and E datasets. The AUPR and AUC scores for S_p, S_d, and S_t test sets are shown in Table 2. The ROC and AURP curves of the first repeat of tenfold cross-validation on four datasets are shown in Fig. 3. The average AUC and average AUPR of our hybrid model in the first repeat of tenfold CV are the mean-AUC and mean-AUPR.

Baseline approaches

We compare six state-of-the-art DTI prediction methods, including DDR, DNILMF, NRLMF, KronRLS-MKL, BLM-NII, and COSINE, to our hybrid model for prediction performance on NR, GPCR, IC, and E datasets with three various scenarios of CV.

DDR

In DDR, a heterogeneous network is used to represent not only the known DTIs, but also multiple similarities between drugs and targets. The DDR method combines different similarities through a nonlinear fusion method. As a preprocessing step, DDR selects a subset of similarities in a heuristic method to produce the optimal combination of similarities before fusion. After that, manual extraction of different graph-based features is performed from the DTI heterogeneous graph. Ultimately, random forest (RF) is used to predict DTIs from feature matrices.

KronRLS-MKL

The first step is to combine multiple drug kernels and target kernels to obtain the final drug kernel and target kernel. KronRLS analyzes the entire drug–target space and uses the Kronecker product algebraic properties to do so, without explicitly calculating the pairwise kernels. Ultimately, it predicts DTIs using Kronecker's regularized least squares.

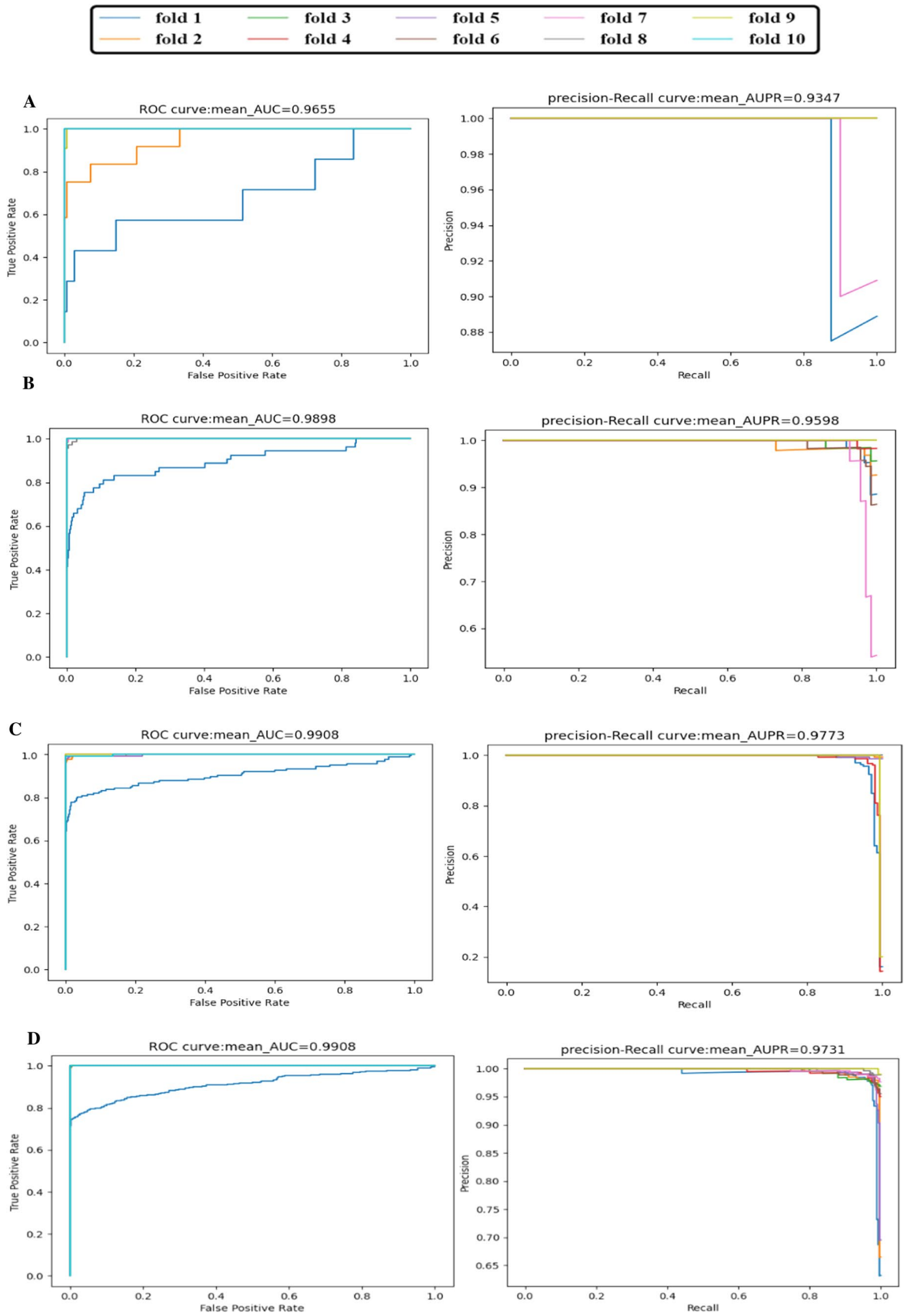
NRLMF

In NRLMF, features of drugs and features of targets are modeled as latent vectors in latent space that are shared in a low-dimensional manner. The interaction probability between each drug and target is estimated using a logistic function of their latent vectors. In addition, the neighborhod regularization technique is used to enhance the ability of the model to predict DTIs based on the local structure of the DTIs data.

BLM-NII

A BLM-NII approach integrates neighbor-based interaction-profile inferring (NII) with the bipartite local models (BLM) to provide DTI predictions that are based on the RLS classifier and GIP kernel.

In our paper, we demonstrate that our hybrid model, using five repeats of tenfold CV, generates better AUPR results than other methods. According to Fig. 4, based on AUPR metric, under the scenario CV of S_p, the hybrid model on every four datasets conduct better than DDR that is the most


fold 1
fold 2
fold 3
fold 4
fold 5
fold 6
fold 7
fold 8
fold 9
fold 10
A
ROC curve:mean_AUC=0.9655
True Positive Rate
False Positive Rate
precision-Recall curve:mean_AUPR=0.9347
Precision
Recall
B
ROC curve:mean_AUC=0.9898
True Positive Rate
False Positive Rate
precision-Recall curve:mean_AUPR=0.9598
Precision
Recall
C
ROC curve:mean_AUC=0.9908
True Positive Rate
False Positive Rate
precision-Recall curve:mean_AUPR=0.9773
Precision
Recall
D
ROC curve:mean_AUC=0.9908
True Positive Rate
False Positive Rate
precision-Recall curve:mean_AUPR=0.9731
Precision
Recall

◂**Fig. 3** For each dataset, the ROC curves and precision–recall curves for the first repeat of tenfold CV are shown. **a** The precision–recall curve and ROC curve for the NR dataset; **b** The precision–recall and ROC curves for the GPCR dataset; **c** The precision–recall and ROC curves for the IC dataset; **d** The precision–recall and ROC curves for the E dataset

appropriate baseline method. Our hybrid model accomplishes results that are 10%, 15%, 6%, and 5% better than DDR for the NR, GPCR, IC, and E datasets, respectively. On all datasets, the our hybrid model outperforms all other approaches other than the DDR approach under of AUPR metric and the cross-validation of S_d, S_t.

Discussion

This paper presents a hybrid DTI prediction model with the help aSDAE and matrix factorization. Our hybrid model learns effective latent factors from both drug–target interaction matrix and side information for drugs and targets. As part of the evaluation of the proposed work, we have published results that show our hybrid model outperforms other state-of-the-art methods on a set of datasets, with different CV settings, as well as using AUPR and AUC as performance metrics. We can observe from Table 2 that DDR and our hybrid model achieve better performance than the rest of the approaches when known DTIs are missing in the training data. Additionally, it shows how effective it can be to incorporate additional side information. Our method utilizes drug fingerprints as additional side information of drug and protein sequence composition descriptors as additional side information of targets.

We observed that the DDR method was the best second method for predicting DTI in S_p cross-validation settings and the best first method in S_d and S_t cross-validation settings, using the AUPR metric on the different datasets.

DDR employs a heterogeneous drug–target graph containing information about various similarity sets between drugs and similarity sets between protein targets. Compared to our hybrid model, the DDR produces better results in S_d and S_t. One reason may be the use of the nonlinear similarity fusion method to merge various similarities between drugs and targets, thereby smoothing their prediction and presumably ensuring greater accuracy of their prediction by relying on neighbor information based on the idea that similarity improves accuracy. So, the DDR model performs better results in both S_d and S_t cross-validation settings.

Our hybrid model outperforms methods based on MF (NRLMF, DNILMF), particularly in AUPR. In traditional MF models, latent factors are learned linearly, whereas our hybrid model uses a sigmoid activation function to learn nonlinear latent factors. As a result, our proposed method

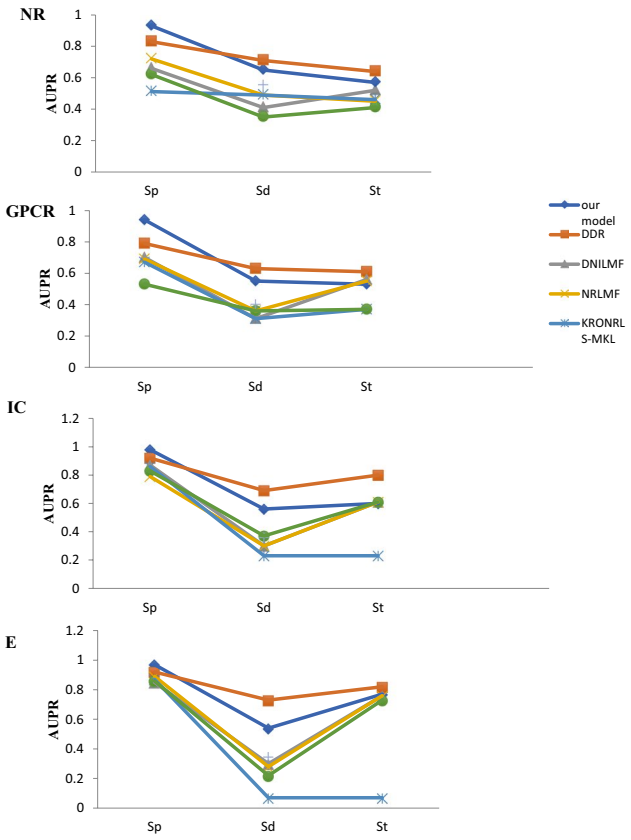


Fig. 4 Comparison results of hybrid method with the six states-of-the-art methods DDR, DNILMF, NRLMF, KRONRLS-MKL, BLM-NII, and COSINE based on AUPR scores, five repeats of tenfold CV. On all datasets, S_p, S_d, and S_t settings are used to obtain results. Based on the best published parameters, the DDR, DNILMF, NRLMF, KRONRLS-MKL, BLM-NII, and COSINE results were generated

learns sufficient and powerful features by using denoising autoencoders to predict true DTIs. Using autoencoders in the hybrid model also has the advantage of filling in all vectors that are not present in the training data, which is why it is superior to the MF method. Furthermore, deep structures can enhance the feature quality of side information. Thus, from Table 2, we can see that our hybrid model validates the strengths of the latent factor vectors learned by aSDAE models. Therefore, the AUPR metric demonstrates the effectiveness of our hybrid model. The hybrid model outperforms the MF by a large gap, indicating that neural networks have a strong potential to learn nonlinear representations, whereas MF models only on linear features.

Therefore, our hybrid model can integrate both the DTI matrix and the side information well because we seamlessly integrate aSDAE models for the side information and matrix factorization for the DTI matrix, and learn a more powerful latent factor for each drug and target, and hence, provide a much more precise prediction.

Conclusion

We introduced a new matrix factorization model with a deep structure to predict DTIs that combines aSDAE deep neural network with matrix factorization. Our proposed approach included two steps. Initially, a pre-processing step was done for replacing miss values in the sparse drug–target interaction matrix with continuous values in the range 0 to 1. Preprocessing did this work by helping drug and target similarity matrix. In the second step, we presented a hybrid model based on stack denoising autoencoders and matrix factorizations to create an unsupervised deep-learning method. In our hybrid model, effective latent factors have been learned from the DTI matrix, side information of drugs, and side information of targets. The aSDAE was built on SDAE in order to learn latent factors by incorporating side information. Using cross-validation methods S_p, S_d, and S_t on all datasets, and evaluating performance using different metrics, our hybrid model provides much superior outcomes than other state-of-the-art approaches. In future work, we will develop our models with other deep-learning models like recurrent neural networks and convolutional neural networks to improve their performance. Also the cross-validation setting S4 [41] is known to be challenging since the drugs and targets used in training do not appear in the test set. In terms of predicting interactions under S4, we believe that hybrid MF approaches with deep learning can provide useful deep representations of drugs, targets, and interactions for incraseing of accuracy of DTI prediction. Future work will confirm this.

Acknowledgements Not applicable.

Author contributions SZS developed and implemented the method, executed the experiments, and wrote the manuscript. MAZCH, MT, and SGH conceptualized the study, interpreted the results, administered the project, supervised the work, and edited the.

Funding No funding was received to assist with the preparation of this manuscript.

Data availability The datasets used in this project can be found in http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose manuscript. The manuscript has been read and approved by all authors.

Consent for publication Not applicable.

Ethical approval Not applicable.

References

1. Wang H, Wang J, Dong C et al (2020) A novel approach for drug-target interactions prediction based on multimodal deep autoencoder. Front Pharmacol 10:1592. https://doi.org/10.3389/FPHAR.2019.01592/BIBTEX
2. Sajadi SZ, Zare Chahooki MA, Gharaghani S, Abbasi K (2021) AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders. BMC Bioinformatics 22:204. https://doi.org/10.1186/s12859-021-04127-2
3. Chen R, Liu X, Jin S, et al Machine learning for drug-target interaction prediction. mdpi.com. https://doi.org/10.3390/molecules23092208
4. Zhang W, Lin W, Zhang D et al (2018) Recent advances in the machine learning-based drug-target interaction prediction. Curr Drug Metab 20:194–202. https://doi.org/10.2174/1389200219666180821094047
5. Ezzat A, Wu M, Li XL, Kwoh CK (2019) Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. Brief Bioinform 20:1337–1357. https://doi.org/10.1093/BIB/BBY002
6. Chen YZ, Zhi DG (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. Proteins 43:217–226. https://doi.org/10.1002/1097-0134
7. Periole X, Knepp AM, Sakmar TP et al (2012) Structural determinants of the supramolecular organization of G protein-coupled receptors in bilayers. J Am Chem Soc 134:10959–10965. https://doi.org/10.1021/JA303286E/SUPPL_FILE/JA303286E_SI_001.PDF
8. Opella SJ (2013) Structure determination of membrane proteins by nuclear magnetic resonance spectroscopy. Palo Alto Calif 6:305–328
9. Wen M, Zhang Z, Niu S et al (2017) Deep-learning-based drug-target interaction prediction. J Proteome Res 16:1401–1409. https://doi.org/10.1021/ACS.JPROTEOME.6B00618/ASSET/IMAGES/LARGE/PR-2016-00618X_0006.JPEG
10. Abbasi K, Razzaghi P, Poso A et al (2020) Deep learning in drug target interaction prediction: current and future perspectives. Curr Med Chem 28:2100–2113. https://doi.org/10.2174/0929867327666200907141016
11. Pan X, Fan YX, Yan J, Bin SH (2016) IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics 17:1–14. https://doi.org/10.1186/S12864-016-2931-8/TABLES/5
12. Deng L, Fan C, Zeng Z (2017) A sparse autoencoder-based deep neural network for protein solvent accessibility and contact number prediction. BMC Bioinformatics 18:211–220. https://doi.org/10.1186/S12859-017-1971-7/FIGURES/6
13. Fu L, Peng Q (2017) A deep ensemble model to predict miRNA-disease association. Sci Rep 7:1–13
14. Gligorijević V, Barot M, Bonneau R (2018) deepNF: deep network fusion for protein function prediction. Bioinformatics 34:3873–3881. https://doi.org/10.1093/BIOINFORMATICS/BTY440
15. Zeng X, Zhu S, Liu X et al (2019) deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics 35:5191–5198. https://doi.org/10.1093/BIOINFORMATICS/BTZ418
16. Hu PW, Chan KCC, You ZH (2016) Large-scale prediction of drug-target interactions from deep representations. Proceedings of the International Joint Conference on Neural Networks 2016-Octob:1236–1243. https://doi.org/10.1109/IJCNN.2016.7727339

17. Öztürk H, Özgür A, Ozkirimli E (2018) DeepDTA: deep drug-target binding affinity prediction. Bioinformatics 34:i821–i829. https://doi.org/10.1093/bioinformatics/bty593
18. Lee I, Keum J, Nam H (2019) DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. PLoS Comput Biol 15:e1007129. https://doi.org/10.1371/journal.pcbi.1007129
19. Yasuo N, Nakashima Y, Sekijima M (2019) CoDe-DTI: collaborative deep learning-based drug-target interaction prediction. proceedings-2018 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2018 792–797. https://doi.org/10.1109/BIBM.2018.8621368
20. Zheng X, Ding H, Mamitsuka H, Zhu S (2013) Collaborative matrix factorization with multiple similarities for predicting drug-Target interactions. Proceedings of the ACM SIGKDD Int Conf Knowl Discov Data Min Part 1288:1025–1033. https://doi.org/10.1145/2487575.2487670
21. Liu Y, Wu M, Miao C et al (2016) Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. PLoS Comput Biol 12:1–26. https://doi.org/10.1371/journal.pcbi.1004760
22. Dong X, Yu L, Wu Z, Sun Y, Yuan L, & Zhang F (2017) A Hybrid Collaborative Filtering Model with Deep Structure for Recommender Systems. Proceedings of the AAAI Conference on Artificial Intelligence. https://ojs.aaai.org/index.php/AAAI/article/view/10747. Accessed 14 May 2022
23. Olayan RS, Ashoor H, Bajic VB (2018) DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. Bioinformatics 34:1164–1173. https://doi.org/10.1093/BIOINFORMATICS/BTX731
24. Hao M, Bryant SH, Wang Y (2017) Predicting drug-target interactions by dual-network integrated logistic matrix factorization. Sci Rep 17:1–11. https://doi.org/10.1038/srep40376
25. Nascimento ACA, Prudêncio RBC, Costa IG (2016) A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinform 17:1–16. https://doi.org/10.1186/S12859-016-0890-3/FIGURES/4
26. Mei JP, Kwoh CK, Yang P et al (2013) Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics 29:238–245. https://doi.org/10.1093/BIOINFORMATICS/BTS670
27. Lim H, Gray P, Xie L, Poleksic A (2016) Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. Scientific 2016 6: 1–11. Doi: https://doi.org/10.1038/srep38860
28. Hattori M, Tanaka N, Kanehisa M, Goto S (2010) SIMCOMP/SUBCOMP: chemical structure search servers for network analyses. Nucleic Acids Res 38:W652–W656. https://doi.org/10.1093/NAR/GKQ367
29. Smite TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
30. Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42:30–37. https://doi.org/10.1109/MC.2009.263
31. Chen M, Li Y, Zhou X (2020) Autoencoders for drug-target interaction prediction. https://doi.org/10.21203/rs.3.rs-76683/v1
32. Bahi M (2018) Deep semi-supervised learning for DTI prediction using large datasets and H2O-spark platform. ieeexplore.ieee.org. Doi:https://doi.org/10.1109/ISACV.2018.8354081
33. Zhang S, Yao L, Sun A, Tay Y (2019) Deep learning based recommender system: a survey and new perspectives. ACM Comput Surv. https://doi.org/10.1145/3285029
34. Vincent P, Larochelle H, Bengio Y, Manzagol PA (2008) Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning 1096–1103. https://doi.org/10.1145/1390156.1390294
35. Ahmadibeni A (2020) Aerial Vehicles Automated Target Recognition of Synthetic SAR Imagery Using Hybrid Stacked Denoising Autoencoders. Dissertation, Tennessee State University
36. Ezzat A, Zhao P, Wu M et al (2017) Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinf 14:646–656. https://doi.org/10.1109/TCBB.2016.2530062
37. Lunnon WF, Brunvoll J, Cyvin SJ et al (1988) SMILES, a chemical language and information system: 1: introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36. https://doi.org/10.1021/CI00057A005
38. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32:1466–1474. https://doi.org/10.1002/JCC.21707
39. Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29:960–962. https://doi.org/10.1093/BIOINFORMATICS/BTT072
40. Yamanishi Y, Araki M, Gutteridge A et al (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24:i232–i240. https://doi.org/10.1093/BIOINFORMATICS/BTN162
41. Pahikkala T, Airola A, Pietilä S et al (2015) Toward more realistic drug-target interaction predictions. Brief Bioinform 16:325–337. https://doi.org/10.1093/bib/bbu010
42. Raghavan V, Bollmann P, Jung GS (1989) A critical investigation of recall and precision as measures of retrieval system performance. ACM Trans Inform Syst (TOIS) 7:205–229. https://doi.org/10.1145/65943.65945
43. Davis J, Goadrich M (2006) The relationship between precision-recall and ROC curves. ACM Int Conf Proc Ser 148:233–240. https://doi.org/10.1145/1143844.1143874

Authors and Affiliations

Seyedeh Zahra Sajadi[1] · Mohammad Ali Zare Chahooki[1] · Maryam Tavakol[2] · Sajjad Gharaghani[3]

[1] Department of Computer Engineering, Yazd University, Yazd, Iran

[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands

[3] Laboratory of Bioinformatics and Drug Design (LBD), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran