Autoencoders for Drug-Target Interaction Prediction

Ming Chen (✉ chm@zju.edu.cn)

Hithink RoyalRlush Information Network Co., Ltd.

Xiuze Zhou

Hithink RoyalFlush Information Network Co., Ltd. <https://orcid.org/0000-0002-0717-6936>

Research article

Keywords: Drug-target interactions, Neural network, Denoising AutoEncoder, Similarity information

Posted Date: October 6th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-76683/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Autoencoders for Drug-Target Interaction Prediction

Ming Chen^{*} and Xiuze Zhou

^{*}Correspondence: chm@zju.edu.cn
Hithink RoyalFlush Information
Network Co., Ltd., 310023
Hangzhou, China
Full list of author information is
available at the end of the article

Abstract

Background: Because it is so laborious and expensive to experimentally identify Drug-Target Interactions (DTIs), only a few DTIs have been verified. Computational methods are useful for identifying DTIs in biological studies of drug discovery and development.

Results: For drug-target interaction prediction, we propose a novel neural network architecture, DAEi, extended from Denoising AutoEncoder (DAE). We assume that a set of verified DTIs is a corrupted version of the full interaction set. We use DAEi to learn latent features from corrupted DTIs to reconstruct the full input. Also, to better predict DTIs, we add some similarities to DAEi and adopt a new nonlinear method for calculation. Similarity information is very effective at improving the prediction of DTIs.

Conclusion: Results of the extensive experiments we conducted on four real data sets show that our proposed methods are superior to other baseline approaches.

Availability: All codes in this paper are open-sourced, and our projects are available at: <https://github.com/XiuzeZhou/DAEi>.

Keywords: Drug-target interactions; Neural network; Denoising AutoEncoder; Similarity information

1 Introduction

To understand the protein functions and molecular mechanisms of cells, it is important to study Drug-Target Interactions (DTIs) [1]. Although an enormous amount of research has resulted in the discovery of new drugs, the number of approved drugs, which remains modest [2], is far from meeting the rapidly growing demand for identifying DTIs. Thus, there is an urgent need for effective and efficient computational methods to solve this problem.

Great advances in machine learning have contributed significantly to biological development [3, 4]. For example, Support Vector Machines (SVMs), one of the most widely applied supervised learning methods, is adopted for drug discovery [5, 6]. Random Forest, a popular tree-based ensemble machine learning tool with highly adaptive data, is applied to predict protein-protein interaction sites [7, 8]. k-Nearest Neighbor (kNN), a simple and commonly used method, is designed to improve prediction accuracy in drug discovery [9, 10]. Matrix Factorization (MF), an effective way to learn latent features from matrix data, is extended to various models to predict DTIs [10, 11, 12, 13, 14, 15]. For further experimental investigation, such computational methods are useful for providing some critical and reasonable clues.

Recently, to predict DTIs, some researchers started to explore the application of powerful deep learning models to capture the higher-order relationship from inputs by their hidden layers. For example, to accurately predict new DTIs between approved drugs and targets, Wen *et al.* [16] used Deep Belief Network (DBN) architecture, which is constructed by stacking many Restricted Boltzmann Machines (RBMs). Gao *et al.* [17] combined Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) to obtain meaningful information from protein sequences and drug structures. Zhao *et al.* [18] proposed an end-to-end model, associated with an attention mechanism, to predict the binding affinity of DTIs. Zhao *et al.* [19] proposed a neural network, GANsDTA, which combined two Generative Adversarial Networks (GANs) and a regression network to predict binding affinity. Those deep learning models show great promise for identifying DTIs.

Among all deep learning architectures, autoencoders are one of the easiest and simplest tools to be applied in many fields, such as images and texts, to learn the meaningful features from raw data. For example, Hu *et al.* [20] used the Stacked AutoEncoder (SAE) to learn the deep features from protein sequence descriptors and drug molecular fingerprints, which were then fed to SVMs to make predictions. To fully train rotation forest classifiers, Wang *et al.* [21] used SAE to extract protein features from protein sequences. Yasuo *et al.* [22] adopted a Denoising AutoEncoder (DAE) to capture drug features, which are added to the MF as auxiliary information. But, in applying those models, one major problem is they cannot be applied to the fields without any, or with little, auxiliary information, such as protein sequence and drug molecular structure.

For DTI prediction, unlike using existing autoencoders as just ancillary tools to obtain latent features from protein sequences and drug molecular structures, we use autoencoders as the principal methods for directly predicting DTIs. We assume that the set of verified DTIs is a corrupted version of the full set of DTIs. We applied **AutoEncoder** for drug-target interaction prediction (AEi), and extended **Denoising AutoEncoder** for drug-target interaction prediction (DAEi) to learn latent features from a set of corrupted DTIs for reconstructing its full inputs. AEi and DAEi are reliable and effective for detecting potential interactions between drugs and targets.

Because DAEi has a simple and flexible structure, additional information can easily be incorporated. To further improve identification performance, some similarity information about drug-drug and target-target is added to our model. Similarity information improves the ability of the models to identify DTIs [13, 23]. However, to calculate similarities, most previous methods adopted a linear method, which ignores some nonlinear and complicated relationships between drug-drug and target-target. Different from existing similarity metrics, we design a nonlinear technique to effectively extract the proper relationships.

Finally, the performance of our models was empirically evaluated using several state-of-the-art methods and four benchmark data sets: Enzymes, Ion channels, G-Protein-Coupled Receptors (GPCRs), and Nuclear receptors. Results of extensive experiments show that our proposed models considerably outperform the baseline approaches for all data sets in terms of the two most used assessment methods: Area Under the Precision-Recall curve (AUPR) and the Area Under the receiver operator characteristic Curve (AUC). In addition, because our models have a very simple and effective framework, they can be easily extended to further research.

The rest of the paper is organized as follows: Section 2 briefly reviews the background and some related work. Section 3 presents our proposed models in detail. Section 4 describes the experimental results for several data sets to show the performance of our models. Section 5 gives the conclusion and provides future directions.

2 Related Work

First, we define the problem. Then, to solve the problem, we simply introduce MF and Collaborative Matrix Factorization (CMF), which extends MF by combining similarity information to achieve better performance.

2.1 Problem Definition

Given a set of drugs, $\mathbf{D} = \{d_1, \dots, d_n\}$, and a set of targets, $\mathbf{T} = \{t_1, \dots, t_m\}$, where n and m are the number of drugs and targets, respectively, their interactions, $\mathbf{Y} \in \mathbb{R}^{n \times m}$, is defined as follows:

$$y_{d,t} = \begin{cases} 1, & \text{if drug } d \text{ has an interaction with target } t \\ 0, & \text{otherwise} \end{cases}$$

For typical DTIs, DTI prediction is viewed as a binary classification problem. When given a drug and a target, the goal of the model is to use learned latent features to accurately predict the possible values of their interaction. In general, the prediction of DTIs is defined as follows:

$$\hat{y}_{d,t} = \mathcal{F}_{\Theta}(d, t),$$

where $\hat{y}_{d,t}$ denotes the predicted interaction between drug, d , and target, t ; Θ is the set of learning parameters of models for training; $\mathcal{F}$ is the function of the models for predicting DTIs.

Let $\ell(\cdot)$ denote a loss function, which measures the distance between the true label, $y_{d,t}$, and the predicted label, $\hat{y}_{d,t}$; $\Omega(\cdot)$ denotes a regularization term, which prevents the model from over-fitting. The objective function is defined as follows:

$$\mathcal{L} = \sum_{(d,t) \in \mathbf{V}} \ell(y_{d,t}, \hat{y}_{d,t}) + \lambda \Omega(\Theta), \quad (1)$$

where $\mathbf{V}$ denotes the set of instances whose interactions have been experimentally verified; λ denotes a regularization coefficient.

2.2 MF and CMF

Let $\mathbf{P} \in \mathbb{R}^{n \times k}$ and $\mathbf{Q} \in \mathbb{R}^{m \times k}$ denote the latent features matrices of drugs and targets, respectively, where k is the number of latent features. $\mathbf{p}_d \in \mathbb{R}^{1 \times k}$ and $\mathbf{q}_t \in \mathbb{R}^{1 \times k}$ denote the latent features of drug, d , and target, t , respectively. To predict interactions, MF uses a dot product, which linearly combines latent features. Then, *square loss* is used to measure its loss. The objective function of MF is defined as follows:

$$\mathcal{L}_1 = \sum_{(d,t) \in \mathbf{V}} (y_{d,t} - \mathbf{p}_d \mathbf{q}_t^T)^2 + \lambda (\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2), \quad (2)$$

where λ denotes a regularization coefficient, and $\|\cdot\|_F^2$ denotes the Frobenius norm.

To better predict DTIs, Zheng *et al.* [13] proposed a method, CMF, which adds multiple similarities to improve MF. CMF assumes that similar drugs will interact with similar targets, and vice versa. Thus, the feature vectors of similar drugs/targets should be near to each other. The objective function of CMF is defined as follows:

$$\begin{aligned} \mathcal{L}_2 = \sum_{(d,t) \in \mathbf{V}} (y_{d,t} - \mathbf{p}_d \mathbf{q}_t^T)^2 + \lambda (\|\mathbf{P}\|_F^2 + \|\mathbf{Q}\|_F^2) \\ + \lambda_d \|\mathbf{S}^d - \mathbf{P} \mathbf{P}^T\|_F^2 + \lambda_t \|\mathbf{S}^t - \mathbf{Q} \mathbf{Q}^T\|_F^2 \end{aligned} \quad (3)$$

where λ_d and λ_t denote regularization coefficients for drug and target similarities, respectively; $\mathbf{S}^d \in \mathbb{R}^{n \times n}$ denotes the similarity matrix for drugs, and $\mathbf{S}^t \in \mathbb{R}^{m \times m}$ denotes the similarity matrix for targets.

Compared with MF, CMF improves predictive performance for DTIs by incorporating two similarity matrices as regularization terms. Similarity provides additional information about drug-drug and target-target relationships. Through that relationship information, more accurate representations are captured.

3 Methods

First, we introduce in detail the AEi framework applied to DTIs. Then, to predict DTIs, we describe DAEi, which learns the correlations between drugs and targets by training on a corrupted version of the known interaction set. Finally, for further development, two similarity matrices are added to our model and a nonlinear method is designed. The main goal of AEi and DAEi is to predict all unknown interactions.

3.1 AEi for DTIs

Autoencoders, one of the most popular deep neural networks, are widely used for classification and learning unsupervised features by reconstructing the original input [24, 25]. A classical autoencoder is typically implemented as a one-hidden layer neural network. From the point of view of the drug, the set of all targets that interacts with each drug as an input for AEi, we call drug AEi (d-AEi). For example, take d-AEi; then, target AEi (t-AEi) is similar. The framework of d-AEi is shown in Figure 1.

Given a vector $\mathbf{y}_d \in \mathbb{R}^m$ of drug, d , as input, the vector is mapped to a hidden representation $\mathbf{z} \in \mathbb{R}^k$ through the following mapping function:

$$\mathbf{z} = a(\mathbf{W}^T \mathbf{y}_d + b),$$

where $\mathbf{W} \in \mathbb{R}^{m \times k}$, $b \in \mathbb{R}^k$, and $a(\cdot)$ denote weight, bias, and activation function of the encoder, respectively.

The purpose of the autoencoder is to reconstruct input, $\mathbf{y}_d$. Thus, the hidden representation, $\mathbf{z}$, is mapped back to a reconstructed vector, $\hat{\mathbf{y}}_d \in \mathbb{R}^m$, defined as follows:

$$\hat{\mathbf{y}}_d = a'(\mathbf{W}' \mathbf{z} + b'),$$

Figure 1: Framework of d-AEi

where $\mathbf{W}' \in \mathbb{R}^{m \times k}$, $b' \in \mathbb{R}^m$, and $a'(\cdot)$ denote weight, bias, and activation function of the decoder, respectively.

The learning parameters of the autoencoder are trained by minimizing the average reconstruction error, defined as follows:

$$\mathcal{L}_3 = \frac{1}{n} \sum_{d=1}^n \ell(\mathbf{y}_d - \hat{\mathbf{y}}_d) + \lambda \left(\|\mathbf{W}\|_F^2 + \|\mathbf{W}'\|_F^2 \right), \quad (4)$$

where $\ell(\cdot)$ denotes a loss function, and *square loss* and *cross entropy loss* are two of the most commonly used functions to measure loss in training. *square loss* is chose as our loss function.

3.2 DAEi for DTIs

d-AEi has two main shortcomings: 1) It simply reconstructs its input, ignoring the possible noise of the data; 2) It models on a set of targets, which only consider the association between the targets and ignores their interactions with drugs. To alleviate these problems and achieve an enhanced prediction of DTIs, we propose a DAEi model extended from DAE, which is robust for reconstructing original data.

DAEi learns representation from corrupted input and reconstructs the input from that representation. We associate a latent feature of the drug with all targets (d-DAEi) and associate a latent feature of the target with all drugs (t-DAEi). DAEi learns from both drugs and targets – not separately from them. Figure 2 shows the network architecture of d-DAEi, which consists of three layers: input, hidden, and output. Next, we illustrate d-DAEi in detail; t-DAEi is similar to d-DAEi.

Figure 2: Framework of d-DAEi

Input Layer. In the input layer, instead of the traditional approach of modeling separately, we simultaneously combine drugs and targets. A drug, d and its interactions are recorded with all targets, $\mathbf{y}_d \in \mathbb{R}^m$, i.e. in the d -th row of the matrix, $\mathbf{Y}$. $\tilde{\mathbf{y}}_d$, the inputs of DAEi are the corrupted vector of $\mathbf{y}_d$. There are two popular ways to corrupt the original inputs: 1) Randomly discard some non-zero values of the inputs; 2) Add Gaussian noise to the inputs. We chose to add Gaussian noise to $\mathbf{y}_d$ to generate its corrupted vector, defined as follows:

$$\tilde{\mathbf{y}}_d = \mathbf{y}_d + \epsilon_d,$$

where $\epsilon_d \sim N(0, \sigma_d)$ denotes the Gaussian noise.

Hidden Layer. The role of the hidden layer is to translate drug/target into latent representation space. Also, the hidden layer has a propensity for nonlinear representation. An interaction function for modeling $\tilde{\mathbf{y}}_d$ and d , such as concatenation, element-wise product, and element-wise sum, is critical for the methods to learn their relationships. Element-wise sum is chose as our interaction function. The function of the hidden layer is defined as follows:

$$\mathbf{z} = a(\mathbf{W}^T \tilde{\mathbf{y}}_d + \mathbf{V}_d + b),$$

where $\mathbf{W} \in \mathbb{R}^{m \times k}$, $\mathbf{b} \in \mathbb{R}^k$, $\mathbf{z} \in \mathbb{R}^k$ and $a(\cdot)$ denote weight, bias, output, and activation function of the hidden layer, respectively. $\mathbf{V}_d \in \mathbb{R}^k$ is the latent feature of drug, d . In DAEi, we chose the *sigmoid* function as the activation function.

Output Layer. The goal of the output layer is to map back the latent representation to the original input space to reconstruct the input vector, defined as follows:

$$\hat{\mathbf{y}}_d = f(\mathbf{W}'\mathbf{z} + \mathbf{b}'),$$

where $\mathbf{W}'$, $\mathbf{b}'$, $\mathbf{z}$ and $f(\cdot)$ denote weight, bias, output, and map function of the output layer, respectively. And the *identity* function is chose as the map function of DAEi.

Finally, the objective function of DAEi is defined as follows:

$$\mathcal{L}_4 = \frac{1}{n} \sum_{d=1}^n \ell(\tilde{\mathbf{y}}_d - \hat{\mathbf{y}}_d) + \lambda \left(\|\mathbf{W}\|_F^2 + \|\mathbf{W}'\|_F^2 + \|\mathbf{V}\|_F^2 \right). \quad (5)$$

3.3 Similarity Information

Drug-drug and target-target similarities provide important information: two similar things have similar characteristics. An interaction of one drug that has been experimentally verified with a target will interact with a similar target, and vice versa. Therefore, to improve our models, we added some similarity information.

Existing methods use a dot product, which linearly combines latent features to calculate the similarities between drug d and i , and between target t and j , as shown in the following:

$$s_{d,i} = \mathbf{p}_d \mathbf{p}_i^T, \quad s_{t,j} = \mathbf{q}_t \mathbf{q}_j^T.$$

But, a linear combination for the latent features between drugs and targets might neglect their possible nonlinear relationships. To better capture drug-drug and target-target relationships, we propose a nonlinear method. To calculate similarities, we use an exponential function defined as follows:

$$s_{d,i} = \exp\left(-\|\mathbf{p}_d - \mathbf{p}_i\|_F^2\right), \quad (6)$$

$$s_{t,j} = \exp\left(-\|\mathbf{q}_t - \mathbf{q}_j\|_F^2\right). \quad (7)$$

Then *square loss* is used to measure the distance between the predicted and true values as shown in the following:

$$\mathcal{L}_s = \lambda_d \left\| \mathbf{S}^d - \hat{\mathbf{S}}^d \right\|_F^2 + \lambda_t \left\| \mathbf{S}^t - \hat{\mathbf{S}}^t \right\|_F^2 \quad (8)$$

where $\hat{\mathbf{S}}^d$ and $\hat{\mathbf{S}}^t$ are the predicted similarity matrices of drugs and targets, respectively.

Finally, DAEi is developed to DAEi with similarity information (DAEi-s), whose objective function is defined as follows:

$$\mathcal{L}'_4 = \mathcal{L}_4 + \mathcal{L}_s. \quad (9)$$

Table 1: Statistics for all data sets.

Data sets	# of Drugs	# of Targets	# of Interactions
Enzymes	445	664	2926
Ion Channels	210	204	1476
GPCRs	223	95	635
Nuclear Receptors	54	26	90

4 Experiments

First, we describe the data sets used in our experiments, and their similarity matrices. Second, we introduce the metric we used to evaluate our models. Then, for all methods, we present some baseline approaches compared with our model and parameter settings. Finally, to illustrate the performance of our models, we compare our models with baselines using extensive experiments and discuss the influence of similarity information, the nonlinear method, and a key parameter, the hidden layer size.

4.1 Experimental Setting

Data Sets. To evaluate our proposed methods in respect to the prediction of DTIs, we use the following four benchmark data sets^[1]: Nuclear Receptors, GPCRs, Ion Channels and Enzymes, collected by Yamanishi *et al.* [26] from KEGG BPITE [27], BRENDA [28], SuperTarget [29], and DrugBank [30], respectively. Each data set contains three types of information: 1) verified DTIs; 2) drug-drug similarity; and 3) target-target similarity. Listed in Table 1 are some statistics about the verified DTIs in all the data sets.

The similarity of the chemical structure between compounds was computed by SIMCOMP [31], which measures the similarity score between two compound structures. Amino acid sequences of the target proteins were obtained from the KEGG GENES database [27]. The sequence similarity between two proteins was computed using a normalized version of the Smith-Waterman score [32].

Evaluation Metrics. Following previous studies [10, 12, 13, 33], to evaluate the performance of the DTI prediction methods, we repeated 10-fold Cross-Validation (CV) five times. In CV, the set of all DTIs ($\mathbf{Y}$) was randomly split into ten folds, nine of which were used for training and one used for testing.

In this paper, we consider two popular metrics: AUC and AUPR. An AUC score is estimated in each repetition of CV; finally, the average score over all five repetitions is determined. The AUPR score is estimated in the same way. Because AUPR punishes highly ranked false positives much more than AUC [34], AUPR is chose as our major evaluation metric in this paper.

Baseline Approaches. To show the effectiveness of our models, we compared our models with four baseline approaches. Each baseline has its unique characteristics and function and represents a particular capability to model. Those approaches are briefly introduced as follows:

- **PMF** [14]. Probabilistic Matrix Factorization (PMF), one of the most widely applied in the prediction of DTIs, effectively learns latent features from matrix data;

^[1]<http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/>

Table 2: AUC and AUPR values of all methods on four data sets.

		PMF	CMF	RBM	BRDTI	CnnDTI	d-AEi	t-AEi	d-DAEi	t-DAEi
Enzymes	AUC	0.9290	0.9605	0.9470	0.9668	0.9801	0.9575	0.9656	0.9904	0.9883
	AUPR	0.7867	0.8514	0.7703	0.9258	0.9188	0.8368	0.8643	0.9665	0.9455
Ion Channels	AUC	0.9398	0.9766	0.9448	0.9771	0.9754	0.9881	0.9792	0.9928	0.9920
	AUPR	0.9253	0.9333	0.8223	0.9507	0.9487	0.9532	0.9132	0.9749	0.9746
GPCRs	AUC	0.9436	0.9575	0.9150	0.9550	0.9513	0.9668	0.9305	0.9850	0.9834
	AUPR	0.8964	0.9024	0.7324	0.9105	0.9081	0.8837	0.7400	0.9449	0.9462
Nuclear Receptors	AUC	0.8941	0.9204	0.9173	0.9163	0.8965	0.9604	0.9216	0.9590	0.9752
	AUPR	0.8943	0.9003	0.8059	0.9111	0.9034	0.9028	0.6487	0.9271	0.9554

- **CMF** [13]. The state-of-the-art MF based method for DTI prediction simultaneously learns latent features from DTIs, drug-drug and target-target similarities, and uses dot product to calculate similarities;

- **RBM** [35]. Restricted Boltzmann Machine (RBM), regarding the state-of-the-art shallow neural network-based method for DTI prediction, its visible units encode observed types of DTIs, and its hidden units represent latent features describing DTIs;

- **BRDTI** [33]. The state-of-the-art Bayesian ranking method for DTI prediction, which extends Bayesian Personalized Ranking (BPR) [36] by adding similarity information and target bias;

- **CnnDTI** [37]. The state-of-the-art deep neural network-based method for DTI prediction resembles the LeNet-5 framework and CNN on drug descriptor and target protein sequences.

Parameter Settings. To make a fair comparison, the corresponding parameters in all models are set to the same value. We set the learning rate in all methods at 0.001; regularization coefficient (λ) in all methods except RBM at 10^{-6} ; the number of latent feature (k) in PMF, CMF, and BRDTI, hidden size in RBM and DAE, embedding size of CnnDTI at 1024, 256, 256, and 128 for Enzymes, Ion Channels, GPCRs, and Nuclear Receptors, respectively; batch size in RBM, CnnDTI, and our models at 256; regularization coefficients for similarity (λ_d and λ_t) in CMF, BRDTI, and DAEi models at 10^{-6} ; epochs of neural network frameworks at 100.

4.2 Results and Analysis

Overall Performance. First, we conducted some experiments, using all data sets, to determine the overall performance of our models. Table 2 shows the average AUC and AUPR values obtained by all methods on four data sets. The best results are shown in bold.

From the results shown in Table 2, the following can be observed: (1) In terms of both AUC and AUPR, DAEi models significantly outperform other baselines, indicating that DAEi models effectively predict DTIs; (2) In terms of AUC, d-DAEi performs better than t-DAEi on Enzymes, Ion Channels, and GPCRs data sets. In terms of AUPR, t-DAEi performs better than d-DAEi on GPCRs and Nuclear Receptors data sets, possibly because the number of drugs and targets affects DAEi models. Regarding a data set, the set with a larger number of drugs/targets is more suitable for learning meaningful features; (3) On the same data set, results between d-AEi and t-AEi differ significantly, especially on the GPCRs and Nuclear

Table 3: Effect of similarity information and nonlinear calculation method.

		0	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
Enzymes	Linear	0.9665	0.9683	0.9673	0.9637	0.9653	0.9708	0.9644
	Nonlinear	0.9665	0.9682	0.9685	0.9712	0.9687	0.9623	0.9673
Ion Channels	Linear	0.9749	0.9808	0.9815	0.9795	0.9784	0.9777	0.9771
	Nonlinear	0.9749	0.9819	0.9791	0.9799	0.9804	0.9792	0.9789
GPCRs	Linear	0.9449	0.9461	0.9441	0.9426	0.9366	0.9407	0.9406
	Nonlinear	0.9449	0.9545	0.9459	0.9468	0.9391	0.9444	0.9416
Nuclear Receptors	Linear	0.9271	0.9298	0.9315	0.9238	0.9233	0.9246	0.9280
	Nonlinear	0.9271	0.9355	0.9450	0.9311	0.9286	0.9283	0.9292

Receptors data sets in terms of the AUPR metric. Possibly, the input vector with a large sparsity difference results in an instable model; (4) Gaps among the DAEi models, CnnDTI, the best baseline method are significant: in terms of AUPR, DAEi models achieve results for Enzyme, Ion Channel, GPCR and Nuclear Receptor, which are 5.2%, 2.8%, 4.2% and 5.8% higher than CnnDTI, indicating that our architectures, modeled on the set of drugs and targets, are more reasonable than CnnDTI, which is modeled on only a single drug and single target; (5) RBM has a similar neural network with the DAEi models; however RBM is much worse than the d-AEi model, especially in terms of AUPR. Possibly, one reason is that it has only a binary representation, which limits its performance; another reason might be that it does not consider any relationship between drugs and targets; (6) Compared with the PMF and DAEi models, the DAEi models achieve better results with a large gap, which indicates that neural networks have a powerful ability to learn nonlinear representations; whereas, PMF models only on linear features. Finally, it can be concluded that our proposed methods learn sufficient and effective features by neural networks to detect true DTIs.

Effect of Similarity and Nonlinear Calculation. Next, we investigate how similarity information, which provides additional information to build more accurate drug-drug and target-target relationships, affects our models. In this experiment, we set $\lambda_d = \lambda_t$, and select their values from $\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$. We can view DAEi as special cases of DAEi-s when $\lambda_d = \lambda_t = 0$. Then, a linear calculation method, as adopted in CMF, is used for comparison with our nonlinear calculation method. In this experiment, AUPR is chose as our evaluation metric. Table 3 shows the average AUPR values of similarity information and calculation methods of d-DAEi for all data sets.

It is seen from Table 3 that, in most cases of all the data sets (21 cases out of 24), our method with a nonlinear calculation method always achieves higher AUPR values than the linear calculation method, which indicates that nonlinear technology has a much stronger ability to extract the proper information from similarity information. Also, for the regularization coefficients, small values (smaller than 10^{-4}) improve the experimental results slightly; whereas, large values (larger than 10^{-4}) adversely affect the experimental results. Thus, for good performance, choosing an appropriate value for the regularization coefficient is critical.

Effect of Hidden size (k). Finally, we illustrate the effect that hidden layer size (embedding size or the number of latent features) has on DAEi models. Hidden layer size is a key factor for DAEi models to learn meaningful latent features from

Table 4: Effect of hidden size of DAEi models on all data sets.

		8	16	32	64	128	256	512	1024
Enzymes	d-DAEi	0.6811	0.7187	0.7551	0.7939	0.8570	0.9270	0.9614	0.9665
	t-DAEi	0.6637	0.6601	0.6820	0.7469	0.8254	0.9056	0.9381	0.9455
Ion Channels	d-DAEi	0.7969	0.8611	0.9031	0.9255	0.9579	0.9799	0.9741	0.9760
	t-DAEi	0.7868	0.8600	0.8970	0.9202	0.9565	0.9753	0.9746	0.9666
GPCRs	d-DAEi	0.6382	0.7822	0.8369	0.8740	0.9230	0.9449	0.9441	0.9296
	t-DAEi	0.6124	0.7439	0.8279	0.8853	0.9318	0.9497	0.9462	0.9450
Nuclear Receptors	d-DAEi	0.6339	0.8461	0.9183	0.9201	0.9271	0.9281	0.9293	0.9273
	t-DAEi	0.6553	0.8758	0.9217	0.9226	0.9554	0.9436	0.9373	0.9194

raw inputs. In this experiment, we selected the hidden layer size from {8, 16, 32, 64, 128, 256, 512, 1024} for all data sets. Other parameters are set the same as in the previous experiments. In this experiment, AUPR is chose as our evaluation metric. Table 4 shows the average AUPR values of the DAEi models on four data sets.

As shown in Table 4, there is a common trend across all the data sets. As the value of k increases, the AUPR value increases very quickly and then decreases slowly. When k is small (less than 128), models are in a state of under-fitting, where the model is too simple to achieve optimal results with test data; when k is too large, models are in a state of over-fitting, where the model is too complex to learn from the highly sparse interaction matrix. Those observations indicate that choosing the appropriate value (about 1 to 2 times the number of drugs/targets in each data set) for hidden size (the number of latent features or embedding size) is essentially a tradeoff between under-fitting and over-fitting.

5 Conclusion

In drug discovery and development, effective and efficient computational methods are important to detect potential DTIs. To predict DTIs, we proposed two novel methods, AEi and DAEi, which were developed from autoencoder and DAE, respectively. DAEi models, which reconstruct the full input from the set of their corrupted DTIs, are effective in identifying potential interactions for drugs and targets. After some similarities are added to our models, they achieve better performance. Experimental results show that our models considerably outperform baseline approaches.

For future work, first, to develop more effective methods for larger data sets, we plan to further investigate the integration of other deep learning models into our methods. Second, in our experiments, we found that our models did not perform well in predicting new drugs and targets; therefore, to improve the performance of our models in new drug and target tasks, it will be necessary to incorporate some additional information, such as amino acid sequences of target proteins and drug molecular structure. Finally, other possible directions will be to explore extending our models by adding more hidden layers, and extending more autoencoders, such as Stacked AutoEncoder (SAE) and Variational AutoEncoder (VAE), to predict DTIs.

Acknowledgements

The authors thank Michael McAllister for proofreading this paper.

Abbreviations

DTIs: Drug-Target Interactions; AEi: AutoEncoder for drug-target interaction prediction; DAEi: Denoising AutoEncoder for drug-target interaction prediction; AUC: Area Under the receiver operator characteristic Curve; AUPR: Area Under the Precision-Recall curve

Availability of data and materials

Our projects are available at: <https://github.com/XiuzeZhou/DAEi>.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

MC initialized the research project and designed the experiments; XZ performed the experiments and wrote the paper; All authors read and approved the final manuscript.

References

1. Cheng, F., Liu, C., Jiang, J., Lu, W., Li, W., Liu, G., Zhou, W., Huang, J., Tang, Y.: Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Computational Biology* **8**(5), 1002503 (2012)
2. Hao, M., Bryant, S.H., Wang, Y.: Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Scientific Reports* **7**(1), 1–11 (2017)
3. Weng, Y., Lin, C., Zeng, X., Liang, Y.: Drug target interaction prediction using multi-task learning and co-attention. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 528–533 (2019). IEEE
4. Öztürk, H., Özgür, A., Ozkirimli, E.: Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**(17), 821–829 (2018)
5. Heikamp, K., Bajorath, J.: Support vector machines for drug discovery. *Expert Opinion on Drug Discovery* **9**(1), 93–104 (2014)
6. Korkmaz, S., Zararsiz, G., Goksuluk, D.: Drug/nondrug classification using support vector machines with various feature selection strategies. *Computer Methods and Programs in Biomedicine* **117**(2), 51–60 (2014)
7. Šikić, M., Tomić, S., Vlahoviček, K.: Prediction of protein–protein interaction sites in sequences and 3d structures by random forests. *PLoS Computational Biology* **5**(1), 1000278 (2009)
8. Wei, Z.-S., Yang, J.-Y., Shen, H.-B., Yu, D.-J.: A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE Transactions on Nanobioscience* **14**(7), 746–760 (2015)
9. Shi, J.-Y., Li, J.-X., Lu, H.-M., Zhang, Y.: Predicting drug-target interactions between new drugs and new targets via pairwise k-nearest neighbor and automatic similarity selection. In: International Conference on Intelligent Science and Big Data Engineering, pp. 477–486 (2015). Springer
10. Ezzat, A., Zhao, P., Wu, M., Li, X.-L., Kwok, C.-K.: Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **14**(3), 646–656 (2016)
11. Gönen, M.: Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics* **28**(18), 2304–2310 (2012)
12. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.-L.: Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Computational Biology* **12**(2), 1004760 (2016)
13. 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, pp. 1025–1033 (2013)
14. Cobanoglu, M.C., Liu, C., Hu, F., Oltvai, Z.N., Bahar, I.: Predicting drug–target interactions using probabilistic matrix factorization. *Journal of Chemical Information and Modeling* **53**(12), 3399–3409 (2013)
15. Li, L., Cai, M.: Drug target prediction by multi-view low rank embedding. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2017)
16. Wen, M., Zhang, Z., Niu, S., Sha, H., Yang, R., Yun, Y., Lu, H.: Deep-learning-based drug–target interaction prediction. *Journal of Proteome Research* **16**(4), 1401–1409 (2017)
17. Gao, K.Y., Fokoue, A., Luo, H., Iyengar, A., Dey, S., Zhang, P.: Interpretable drug target prediction using deep neural representation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, vol. 2018, pp. 3371–3377 (2018)
18. Zhao, Q., Xiao, F., Yang, M., Li, Y., Wang, J.: Attentiondta: prediction of drug–target binding affinity using attention model. In: 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 64–69 (2019). IEEE
19. Zhao, L., Wang, J., Pang, L., Liu, Y., Zhang, J.: Gansdta: Predicting drug-target binding affinity using gans. *Frontiers in Genetics* **10** (2019)
20. Hu, P.-W., Chan, K.C., You, Z.-H.: Large-scale prediction of drug-target interactions from deep representations. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 1236–1243 (2016). IEEE
21. Wang, L., You, Z.-H., Chen, X., Xia, S.-X., Liu, F., Yan, X., Zhou, Y., Song, K.-J.: A computational-based method for predicting drug–target interactions by using stacked autoencoder deep neural network. *Journal of Computational Biology* **25**(3), 361–373 (2018)
22. Yasuo, N., Nakashima, Y., Sekijima, M.: Code-dti: Collaborative deep learning-based drug-target interaction prediction. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 792–797 (2018). IEEE
23. Yuan, Q., Gao, J., Wu, D., Zhang, S., Mamitsuka, H., Zhu, S.: Druge-rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics* **32**(12), 18–27 (2016)

24. Rifai, S., Vincent, P., Muller, X., Glorot, X., Bengio, Y.: Contractive auto-encoders: Explicit invariance during feature extraction. In: Proceedings of the 28th International Conference on Machine Learning (2011)
25. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: Greedy layer-wise training of deep networks. In: Advances in Neural Information Processing Systems, pp. 153–160 (2007)
26. 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* **24**(13), 232–240 (2008)
27. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Research* **34**(suppl.1), 354–357 (2006)
28. Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., Schomburg, D.: Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Research* **32**(suppl.1), 431–433 (2004)
29. Günther, S., Kuhn, M., Dunkel, M., Campillos, M., Senger, C., Petsalaki, E., Ahmed, J., Urdiales, E.G., Gewiss, A., Jensen, L.J., *et al.*: Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Research* **36**(suppl.1), 919–922 (2007)
30. Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., Tzur, D., Gautam, B., Hassanali, M.: Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research* **36**(suppl.1), 901–906 (2008)
31. Hattori, M., Okuno, Y., Goto, S., Kanehisa, M.: Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *Journal of the American Chemical Society* **125**(39), 11853–11865 (2003)
32. Smith, T.F., Waterman, M.S., *et al.*: Identification of common molecular subsequences. *Journal of Molecular Biology* **147**(1), 195–197 (1981)
33. Peska, L., Buza, K., Koller, J.: Drug-target interaction prediction: A bayesian ranking approach. *Computer Methods and Programs in Biomedicine* **152**, 15–21 (2017)
34. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240 (2006)
35. Wang, Y., Zeng, J.: Predicting drug-target interactions using restricted boltzmann machines. *Bioinformatics* **29**(13), 126–134 (2013)
36. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, 452–461 (2012)
37. Hu, S., Zhang, C., Chen, P., Gu, P., Zhang, J., Wang, B.: Predicting drug-target interactions from drug structure and protein sequence using novel convolutional neural networks. *BMC Bioinformatics* **20**(25), 1–12 (2019)

Figures

Figure 1

Framework of d-AEi

Figure 1

Framework of d-AEi

Figure 1

Framework of d-AEi

Figure 2

Framework of d-DAEi

Figure 2

Framework of d-DAEi

Figure 2

Framework of d-DAEi