

A similarity-based deep learning approach for determining the frequencies of drug side effects

Haochen Zhao, Shaokai Wang, Kai Zheng, Qichang Zhao, Feng Zhu and Jianxin Wang

Corresponding author: Jianxin Wang, School of Computer Science and Engineering and Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha 410083, China. E-mail: jxwang@mail.csu.edu.cn

Abstract

The side effects of drugs present growing concern attention in the healthcare system. Accurately identifying the side effects of drugs is very important for drug development and risk assessment. Some computational models have been developed to predict the potential side effects of drugs and provided satisfactory performance. However, most existing methods can only predict whether side effects will occur and cannot determine the frequency of side effects. Although a few existing methods can predict the frequency of drug side effects, they strongly depend on the known drug-side effect relationships. Therefore, they cannot be applied to new drugs without known side effect frequency information. In this paper, we develop a novel similarity-based deep learning method, named SDPred, for determining the frequencies of drug side effects. Compared with the existing state-of-the-art models, SDPred integrates rich features and can be applied to predict the side effect frequencies of new drugs without any known drug-side effect association or frequency information. To our knowledge, this is the first work that can predict the side effect frequencies of new drugs in the population. The comparison results indicate that SDPred is much superior to all previously reported models. In addition, some case studies also demonstrate the effectiveness of our proposed method in practical applications. The SDPred software and data are freely available at <https://github.com/zhc940702/SDPred>, <https://zenodo.org/record/5112573> and <https://hub.docker.com/r/zhc940702/sdpred>.

Key words: drug-side effect frequencies; deep learning; multi-similarities

Haochen Zhao is a PhD candidate in the School of Computer Science and Engineering, Central South University, China. His current research interests include machine learning, deep learning and bioinformatics.

Shaokai Wang is a PhD candidate in the School of David R. Cheriton School of Computer Science, University of Waterloo, Canada. His current research interests include machine deep learning, proteomics and bioinformatics.

Kai Zheng is a PhD candidate in the School of Computer Science and Engineering, Central South University, China. His current research interests include machine learning, deep learning and bioinformatics.

Qichang Zhao is a PhD candidate in the School of Computer Science and Engineering, Central South University, China. His current research interests include machine learning, deep learning and bioinformatics.

Feng Zhu is a Professor of the College of Pharmaceutical Sciences in Zhejiang University, China. His research laboratory (<https://idrblab.org/>) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry.

Jianxin Wang is the Dean and a Professor in the School of Computer Science and Engineering, Central South University, China. His current research interests include bioinformatics and algorithms.

Submitted: 9 August 2021; **Received (in revised form):** 29 September 2021

© The Author(s) 2021. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

The side effects of drugs refer to the additional adverse reactions that patients produce when standard doses are used to prevent, diagnose and treat diseases [1, 2]. Drug side effects are a heavy burden for patients and a serious obstacle to the development of new drugs, which has caused serious clinical burdens and economic losses [3, 4]. According to an analysis of more than 1 million medical insurance beneficiaries' visits [5], about one-seventh of the visits are related to the drug side effects. In the USA alone, severe drug side effects (drug toxicity) may cause more than 100 000 deaths each year [6]. Drug side effects are generally discovered by researchers through pharmacological experiments or clinical observations. These methods usually consume a lot of time and energy, and some minor or rare side effects are not easy to be discovered [7, 8]. The side effects of many drugs are not determined until many years after they have been on the market [9]. For example, an appetite suppressant called Fen-Phen withdrew from the market after the death of many patients who takes the suppressant [10]. Therefore, it is of great practical significance to analyze and predict the side effects of the drugs by means of bioinformatics [11].

In recent years, many computational models have been developed to predict the side effects of drugs based on drug-related databases [12–18]. However, most methods only explore whether a drug has one or more side effects and cannot determine the frequency of side effects, which is the central issue in drug risk-benefit assessment [19]. Accurate estimation of the frequencies of drug side effects is not only critical to patient care in clinical practice but also important to pharmaceutical companies because it reduces the risk of withdrawing drugs from the market [20, 21]. Although a few methods have been proposed to predict the frequencies of drug side effects, they all rely heavily on known drug-side effect associations or frequencies. For example, based on the known drug-side effect frequency information, Galeano *et al.* [22] constructed a drug-side effect adjacency matrix and proposed a new matrix factorization model to predict the frequencies of potential drug side effects. The model achieves good predictive performance, but when the given sample is a new drug without information on side effects, the method that relies on known drug-side effect frequencies will not be able to predict its potential side effects. In addition, Zhao *et al.* [23] developed a deep learning framework to predict the side effect frequencies of drugs by integrating chemical structure similarity, known drug-side effect frequency scores, side effect semantic similarity and pre-trained word vector representations. The model's core is to construct a drug-side effect bipartite graph and learn the feature representations of the node in the graph from the node's direct neighbors based on the attention mechanism. However, drugs that do not belong to the training data set have no neighbor nodes in the constructed heterogeneous graph, so the model cannot predict the side effects of new drugs.

As discussed above, most previous methods focus on predicting the presence or absence of drug side effects, and a few methods for predicting the frequencies of drug side effects but cannot be applied to new drugs. Therefore, there is a strong need for methods that can predict the side effects frequencies of new drugs. In this paper, we develop a novel multi-task learning framework, named SDPred, for drug-side effect associations and frequencies by integrating the multi-correlation between embedding of drugs and side effects. More specifically, we first collect the raw features of drugs and side effects based on the multiple types of similarity information, respectively.

Then, we perform projection operations on multiple similarity vectors of the drugs and side effects to ensure the consistency of the dimensions. Third, we use the outer product operation for each drug similarity vector and side effect similarity vector under the projection layer, explicitly capturing the pairwise correlations between the drug and side effect vectors. Fourth, we use a Convolution Neural Network (CNN) module to learn the high-order correlations among vector dimensions from locally to globally in a hierarchical way. Finally, we combine the drug embeddings, side effect embeddings and high-order correlation embeddings after the Multi-Layer Perceptron (MLP) and CNN module to predict the potential drug-side effect associations and frequencies. Compared with other state-of-the-art methods by our experimental results, our method can predict the side effect frequencies of the drugs without any known association or frequency information and achieve better performance on the datasets. Moreover, we perform ablation experiments and some case studies to illustrate the predictive capability of our model. All results show that our method could be an efficient tool to identify and discover potential side effect associations and frequencies of novel and known drugs.

Materials and Methods

Datasets

The dataset of the drug-side effect frequencies is first used in Galeano *et al.*'s study [22]. The frequency values of the drug-side effects are divided into five classes: very rare (frequency = 1), rare (frequency = 2), infrequent (frequency = 3), frequent (frequency = 4) and very frequent (frequency = 5). To collect the drug-related information, we download the names, compound IDs and Anatomical Therapeutic Chemical (ATC) codes of the drugs in Galeano *et al.*'s study [22] from the STITCH database [24], and then map the drugs to the Drugbank database [25] based on the names and ATC codes. After removing some drugs with no matching names and ATC codes in the above databases, we finally get 757 drugs and their corresponding Simplified Molecular Input Line Entry Specification (SMILES) sequences, ATC codes and targets. In total, the benchmark dataset DS₁ has 37 366 frequency items that containing 757 drugs and 994 side effects.

In this paper, two application scenarios are considered: (i) Discovering the missing association and frequency between known drugs and side effects, where the drugs and side effects are known, but their known association and frequency information is not complete. (ii) Discovering the side effects for new drugs, where the drugs are unknown, and there are no records of any side effects related to the drug. For the first scenario, the 10-fold cross-validation framework is used to test the performance of our model. For the second scenario, the independent test and de novo test are used to test the performance of our model. In the independent test, we introduce Zhao's dataset [18] that built on the SIDER2 [26] database to test the model. We find the overlapping drugs of the benchmark dataset and Zhao's dataset and then collect their side effect frequency information from the benchmark dataset to construct the training dataset DS₂. The rest of the benchmark dataset, named DS₃, is used as the independent test dataset. The statistical information of the three datasets is listed in Table 1.

Our model has two subtasks, one focuses on identifying the associations of the drug-side effects, and the other focuses on predicting the frequencies of the drug-side effects. Therefore, the positive and negative datasets are also needed to train the

Table 1. The details of the three datasets

Datasets	Drugs	Side effects	Frequencies
DS ₁	757	994	37,366
DS ₂	588	994	31,431
DS ₃	169	994	5,935

prediction model. Let n and m be the number of different drugs and side effects, respectively, $S_{\text{Drug}} = \{d_1, d_2, \dots, d_n\}$ represent the set of all n different drugs and $S_{\text{side-effect}} = \{s_1, s_2, \dots, s_m\}$ represent the set of all m different side effects. The Cartesian product $S_{\text{Drug}} \times S_{\text{side-effect}}$ contained all pairs of drugs and side effects, in which all known drug-side effect frequency pairs are positive samples and the rest pairs were not labeled. We select all positive samples to construct the Positive Sample Set (PSS) and sample a matching number of unlabeled pairs randomly to construct the negative sample set (NSS). For the convenience of the later sections, we construct two adjacency matrices to represent the two different relationships of drugs and side effects. For the drug-side effects associations, we construct an $m \times n$ dimensional matrix DSA as follows:

$$DSA(i, j) = \begin{cases} 1 & \text{if } d_i \text{ is related to } s_j \text{ in the dataset} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For the drug-side effect frequencies, we also construct an $m \times n$ dimensional matrix DSF . The value of the element's value at the corresponding position of matrix DSF is set to the frequency value, otherwise 0.

Drug similarity

Here, we construct ten matrices to represent the similarity between the drugs, respectively, based on chemical-chemical associations, chemical structures, target proteins, pre-trained word vectors, known drug-side effect associations and frequencies (see Figure 1).

Chemical-chemical association

It has been reported that interactive mode between chemicals is important information for drug-related prediction problems [27]. Based on the fact that two chemicals with an interactive relationship may have some common functions [28], we find the compound IDs of the drugs and search the compound-compound association scores from the STITCH database. Then, we can collect five chemical-chemical association scores named 'Similarity', 'Experimental', 'Database', 'Text mining' and 'Combined score' and use five adjacency matrices ($SMD_{\text{Similarity}}$, $SMD_{\text{Experimental}}$, SMD_{Database} , SMD_{Text} , SMD_{Combined}) $\in \mathbb{R}^{n \times n}$ to represent these five association scores between the drugs, respectively. Since all chemical-chemical association scores in STITCH range from 1 to 1000, we divide all scores by 1000 to ensure that the similarity value of drugs is between 0 and 1. In addition, we set the similarity value of each drug to itself as 1.

Drug structure similarity

The chemical substructure of drugs has been proved to be an effective feature for the prediction of the drug's side effects [29]. We collect the SMILES sequence of each drug from the STITCH database and input the obtained sequence into the RDKit [30], which is an open-source toolkit that uses machine learning

methods to generate compound descriptors and fingerprints. Then, each drug can be represented by a 2048-dimensional fingerprint vector and each dimension of the vector giving the presence (1) or absence (0) of a particular functional group in the molecule. The structure similarities between the two drugs d_i and d_j are computed according to the Jaccard score:

$$SMD_{\text{Structure}}(i, j) = \text{Jaccard score}(d_i, d_j) = \frac{|FV_i \cap FV_j|}{|FV_i \cup FV_j|}, \quad (2)$$

where FV_i and FV_j represent the fingerprint vector of d_i and d_j , respectively, and $SMD_{\text{Structure}}(i, j)$ represents the fingerprint similarity of drug d_i and d_j .

Drug target similarity

The target protein information of the drug is obtained from the DrugBank database. Each drug can be represented by an 847D target feature vector, and each dimension of the feature vector represents a protein. If the drug targets a protein, the corresponding dimension value of the feature vector is set to 1, otherwise, it is set to 0. The target similarities between two drugs d_i and d_j are computed according to the cosine similarity coefficient:

$$SMD_{\text{Target}}(i, j) = \frac{\sum_{k=1}^{847} TV_k^i \times TV_k^j}{\sqrt{\sum_{k=1}^{847} (TV_k^i)^2} \times \sqrt{\sum_{k=1}^{847} (TV_k^j)^2}}, \quad (3)$$

where TV^i and TV^j represent the target vector of d_i and d_j , respectively, TV_k^i and TV_k^j represent the k -th dimension of the TV^i and TV^j , respectively, and SMD_{Target} is an $n \times n$ matrix storing the drug target similarity.

Drug word similarity

Here, we use an unsupervised machine learning method [31], called Mol2vec, to learn the representations of drug molecular substructures. The idea of the model is similar to the Word2vec [32] model in the field of natural language process. Mol2vec can learn the vector representation of the drug substructures and then encode the drug into a vector by summing the vectors of each substructure of the drug. We feed the SMILES sequences of the drugs into the Mol2vec model and then get the 100-dimensional pre-trained word vector for each drug. We use the cosine similarity coefficient to calculate the word vector similarity between two drugs based on the pre-trained word vectors as follows:

$$SMD_{\text{Word}}(i, j) = \frac{\sum_{k=1}^{100} WV_k^i \times WV_k^j}{\sqrt{\sum_{k=1}^{847} (WV_k^i)^2} \times \sqrt{\sum_{k=1}^{847} (WV_k^j)^2}}, \quad (4)$$

where WV^i and WV^j represent the word vector of d_i and d_j , respectively, WV_k^i and WV_k^j represent the k th dimension of the WV^i and WV^j , respectively, and SMD_{Word} is an $n \times n$ matrix storing the drug word similarity.

Drug interaction profile similarity

Based on the assumption that similar drugs tend to show similar interaction and non-interaction patterns with side effects, the drug association profile similarity matrix SMD_{DIPA} and frequency profile similarity matrix SMD_{DIPF} are obtained based on

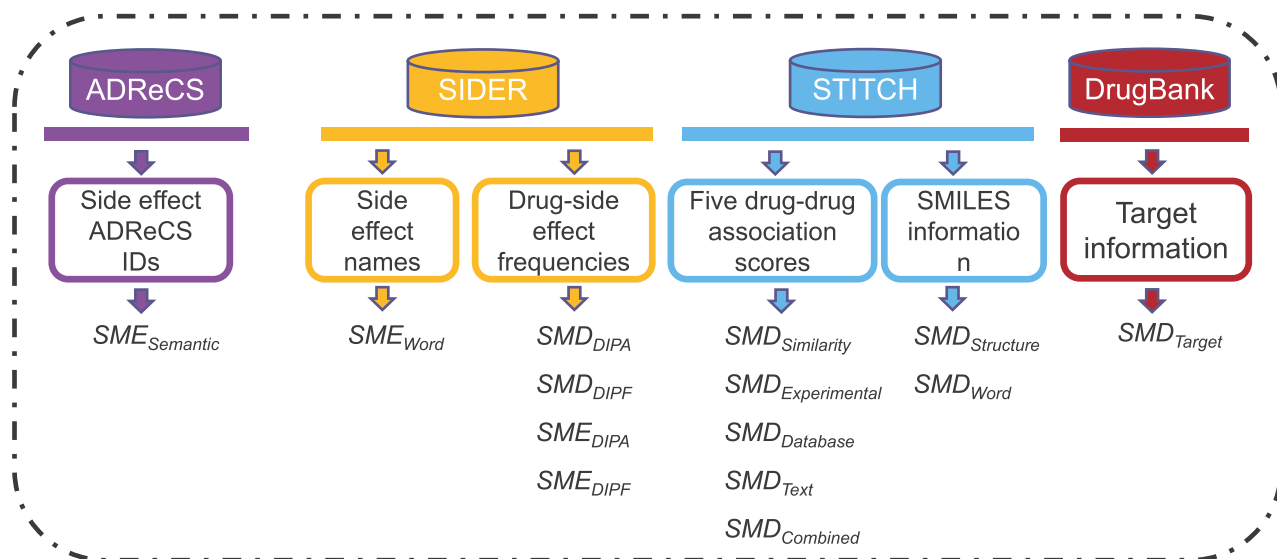


Figure 1. Data collection flowchart. Ten drug similarity matrices and four side effect similarity matrices are created from the ADReCS, SIDER, STITCH and DrugBank databases, where $SME_{Semantic}$ is constructed based on the ADReCS; SME_{Word} , SME_{DIPF} , SME_{DIPA} , SMD_{DIPA} and SMD_{DIPF} are constructed based on the SIDER; $SMD_{Structure}$, SMD_{Word} , $SMD_{Similarity}$, $SMD_{Experimental}$, $SMD_{Database}$, SMD_{Text} and $SMD_{Combined}$ are constructed based on the STITCH and SMD_{Target} is constructed based on the DrugBank.

the known drug-side effect associations and frequencies, respectively. The construction procedures of SMD_{DIPA} and SMD_{DIPF} are described as follows. Firstly, we define PA^{d_i} and PF^{d_i} to represent the association and frequency profile of d_i , respectively. PA^{d_i} is a binary vector encoding the presence or absence of association between drug d_i and all side effects, i.e. the i th row of DSA. PF^{d_i} is a vector encoding the presence or absence of frequency between drug d_i and all side effects, i.e. the i th row of DSF. Then, we introduce the cosine similarity coefficient for the association and frequency profile of drugs to calculate the similarity scores between drugs, respectively.

$$SMD_{DIPA}(i, j) = \frac{\sum_{k=1}^m PA_k^{d_i} \times PA_k^{d_j}}{\sqrt{\sum_{k=1}^m (PA_k^{d_i})^2} \times \sqrt{\sum_{k=1}^m (PA_k^{d_j})^2}}, \quad (5)$$

where $PA_k^{d_i}$ and $PA_k^{d_j}$ represent the k th dimension of the PA^{d_i} and PA^{d_j} , respectively.

$$SMD_{DIPF}(i, j) = \frac{\sum_{k=1}^m PF_k^{d_i} \times PF_k^{d_j}}{\sqrt{\sum_{k=1}^m (PF_k^{d_i})^2} \times \sqrt{\sum_{k=1}^m (PF_k^{d_j})^2}}, \quad (6)$$

where $PF_k^{d_i}$ and $PF_k^{d_j}$ represent the k th dimension of the PF^{d_i} and PF^{d_j} , respectively.

Side effect similarity

Here, we construct four matrices to represent the similarity between the side effects, respectively, based on the semantic value, pre-trained word vectors and known drug-side effect associations and frequencies (see Figure 1).

Side effect semantic similarity

Here, we calculate the side effect semantic similarity by using the existing measurement [23]. For each side effect, we construct a directed acyclic graph (DAG), which contains all the semantic descriptors related to the side effects. Figure 2 shows the DAGs of two side effects ‘Cerebral Infarction (CI)’ and ‘Angina Pectoris (AP)’. The DAG of a side effect such as ‘AP’ is denoted as $DAG(AP) = (T_{AP}, E_{AP})$, where T_{AP} is a set that includes all the ancestor descriptions of ‘AP’ and itself, and E_{AP} is a set of edges connecting these descriptions. The contribution of descriptor t in $DAG(AP)$ to the ‘AP’ is calculated by:

$$D_{AP}(t) = \begin{cases} 1, & \text{if } t = AP \\ \max(\theta \times D_{AP}(t^*) \mid t^* \in \text{children of } t), & \text{otherwise} \end{cases}, \quad (7)$$

where θ is a semantic contribution factor for the edges linking node t with its child t^* . As suggested in the study [23], it is set to 0.5. Then, the semantic value of ‘AP’ can be obtained by summing the contribution from all descriptors in $DAG(AP)$. The semantic similarity between the side effects s_i and s_j are calculated based on the semantic value as follows:

$$SME_{Semantic}(i, j) = \frac{\sum_{t \in (T_{s_i} \cap T_{s_j})} (D_{s_i}(t) + D_{s_j}(t))}{\sum_{t \in T_{s_i}} D_{s_i}(t) + \sum_{t \in T_{s_j}} D_{s_j}(t)}, \quad (8)$$

where $SME_{Semantic}$ is an $m \times m$ matrix storing the side effect semantic similarity.

Side effect word similarity

The side effect word vector is calculated using the method proposed by Zhao et al. [23]. Each side effect can be represented as a pre-trained 300D word vector. Then, the cosine correlation coefficient is used to measure the side effect word vector

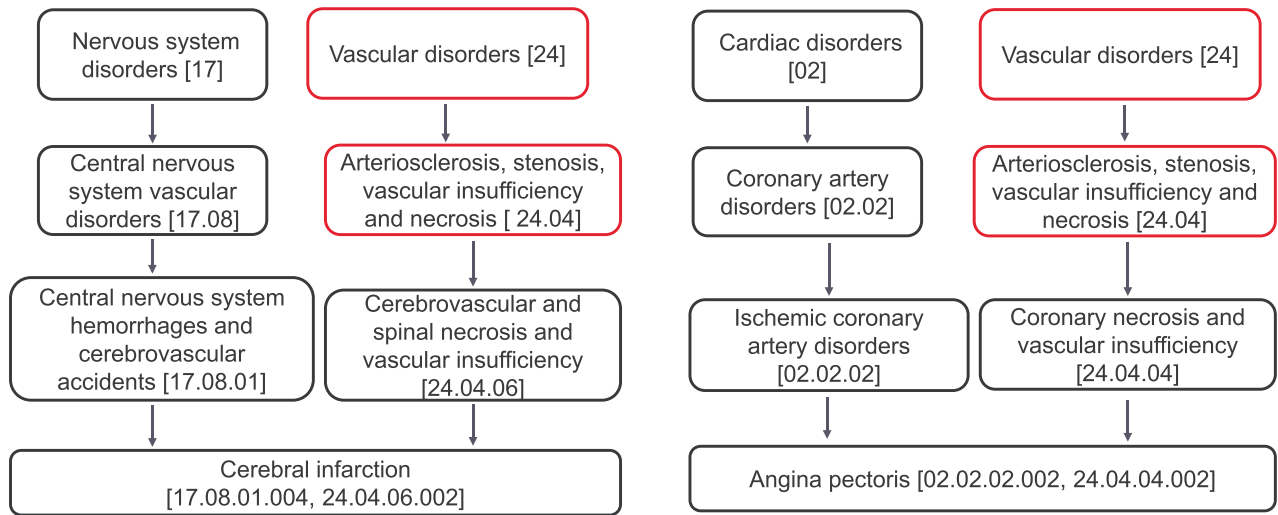


Figure 2. The side effect DAGs of CI and AP.

similarity of two side effects s_i and s_j , which is defined as:

$$SME_{\text{Word}}(i, j) = \frac{\sum_{k=1}^{300} SWV_k^i \times SWV_k^j}{\sqrt{\sum_{k=1}^{300} (SWV_k^i)^2} \times \sqrt{\sum_{k=1}^{300} (SWV_k^j)^2}}, \quad (9)$$

where SWV^i and SWV^j represent the word vector of s_i and s_j , respectively, SWV_k^i and SWV_k^j represent the k th dimension of the SWV^i and SWV^j , respectively, and SME_{Word} is an $m \times m$ matrix storing the side effect word similarity.

Side effect interaction profile similarity

The construction of the interaction feature similarity matrices for side effects is similar to drugs. We define PA^{s_i} and PF^{s_j} to represent the association and frequency profile of s_j , respectively. PA^{s_j} is a binary vector encoding the presence or absence of association between s_j and all drugs, i.e. the i th column of DSA. PF^{s_j} is a vector encoding the presence or absence of frequency between s_j and all drugs, i.e. the i th column of DSF. The side effect association profile similarity matrix SME_{DIPA} and frequency profile similarity matrix SME_{DIPF} can be calculated as follows:

$$SME_{\text{DIPA}}(i, j) = \frac{\sum_{k=1}^n PA_k^{s_i} \times PA_k^{s_j}}{\sqrt{\sum_{k=1}^n (PA_k^{s_i})^2} \times \sqrt{\sum_{k=1}^n (PA_k^{s_j})^2}}, \quad (10)$$

where $PA_k^{s_i}$ and $PA_k^{s_j}$ represent the k th dimension of the PA^{s_i} and PA^{s_j} , respectively.

$$SME_{\text{DIPF}}(i, j) = \frac{\sum_{k=1}^n PF_k^{s_i} \times PF_k^{s_j}}{\sqrt{\sum_{k=1}^n (PF_k^{s_i})^2} \times \sqrt{\sum_{k=1}^n (PF_k^{s_j})^2}}, \quad (11)$$

where $PF_k^{s_i}$ and $PF_k^{s_j}$ represent the k th dimension of the PF^{s_i} and PF^{s_j} , respectively.

SDPred

The key to developing a model that can accurately predict drug-side effect associations and frequencies is in (i) how to represent a drug and a side effect and (ii) how to model their interaction based on the representation. After obtaining the multiple similarities of drugs and side effects, we use the deep learning framework to build a novel multi-task model to predict the associations and frequencies between drugs and side effects. The model can be described as six steps (see Figure 3): (i) Collecting the features of the drugs and side effects based on the multiple types of similarities; (ii) Projecting the similarity vectors of different types of drugs and side effects into the vector space of the same dimension to represent the drugs and side effects; (iii) Using the outer product operations on each drug and side effect vector to obtain the interaction maps; (iv) Using the CNN module to extract the interaction embeddings from the interaction maps; (v) Integrating multi-type interaction embeddings, drug and side effect updated vectors to learn the drug-side effect pair representation vectors and (vi) Predicting the association scores of the drug-side effect pairs. If the drug-side effect pairs are predicted to be associated, predicting the frequency score of the drug-side effect pairs. Next, we take drug-side effect pair $d_i - s_j$ as an example and discuss the specific implementation details of each step.

In step 1, we define the drug similarity set $S_{\text{drug}}^{\text{sm}} = \{SMD_{\text{Combined}}, SMD_{\text{Text}}, SMD_{\text{Similarity}}, SMD_{\text{Experimental}}, SMD_{\text{Database}}, SMD_{\text{Structure}}, SMD_{\text{Target}}, SMD_{\text{Word}}, SMD_{\text{DIPA}}, SMD_{\text{DIPF}}\} \in \mathbb{R}^{n \times n}$ and side effect similarity set $S_{\text{side-effect}}^{\text{sm}} = \{SME_{\text{Semantic}}, SMD_{\text{Word}}, SMD_{\text{DIPA}}, SMD_{\text{DIPF}}\} \in \mathbb{R}^{m \times m}$ to represent multiple types of the drug and side effect similarities, respectively. Then, we collect the features of the drug d_i and side effect s_j based on $S_{\text{drug}}^{\text{sm}}$ and $S_{\text{side-effect}}^{\text{sm}}$, respectively. Concretely, for the k th drug similarity type (the k th element in $S_{\text{drug}}^{\text{sm}}$), the drug d_i can be described as a similarity vector $X_{d_i}^k$ as follows:

$$X_{d_i}^k = S_{\text{drug}}^{\text{sm}}(k)[i] = (v_1^k, \dots, v_i^k, \dots, v_h^k), \quad (12)$$

where $S_{\text{drug}}^{\text{sm}}(k)[i]$ represents the i th row of the k th element in $S_{\text{drug}}^{\text{sm}}$ and h represents the dimension of $S_{\text{drug}}^{\text{sm}}(k)[i]$. Similarly, for the l th side effect similarity type (the l th element in $S_{\text{side-effect}}^{\text{sm}}$), the side

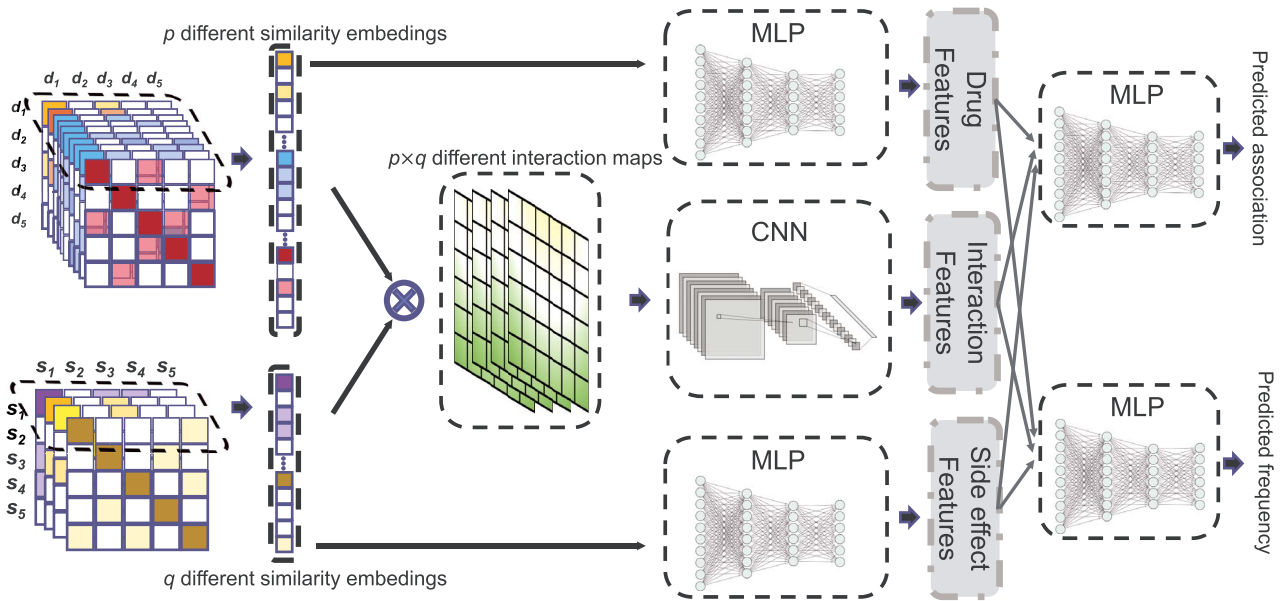


Figure 3. The architecture of SDPred.

effect s_j can be described as a similarity vector $X_{s_j}^l$ as follows:

$$X_{s_j}^l = S_{\text{side-effect}}^{\text{sm}}(l)[j] = (u_1^l, \dots, u_j^l, \dots, u_e^l), \quad (13)$$

where $S_{\text{side-effect}}^{\text{sm}}(l)[j]$ represents the j th row of the l th element in $S_{\text{side-effect}}^{\text{sm}}$ and e represents the dimension of $S_{\text{side-effect}}^{\text{sm}}(l)[j]$.

In step 2, according to the fact that the drugs and side effects have rich features and considering the heterogeneity of drugs and side effects, we project each drug and side effect similarity vector into the same vector space. The k th similarity vector of d_i and the l th similarity vector of s_j capture the k th interaction relationships between d_i and other drugs and l th interaction relationships between s_j and other side effects, respectively. Therefore, we design p and q (in this paper, $p=10$, $q=4$) similarity-specific transformation matrices for the drugs and side effects, respectively, to extract different features that correspond to particular interaction relationships, i.e. $(P_1, P_2, \dots, P_p) \in \mathbb{R}^{h \times r}$ and $(Q_1, Q_2, \dots, Q_q) \in \mathbb{R}^{e \times r}$, where r is a hyperparameter to represent the transformed dimensions of each similarity-specific vector. For the drug d_i , its k th drug similarity-specific vector H_k^i can be calculated as follows:

$$H_k^i = P_k X_{d_i}^k. \quad (14)$$

Similarly, for side effect s_j , its l th side effect similarity-specific vector E_l^j can be calculated as follows:

$$E_l^j = Q_l X_{s_j}^l. \quad (15)$$

Several recent efforts on neural network models [33, 34] have demonstrated that better prediction performance can be obtained by learning the interaction function from data. In step 3, we use the outer product operations between each drug similarity-specific vector and side effect similarity-specific vector. For the k th drug similarity-specific vector H_k^i and l th side effect similarity-specific vector E_l^j , the interaction map

Intermap $_{i,j}^{k,l}$ can be calculated as follows:

$$\text{Intermap}_{i,j}^{k,l} = H_k^i \otimes E_l^j, \quad (16)$$

where \otimes represents the outer product operation and Intermap $_{i,j}^{k,l}$ is a $r \times r$ matrix. Here, we can get $p \times q$ different interaction maps to represent drug-side effect pairs $d_i - s_j$. This is the core design of the SDPred framework to ensure that the model can capture the interaction relationships between drugs and side effects without any known drug-side effect associations or frequencies. The outer product operation is more meaningful than the simple concatenation operation because it only retains the original information in embeddings without modeling any interaction [35].

In step 4, we treat the pairwise interactions encoded in the multiple interaction maps as the local features of multiple images. As we all known, CNN has achieved great success in the field of computer vision. Here, we use a CNN module to extract the interaction embeddings between drugs and side effects from the multiple interaction maps. In the CNN module, each CNN layer includes three operations: (i) convolution operation; (ii) batch normalization operation and (iii) activation operation. The convolution operation can be described as follows:

$$x_{\text{Con}}^i = w_{\text{Con}}^i \circ t_{\text{Con}}^{i-1} + b_{\text{Con}}^i, \quad (17)$$

where w_{Con}^i represents the weight vector of the i th layer convolution kernel, t_{Con}^{i-1} represents the output of $(i-1)$ th layer, the arithmetic symbol \circ represents the convolution operation, and b_{Con}^i represents the offset vector of the i th layer. The Batch Normalization (BN), as an important achievement in the field of deep learning in recent years, has been widely proven its effectiveness and importance [36, 37]. The process of the BN of

the i th layer is described as follows:

$$x_{BN}^i = \text{BN}(x^i) = \frac{\alpha}{\sqrt{\text{Var}[x^i] + \epsilon}} \cdot x^i + \left(\beta - \frac{\alpha E[x^i]}{\sqrt{\text{Var}[x^i] + \epsilon}} \right), \quad (18)$$

where α , β and ϵ are parameters to be learned, $\text{Var}[\cdot]$ is the variance operation and $E[\cdot]$ is the mean operation. In this paper, we use the rectified linear unit (ReLU) function to perform the activation operation, which sets negative value to zero

$$x_{ReLU}^i = \text{ReLU}(x_{BN}^i) = \begin{cases} 0, & \text{if } x^i < 0 \\ x_{BN}^i, & \text{otherwise} \end{cases}. \quad (19)$$

The CNN module in SDPred has six hidden CNN layers, where the number of feature maps is 32, the stride is 2 and the convolution kernel is (2×2) in each hidden layer.

To obtain the high-order nonlinear relationships between the drug features and the side effect features, respectively. In step 5, we use the MLP module to extract the embeddings from features of drugs and side effects separately. The forward propagation process of the MLP module in this paper can be defined as follows:

$$t_{MLP}^1 = \text{ReLU}(w_{MLP}^1 t_{MLP}^0 + b_{MLP}^1) \quad (20)$$

$$t_{MLP}^2 = \text{ReLU}(w_{MLP}^2 t_{MLP}^1 + b_{MLP}^2) \quad (21)$$

$$\dots \quad (22)$$

$$t_{MLP}^g = \text{ReLU}(w_{MLP}^g t_{MLP}^{g-1} + b_{MLP}^g), \quad (23)$$

where g is the index of a hidden layer, w_{MLP}^1 , w_{MLP}^2 and w_{MLP}^g are the weight matrices, and b_{MLP}^1 , b_{MLP}^2 and b_{MLP}^g are the biases. We obtain the d_i and s_j updated features $X_{d_i}^{\text{sum}}$ and $X_{s_j}^{\text{sum}}$ by splicing all similarity-specific vectors of the d_i and s_j and input them into two similar MLP modules, respectively. Then, we concatenate the multiple types of interaction embeddings after the flattening operation, d_i and s_j updated features to learn the representation vector of the $d_i - s_j$ pair

$$X_{d_i-s_j}^{\text{sum}} = \left(X_{d_i}^{\text{sum}} \parallel X_{\text{intermap}} \parallel X_{s_j}^{\text{sum}} \right), \quad (24)$$

where $X_{d_i-s_j}^{\text{intermap}}$ is the interaction embeddings after the flattening operation.

In step 6, we adopt two MLP modules for the prediction tasks. The first MLP module is predicting the association scores between the d_i and s_j . If the d_i and s_j are predicted to be associated, the second MLP module is used to predict the frequency scores of the $d_i - s_j$ pair. All MLP modules used in this article have the same structure. Each module has three fully connected hidden layers, and the number of neurons in each hidden layer is r .

Model training

Our goal is to minimize the difference between the predicted score and the known drug-side effect associations and frequencies. Therefore, we use the mean square loss function to train the model. For the task of determining the associations between the drugs and side effects, the loss function is defined as follows:

$$\text{Loss}_1 = \sum_{i=1}^{M_1+M_2} (\text{Pre}_i^1 - Y_i^1)^2, \quad (25)$$

where M_1 and M_2 represent the number of positive and negative samples, respectively, the Pre_i^1 and Y_i^1 represent the true association label and predicted label of the sample i in the PSS and NSS. For the task of determining the frequencies between the drugs and side effects, the loss function is defined as follows:

$$\text{Loss}_2 = \sum_{i=1}^{M_1} (\text{Pre}_i^2 - Y_i^2)^2, \quad (26)$$

where Pre_i^2 and Y_i^2 represent the true frequency label and predicted label of the sample i in the PSS. Some studies have shown that regularization can effectively prevent model overfitting; we employ the L_2 regularization to our loss function. In addition, we only predict the frequency scores of drug-side effect pairs that are predicted to have association relationships. Therefore, we dynamically switch between one of two objective training losses based on whether there are available training examples (in the batch and for the current iteration) for each task. We train on two objectives jointly and the overall loss function becomes

$$\text{Loss}_{\text{total}} = \begin{cases} \text{Loss}_1 + \mu \sum_{i=1}^{M_3} \|\theta_i\|_2, & \text{if the sample is positive} \\ \text{Loss}_1 \times \text{Loss}_2 + \mu \sum_{i=1}^{M_4} \|\theta_i\|_2, & \text{otherwise} \end{cases}, \quad (27)$$

where M_3 and M_4 represent the number of the parameters, and μ is a positive constant that determines the degree of regularization influence. We train the model for a maximum of 100 epochs and plot the training loss curve in Figure 4. To avoid the overfitting problems, we randomly drop out hidden units in each fully connected layer.

There are some important hyperparameters in our model, namely the learning rate, the batch size, the dropout rate, r and μ . These hyperparameters are determined by grid-search on DS_1 . In grid-search, the learning rate is in $[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$, the batch size is in $[8, 16, 32, 64, 128, 256]$, the dropout rate is in $[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]$, r is in $[8, 16, 32, 64, 128]$ and μ is in $[1e-1, 1e-2, 1e-3, 1e-4, 1e-5]$. In general, the learning rate directly determines the performance, and the batch size is correlated with the learning rate. Therefore, we first determine the learning rate and the batch size in a grid-search. After the learning rate and batch size are fixed, we select the dropout rate and μ to improve the robustness of our model and determine the value of r . In total, the optimized learning rate, batch size, dropout rate, r and μ are $1e-4$, 128, 0.5, 32 and $1e-5$, respectively, and the detailed experimental results of hyper-parameters are shown in Supplementary Tables S1–S5, see Supplementary Data available online at <http://bib.oxfordjournals.org/>. In addition, we use the Adam optimization algorithm [38] to update the parameters of the model, which has shown excellent performance in deep learning tasks. We save the model with the current best performance before further training. If the best performance has

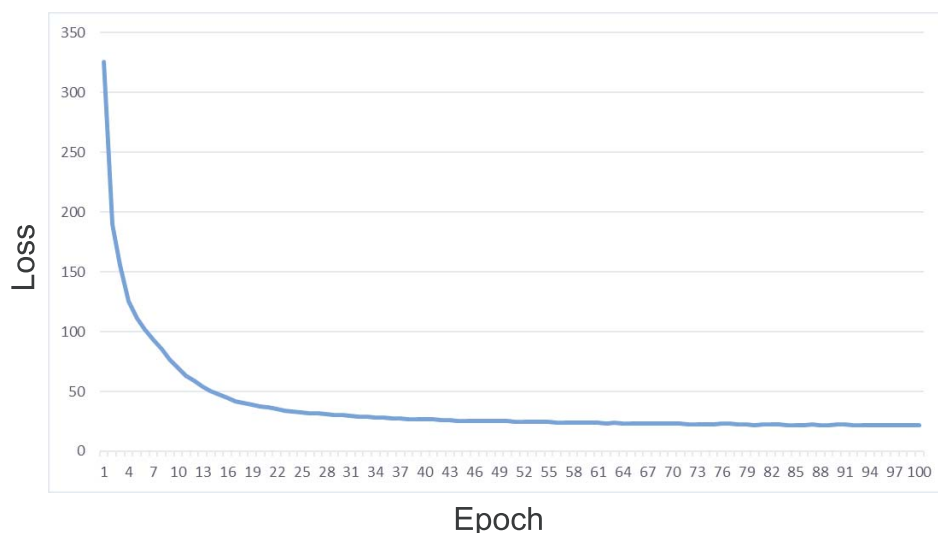


Figure 4. The loss function of our model at 100 epochs.

Table 2. Performance comparison of different methods on the DS_1

Model name	AUROC	AUPRC	RMSE	MAE
Galeano's method	0.9161	0.9165	1.3212	0.9642
MGPred	0.9234	0.9125	0.6510	0.4891
SDPred	0.9452	0.9367	0.5794	0.4212

not improved after 5 consecutive steps, the training will early stop.

Results

Evaluation metrics

To evaluate the performance of SDPred, we perform three kinds of computational experiments, including the 10-fold cross-validation, independent test, and de novo test. When the 10-fold cross-validation is implemented, each drug-side effect pair in PSS and NSS are divided randomly into 10 subsets, respectively. In the k th fold, the k th positive and negative subset are set as the testing set for model testing and the remaining nine positive and negative subsets are set as the training set for model training. The de novo test is designed to explore the performance of the model under cold boot conditions. As far as we know, none of the existing models for predicting the side effect frequencies of drugs can be applied to new drugs. In the de novo test, each drug in the dataset is left out in turn as the test sample and all known associations and frequencies between the drug and all existing side effects are deleted. We predict the association and frequency scores between the test drug and all side effects in the datasets. In addition, in the 10-fold cross-validation, the ten different drug similarities and four different side effect similarities are applied to construct the interaction maps. However, there is no information about the side effects of new drugs. Therefore, we remove two drug similarities matrices (SMD_{DIPA} and SMD_{DIPF}) and two side effect similarities matrices (SME_{DIPA} and SME_{DIPF}) in the de novo test and independent test.

Our model can simultaneously predict the presence/absence associations and frequencies of drug side effects. For the task of predicting the drug-side effect associations, we use the Area

Under the Receiver Operating Characteristic Curve (AUROC) and The Area Under the Precision-Recall Curve (AUPRC) to assess the performance of our model and other methods in comparison and a higher rank for the positive samples indicated better the prediction performance of the method. For the task of predicting the side effect frequencies of drugs, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to assess the performance of our model and other methods in comparison and a lower error between the predicted value and true value for positive samples indicated better the prediction performance of the method.

Performance comparison with other methods

To evaluate the performance of the SDPred model, we compare it with two state-of-the-art methods based on the benchmark dataset DS_1 , including Galeano's method and MGPred. For comparison fairness, the hyperparameters in Galeano's method and MGPred are set according to the optimal value as suggested by the authors. Table 2 and Figure 5 show the comparison results of Galeano's method, MGPred and SDPred. Our method gets 2.3, 2.7, 11.0 and 13.9% improvements in AUROC, AUPRC, RMSE and MAE over the second best MGPred. As shown by the results, even though MGPred utilizes the neural network, it only collects information from three angles and does not reasonably extract the interaction information between drugs and side effects, so its performance is not as good as our methods. In addition, a Wilcoxon test to evaluate the prediction results of all drugs in the benchmark dataset reveals that SDPred significantly outperforms the other methods. These results are observed using a P-value threshold of 0.05, with SDPred showing better performance in terms of both AUROCs, AUPRCs, RMSEs and MAEs (see Table 3).

De novo test

SDPred is an end-to-end deep learning method that can do two jobs: predicting potential side effect frequencies of an approved drug and predicting side effect frequencies of a new drug candidate. For the second work, we use the independent test and de novo test to evaluate the prediction performance of the model. In the independent test, the validation of the side effect

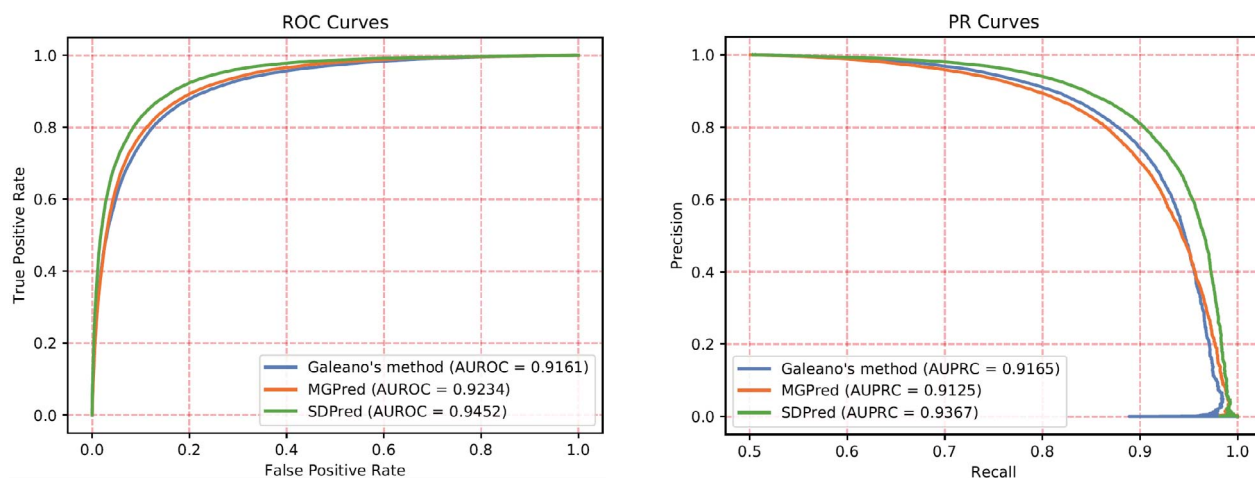


Figure 5. The ROC and PR curves of different models.

Table 3. Results of Wilcoxon test on SDPred and two other contrast methods for 757 drugs on the DS_1

SDPred and another method	Galeano's method	MGPred
P-value of AUROC	4.18×10^{-69}	2.54×10^{-21}
P-value of AUPRC	4.91×10^{-43}	2.51×10^{-62}
P-value of RMSE	5.67×10^{-119}	1.89×10^{-05}
P-value of MAE	5.12×10^{-117}	2.30×10^{-09}

Table 4. AUROC, AUPRC, RMSE and MAE values of SDPred in independent and de novo tests

Tests	AUROC	AUPRC	RMSE	MAE
Independent test	0.8457	0.8095	0.8391	0.6057
De novo test	0.8937	0.4285	0.7553	0.6092

frequency prediction performance is performed based on the independent test datasets DS_2 and DS_3 . Since Galeano's method and MGPred are strongly dependent on the known drug-side effect association and frequency information, the performance of the models cannot be tested using a completely independent dataset. Therefore, we only evaluate the performance of SDPred and the results are listed in Table 4. In the de novo test, the validation of the side effect frequency prediction performance is performed based on the benchmark dataset DS_1 . Figure 6A–D, respectively, show the distribution of RMSE, MAE, AUPRC and AUROC values of each drug in the de novo test.

Analysis of the contribution of each similarity

To build the SDPred, we introduce ten different drug similarities and four different side effect similarities. The feature information of new side effects and drugs may not be complete. The determination of the similarity with the most contribution for SDPred is an interesting problem. Accordingly, each drug or side effect similarity is singled out, and the nine drug or three side effect remaining similarities are used to extract drug features via the outer product, thereby constructing an SDPred. Thus, 14 SDPred models based on different similarity combinations are obtained, and the same hyperparameters mentioned in the Method section are used. These classifiers are also evaluated

Table 5. Performance of SDPred when one drug similarity is removed

Excluded similarity	AUROC	AUPRC	RMSE	MAE
SMD _{Similarity}	0.9385	0.9362	0.6051	0.4436
SMD _{Experimental}	0.9410	0.9374	0.5837	0.4363
SMD _{Database}	0.9406	0.9381	0.6013	0.4433
SMD _{Text}	0.9446	0.9408	0.5771	0.4281
SMD _{Combined}	0.9397	0.9371	0.5887	0.4319
SMD _{Structure}	0.9386	0.9356	0.5834	0.4404
SMD _{Target}	0.9419	0.9391	0.5931	0.4341
SMD _{Word}	0.9429	0.9391	0.5766	0.4267
SMD _{DIPA}	0.9406	0.9376	0.5959	0.4391
SMD _{DIPF}	0.9314	0.9269	0.6281	0.4545

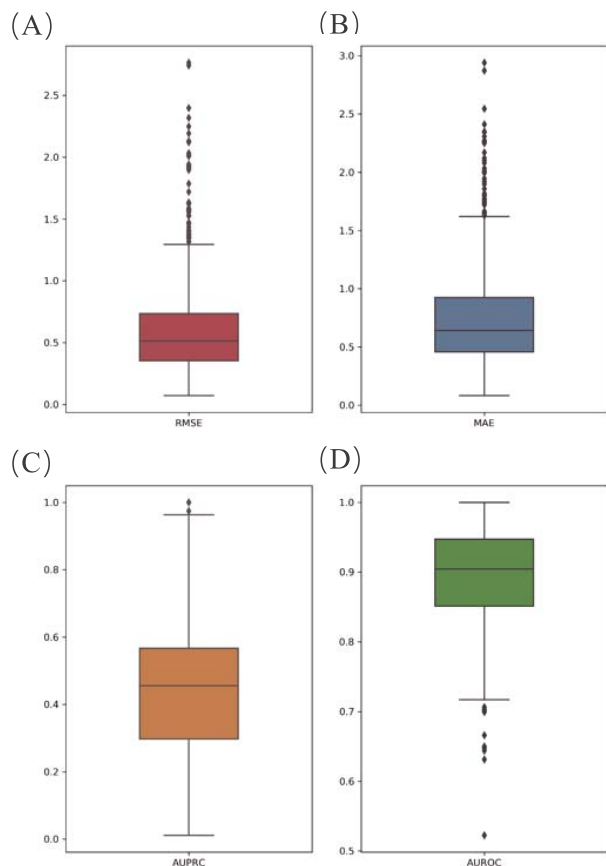
via 10-fold cross-validation based on the benchmark dataset DS_1 and the performance evaluation results are listed in Tables 5 and 6. The performance of the SDPred with few similarity information is lower than the original SDPred with all 14 different similarities, suggesting that different types of similarity information provide less or more contributions. After careful checking, for the drug similarities, the SDPred without SMD_{DIPF} yielded the lowest AUROC and AUPRC, but the highest RMSE and MAE. For the side effect similarities, the SDPred without SMD_{DIPA} yielded the lowest AUROC and AUPRC, but the highest RMSE and MAE. Therefore, such two types of similarities provide the most contribution. On the other hand, these results also prove that our model can achieve good predictive performance when less information about drugs and side effects is known.

Ablation experiment

For improving the performance of models in predicting drug side effect frequencies, our proposed model introduces the outer product operations between each drug similarity-specific vector and side effect similarity-specific vector and generates multiple interaction maps. To illustrate the effect of the outer product operation and interaction maps, we perform the ablation experiments using the 10-fold cross-validation. We compare our model with SDPred only using interaction embeddings (SDPred-Inter) and a simple MLP model with no interaction embeddings (SDPred-MLP) based on the benchmark dataset DS_1 . Table 7

Table 6. Performance of SDPred when one side effect similarity is removed

Excluded similarity	AUROC	AUPRC	RMSE	MAE
SME _{Semantic}	0.9388	0.9367	0.6065	0.4462
SME _{Word}	0.9394	0.9355	0.6044	0.4450
SME _{DIPA}	0.9303	0.9266	0.6101	0.4484
SME _{DIPF}	0.9409	0.9366	0.5788	0.4312

**Figure 6.** The distribution of RMSE, MAE, AUPRC and AUROC values of each drug in the de novo test.**Table 7.** Performance comparison of ablation experiment on DS₁

Model name	AUROC	AUPRC	RMSE	MAE
SDPred-MLP	0.9135	0.9129	0.6784	0.5067
SDPred-Inter	0.9416	0.9378	0.5872	0.4232
SDPred	0.9452	0.9426	0.5794	0.4212

shows the comparison results and we can find SDPred outperforms other methods, verifying the positive effect of using outer product above the embedding layer.

Case studies

To evaluate the capability of SDPred for predicting side effect frequencies of drugs in practical applications, we perform case studies of three drugs, including escitalopram, paroxetine and gabapentin and then analyzed their top 10 candidate side effects.

Table 8. The top 10 ranks of predictive side effects for drug escitalopram

Rank	Side effect	Frequency score	Evidence
1	influenza	3.6469	[41]
2	urinary retention	3.1214	SIDER
3	salivary hypersecretion	2.0950	NA
4	upper respiratory tract infection	3.6561	[42]
5	thrombocytopenia	2.7619	SIDER
6	gait disturbance	3.1590	SIDER
7	hyperaesthesia	2.8634	NA
8	dysarthria	3.1810	SIDER
9	diplopia	3.1683	SIDER
10	ecchymosis	2.8793	SIDER

Table 9. The top 10 ranks of predictive side effects for drug gabapentin

Rank	Side effect	Frequency score	Evidence
1	dehydration	2.8586	OFFSIDES
2	ataxia	3.3720	OFFSIDES
3	diplopia	3.1268	[43]
4	musculoskeletal stiffness	3.1067	OFFSIDES
5	vaginal infection	3.0857	NA
6	muscle contractions involuntary	2.8557	NA
7	gastroesophageal reflux disease	3.0137	NA
8	abdominal distension	2.9997	OFFSIDES
9	neuralgia	2.6622	OFFSIDES
10	delusion	2.6524	SIDER

The samples in the PSS and NSS are conducted to train the model, and the remaining drug-side effect pairs in the benchmark dataset are used as the test data. We list the escitalopram (Drugbank ID: DB01175), gabapentin (Drugbank ID: DB00996) and paroxetine (Drugbank ID: DB00715) top 10 candidate side effects and top 10 drug-side effect associations according to the prediction scores and confirmed them by the SIDER database [39], OFFSIDES database [40] and recently published experimental studies. The SIDER database contains information on marketed medicines and their recorded adverse drug reactions. The OFFSIDES database contains complementary information with the SIDER database and adds predictions for protein targets and drug indications. The top 10 ranked candidate side effects for each drug are given in Tables 8–10. We can find that SDPred has good predictive power for potential drug-related side effects and can narrow the scope of candidates for biological experiments.

Discussion and Conclusion

While drugs are treating patients' diseases, they may also have side effects that endanger the lives and health of patients. Reasonable analysis and prediction for the side effect frequencies of drugs can not only protect the health of patients but also have important implications for drug development. In this paper,

Table 10. The top 10 ranks of predictive side effects for drug paroxetine

Rank	Side effect	Frequency score	Evidence
1	psoriasis	2.4595	[44]
2	muscle twitching	3.0244	[45]
3	apathy	2.8770	OFFSIDES
4	flat affect	2.8852	OFFSIDES
5	hyperreflexia	2.5753	NA
6	lacrimal disorder	2.6172	NA
7	aggression	3.3847	SIDER
8	cholecystitis	2.3059	NA
9	hypotonia	2.6206	OFFSIDES
10	disorientation	3.2895	SIDER

we develop a new deep learning method of integrating multiple similarities of drugs and side effects for predicting the side effect frequencies of drugs. SDPred introduces the outer product operation for integrating multiple similarities and extracting useful features. Existing methods predict new associations and frequencies based on the known side effect information of drugs. However, the relationships between side effects and new drugs are unclear. It is difficult to capture sufficient association or frequency information, leading to insufficiently trained classification methods. In contrast, SDPred is a new end-to-end method based on multiple similarities and shares the advantages of both the matrix decomposition methods and deep learning methods and do not entirely dependent on the known relationships between drugs and side effects. The results of cross-validation and de novo experiments indicate that our method is capable of predicting side effect frequencies of both existing and novel drugs.

Although SDPred has demonstrated effective performances in predicting side effect frequencies of drugs, there is still room for improvement. Firstly, SDPred relies heavily on the similarity information, but the information of new side effects and drugs may not be complete. Secondly, the number of training samples is crucial for prediction and feature selection. But collecting side effect frequency information is difficult and the missing and noisy data could bring a negative impact on side effect frequency prediction. In the future, we plan to integrate gene expression information into our model and employ the oversampling algorithm to solve the problem of the insufficient number of positive samples.

Key points

- We propose a novel similarity-based deep learning architecture, named SDPred, for drug side effect frequency prediction by integrating the multiple similarities between pairwise drugs and pairwise side effects.
- The outer product operation is the core design of SDPred to ensure that the model can capture the interaction relationships between drugs and side effects without any known drug side effect associations or frequencies.
- SDPred is the first model that successfully addresses the problem of predicting the side effect frequencies of new drugs without known side effect frequency information.

Supplementary Data

Supplementary data are available online at <http://bib.oxfordjournals.org/>.

Funding

NSFC-Zhejiang Joint Fund for the Integration of Industrialization and Informatization under grant No. U1909208; National Natural Science Foundation of China (No. 61772552 and No. 61832019); 111 Project (No. B18059); Hunan Provincial Science and Technology Program (No. 2018WK4001); Scientific Research Fund of Hunan Provincial Education Department (No. 18B469).

References

1. Etkin NL. "side effects": cultural constructions and reinterpretations of western pharmaceuticals. *Med Anthropol Q* 1992;6(2):99–113.
2. Guo X, Zhou W, Yu Y, et al. A novel triple matrix factorization method for detecting drug-side effect association based on kernel target alignment. *Biomed Res Int* 2020;2020.
3. Berry DC, Knapp P, Raynor DK. Provision of information about drug side-effects to patients. *The Lancet* 2002; 359(9309):853–4.
4. Schwartz LM, Woloshin S, Welch HG. Using a drug facts box to communicate drug benefits and harms two randomized trials: Two randomized trials. *Ann Intern Med* 2009;150(8):516–27.
5. Da Silva BA, Krishnamurthy M. The alarming reality of medication error: a patient case and review of pennsylvania and national data. *Journal of community hospital internal medicine perspectives* 2016;6(4):31758.
6. Cappell MS, Schein JR. Diagnosis and treatment of non-steroidal anti-inflammatory drug-associated upper gastrointestinal toxicity. *Gastroenterol Clin North Am* 2000;29(1): 97–124.
7. Tatonetti NP. The next generation of drug safety science: coupling detection, corroboration, and validation to discover novel drug effects and drug–drug interactions. *Clinical Pharmacology & Therapeutics* 2018;103(2):177–9.
8. Voskens CJ, Goldinger SM, Loquai C, et al. The price of tumor control: an analysis of rare side effects of anti-ctla-4 therapy in metastatic melanoma from the ipilimumab network. *PloS one* 2013;8(1):e53745.
9. Rodriguez EM, Staffa JA, Graham DJ. The role of databases in drug postmarketing surveillance. *Pharmacoepidemiol Drug Saf* 2001;10(5):407–10.
10. Blanck HM, Khan LK, Serdula MK. Prescription weight loss pill use among americans: patterns of pill use and lessons learned from the fen-phen market withdrawal. *Prev Med* 2004;39(6):1243–8.
11. Xia X. Bioinformatics and drug discovery. *Curr Top Med Chem* 1709–1726;17(15):2017.
12. Zhang W, Liu F, Luo L, et al. Predicting drug side effects by multi-label learning and ensemble learning. *BMC bioinformatics* 2015;16(1):1–11.
13. Ding Y, Tang J, Guo F. Identification of drug-side effect association via semisupervised model and multiple kernel learning. *IEEE J Biomed Health Inform* 2018;23(6): 2619–32.

14. Ding Y, Tang J, Guo F. Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 2019;**325**:211–24.
15. Yamanishi Y, Pauwels E, Kotera M. Drug side-effect prediction based on the integration of chemical and biological spaces. *J Chem Inf Model* 2012;**52**(12):3284–92.
16. Liang H, Chen L, Zhao X, et al. Prediction of drug side effects with a refined negative sample selection strategy. *Comput Math Methods Med* 2020;**2020**.
17. Zhao X, Chen L, Guo Z-H, et al. Predicting drug side effects with compact integration of heterogeneous networks. *Current Bioinformatics* 2019;**14**(8):709–20.
18. Zhao X, Chen L, Lu J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math Biosci* 2018;**306**:136–44.
19. Luteijn JM, White BC, Gunnlaugsdóttir H, et al. State of the art in benefit–risk analysis: Medicines. *Food Chem Toxicol* 2012;**50**(1):26–32.
20. Breckenridge A, Woods K, Raine J. Monitoring the safety of licensed medicines. *Nat Rev Drug Discov* 2005;**4**(7):541–3.
21. Martin L, Hutchens M, Hawkins C, et al. How much do clinical trials cost. *Nat Rev Drug Discov* 2017;**16**(6):381–2.
22. Galeano D, Li S, Gerstein M, et al. Predicting the frequencies of drug side effects. *Nat Commun* 2020;**11**(1):1–14.
23. Zhao H, Zheng K, Li Y, et al. A novel graph attention model for predicting frequencies of drug-side effects from multi-view data. *Brief Bioinform* 07 2021a. <https://doi.org/10.1093/bib/bbab239>.
24. Kuhn M, vonMering C, Campillos M, et al. Stitch: interaction networks of chemicals and proteins. *Nucleic Acids Res* 2007;**36**(suppl_1):D684–8.
25. Wishart DS, Feunang YD, Guo AC, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res* 2018;**46**(D1):D1074–82.
26. Kuhn M, Campillos M, Letunic I, et al. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**(1):343.
27. Zhao H, Li Y, Wang J. A convolutional neural network and graph convolutional network-based method for predicting the classification of anatomical therapeutic chemicals. *Bioinformatics* 2021b;**03**. <https://doi.org/10.1093/bioinformatics/btab204>.
28. Chen L, Lu J, Zhang J, et al. Predicting chemical toxicity effects based on chemical-chemical interactions. *PLoS One* 2013;**8**(2):e56517.
29. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC bioinformatics* 2011;**12**(1):1–13.
30. Lovrić M, Molero JM, Kern R. Pyspark and rdkit: moving towards big data in cheminformatics. *Molecular informatics* 2019;**38**(6):1800082.
31. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**(1):27–35.
32. Ma L, Zhang Y. Using word2vec to process big text data. In: *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, 2015, 2895–7.
33. Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. Latent relational metric learning via memory-based attention for collaborative ranking. In *Proceedings of the 2018 World Wide Web Conference*, pages 729–739, 2018.
34. Bai T, Wen J-R, Zhang J, et al. A neural collaborative filtering model with interaction-based neighborhood. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, 1979–82.
35. He X, Chua T-S. Neural factorization machines for sparse predictive analytics. In: *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2017, 355–64.
36. Huang N, Nie F, Ni P, et al. Neuralpolish: a novel nanopore polishing method based on alignment matrix construction and orthogonal bi-gru networks. *Bioinformatics* 2021. <https://doi.org/10.1093/bioinformatics/btab354>.
37. Ni P, Huang N, Zhang Z, et al. Deepsignal: detecting dna methylation state from nanopore sequencing reads using deep-learning. *Bioinformatics* 2019;**35**(22):4586–95.
38. Jais IKM, Ismail AR, Nisa SQ. Adam optimization algorithm for wide and deep neural network. *Knowledge Engineering and Data Science* 2019;**2**(1):41–6.
39. Kuhn M, Letunic I, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic Acids Res* 2016;**44**(D1):D1075–9.
40. Tatonetti NP, Patrick PY, Daneshjou R, et al. Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**(125):125ra31–1.
41. Emslie GJ, Ventura D, Korotzer A, et al. Escitalopram in the treatment of adolescent depression: a randomized placebo-controlled multisite trial. *J Am Acad Child Adolesc Psychiatry* 2009;**48**(7):721–9.
42. Bielski RJ, Bose A, Chang C-C. A double-blind comparison of escitalopram and paroxetine in the long-term treatment of generalized anxiety disorder. *Ann Clin Psychiatry* 2005;**17**(2):65–9.
43. Wilson EA, Sills GJ, Forrest G, et al. High dose gabapentin in refractory partial epilepsy: clinical observations in 50 patients. *Epilepsy Res* 1998;**29**(2):161–6.
44. Osborne SF, Stafford L, David KG, Orr. Paroxetine-associated psoriasis. *Am J Psychiatry* 2002;**159**(12):2113–3.
45. Milliken C, Cooper SJ. Withdrawal symptoms from paroxetine. *Human Psychopharmacology: Clinical and Experimental* 1998;**13**(3):217–9.