OXFORD

Briefings in Bioinformatics, 2022, **23(2)**, 1–13

https://doi.org/10.1093/bib/bbab586

Problem Solving Protocol

DSGAT: predicting frequencies of drug side effects by graph attention networks

Xianyu Xu, Ling Yue, Bingchun Li, Ying Liu, Yuan Wang, Wenjuan Zhang and Lin Wang

Corresponding authors: Lin Wang, College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin 300457, China. Tel.: +86-22-60600986; Fax: +86-22-60600022. E-mail: linwang@tust.edu.cn; Wenjuan Zhang, College of General Education, Tianjin Foreign Studies University, Tianjin 300204, China. Tel.: +86-22-63353131; Fax: +86-22-23285749. E-mail: juanzi_1982_49@126.com

Abstract

A critical issue of drug risk–benefit evaluation is to determine the frequencies of drug side effects. Randomized controlled trail is the conventional method for obtaining the frequencies of side effects, while it is laborious and slow. Therefore, it is necessary to guide the trail by computational methods. Existing methods for predicting the frequencies of drug side effects focus on modeling drug–side effect interaction graph. The inherent disadvantage of these approaches is that their performance is closely linked to the density of interactions but which is highly sparse. More importantly, for a cold start drug that does not appear in the training data, such methods cannot learn the preference embedding of the drug because there is no link to the drug in the interaction graph. In this work, we propose a new method for predicting the frequencies of drug side effects, DSGAT, by using the drug molecular graph instead of the commonly used interaction graph. This leads to the ability to learn embeddings for cold start drugs with graph attention networks. The proposed novel loss function, i.e. weighted ε-insensitive loss function, could alleviate the sparsity problem. Experimental results on one benchmark dataset demonstrate that DSGAT yields significant improvement for cold start drugs and outperforms the state-of-the-art performance in the warm start scenario. Source code and datasets are available at https://github.com/xxy45/DSGAT.

Keywords: side effect frequency, chemical structure, cold start, deep learning, graph attention network

Introduction

Drug risk–benefit evaluation refers to the assessment between the therapeutic benefits that patients can obtain after using the drug and the risks they bear, such as the benefit–risk assessment for remdesivir and lopinavir-ritonavir in treatment of COVID-19 [1, 2]. The central issue of this assessment is to determine the frequencies of drug side effects [3]. Now the standard method for frequency acquisition is randomized controlled trail, i.e. the study subjects are randomly grouped, and different interventions are implemented in different groups to contrast the different effects [4].

However, due to the time window and limited sample size, some side effects were not discovered in the clinical trail and appeared after many years on the market [5]. Therefore, drug side effects are still the main cause of illness and death in medical health [6]. Meanwhile, side effects are the leading factors in the delisting of drugs, which brought about the failure of drug research and development and the loss of huge funds [7, 8]. Some computational methods [9–17] were adopted to predict the side effects of a given drug, but most of these methods can only predict the presence or absence of side effects, while cannot predict their frequencies, which limits the application of these methods in drug risk–benefit evaluation.

Galeano *et al.* proposed the first computational method that predicts the frequencies of drug side effects by using non-negative matrix factorization [18]. Their approach used solely drug–side effect frequency matrix, and so cannot be adopted to predict side effect frequencies for cold start drugs. Recently, Zhao *et al.* presented a graph

Xianyu Xu is a student at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. His current research interests include bioinformatics and machine learning.

Ling Yue is a student at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. Her current research interests include bioinformatics and algorithms.

Bingchun Li is a student at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. Her current research interests include bioinformatics and machine learning.

Ying Liu is a professor at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. Her current research interests are big data analysis and intelligent information processing.

Yuan Wang is an associate professor at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. Her current research interests are computational biology and natural language processing.

Wenjuan Zhang is an assistant professor at the College of General Education, Tianjin Foreign Studies University, Tianjin, China. Her current research interests are bioinformatics and algorithms.

Lin Wang is an associate professor at the College of Artificial Intelligence, Tianjin University of Science and Technology, Tianjin, China. His current research interests include bioinformatics, big data analysis and machine learning.

Received: October 6, 2021. **Revised:** December 1, 2021. **Accepted:** December 20, 2021

attention model for predicting frequencies of drug–side effects from multi-view data, including the similarity, available drug–side effect frequencies and word embedding [19]. Their method learns the drug embedding and side effect embedding from drug–side effect interaction graph. Though their method predicted already known drug–side effect frequencies with a lower RMSE, it often introduces false positives in practical usage, i.e. unknown drug–side effect associations tend to be predicted as the frequent class.

Here we propose a novel deep learning model named DSGAT for drug–side effect frequency prediction by using graph attention networks (GAT) [20]. DSGAT learns the drug representation from the molecular graph, and has the ability to predict the frequencies of side effects for cold start drugs that do not appear in the training data. Experimental results on the benchmark dataset show that DSGAT yields the significant performance improvement for cold start drugs and exceeds the other competing methods in warm start scenario. Data distribution analysis and independent test all illustrate that DSGAT could not only accurately discriminate true drug–side effect associations from unknown associations with high AUC values, but also correctly predict frequencies of side effects with relative high Spearman's correlation coefficients (SCCs).

Material and methods

Dataset

Here, we used the benchmark dataset used in Galeano *et al.* [18] and Zhao *et al.* [19] to validate the utility of our drug–side effect frequency prediction method. It includes 750 drugs and 994 side effects and 37 071 known frequency items that were derived from SIDER database version 4.1 [21]. The frequency items of drug side effects are categorized into five classes and are consequently encoded with integers: very rare (frequency = 1), rare (frequency = 2), infrequent (frequency = 3), frequent (frequency = 4) and very frequent (frequency = 5). Among the frequency items, the percentages of very rare, rare, infrequent, frequent and very frequent terms are 3.21%, 11.29%, 26.92%, 47.46% and 11.12%, respectively. We represent the frequencies between drugs and side effects by a rating matrix M, where a nonzero rating value indicates a known frequency for a specific drug–side effect pair and 0 otherwise. The rating matrix M is extremely sparse and nonzero elements account for only 4.97%. Our goal is to predict drug–side effect frequency values with using GAT and matrix factorization.

DSGAT

Here we present a model named DSGAT that tackles the drug-side effect frequency task by adopting the popular encoder-decoder framework (Figure 1). We assume that similar drugs may cause similar side effects and that similar side effects may appear for similar drugs. Here, we try to extract complicated relationship from side effect similarity graph via embedding method GAT. As to the drug, we can obtain its representation from the chemical molecular graph via GAT, too. Then, we consider using the matrix factorization as the decoder.

Graph construction

Each drug has its own unique chemical structure, which can be naturally represented as a graph, where the vertices and edges represent chemical atoms and chemical bonds, respectively. The intrinsic chemical structures of drugs have been employed to predict drug–drug interactions [22], drug–target interactions [23] and drug responses [24, 25]. For each drug d_i ($i = 1, 2, \ldots, m$), its graph representation can be denoted as $G_i = (H_i, A_i)$, where m is the number of drugs used in our study, $H_i \in \mathbb{R}^{K_i \times F}$ and $A_i \in \mathbb{R}^{K_i \times K_i}$ are the feature matrix and adjacent matrix of the drug d_i. K_i is the number of atoms in drug d_i, and F is the number of atom features. The properties of each atom in the compound are expressed as a 109 dimensional multi-hot vector ($F = 109$), including chemical and topological attributes, that is, atom type, degree, number of adjacent hydrogens, valence, formal charge, hybridization and whether it is in aromatic hydrocarbon. Specifically, each attribute is encoded with a one-hot vector, and the atom features are achieved by concatenating the multiple one-hot vectors. We obtain the Simplified Molecular Input Line Entry Specification notation of a given drug from DrugBank [26] and PubChem [27], which is then fed into the Open-Source Cheminformatics (RDKit) package in python to compute the atom attributes.

For any two side effects, such as e_i and e_j, we extract their corresponding frequency profiles from drug–side frequency matrix M, i.e. the ith and jth columns of M, and then compute their cosine similarity as follows:

$$\mathbf{Sim}\left(e_i, e_j\right) = \frac{M(i)^T M(j)}{\|M(i)\| \, \|M(j)\|}, \tag{1}$$

where $M(i)$ and $M(j)$ represent the ith and jth columns of M. It is noted that, as to the similarity computation, the side effect frequencies in the test set are set as 0 values, and thus the problem of data leakage is avoided. To avoid time-consuming computation and noise introduction, we use k-nn neighbor graph ($k = 10$ for warm start scenario and $k = 5$ for cold start scenario), which is a sparse graph, to represent the connections between k-nearest neighbors of each side effect and the side effect itself as follows:

$$Ae\left(e_i, e_j\right) = \begin{cases} 1 \; if \; e_j \in N\left(e_i\right) \\ 0 \;\; \text{otherwise} \end{cases}, \tag{2}$$

where $N(e_i)$ is a set of side effect e_i's neighbors whose similarities to e_i, i.e. $\mathbf{Sim}(e_i, e_j)$, are in the top k. Each of the 994 side effects in our dataset is annotated across the top two description levels of the MedDRA terminology hierarchy [28], namely level 1 (i.e. System Organ Class)

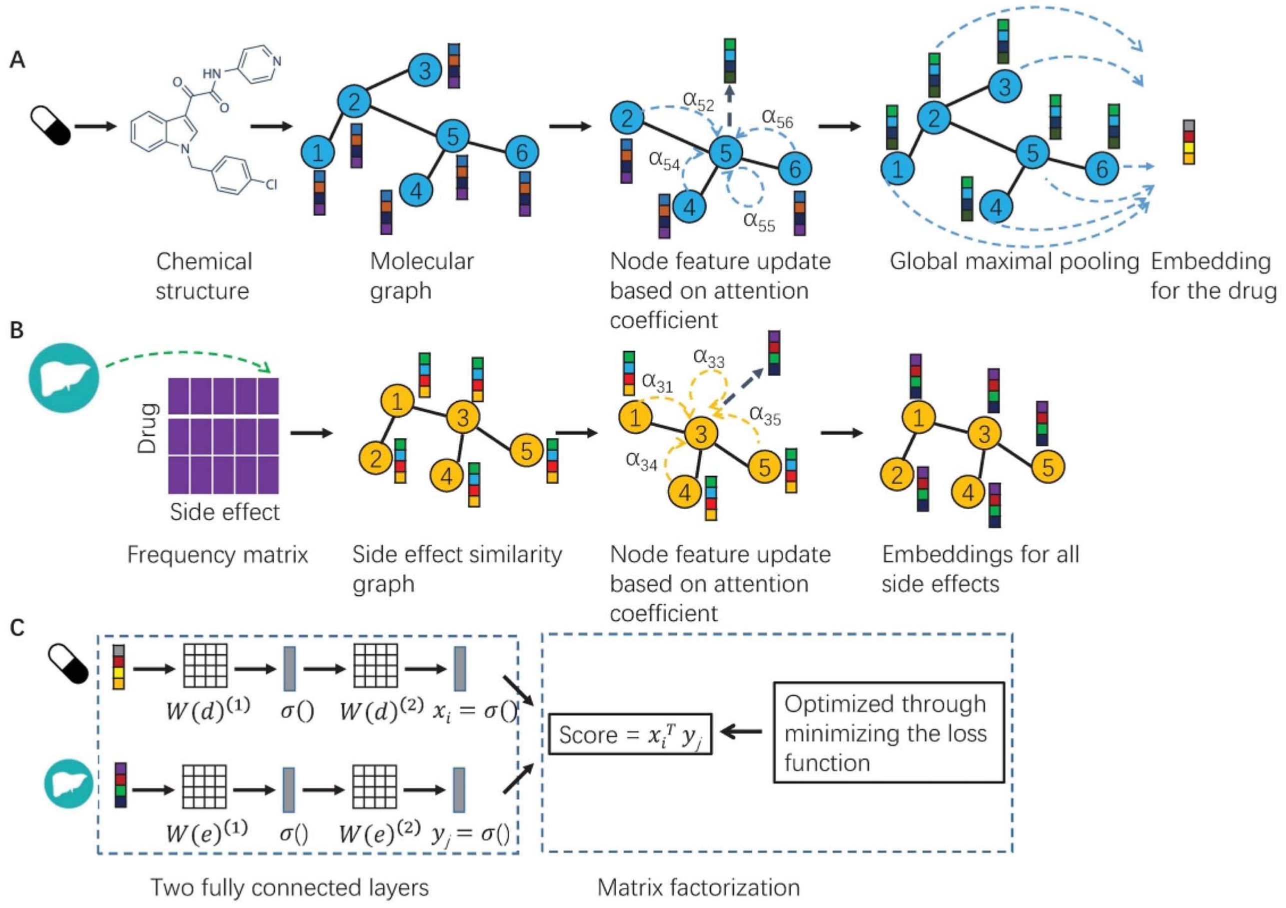


Figure 1. The flowchart of our model DSGAT. (**A**) For each drug, based on the chemical structure, we obtain the corresponding molecular graph by using RDKit package. The nodes and edges represent chemical atoms and chemical bonds, respectively. The properties of each atom in a compound are encoded as a multi-hot vector, which represents chemical and topological attributes. Then, a three-layer GAT network is applied to the molecular graph to obtain the embeddings of all atoms. After that, we get the embedding for each drug by using global maximal pooling. (**B**) We measure the similarity between any two side effects based on the cosine similarity of their corresponding frequency profiles and use *k*-nearest neighbors to construct the side effect graph. Each node of the graph, the side effect, is encoded by the first two description levels of the MedDRA terminology hierarchy. Then, a three-layer GAT network is applied to the side effect graph to obtain the embeddings of all side effects. (**C**) Once obtaining the embeddings for any pair of drug and side effect, we project these embeddings into a shared space by adding two fully connected layers, respectively. Then, we use matrix factorization as the decoder, i.e. using the inner product of these projections as the prediction score, which is optimized through minimizing the loss function.

and level 2 (i.e. High-Level Group Term). MedDRA categories for each given side effect are encoded with a 243 dimensional multi-hot feature vector. In total, the side effect similarity graph can be described as $G_e = (H_e, A_e)$, where $H_e \in \mathbb{R}^{K_e \times F_e}$ and $A_e \in \mathbb{R}^{K_e \times K_e}$ are the feature matrix and adjacent matrix. K_e=994 is the number of side effects in this study, and F_e=243 is the number of side effect features.

Model architecture

We use GAT to learn the representation of the molecular graph [29–31]. GAT introduces a self-attention mechanism in the propagation process, that is, the node representation is updated according to the attention of each node on its neighboring nodes. Given the molecular graph containing K atoms and the F dimension feature h_i of atom u_i ($i = 1, 2, \ldots, K$), for adjacent atoms u_j of atom u_i [i.e. $j \in N(i)$], we could first use attention mechanism to compute the importance of atom u_j to atom u_i as follows:

$$e_{ij} = \mathbf{LeakyReLU}\left(a^T\left[Wh_i \middle\| Wh_j\right]\right), \tag{3}$$

$$\alpha_{ij} = \mathbf{Softmax}_j\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k\in N(i)} \exp\left(e_{ik}\right)}, \tag{4}$$

where W is a linear transformation matrix with dimensions of $F \times F'$, $a \in \mathbb{R}^{2F'}$ is the weight vector of a single-layer feedforward neural network and $Wh_i \| Wh_j$ means the concatenation of Wh_i and Wh_j, which is the input of the single-layer neural network. The output is then passed through is a nonlinear activation function LeakyReLU, and the Softmax is introduced to normalize the outputs across all adjacent atoms u_j of u_i. Notably, adjacent atoms of u_i also contain u_i, i.e. $i \in N(i)$. For each atom u_i, after obtaining the normalized attention coefficients α_{ij} between adjacent atoms u_j and u_i, we could get the F'

dimension embedding h_i' of atom u_i as follows:

$$h_i' = \sigma\left(\sum\nolimits_{j\in N(i)} \alpha_{ij} W h_j\right), \tag{5}$$

where σ is the nonlinear activation function, such as Relu. Furthermore, if there are L attentions, the vectors generated by the L attentions need to be concatenated as follows:

$$h_i' = \Big\|_{l=1}^{L} \sigma\left(\sum\nolimits_{j\in N(i)} \alpha_{ij}^{l} W^{l} h_j\right), \tag{6}$$

where α_{ij}^{l} and W^l are attention coefficient and linear transformation for the lth attention mechanism. But, if it is the last GAT layer, we just use single-head attention as Equation (5).

In this article, we set the number of attention heads $L = 10$, and use a three-layer GAT network to obtain the embeddings of atoms in a given drug and then adopt global max pooling to get the embedding of the drug molecular graph. Similarly, we employ a three-layer GAT network on the side effect similarity graph to obtain the embeddings of side effects. Here, we use p_i to denote the embedding of drug d_i $(i = 1, \dots, m)$, and q_j to denote the embedding of side effect s_j $(j = 1, \dots, n)$. Then, two two-layer fully connected networks are used on p_i and q_j, respectively, to project the embeddings of drugs and side effects into a common space as follows:

$$x_i = W(d)^{(2)}\sigma\left(W(d)^{(1)}p_i + b(d)^{(1)}\right) + b(d)^{(2)}, \tag{7}$$

$$y_j = W(e)^{(2)}\sigma\left(W(e)^{(1)}q_j + b(e)^{(1)}\right) + b(e)^{(2)}, \tag{8}$$

where $W^{(i)}$ and $b^{(i)}$ are the weight and bias of ith layer and σ represents activation function Relu. At last, we consider using the matrix factorization [32] to get the preference matrix $S \in \mathbb{R}^{m\times n}$ as

$$S_{ij} = {x_i}^T y_j. \tag{9}$$

Weighted ε-insensitive loss function

We adopt the loss function that is to minimize the Frobenius norm of the difference between preference matrix and label matrix. Compared to the number of drug–side effect pairs, the number of known frequency is far less, which makes the frequency matrix M extremely sparse. For the dataset we use, there are only 37 071 discovered drug–side effect frequency pairs that cover 750 drugs and 994 side effects. To overcome this, for unknown drug–side effect associations, we control the margin between the predicted scores and the labels with a fixed variable ε= 0.5. Besides, the known drug–side effect associations are more trustworthy and important than the unknown associations for improving prediction performance. We set a tuning weight α= 0.03 for the loss of the unknown drug–side effect associations. In summary, we could get our weighted ε-insensitive loss function as follows:

$$\left\|I^{\Omega}\circ(S-M)\right\|_F^2 + \alpha\left\|I^{o}\circ(S-\varepsilon A)\right\|_F^2, \tag{10}$$

where I^{Ω} and I^{o} are mapping functions to distinguish between the known and unknown items in the rating matrix M, i.e. ${I_{ij}}^{\Omega}$=1 if M_{ij}>0 or ${I_{ij}}^{\Omega}$=0 otherwise; ${I_{ij}}^{O}$=1 if M_{ij}=0 or ${I_{ij}}^{O}$=0 otherwise, S is the predicted score matrix, A is a matrix of all ones, $\|\cdot\|_F$ is the Frobenius norm and $\circ$ denotes the Hadamard product of two matrices.

Optimization

We adopt Adam optimization method to optimize the loss function Equation (10) by PyTorch. For Adam, we empirically set the learning rate lr = 0.0001 and weight decay wd = 0.001 by cross-validation (CV).

CV on the benchmark dataset

In order to evaluate the performance of DSGAT in the benchmark dataset, we use 10-fold CV under two settings, i.e. warm start scenario and cold start scenario. First, for the dataset, the frequencies of all known drug–side effect pairs are randomly divided into 10-folds, which are almost the same size. For each fold, we set it as the test set, and the remaining 9-folds are used as the training set. We denote this CV setting as CVS1. Second, we conduct *de novo* drug–side effect frequency prediction to verify the ability of DSGAT in predicting the frequencies of potential side effects of new drugs. Specifically, all drugs were randomly divided into 10 subsets of almost the same size. In the CV test, the drug–side effect frequencies of one subset are taken as the test set, and the drug–side effect frequencies of the other nine subsets form the training set. We represent this CV setting as CVS2.

The accuracy of a prediction model is measured from two aspects, i.e. performance for identifying drug–side effect association and performance for frequency prediction. As to association evaluation, five metrics, namely area under the PR curve (AUPR), area under the receiver operating characteristic curve (AUC), normalized discounted cumulative gain (NDCG@N), Precision@N and Recall@N, are adopted to evaluate the performance. The first two indicators are widely used to assess ranking accuracy. Because we are usually interested in some top-ranked side effects, we use NDCG@N, precision@N and recall@N to evaluate the top N recommendation performances as follows:

$$\mathrm{NDCG@}N = Z_N \sum\nolimits_{i=1}^{N} \frac{r(i)}{\log_2(i+1)}, \tag{12}$$

$$\mathrm{Precision@}N = \frac{\mathrm{TP}}{N}, \tag{13}$$

$$\mathrm{Recall@}N = \frac{\mathrm{TP}}{T}, \tag{14}$$

where $r(i) \in \{0, 1\}$ is the relevance score, Z_N is the normalization, TP is the number of positive samples in the top N recommendations and T is the number of positive samples in the test set. In our experiment, we set $N = 10$ for NDCG and $N = \{1, 15\}$ for precision and recall. As to each indicator, we first calculate the indicator value for each drug on the test set and then report the average indicator of all drugs. Specifically, for each given drug, its side effects with known frequencies in the test set and its unknown side effects in the rating matrix M are deemed as positive and negative labels, respectively. The average AUPR of all drugs is also defined as mean average precision (MAP). We compute a final metric score that is the average over 10-fold CVs.

As to frequency prediction, we use SCC, root mean square error (RMSE) and mean absolute error (MAE) as the evaluation metrics. The definitions of RMSE and MAE are as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{d,e} \left(S_{d,e} - M_{d,e}\right)^2}{t}}, \tag{15}$$

$$\mathrm{MAE} = \frac{\sum_{d,e} \mid S_{d,e} - M_{d,e} \mid}{t}, \tag{16}$$

where t is the total number of known drug–side effect frequency pairs being tested, $S_{d,e}$ is the predicted score and $M_{d,e}$ is the ground truth.

Results

Similar side effects have similar frequency profiles across drugs

We encoded MedDRA categories for each side effect in our study with a 243-dimensional multi-hot vector. Here, we defined the category similarity (cs) between two side effects by calculating the Jaccard coefficient between their multi-hot encodings. Furthermore, for any two side effects, we calculated the cosine similarity between their frequency profiles across drugs as Equation (1). Figure 2 shows that the higher the cs values of side effect pairs, the higher their frequency profile similarity values. For side effect pairs with $\mathrm{cs} > 0.5$, $0.2 < \mathrm{cs} \leq 0.5$ and $\mathrm{cs} \leq 0.2$, the arithmetic mean values of their frequency profile similarity are 0.196, 0.144 and 0.111, respectively. These results show that side effect pairs having similar MedDRA categories tend to have similar frequency profiles across drugs.

Comparison with baselines in the warm start scenario

We compared DSGAT with the following competing methods. The optimal or default parameters of each method were considered in the comparing experiments.

- Galeano's method [18] is a non-negative matrix factorization model and first considered the drug–side effect frequency prediction problem. We used the best parameters reported in the original paper.
- MGPred [19] is a deep learning-based architecture for predicting frequencies of drug side effects, which integrates multi-view data. The best parameter settings are derived from the original paper.
- IGMC is a graph neural network (GNN)-based model and does not use any side information [33]. We modeled the drug–side effect frequency estimation as a recommendation system and used the cutting-edge method inductive graph-based matrix completion (IGMC) as a contrast. IGMC retrieves surrounding local subgraphs of drug–side effect pairs from the bipartite graph formed by the frequency matrix and then trains a GNN solely on the subgraphs and maps these subgraphs to their corresponding frequency items. We used the default parameters in IGMC for drug–side effect frequency prediction task.
- Ridge regression is a linear regression model with L2 regularization term. To demonstrate the superiority of molecular graph representations for drugs, we used the 100-dimensional vector representations of drugs obtained from mol2vec [34] as drug features, which were concatenated with MedDRA category features of side effects and then fed into ridge regression model. In the training process, we used the known drug–side effect frequency pairs as positive samples and randomly selected the same number of unknown pairs as negative samples. The ridge regression was performed with sklearn library [35]. We tried the different settings of the hyperparameter, i.e. L2 penalty coefficient, α, from {0.1, 0.5, 1, 5} and found that the results are basically the same. Here, we set $\alpha = 0.5$ in the comparing experiments.
- XGBoost is a tree-based regressor [36] whose input is the same as the ridge regression model. We set the number of trees from {100, 200, 500}, learning rate from {0.001, 0.01, 0.1}, L1 and L2 regularization coefficients from {0.001, 0.01, 0.1} and selected the best hyperparameters for model comparison.

We first evaluated the ability of DSGAT under the CV setting CVS1. Table 1 shows the comparison results obtained by various methods. DSGAT outperforms five other competing methods with a relatively large margin by achieving a MAP value of 0.251 as compared to 0.220 of Galeano's model, 0.135 of MGPred, 0.119 of IGMC, 0.041 of ridge regression and 0.048 of XGBoost. This means that DSGAT can better identify drug–side effect associations. This conclusion is also consistent considering other metrics such as AUC, NDCG@10, precision@1, precision@15, recall@1 and recall@15. As to the metrics referring to frequency prediction, such as Spearman's correlation, RMSE and MAE, DSGAT performs better than Galeano's model but worse than MGPred and IGMC. However, it should be noted that MGPred and IGMC have obtained lower MAP and AUC, which means that these methods cannot accurately identify drug–side effect associations and tend to introduce false positives in practical usage.

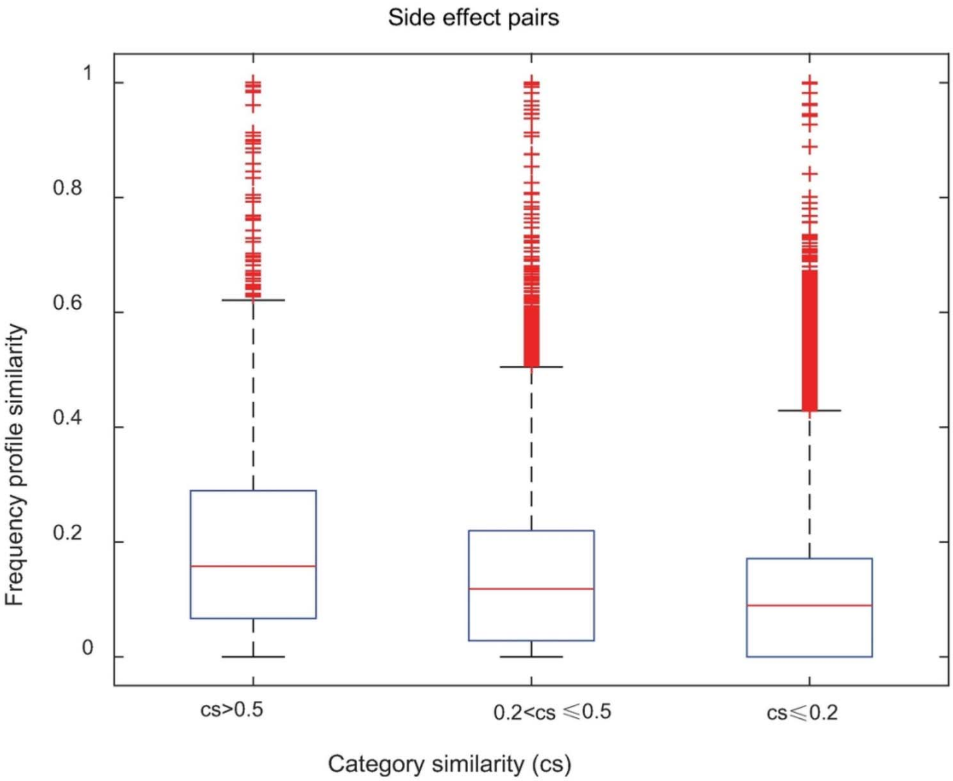


Figure 2. Box plots of side effect pairs concerning the cosine similarity between their associated frequency profiles across drugs. The box plots divide side effect pairs into three parts by cs, i.e. highly similar (cs > 0.5), moderately similar (0.2 < cs ≤ 0.5) and less similar (cs ≤ 0.2). By using one-tailed Wilcoxon rank-sum test, the differences in the distributions of frequency similarity values are statistically significant: highly similar versus moderately similar ($P < 1.04 \times 10^{-57}$) and moderately similar versus less similar ($P < 2.36 \times 10^{-211}$).

Table 1. Comparison results on the benchmark dataset under the CV setting CVS1

Method	MAP	AUC	NDCG@10	precision@1	precision@15	recall@1	recall@15	Spearman’s correlation	RMSE	MAE
Galeano’s model	0.220	0.907	0.447	0.265	0.112	0.085	0.390	0.494	1.285	0.948
MGPred	0.135	0.771	0.350	0.233	0.075	0.057	0.228	0.723	0.663	0.506
IGMC	0.119	0.745	0.312	0.209	0.068	0.052	0.186	**0.750**	**0.618**	**0.455**
Ridge regression	0.041	0.761	0.144	0.050	0.028	0.020	0.118	0.291	1.727	1.482
XGBoost	0.048	0.801	0.156	0.048	0.033	0.020	0.127	0.264	1.545	1.260
DSGAT	**0.251**	**0.922**	**0.497**	**0.312**	**0.128**	**0.087**	**0.434**	0.566	1.044	0.764

The bold values mean the best values achieved for each metric.

This point could be further clarified in the following analysis.

We discretized preference scores of drug–side effect items to determine their frequency classes for DSGAT, Galeano’s model, IGMC and MGPred. The preference scores of the known drug–side effect frequencies were obtained by using 10-fold CV under CVS1, while the preference scores of the unknown associations were achieved from one of the 10-fold CVs. Then, the kernel density estimation was adopted to estimate a probability density function (pdf) of preference scores for each frequency class. Finally, the thresholds of the classification decision were defined by using the pdfs and the maximum likelihood. As to DSGAT, the obtained boundary between unknown class 0 and frequency class 1 (0∼1) was 0.91, and the boundaries for classes 1∼2, 2∼3, 3∼4 and 4∼5 were 1.63, 2.46, 3.28 and 3.99, respectively. Figure 3 shows the case studies for the drug zaleplon and the side effect sinusitis by using DSGAT, where no more than five associations were randomly selected from each frequency class. Figure 4**A** shows the pdfs and accuracy percentages for each class for DSGAT. For each frequency class, the accuracy was defined as the number of items predicted to be a specific class divided by the number of true items. For Galeano’s model, the frequency class thresholds were 0.43, 1.26, 2.43, 3.25 and 3.93, which were taken from the originally published paper. For the infrequent and frequent classes (frequencies = 3 and 4), DSGAT outperforms Galeano’s model with a relatively large margin by obtaining accuracy values of 49.0% and 44.0%, as compared to 39.4% and 31.1% of Galeano’s model (Figure 4**B**). It is noted that items of frequency = 3 or 4 account for the majority of the known frequency items (74.38%). For IGMC, the pdf for

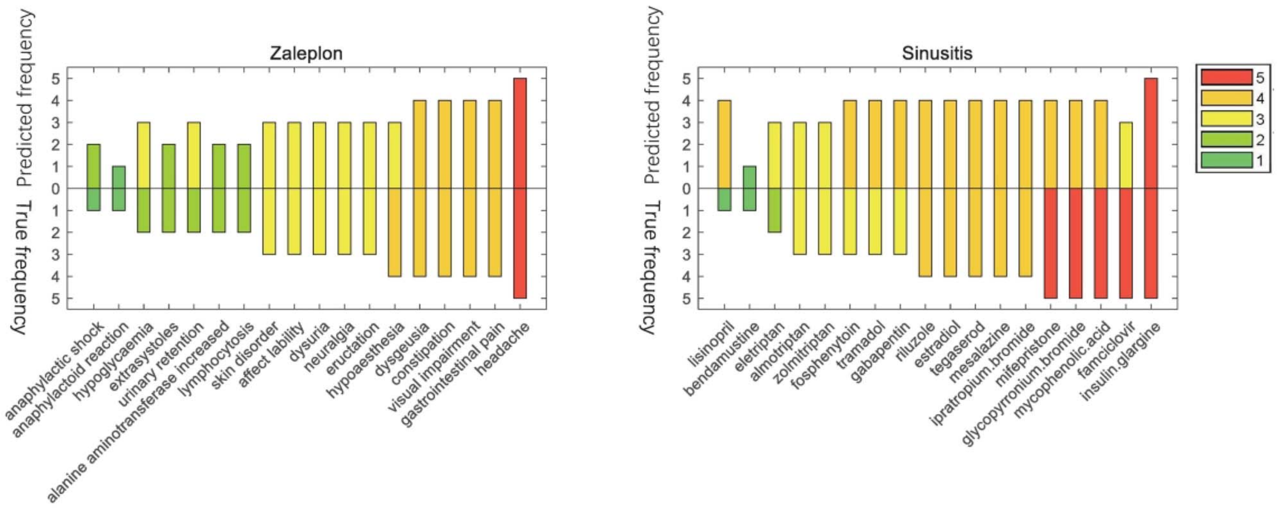


Figure 3. Case studies for zaleplon and sinusitis by using DSGAT. For the given drug (or side effect), we randomly chose no more than five side effect (or drug) associations from each frequency class.

unknown association class (frequency = 0) is similar to the pdf for frequent class (frequency = 4), which coincides with the result that most of the unknown associations (79.01%) were predicted as the infrequent or frequent class (Figure 4**C**). Here, the frequency class thresholds were set as 0, 2.07, 2.85, 3.55 and 4.16. For MGPred, the pdf for frequency = 0 is also similar to the pdf for frequency = 4, where most of the unknown associations (69.11%) were predicted as the infrequent or frequent class (Figure 4**D**). The frequency class thresholds were set as 0, 2.26, 2.74, 3.45 and 4.01 for MGPred. For drug–side effect rating matrix, nonzero elements take up only 4.97%. The case that most unknown associations are predicted to be infrequent or frequent class limits the use of IGMC and MGPred in the actual drug–side effect frequency prediction task.

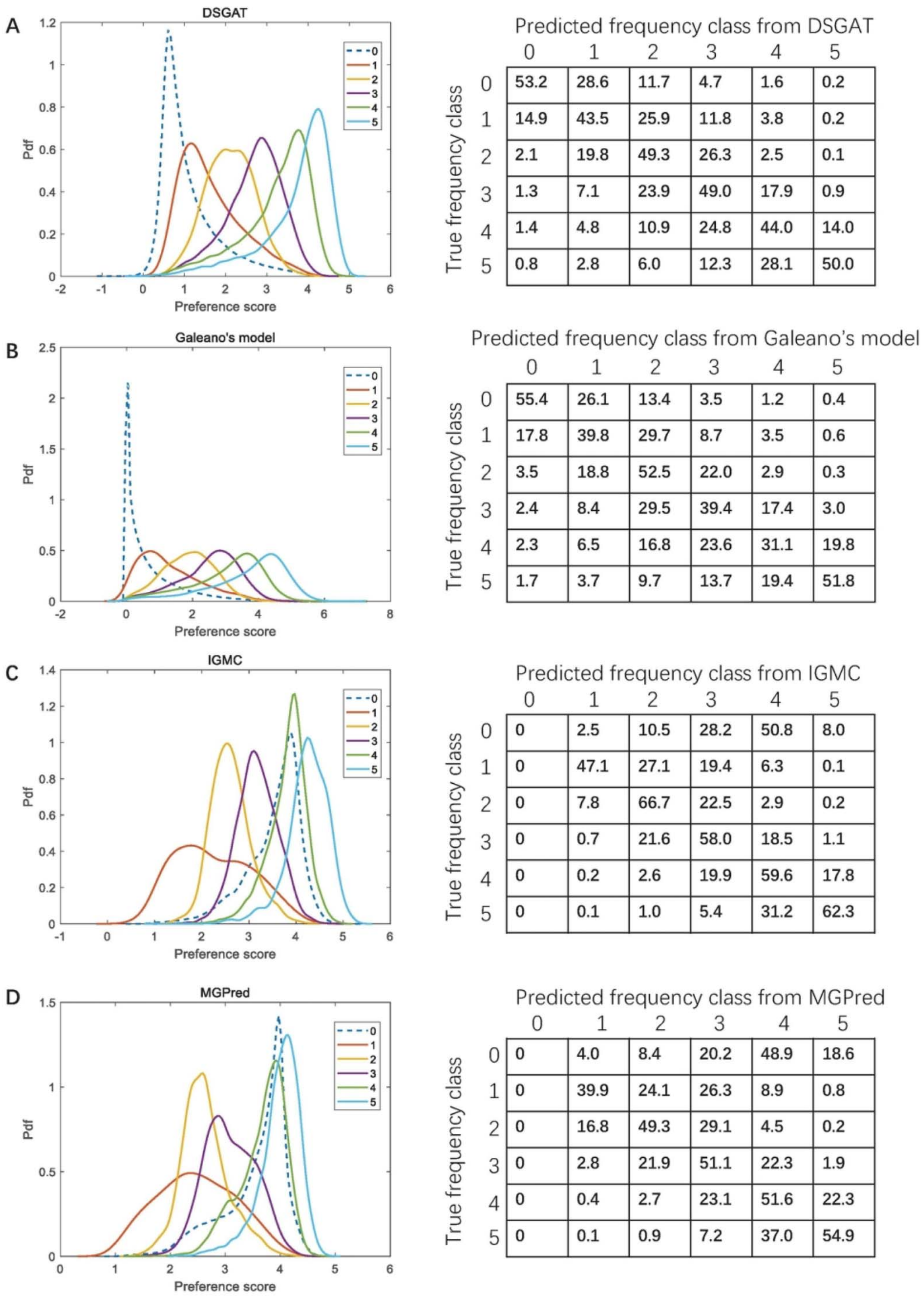
Cold start drugs

We then validated the performance of DSGAT for *de novo* drug–side effect frequency prediction under the CV setting CVS2 (Table 2). DSGAT yielded the most superior performance as to the association prediction metrics such as MAP = 0.404 and AUC = 0.878. Meanwhile, DSGAT achieved the best Spearman's correlation = 0.438, though RMSE and MAE were worse than those of MGPred and IGMC. In fact, for a given drug, we are more concerned with the order of recommended side effects rather than the magnitude of the predicted scores. Galeano's method cannot be applicable to cold start drugs. It approximates the frequency matrix M with the product of two low-rank matrices as follows: $M \approx WH$. For new drugs, their corresponding rows in M are all zero vectors, and the matrix decomposition model makes the rows in W for the new drugs all zeros, and consequently, the preference scores of the new drugs are completely zeros.

We further checked whether the chemical similarity between the drugs would affect our assessment of the performance of DSGAT in some way. To this end, we measured the performance of our method for drug training sets with different ranges of chemical similarity with the cold start drugs. Specifically, we randomly selected 75 drugs as a test set and used the remaining 675 drugs as a training set. The chemical similarity between any two drugs was computed by the Dice similarity from their Morgan fingerprints using the RDKit package. We filtered out the drugs in the training set whose chemical similarity to any drug in the test set exceeds a certain threshold and used the remaining training set for model training. The threshold is varied from 1 to 0.3 with the step size of 0.1. Consequently, the numbers of drugs in the training set are 675, 667, 652, 621, 576, 508, 295 and 70. Notably, GAT relies on pairs of adjacent nodes during training instead of relying on specific network structure, so it can be used for inductive tasks. Figure 5 shows the trends of AUC and Spearman's correlation with respect to the threshold, and the results are robust to the variation of the chemical similarity. Specifically, when the threshold was set as 0.3, the values of AUC and Spearman's correlation were 0.837 and 0.382, respectively.

Independent test

We first show the prediction ability of DSGAT in the warm start scenario. Among the 750 drugs and 994 side effects in the benchmark dataset, there are 9288 post-marketing side effect associations that are labeled as 'post-marketing' in SIDER and are reported after the drug entered the market. These post-marketing side effects are not appeared in the rating matrix M and are generally considered to be very rare side effects (frequency = 1) in the population because they have not been found in clinical trials [37, 38]. We used all the known entries (frequency > 0) in the rating matrix M as the training set and compared the prediction results of DSGAT, Galeano's model, IGMC and MGPred on the unknown entries (frequency = 0) of M. Notably, these prediction results



Predicted frequency class from DSGAT

True frequency class	0	1	2	3	4	5
0	53.2	28.6	11.7	4.7	1.6	0.2
1	14.9	43.5	25.9	11.8	3.8	0.2
2	2.1	19.8	49.3	26.3	2.5	0.1
3	1.3	7.1	23.9	49.0	17.9	0.9
4	1.4	4.8	10.9	24.8	44.0	14.0
5	0.8	2.8	6.0	12.3	28.1	50.0

Predicted frequency class from Galeano's model

True frequency class	0	1	2	3	4	5
0	55.4	26.1	13.4	3.5	1.2	0.4
1	17.8	39.8	29.7	8.7	3.5	0.6
2	3.5	18.8	52.5	22.0	2.9	0.3
3	2.4	8.4	29.5	39.4	17.4	3.0
4	2.3	6.5	16.8	23.6	31.1	19.8
5	1.7	3.7	9.7	13.7	19.4	51.8

Predicted frequency class from IGMC

True frequency class	0	1	2	3	4	5
0	0	2.5	10.5	28.2	50.8	8.0
1	0	47.1	27.1	19.4	6.3	0.1
2	0	7.8	66.7	22.5	2.9	0.2
3	0	0.7	21.6	58.0	18.5	1.1
4	0	0.2	2.6	19.9	59.6	17.8
5	0	0.1	1.0	5.4	31.2	62.3

Predicted frequency class from MGPred

True frequency class	0	1	2	3	4	5
0	0	4.0	8.4	20.2	48.9	18.6
1	0	39.9	24.1	26.3	8.9	0.8
2	0	16.8	49.3	29.1	4.5	0.2
3	0	2.8	21.9	51.1	22.3	1.9
4	0	0.4	2.7	23.1	51.6	22.3
5	0	0.1	0.9	7.2	37.0	54.9

Figure 4. pdf of preference scores and accuracy percentages for each frequency class. (**A**) Preference scores were obtained from the predictions in the 10-fold CVs for DSGAT. Note that frequency class 0 means the unknown association class. For each frequency class, among the true items, the percentage of items predicted to be a specific class were computed. (**B**–**D**) Images represent preference scores that were from Galeano's model, IGMC and MGPred, respectively.

also contain the results for post-marketing side effects. Figure 6 shows the pdfs of prediction scores for unknown class (frequency = 0) and post-marketing side effects. Besides, we also show the pdfs for very rare side effects (frequency = 1) in the rating matrix *M* by using 10-fold CV under CVS1. As illustrated in the Figure 6, for IGMC and MGPred, most of the post-marketing side effects (74.93% and 77.56%, respectively) were incorrectly predicted as infrequent (frequency = 3) or frequent (frequency = 4) class, which resulted in poor generalization performance for these two methods. For DSGAT and Galeano's model, 29.01% and 27.39% of the post-marketing side effects were correctly predicted as the very rare class. In terms of predicting whether there will be side effects, for DSGAT

Table 2. Comparison results on the benchmark dataset under the CV setting CVS2

Method	MAP	AUC	NDCG@10	precision@1	precision@15	recall@1	recall@15	Spearman's correlation	RMSE	MAE
Galeano's model	–	–	–	–	–	–	–	–	–	–
MGPred	0.224	0.738	0.756	0.621	0.318	0.025	0.167	0.421	0.896	0.668
IGMC	0.201	0.766	0.759	0.680	0.261	0.030	0.137	0.373	**0.864**	**0.620**
Ridge regression	0.141	0.764	0.562	0.269	0.170	0.010	0.091	0.174	2.325	2.078
XGBoost	0.135	0.719	0.581	0.331	0.197	0.014	0.099	0.317	3.541	3.414
DSGAT	**0.404**	**0.878**	**0.815**	**0.699**	**0.512**	**0.031**	**0.269**	**0.438**	1.469	1.175

The bold values mean the best values achieved for each metric.

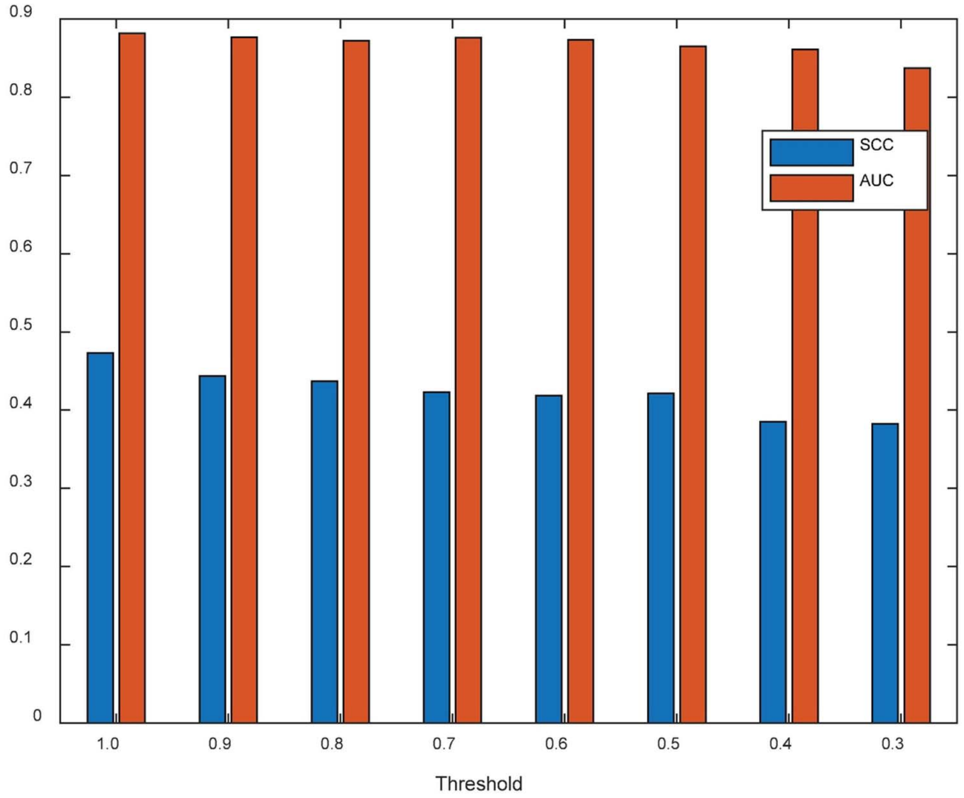


Figure 5. Variation of the performance of DSGAT, which was measured by AUC and SCC with the different settings of threshold for chemical similarity.

and Galeano's model, 84.07% and 78.01% of the post-marketing side effects were correctly identified to be occurred, respectively.

Based on the good performance of DSGAT on cold start drugs, we used nine new drugs from Galeano *et al*. [18] as the independent test set. Besides 370 known side effect associations having frequency information from SIDER database, we also integrated 93 post-marketing side effect associations, which are labeled as 'post-marketing' in SIDER. Similarly, post-marketing side effects are considered to be in very rare class (frequency = 1). We performed cold start prediction on the nine drugs, and for each drug, the prediction scores for the 994 side effects are shown in Figure 7. For each drug, the 994 side effects were categorized into six classes, i.e. the five known frequency classes and the one unknown association class. Figure 7 illustrates that DSGAT can not only distinguish the true drug–side effect associations from the unknown associations but can also correctly identify the true frequency. Specifically, the prediction performances are: AUC = 0.815 and Spearman's correlation = 0.429 for everolimus, AUC = 0.957 and Spearman's correlation = 0.415 for fidaxomicin, AUC = 0.839 and Spearman's correlation = 0.538 for gadoteridol, AUC = 0.837 and Spearman's correlation = 0.616 for esomeprazole, and AUC = 0.828 and Spearman's correlation = 0.701 for balsalazide.

Architecture and hyperparameter analysis

In this section, we will analyze some of the architecture components and hyperparameters in our model to show their influence. For one of 10-fold CVs under CVS1, we calculated the performance metrics for the test set after each epoch of training data.

Number of GAT layers

One-layer GAT only considers using the information of neighbors, while multiple-layer GAT tends to bring

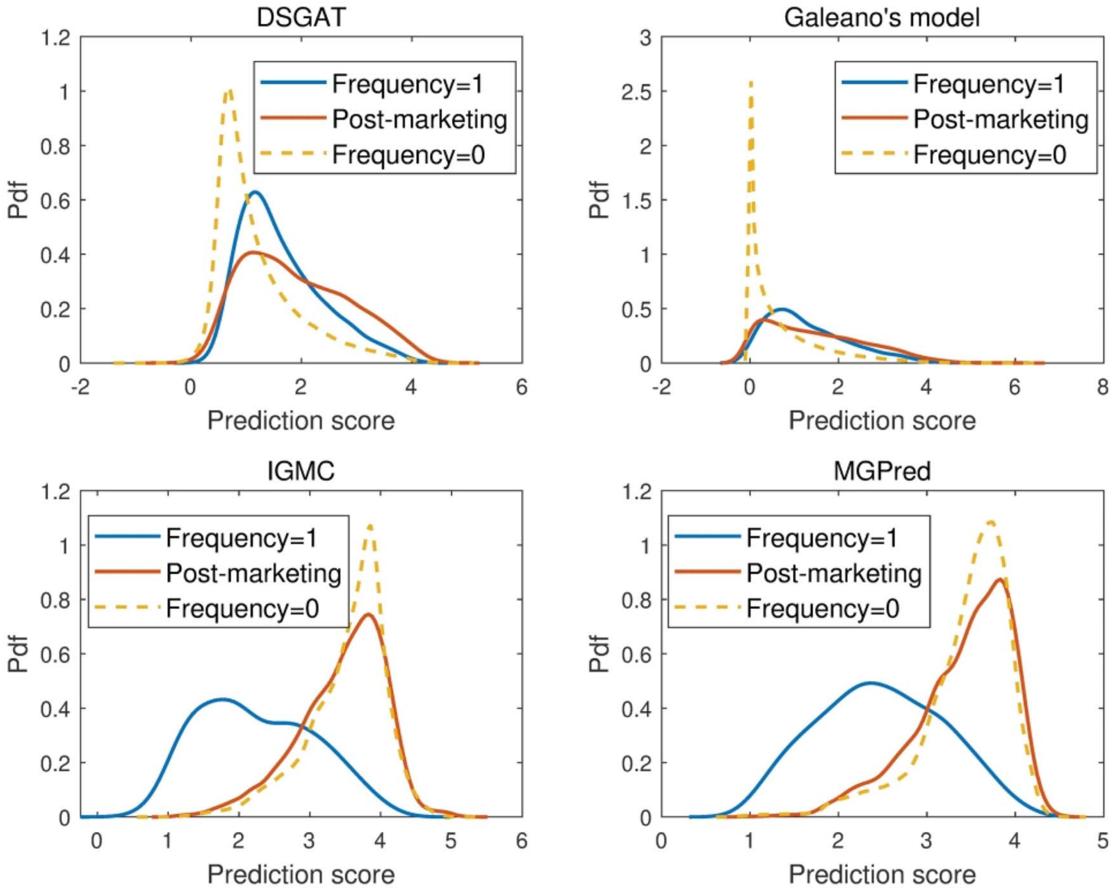


Figure 6. pdf of prediction scores for unknown association class (frequency = 0), very rare class (frequency = 1) and post-marketing side effects.

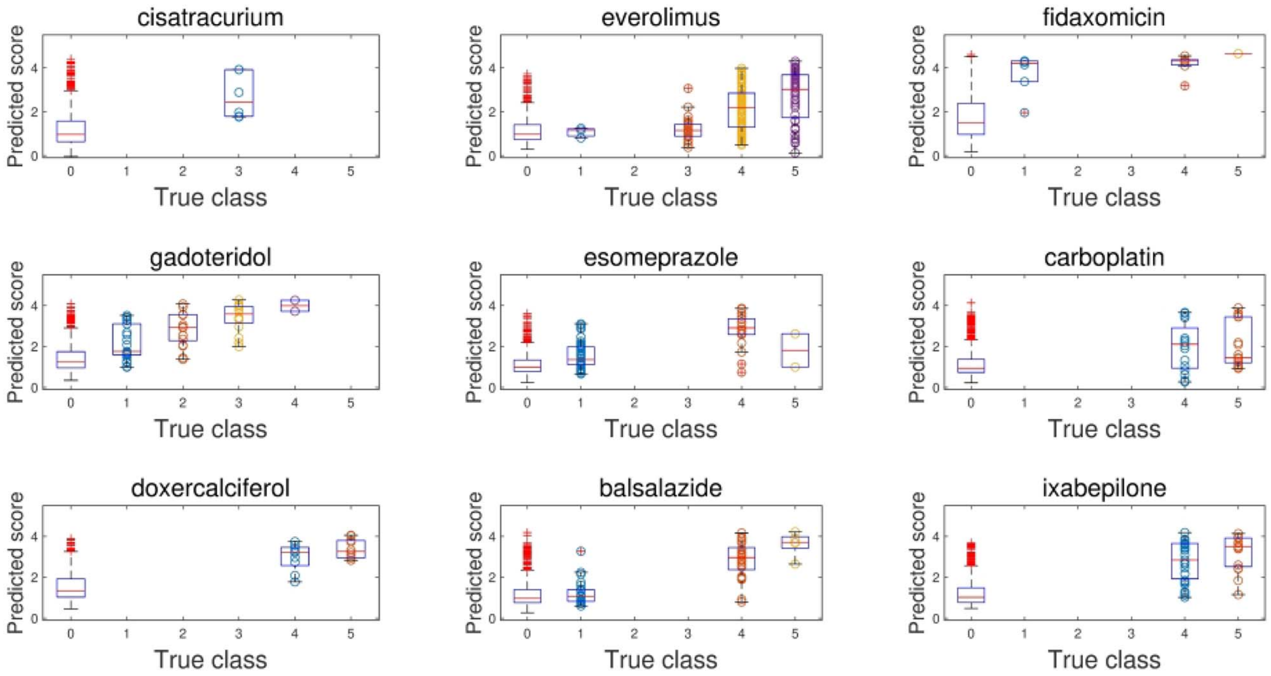


Figure 7. Box plots with scatter plots of predicted scores for drug–side effect items. The class 0 denotes the unknown association class, and we do not draw a scatter plot for class 0 as there are too many unknown side effects for a given drug.

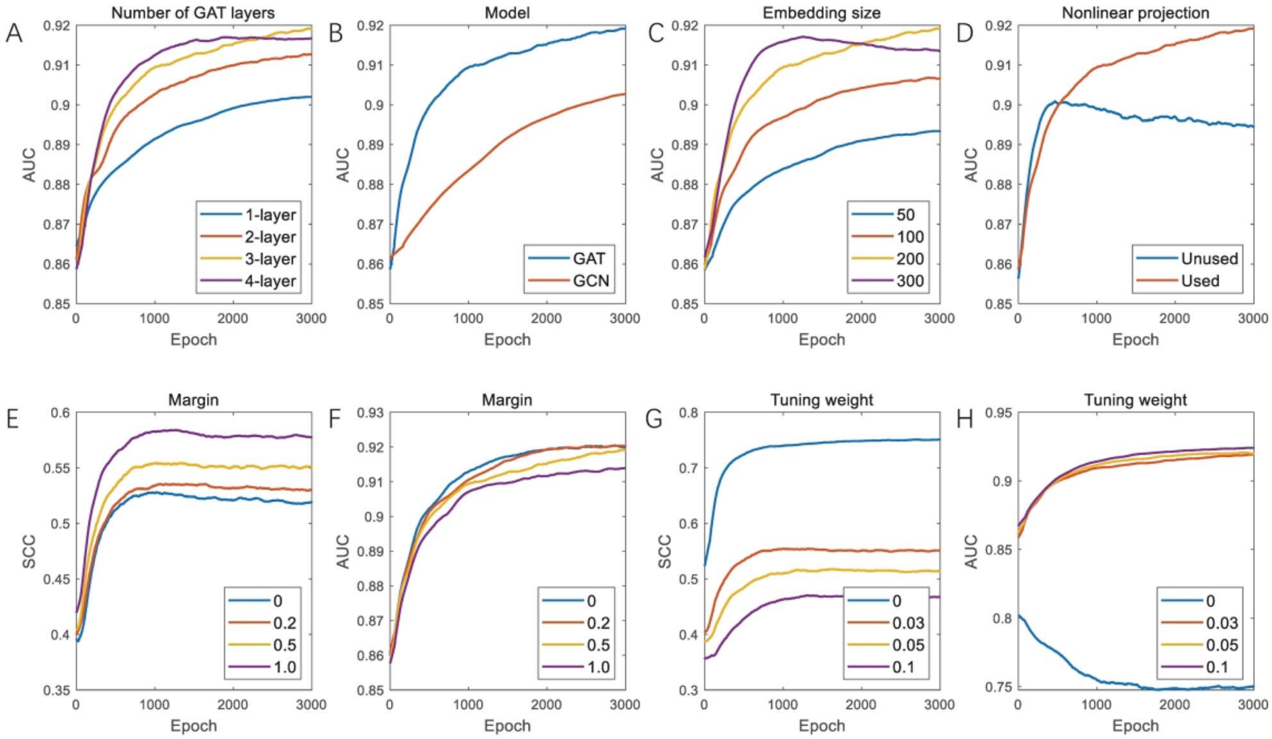


Figure 8. Parameter analysis. The trends of AUC values and SCCs for the test set with respect to the training epoch. (**A–D**) Images represent AUC values as to the number of GAT layers, the model, the embedding size and nonlinear projection. (**E–H**) Images represent SCCs and AUC values as to the margin and the tuning weight.

in more noises especially for the hand-crafted graphs. We selected the number of GAT layer from {1, 2, 3, 4}. As shown in Figure 8**A**, our method obtained the better AUC when the number of GAT layers was set to 3.

Graph model selection

Graph convolutional network (GCN) and GAT are two commonly used GNN models. We compared three-layer GAT with three-layer GCN, and the results showed that AUC of GAT exceeds that of GCN by 0.0165 (Figure 8**B**).

Embedding size

Embedding size is a key factor for encoder-decoder framework, which could directly affect the performance of the model. If the embedding size is too small, the model would be underfitting, and conversely overfitting occurs when the embedding size is too large. We set the embedding size from {50, 100, 200, 300}. As shown in Figure 8**C**, our model achieved the better AUC when the embedding size was set as 200.

Nonlinear projection

Here, we used GAT to learn embeddings for drugs and side effects, respectively. Because the distributions of drug and side effect embeddings may vary differently, we utilized the non-linear projection, such as two-layer fully connected network, to map the embeddings into a shared space. From the results in Figure 8**D**, we could see that nonlinear projection could improve prediction performance in our dataset compared to not using it.

Margin and tuning weight for unknown associations

For the unknown drug–side effect associations, the margin between the predicted scores and the labels determines the abscissa of the peak of the pdf, i.e. the expectation, of the predicted scores. The tuning weight controls the relative importance of unknown associations in the loss function. In our comparison, we chose the margin from {0, 0.2, 0.5, 1.0} and the tuning weight from {0, 0.03, 0.05, 0.1}. From Figure 8**E**–**H**, we could see that either margin ε= 1.0 or tuning weight α=0 results in a higher Spearman's correlation and a lower AUC, and the properly selected margin ε= 0.5 and tuning weight α=0.03 could obtain a relatively higher Spearman's correlation and a relatively higher AUC.

Conclusion and discussion

In this paper, we propose a drug–side effect frequency prediction method, DSGAT, based on the encoder-decoder framework. First, we represent each drug by its natural molecular graph and apply a three-layer GAT network to obtain the embedding for the drug. Second, using the cosine similarity between frequency profiles of side effects across drugs, we obtain side effect similarity and then use the top k neighbors of each side effect to construct the side effect graph. Subsequently, the side effect embedding is obtained by using a three-layer GAT network on the side effect graph. Finally, we project the drug and side effect embedding into a shared space by two-layer fully connected networks and adopt matrix factorization as the decoder.

In order to fully evaluate the model, we performed 10-fold CVs on the benchmark dataset under the warm start scenario and cold start scenario, respectively. The experimental results confirmed that our method DSGAT produces a significant improvement on cold start drugs and outperforms the state-of-the-art methods in warm start scenario. Notably, MGPred achieved AUC = 0.771 which was different from AUC = 0.931 reported in its original paper. In the original paper of MGPred, the authors treated the drug–side effect association prediction as a binary classification problem. The 37 071 drug–side effect frequency pairs in the rating matrix *M* were used as positive samples, and the randomly selected 37 071 unknown associations were used as negative samples. In each round, one-tenth of the positive samples and one-tenth of the negative samples were used as the test set, and the rest of the data were used as the training set, and the average metrics of the 10 rounds were reported as the final metrics. It is worth noting that the authors set frequency values (1, 2, 3, 4 and 5) to labels 1 regardless of whether they were in the training set or the test set. In such a way, AUC was indeed 0.931. MGPred's authors treated the drug–side effect association prediction problem and drug–side effect frequency prediction problem as two independent tasks. So as to the same drug–side effect pair, there were two distinct scores for association prediction and frequency prediction, respectively. As a contrast, in this paper, we considered drug–side effect association prediction and frequency prediction simultaneously, i.e. using the same training set and having the same prediction scores for the two tasks.

The independent test showed that DSGAT could correctly predict the true side effect frequencies and thus be helpful to guide drug risk–benefit evaluation. Now, our method only uses the chemical structure of the drug, while the integration of heterogeneous data, such as drug targets [26], therapeutic indications [39], perturbation transcriptomics data [40] and perturbation proteomics data [41], may further improve prediction performance. Besides, the roles of the learned embeddings for drugs and side effects from DSGAT are not fully investigated, and we will focus on their biological interpretation in the future.

Key Points

- DSGAT exploits the molecular graph by GAT and has the capability of predicting frequencies of side effects for the cold start drugs that do not appear in the training data.
- Empirical results on the benchmark dataset demonstrate that DSGAT yields significant improvement for cold start drugs and outperforms the state-of-the-art performance in the warm start scenario.
- The proposed novel loss function, i.e. weighted ε-insensitive loss function, could alleviate the sparsity problem and improve the prediction performance.
- Distribution analysis and independent test show that DSGAT can not only identify real drug–side effect associations from unknown associations but can also correctly predict the true frequency and thus could be a practical tool for drug risk–benefit evaluation.

Funding

Tianjin Municipal Science and Technology Bureau (18JCQNJC69500, 20JCZDJC00140, 20YDTPJC00560); Tianjin Education Commission (2018KJ107).

References

1. Davies M, Osborne V, Lane S, *et al.* Remdesivir in treatment of COVID-19: a systematic benefit-risk assessment. *Drug Saf* 2020;**43**(7):645–56.
2. Osborne V, Davies M, Lane S, *et al.* Lopinavir-ritonavir in the treatment of COVID-19: a dynamic systematic benefit-risk assessment. *Drug Saf* 2020;**43**(8):809–21.
3. Godat S, Fournier N, Safroneeva E, *et al.* Frequency and type of drug-related side effects necessitating treatment discontinuation in the Swiss Inflammatory Bowel Disease Cohort. *Eur J Gastroenterol Hepatol* 2018;**30**(6):612–20.
4. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med* 2000;**342**(25):1887–92.
5. Banda JM, Evans L, Vanguri RS, *et al.* A curated and standardized adverse drug event resource to accelerate drug safety research. *Sci Data* 2016;**3**:160026.
6. Pirmohamed M, James S, Meakin S, *et al.* Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ* 2004;**329**(7456):15–9.
7. Hughes JP, Rees S, Kalindjian SB, *et al.* Principles of early drug discovery. *Br J Pharmacol* 2011;**162**(6):1239–49.
8. Zhang W, Xu H, Li X, *et al.* DRIMC: an improved drug repositioning approach using Bayesian inductive matrix completion. *Bioinformatics* 2020;**36**(9):2839–47.
9. Cami A, Arnold A, Manzi S, *et al.* Predicting adverse drug events using pharmacological network models. *Sci Transl Med* 2011;**3**(114):114–27.
10. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics* 2016;**32**(15):2338–45.
11. Cakir A, Tuncer M, Taymaz-Nikerel H, *et al.* Side effect prediction based on drug-induced gene expression profiles and random forest with iterative feature selection. *Pharmacogenomics J* 2021;**21**:673–81.
12. Zhao X, Chen L, Guo ZH, *et al.* Predicting drug side effects with compact integration of heterogeneous networks. *Curr Bioinform* 2019;**14**:709–20.
13. Atias N, Sharan R. An algorithmic framework for predicting side effects of drugs. *J Comput Biol* 2011;**18**:207–18.
14. Huang LC, Wu XG, Chen JY. Predicting adverse side effects of drugs. *BMC Genomics* 2011;**12**:S11.

15. Nguyen PA, Born DA, Deaton AM, *et al.* Phenotypes associated with genes encoding drug targets are predictive of clinical trial side effects. *Nat Commun* 2019;**10**:1579.
16. Ding Y, Tang J, Guo F. Identification of drug-side effect association via semisupervised model and multiple kernel learning. *IEEE J Biomed Health Inform* 2019;**23**(6):2619–32.
17. 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.
18. Galeano D, Li S, Gerstein M, *et al.* Predicting the frequencies of drug side effects. *Nat Commun* 2020;**11**(1):4575.
19. Zhao H, Zhang K, Li Y, *et al.* A novel graph attention model for predicting frequencies of drug–side effects from multi-view data. *Brief Bioinform* 2021;**22**(6):bbab239.
20. Veličković P, Cucurull G, Casanova A, *et al.* Graph Attention Networks. In: *International Conference on Learning Representations (ICLR)*. BC, Canada: Vancouver, 2018.
21. Kuhn M, Letunic I, Jensen LJ, *et al.* The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**(D1): D1075–9.
22. Chen Y, Ma T, Yang X, *et al.* MUFFIN: multi-scale feature fusion for drug-drug interaction prediction. *Bioinformatics* 2021;**37**(17): 2651–8.
23. Torng W, Altman RB. Graph convolutional neural networks for predicting drug-target interactions. *J Chem Inf Model* 2019;**59**(10): 4131–49.
24. Liu Q, Hu Z, Jiang R, *et al.* DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics* 2020;**36**(Suppl_2):i911–8.
25. Zuo Z, Wang P, Chen X, *et al.* SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures. *BMC Bioinformatics* 2021;**22**(1):434.
26. 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.
27. Wang Y, Xiao J, Suzek TO, *et al.* Pubchem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* 2009;**37**(suppl_2):W623–33.
28. Brown EG, Wood L, Wood S. The medical dictionary for regulatory activities (MedDRA). *Drug Saf* 1999;**20**(2):109–17.
29. Jiang D, Wu Z, Hsieh CY, *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Chem* 2021;**13**(1):12.
30. Xiong Z, Wang D, Liu X, *et al.* Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J Med Chem* 2020;**63**(16):8749–60.
31. Withnall M, Lindelöf E, Engkvist O, *et al.* Building attention and edge message passing neural networks for bioactivity and physical-chemical property prediction. *J Chem* 2020;**12**(1):1.
32. Han P, Shang S, Yang P, *et al.* GCN-MF: disease-gene association identification by graph convolutional networks and matrix factorization. In: *Proceedings of the 25nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. USA: Anchorage, Alaska, 2019.
33. Zhang M and Chen Y. Inductive matrix completion based on graph neural networks. In: *International Conference on Learning Representations (ICLR)*. Ethiopia: Addis Ababa, 2020.
34. Jaeger S, Fulle S, Turk S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J Chem Inf Model* 2018;**58**(1): 27–35.
35. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
36. Chen T and Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. California, USA: San Francisco, 2018.
37. Brewer T, Colditz GA. Postmarketing surveillance and adverse drug reactions: current perspectives and future needs. *JAMA* 1999;**281**(9):824–9.
38. Tatonetti NP, Ye PP, Daneshjou R, *et al.* Data-driven prediction of drug effects and interactions. *Sci Transl Med* 2012;**4**(125):125ra31.
39. Davis AP, Grondin CJ, Johnson RJ, *et al.* Comparative toxicogenomics database (CTD): update 2021. *Nucleic Acids Res* 2021;**49**(D1):D1138–43.
40. Subramanian A, Narayan R, Corsello SM, *et al.* A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* 2017;**171**(6):1437–52.e17.
41. Zhao W, Li J, Chen MM, *et al.* Large-scale characterization of drug responses of clinically relevant proteins in cancer cell lines. *Cancer Cell* 2020;**38**(6):829–843.e4.